Scientific Software International

Choosing the type of correlation matrix to analyze

Contents

1. Types of correlation matrices

When one or more of the variables to be analyzed in LISREL are ordinal, it is important to choose the right type of moment matrix to analyze. Because ordinal variables do not have an origin or unit of measurement, the only meaningful moment matrices, when all variables are ordinal, are correlation matrices. PRELIS provides for four choices:

KM (continuous)

A matrix of product-moment (Pearson) correlations based on raw scores; that is, with scores 1,2,3, … on ordinal variables treated as if they come from interval-scaled variables. This is the KM option in PRELIS when all variables are declared continuous.

KM (ordinal)

A matrix of product-moment (Pearson) correlations with observations on ordinal variables replaced by normal scores determined from the marginal distributions. This is the KM option in PRELIS when *ordinal* variables are *normalized*.

OM (ordinal)

A matrix of product-moment (Pearson) correlations with observations on ordinal variables replaced by optimal scores determined for each pair. This is the OM option in PRELIS when *ordinal* variables are declared *ordinal.*

PM (ordinal)

A matrix of polychoric correlations. This is the PM option in PRELIS when *ordinal* variables are declared *ordinal*.

To investigate which of these correlations is “best”, we conducted two small experiments. The first involved only ordinal variables; the second involved both ordinal and continuous variables.

2. A Monte Carlo study of six correlation measures for ordinal variables

Let *x* and *y* bet two ordinal variables with *r* and *s* categories, respectively. The Monte Carlo study involves five steps.

Step 1

Choose the population correlation ρ and thresholds α_1, α_2, …, α_{r-1} for x and β_1, β_2, …, β_{s-1} = $-\infty$; $\left(\alpha_0 = \beta_0 = -\infty; \alpha_r = \beta_s = +\infty\right)$. Compute probabilities $\pi_{ij} = \Pr\left(x = i, y = j\right)$

$$\pi_{ij} = \int_{\alpha_{i-1}}^{\alpha_i} \int_{\beta_{j-1}}^{\beta_j} \phi_2(u,v)du\,dv$$

where ϕ_2 is the standard bivariate normal density with correlation ρ.

Step 2

Generate a random observation *x* = *i*, *y* = *j* with probability π_{ij}. Repeat this *N* times. This gives a contingency table:

$$\begin{bmatrix} n_{11} & n_{12} & \cdots & n_{1s} \\ n_{21} & n_{22} & \cdots & n_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ n_{r1} & n_{r1} & \cdots & n_{rs} \end{bmatrix}$$

Step 3

Compute all six correlation estimates and score 1 for the estimate closest to ρ. The six correlation types are:

KM (continuous)	PCM-RS	Product-moment correlation (raw scores)
*	SPEARMAN	Spearman’s rank correlation
*	KENDALL	Kendall’s tau-b coefficient
OM	CANON	Canonical correlation
KM (ordinal)	PMS-NS	Product-moment correlation (normal scores)
PM	POLYCHOR	Polychoric correlation

Step 4

Repeat Steps 2 and 3 three hundred times.

Step 5

Compute the mean, variance, bias, and mean squares error.

Some typical results

Results from two populations are presented in Tables 1 and 2. Population 1 has rather normal marginal distributions. In population 2, one variable has a U-shaped distribution and other a skewed distribution.

Table 1: Monte Carlo results on six correlation measures for ordinal variables

Population 1

ρ = 0.6

Thresholds for *x*: -0.842 0.524

Thresholds for *y*: -0.852 0.524 1.282

Probabilities

		y			
x	1	2	3	4	
1	0.100	0.090	0.009	0.001	0.200
2	0.090	0.293	0.092	0.025	0.500
3	0.010	0.117	0.099	0.074	0.300
	0.200	0.500	0.200	0.100	1.000

Results

		PMC-RS	SPEARMAN	KENDALL	CANON	PMC-NS	POLYCHOR
N = 100	Nbest	0	8	9	70	11	202
	Mean	0.4843	0.4869	0.3592	0.5122	0.4883	0.5867
	Variance	0.0053	0.0057	0.0078	0.0054	0.0055	0.0073
	Bias	-0.1157	-0.1131	-0.2408	-0.0878	-0.1117	-0.0133
	MSE	0.0187	0.0185	0.0658	0.0131	0.0180	0.0075
N = 100	Nbest	0	0	0	43	0	257
	Mean	0.4971	0.5008	0.3566	0.5091	0.5012	0.6001
	Variance	0.0014	0.0015	0.0019	0.0015	0.0015	0.0019
	Bias	-0.1029	-0.0992	-0.2434	-0.0909	-0.0988	-0.0001
	MSE	0.0120	0.0114	0.0612	0.0098	0.0112	0.0019
N = 100	Nbest	0	0	0	3	0	297
	Mean	0.4985	0.5021	0.3606	0.5062	0.5030	0.6017
	Variance	0.0005	0.0005	0.0007	0.0005	0.0005	0.0007
	Bias	-0.1015	-0.0979	-0.2394	-0.0938	-0.0970	-0.0017
	MSE	0.0108	0.0101	0.0580	0.0093	0.0099	0.0007

Table 2: Monte Carlo results on six correlation measures for ordinal variables

Population 2

ρ = 0.6

Thresholds for *x*: -0.604 -0.111 0.111 0.604

Thresholds for *y*: -1.645 -1.036 0.674 -0.253 0.126 0.674

Probabilities

x	1	2	3	4	5	6	7	
1	0.039	0.059	0.045	0.051	0.036	0.30	0.013	0.273
2	0.007	0.021	0.024	0.036	0.034	0.037	0.025	0.183
3	0.002	0.007	0.009	0.015	0.016	0.021	0.018	0.088
4	0.002	0.009	0.014	0.017	0.032	0.048	0.052	0.183
5	0.001	0.004	0.008	0.021	0.032	0.065	0.143	0.273
	0.059	0.100	0.100	0.150	0.150	0.200	0.250	1.000

Results

		PMC-RS	SPEARMAN	KENDALL	CANON	PMC-NS	POLYCHOR
N = 100	Nbest	0	11	0	85	12	192
	Mean	0.5298	0.5381	0.0363	0.5696	0.5393	0.5972
	Variance	0.0024	0.0024	0.0020	0.0023	0.0023	0.0027
	Bias	-0.0702	-0.0619	-0.5637	-0.0304	-0.0607	-0.0028
	MSE	0.0073	0.0063	0.3197	0.0032	0.0060	0.0027
N = 100	Nbest	0	0	0	63	6	223
	Mean	0.5297	0.5386	0.0382	0.5551	0.5395	0.5972
	Variance	0.0015	0.0016	0.0011	0.0015	0.0015	0.0018
	Bias	-0.0703	-0.0614	-0.5618	-0.0449	-0.0605	-0.0031
	MSE	0.0064	0.0054	0.3167	0.0035	0.0052	0.0018
N = 100	Nbest	0	0	0	40	0	260
	Mean	0.5331	0.5410	0.0431	0.5490	0.5415	0.5989
	Variance	0.0006	0.0006	0.0005	0.0005	0.0005	0.0006
	Bias	-0.0669	-0.0590	-0.5569	-0.0510	-0.0585	-0.0011
	MSE	0.0051	0.0041	0.3107	0.0031	0.0040	0.0006

We have many different populations of the kind illustrated in Tables 1 and 2, varying the number of categories, the cell probabilities, ρ, and the sample sizes. General conclusions that can be drawn from these Monte Carlo experiments are as follows:

- All correlations are biased downwards, but the bias for POLYCHOR (PM) is small and negligible for moderate sample sizes.
- POLYCHOR (PM), PCM-NS (KM – ordinal), and CANON (OM) do not appear to be sensitive to the shape of the marginal distributions.
- POLYCHOR (PM) is generally the best estimator, but the relative performances of PMC-NS (KM – ordinal) and CANON (OM) are improved as the number of categories increases, especially in moderate samples.
- POLYCHOR (PM) is almost always the best correlation in each sample in the sense of being closest to the true ρ. CANON (OM) is mostly second, and PCM-NS (KM – ordinal) is third. KENDALL is always the worst correlation, and SPEARMAN is only marginally better than PMC-RS.
- Only POLYCHOR (PM) appears to be a consistent estimator of ρ. Although variances of all the other correlations are small, their biases do not become small when the sample size increases.

3. An experiment with variables of different scale types

Four hundred observations were generated from a multivariate normal distribution with mean vector zero and covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1.000 & & & & & \\ 0.720 & 1.000 & & & & \\ 0.378 & 0.336 & 1.000 & & & \\ 0.324 & 0.288 & 0.420 & 1.000 & & \\ 0.270 & 0.240 & 0.350 & 0.300 & 1.000 & \\ 0.270 & 0.240 & 0.126 & 0.108 & 0.090 & 1.000 \end{bmatrix}$$

This covariance matrix has been constructed to satisfy, exactly, a factor analysis model with two correlated factors and a clear simple structure (see Jöreskog, 1979). The following values were assigned to the thresholds α_1, α_2, …, α_{k-1} of variables 2, 3, 4, and 6, where k is the number of categories.

Variable 2	Variable 3	Variable 4	Variable 6
(k = 7)	(k = 5)	(k = 3)	(k = 2)
α_1 = -1.64	α_1 = -0.60	α_1 = -0.67	α_1 = -0.25
α_2 = -1.04	α_2 = -0.11	α_2 = -0.67	
α_3 = -0.67	α_3 = -0.11		
α_4 = -0.25	α_4 = -0.60		
α_5 = -0.13			
α_6 = -0.67			

These thresholds correspond to the following marginal probabilities.

Variable 2	Variable 3	Variable 4	Variable 6
π_1 = 0.05	π_1 = 0.273	π_1 = 0.25	π_1 = 0.4
π_2 = 0.10	π_2 = 0.183	π_2 = 0.50	π_2 = 0.6
π_3 = 0.10	π_3 = 0.088	π_3 = 0.25	
π_4 = 0.15	π_4 = 0.183		
π_5 = 0.15	π_5 = 0.273		
π_6 = 0.20			
π_7 = 0.25			

Thus, variable 2 is skewed, variable 3 has a U-shaped distribution, variable 4 is symmetrical, and variable 6 is dichotomous.

These four variables were then transformed to ordinal variables as follows:

If $x \le \alpha_1$, the value 1 was assigned to the observation *x*.

If $\alpha_{i-1} < x \le \alpha_i$, the value *i* was assigned to the observation *x*.

If $\alpha_{k-1} < x$, the value *k* was assigned to the observation *x*.

Variables 1 and 5 were unchanged. Finally, 20 percent of all entries in the data matrix were randomly changed to -9, representing missing observations. The first 20 cases of the data matrix generated in this way are as follows.

-2.14	2	1	-9	-0.60	1
-0.42	7	-9	-9	-0.55	2
-9.00	7	1	3	1.33	2
0.57	6	1	1	-1.96	2
-1.72	-9	5	-9	-0.88	2
-9.00	5	4	2	-1.68	-9
0.57	7	1	2	-0.86	2
0.51	6	4	2	0.88	2
0.85	7	5	-9	-9.00	2
1.90	7	-9	2	1.54	1
-1.13	-9	5	2	2.45	-9

-0.03	2	-9	2	-0.75	-9
-2.20	-9	-9	1	-2.26	1
0.66	-9	2	1	1.06	2
-0.81	6	4	3	-0.28	2
-1.58	-9	1	-9	-0.56	-9
-0.56	-9	1	-9	-2.08	-9
0.95	6	-9	-9	-0.03	1

PRELIS was used to compute four types of correlation matrices using both pairwise and listwise deletion. This yielded eight estimated correlation matrices. Each is compared to the true correlation matrix $\boldsymbol{\Sigma}$ given above.

Results are shown in Table 3. (In Tables 3 and 4, columns for the four types of correlation matrices are labeled PMC – RS, PMC – NS, PMC – OS, and PP – PS, corresponding to the PRELIS options KM (continuous), KN (ordinal), OM, and PM, as defined previously.

Each of the eight correlation matrices was further analyzed with LISREL to fit a restricted (confirmatory) factor analysis model with two correlated factors. The factor loading matrix had three fixed zeroes in each column (see Jöreskog, 1979). The maximum likelihood (ML) method was used in LISREL to fit the model, even though there was no theoretical justification for using ML in this case. The sample size was assumed to be 280 for the four correlation matrices computed under listwise deletion. Sample size affects only the chi-square values.

Results are shown in Table 4.

Comments on Tables 3 and 4

- In terms of bias and mean square error, there seems to be a clear trend in the tables: PMC – RS and PMC – NS are most biased and have the largest mean square error; PMC – OS performs somewhat better; and PP – PS is least biased and has the smallest mean square error. This holds for both pairwise and listwise deletion.
- In this case, pairwise deletion gave better results than listwise deletion. This undoubtedly because 20% of the observations were missing at random, reducing the *effective sample size* under listwise deletion to only 102, while the pairwise sample sizes varied from 240 to 328.
- The correlations in Table 3 were mostly underestimated. As can be seen in Table 4, this results in underestimates of factor loadings and in overestimates of unique variances.
- All eight correlation matrices generated by PRELIS were positive-definite. As seen in Table 4, the ML method gave good results even when some of the variables are ordinal. This demonstrates that, although normal-theory chi-square values and standard errors are not valid, the ML method may still be used to fit the model to the data.
- Table 4 also reports three measures of overall fit: chi-square with eight degrees of freedom, adjusted goodness of fit index (AGFI), and root mean residual (RMSR). In this case, chi-square values and other measures of fit behave quite normally.
- The numbers presented in Tables 3 and 4 represent a single case study from which very well-established conclusions cannot be drawn. To obtain clearer and more exact results, a full-scale Monte Carlo study should be undertaken.

Table 3: Eight different correlation estimates

Simulated data, *N* = 400

Except for the first column, all numbers are deviations from the true value. All numbers have been multiplied by 1000 and rounded.

True Value	Pairwise deletion				Listwise deletion			
	PMC -RS	PMC – NS	PMC – OS	PP - PS	PMC -RS	PMC – NS	PMC – OS	PP - PS
720	-16	-12	23	14	-22	-22	12	-2
378	-59	-57	-33	-34	-69	-59	-34	-42
336	-59	-44	-20	-7	-87	-88	-14	-54
324	-43	-43	-11	-17	-41	-41	-9	-18
288	-29	-17	12	25	-19	-9	94	32
420	-83	-81	-56	-16	8	-1	29	77
270	-34	-34	-34	-34	-51	-51	-51	-51
240	4	13	24	17	-59	-59	-31	-37
350	-49	-44	-20	-29	57	59	92	70
300	-14	-14	19	17	-59	-59	-31	-37
270	-98	-98	-51	-53	-118	-118	-77	-78
240	-1	-31	1	36	-48	-65	-14	-10
126	51	40	63	0	35	35	87	94
108	-125	-125	-76	-84	-159	-159	-49	-37
90	-53	-53	-43	-44	-119	-119	-153	-68
BIAS	-42	-40	-13	-14	-48	-49	-10	-11
MSE	59	57	38	35	74	74	65	54

Table 4: Eight different parameter estimates

Simulated data, *N* = 400

Except for the first column, all numbers are deviations from the true value. All numbers have been multiplied by 1000 and rounded.

True Value	Pairwise deletion				Listwise deletion			
	PMC -RS	PMC – NS	PMC – OS	PP - PS	PMC -RS	PMC – NS	PMC – OS	PP - PS
900	-105	-59	-37	-11	-132	-18	-62	-10
800	33	43	61	6	4	-9	73	33
700	-89	-85	-66	-45	65	62	70	3
600	-44	-45	-10	20	-41	-50	-4	20
500	-1	1	18	-2	9	13	28	-2
300	-66	-74	-33	-11	-100	-110	-50	-46
600	-21	-12	-6	-17	-117	-110	-57	-102
190	178	103	66	22	221	32	108	-30
360	-54	-71	-101	-9	-6	14	-123	-54
510	117	111	88	61	-95	-93	-103	-4
640	51	52	12	-24	48	57	5	-24
750	1	-1	-18	2	-9	-13	-28	2
910	35	39	19	7	50	54	28	25
BIAS	3	0	-1	0	-8	-13	-9	-15
MSE	78	63	51	25	92	60	68	38
CHI-SQR	10.94	8.81	8.21	5.22	7.99	7.82	11.02	6.81
AGFI	0.967	0.973	0.975	0.984	0.939	0.940	0.914	0.946
RMSR	0.031	0.029	0.027	0.022	0.041	0.041	0.049	0.036

References

Jöreskog, K.G. (1979). *Basic ideas on factor and component analysis*. In: K.G. Jöreskog & D. Sörbom: *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt books, 5-20.