# Problems with analysis of correlation matrices

The general rule is that the sample covariance matrix should be analyzed. However, in many behavioral sciences applications, units of measurements in the observed variables have no definite meaning and are often arbitrary or irrelevant. For these reasons, for convenience, and for interpretational purposes, the sample correlation matrix is often analyzed as if it is a covariance matrix. This is a common practice. The usual argument for analyzing a correlation matrix is that one wants a standardized solution. However, this argument is not valid as one can obtain a completely standardized solution even when the covariance matrix is analyzed. The completely standardized solution is obtained by putting SC on the OU command in LISREL syntax or on an Options line in SIMPLIS syntax. A path diagram of the completely standardized solution is obtained by selecting the Standardized Solution when the path diagram is visible.

The analysis of correlation matrices is problematic in several ways. As pointed out by Cudeck (1989), such an analysis may

(a) modify the model being analyzed,

(b) produce incorrect $\chi^2$ and other goodness-of-fit measures, and

(c) give incorrect standard errors.

Problem (a) can occur when the model includes equality constraints or other constrained parameters. For example, if $\lambda_{11}^{(x)}$ and $\lambda_{21}^{(x)}$ are constrained to be equal but the variances $\sigma_{11}^{(x)}$ and $\sigma_{22}^{(x)}$ are not equal, then analysis of the correlation matrix will give estimates of

$$\lambda_{11}^{(x)} / \sqrt{\sigma_{11}^{(x)}} \text{ and } \lambda_{21}^{(x)} / \sqrt{\sigma_{22}^{(x)}}$$

which are not equal. Correlation matrices should not be analyzed if the model contains equality constraints of this kind.

Problem (b) can occur in multiple group problems if the observed variables are standardized within groups and there are equality constraints or other constraints across groups.

Problem (c) can occur when a correlation matrix (KM or PM) is analyzed with any method except WLS using the asymptotic covariance matrix of the correlations as estimated by PRELIS. In particular, standard errors will be wrong if a correlation matrix is analyzed with ML.

The main question is whether the standard errors and $\chi^2$ goodness-of-fit measures produced when correlation matrices are used are asymptotically correct. The exact conditions under which this is the case are extremely complicated and give little practical guidance. However, two crucial conditions are

- that the model is scale invariant
- that the fitted covariance matrix has ones in the diagonal

The second condition can be checked by examining the fitted residuals, which should all be zero in the diagonal.

To clarify the issue further, we distinguish between covariance and correlation structures. In principle, all LISREL models are covariance structures, where the variances of $\boldsymbol{\Sigma}$ as well as the covariances are functions of the parameters. By contrast, in a correlation structure, the diagonal elements of $\boldsymbol{\Sigma}$ are constants independent of parameters. We now distinguish between four possible cases.

**A**: A sample covariance matrix is used to estimate a covariance structure.

**B**: A sample correlation matrix is used to estimate a covariance structure.

**C**: A sample covariance matrix is used to estimate a correlation structure.

**D**: A sample correlation matrix is used to estimate a correlation structure.

Case **A** is the standard case in LISREL. In large samples, standard errors and chi-squares will be correct for any method of estimation under multivariate normality of the observed variables and under non-normality if the asymptotic covariance matrix of the sample variances and covariances is provided.

Case **B** is a very common situation. Asymptotic variances and covariances of sample correlations are not of the same form as those of the sample variances and covariances, so standard errors will in general be incorrect. However, if the two conditions above hold, chi-squares will still be correct.

For the most common type of models, it is possible to obtain correct standard errors by writing the model in a different way. The following example illustrates this for a confirmatory factor analysis model.

The first example produces wrong standard errors in LISREL syntax (**LAWLEY1.LIS**), found in the **LISREL Examples** folder.

```
Lawley Factor Analysis Example. Wrong standard errors.
DA NI=9 NO=72 MA=KM
LA
VIS_PERC CUBES LOZENGES PAR_COMP SEN_COMP WRD_MNG
ADDITION CNT_DOT ST_CURVE
KM=LAWLEY.COR
MO NX=9 NK=3 PH=ST
LK
Visual Verbal Speed
FR LX 1 1 LX 2 1 LX 3 1 LX 4 2 LX 5 2 LX 6 2 LX 7 3 LX 8 3 LX 9 1 LX 9 3
OU
```

Next is the command file (**LAWLEY2.LIS**) that produces the correct standard errors.

```
Lawley Factor Analysis Example. Correct standard errors.
DA NI=9 NO=72 MA=KM
LA
VIS_PERC CUBES LOZENGES PAR_COMP SEN_COMP WRD_MNG
ADDITION CNT_DOT ST_CURVE
KM=LAWLEY.COR
```

```
MO NY=9 NE=9 NK=3 LY=DI,FR GA=FI PH=ST PS=DI TE=ZE
LK
Visual Verbal Speed
FR GA 1 1 GA 2 1 GA 3 1 GA 4 2 GA 5 2 GA 6 2 GA 7 3 GA 8 3 GA 9 1 GA 9 3
CO PS 1 1 = 1 - GA 1 1 ** 2
CO PS 2 2 = 1 - GA 2 1 ** 2
CO PS 3 3 = 1 - GA 3 1 ** 2
CO PS 4 4 = 1 - GA 4 2 ** 2
CO PS 5 5 = 1 - GA 5 2 ** 2
CO PS 6 6 = 1 - GA 6 2 ** 2
CO PS 7 7 = 1 - GA 7 3 ** 2
CO PS 8 8 = 1 - GA 8 3 ** 2
CO PS 9 9 = 1 - GA 9 1**2 - 2*GA 9 1*GA 9 3*PH 3 1 - GA 9 3**2
OU SO
```

The factor loadings with correct standard errors can now be found in the Gamma matrix. It may verified that these two different formulations of the same confirmatory factor analysis model give exactly the same parameter estimates and fit statistics; only the standard errors and *t*-values differ.

In Case **C**, the model is defined as a correlation structure $\mathbf{P}(\boldsymbol{\theta})$, with diagonal elements equal to 1 (*i.e., the diagonal elements are not functions of parameters*), it may be formulated as a covariance structure

$$\boldsymbol{\Sigma} = \mathbf{D}_\sigma \mathbf{P}(\boldsymbol{\theta}) \mathbf{D}_\sigma,$$

where $\mathbf{D}_\sigma$ is a diagonal matrix of population standard deviations $\sigma_1, \sigma_2, ..., \sigma_k$ of the observed variables, which are regarded as free parameters. The covariance structure described in the equation above has parameters $\sigma_1, \sigma_2, ..., \sigma_k, \theta_1, \theta_2, ..., \theta_t$. Such a model may be estimated correctly using the sample covariance matrix **S**. However, the standard deviations $\sigma_1, \sigma_2, ..., \sigma_k$, as well as $\boldsymbol{\theta}$ must be estimated from the data and the estimate of $\sigma_i$ does not necessarily equal the corresponding standard deviation $s_i$ in the sample. When $\mathbf{P}(\boldsymbol{\theta})$ is estimated directly from the sample correlation matrix $\mathbf{R} = (\mathbf{r_{ij}})$, standard errors and $\chi^2$ goodness-of-fit values will in general not be correct.

Consider Case **D**. To obtain correct asymptotic standard errors in LISREL for a correlation structure when the correlation matrix is analyzed, the WLS method must be used with a weight matrix $\mathbf{W}^{-1}$, where $\mathbf{W}$ is a consistent estimate of the asymptotic covariance matrix of *the correlations being analyzed*. Such a $\mathbf{W}$ may be obtained with PRELIS under non-normal theory. PRELIS can estimate such a $\mathbf{W}$ also for a correlation matrix containing polychoric and/or polyserial correlations.

The asymptotic covariance matrix $\mathbf{W}$ produced by PRELIS is a consistent estimate of the covariance matrix of

$$\mathbf{r} = \left(r_{21}, r_{31}, r_{32}, r_{41}, r_{42}, ...\right)$$

The diagonal elements of the correlation matrix $\mathbf{R}$ are not included in this vector. The number of distinct elements in $\mathbf{W}$ are

$$\frac{1}{2}k(k-1)\left[\frac{1}{2}k(k-1)+1\right],$$

where $k = p + q$ is the number of observed variables in the model. In fitting a correlation structure $\mathbf{P}(\boldsymbol{\theta})$ to a correlation matrix using WLS, LISREL minimizes the fit function

$$F(\boldsymbol{\theta}) = (\mathbf{r} - \boldsymbol{\rho})' \mathbf{W}^{-1} (\mathbf{r} - \boldsymbol{\rho}) \qquad (1)$$

where

$$\boldsymbol{\rho}' = \left( \rho_{21}(\boldsymbol{\theta}), \rho_{31}(\boldsymbol{\theta}), \rho_{32}(\boldsymbol{\theta}), \rho_{41}(\boldsymbol{\theta}), \ldots, \rho_{k,k-1}(\boldsymbol{\theta}) \right).$$

This approach assumes that the diagonal elements of $\mathbf{P}(\boldsymbol{\theta})$ are fixed ones and not functions of parameters.

WLS may also be used to fit ordinary LISREL models (*i.e., covariance structures*) to sample correlation matrices (Case **B**). This is especially useful when polychoric and polyserial correlations are analyzed. A small problem arises here because the fit function (1) is not a function of the diagonal elements of $\mathbf{P}(\boldsymbol{\theta})$, and, as a consequence, parameters such as the diagonal elements of $\boldsymbol{\Theta}_{\varepsilon}$ and $\boldsymbol{\Theta}_{\delta}$ cannot be estimated directly. However, they can of course be estimated afterwards.

A better and more general approach is to add the term

$$\sum \left[ 1 - \sigma_{ii}(\theta) \right]^2 \qquad (2)$$

to the fit function (1). The advantages of this approach are

- Estimates of all parameters can be obtained directly even when constraints are imposed
- When the diagonal elements of $\boldsymbol{\Theta}_{\varepsilon}$ and $\boldsymbol{\Theta}_{\delta}$ are free parameters, the WLS solution will satisfy

$$Diag(\hat{\boldsymbol{\Sigma}}) = \mathbf{I}.$$

This approach can be generalized further by replacing the one in (2) with a variance $s_{ii}$ which has been estimated or obtained separately from the correlations. Such variances can be obtained with PRELIS for ordinal variables for which the thresholds are assumed to be equal.