Simplex models

A simplex model is a type of covariance structure which often occurs in longitudinal studies when the same variable is measured repeatedly on the same people over several occasions. The simplex model is equivalent to the covariance structure generated by a first-order non-stationary autoregressive process. Guttman (1954) used the term simplex also for variables which are not ordered through time but by other criteria. One of his examples concerns tests of verbal ability ordered according to increasing complexity. The typical feature of a simplex correlation structure is that the entries in the correlation matrix decrease as one moves away from the main diagonal.

Jöreskog (1970a) formulated various simplex models in terms of the well-known Wiener and Markov stochastic processes. A distinction was made between a perfect simplex and a quasi-simplex. A *perfect simplex* is reasonable only if the measurement errors in the variables are negligible. A *quasi-simplex*, on the other hand, allows for sizable errors in measurement.

Consider *p* fallible variables $y_1, y_2, ..., y_p$. The unit of measurement in the true variables η_i may be chosen to be the same as in the observed variables y_i. The equations defining the model are then

$$y_i = \eta_i + \varepsilon_i$$
$$\eta_i = \beta_i \eta_{i-1} + \zeta_i, \quad i = 2, 3, ..., p,$$

where the ε_i are uncorrelated among themselves and uncorrelated with all the η_i and where ζ_i is uncorrelated with η_{i-1} (where $i = 2, 3, ..., p$).

The parameters of the model are $\omega_1 = Var(\eta_1)$, $\psi_i = Var(\zeta_i)$ (where $i = 2, 3, ..., p$), $\theta_i = Var(\varepsilon_i)$, ($i = 1, 2, 3, ..., p$) and β_2, β_3, …, β_p. Let $\omega_i = Var(\eta_i) = \beta_i^2 \omega_{i-1} + \psi_i$ ($i = 1, 2, 3, ..., p$). Then there is a one-to-one correspondence between the parameters β_2, β_3, …, β_p, ω_1, ψ_2, ψ_3, …., ψ_p and the parameters β_2, β_3, …, β_p, ω_1, ω_2, …., ω_p. The ω's are not parameters in the LISREL model, so in LISREL the first set of parameters must be used. However, for identification purposes it is more convenient to use the second set of parameters. In terms of the ω's, for *p* = 4 measurement occasions, the covariance matrix of y_1, y_2, …, y_p has the form

$$\Sigma = \begin{bmatrix} \omega_1 + \theta_1 & & & \\ \beta_2\omega_1 & \omega_2 + \theta_2 & & \\ \beta_2\beta_3\omega_1 & \beta_3\omega_2 & \omega_3 + \theta_3 & \\ \beta_2\beta_3\beta_4\omega_1 & \beta_3\beta_4\omega_2 & \beta_4\omega_3 & \omega_4 + \theta_4 \end{bmatrix}.$$

It is seen that, although the product $\beta_2\omega_1 = \sigma_{21}$ is identified, β_1 and ω_1 are not separately identified. The product $\beta_2\omega_1$ is involved in the off-diagonal elements in the first column (and row) only. One can multiply β_2 by a non-zero constant and divide ω_1 by the same constant without changing the product. The change induced by ω_1 in σ_{11} can be absorbed in θ_1 in such a way that σ_{11} remains unchanged. Hence θ_1 = $Var(\varepsilon_1)$ is not identified. For η_2 and η_3 we have

$$\omega_2 = \frac{\sigma_{32}\sigma_{21}}{\sigma_{31}}, \quad \omega_3 = \frac{\sigma_{43}\sigma_{32}}{\sigma_{42}},$$

so that ω_2 and ω_3, and hence also θ_2 and θ_3, are identified. With ω_2 and ω_3 identified, β_2 and β_3 are identified by σ_{32} and σ_{43}. The middle coefficient β_3 is overidentified since

$$\beta_3\omega_2 = \frac{\sigma_{31}\sigma_{42}}{\sigma_{41}} = \sigma_{32}.$$

Since both ω_2 and θ_4 are involved in σ_{44} only, these are not identified. Only their sum, σ_{44}, is identified.

This analysis of the identification problem shows that for the "inner" variables y_2 and y_3, the parameters ω_2, ω_3, θ_2, θ_3, and β_3 are identified, whereas there is an indeterminacy associated with each of the "outer" variables y_1 and y_4. To eliminate these indeterminacies one condition must be imposed on the parameters ω_1, θ_1, and β_2, and another on the parameters ω_4 and θ_4. In terms of the original LISREL parameters, β_2, $\psi_1 = \omega_1$, ψ_2, ψ_4, θ_1, and θ_4 are not identified whereas β_3, β_4, ψ_3, θ_2 and θ_3 are identified. One indeterminacy is associated with β_2, ψ_1, ψ_2 and θ_1 and another indeterminacy is associated with ψ_4 and θ_4. The parameters β_2, ψ_1, ψ_2 and θ_1 are only determined by the three equations

$$\begin{aligned} \sigma_{11} &= \psi_1 + \theta_1, \\ \sigma_{21} &= \beta_2\psi_1, \\ \omega_2 &= \beta_2^2\psi_1 + \psi_2, \end{aligned}$$

where ψ_2 is identified. The parameters ψ_4 and θ_4 are only determined by the single equation

$$\sigma_{44} = \beta_4^2 \psi_3 + \psi_4 + \theta_4,$$

where ψ_3 is identified. The most natural way of eliminating the indeterminacies is to set $\theta_1 = \theta_2$ and $\theta_4 = \theta_3$, which makes sense if the *y*-variables are on the same scale. It is not necessary to assume that all error variances are equal, only that the error variances for the first and last variable are each equal to one other error variance. The assumption of equal error variances across all variables is in fact testable with *p* – 3 degrees of freedom.

In the general simplex model with *p* variables, there are 3*p* – 3 independent parameters and the degrees of freedom are $\frac{1}{2} p(p+1) - 3p + 3$. If *p* = 3, this is zero and the model is a tautology. For testing a simplex model, *p* must be at least 4.

The *quasi-simplex* model is a LISREL Submodel 3B with $\mathbf{\Lambda}_y = \mathbf{I}$, $\mathbf{\Theta}_\varepsilon$ diagonal, $\mathbf{\Psi}$ diagonal, and

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ \beta_2 & 0 & 0 & 0 \\ 0 & \beta_3 & 0 & 0 \\ 0 & 0 & \beta_4 & 0 \end{bmatrix}.$$

This specification automatically defines ζ_1 as η_1 so that $\psi_1 = \omega_1$.

The *perfect simplex* is obtained by setting $\mathbf{\Theta}_\varepsilon = \mathbf{0}$ (TE = ZE). This can be tested when $p \geq 3$. The perfect simplex implies that the *partial correlation* $\rho_{ik.j}$ is zero whenever *i* < *j* < *k*. Higher-order partial correlations, with two or more intermediate variables held constant, also vanish.

The data for this example (Humphreys (1968)) is shown in the table below as a correlation matrix. Note that, for standard errors and chi-squares to be correct, the covariance matrix should be analyzed.

Table: Correlations between grade point averages, high school rank, and an aptitude test

	y_0	y_0'	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8
y_0	1.000									
y_0'	0.393	1.000								
y_1	0.387	0.375	1.000							
y_2	0.341	0.298	0.556	1.000						
y_3	0.278	0.237	0.456	0.490	1.000					
y_4	0.270	0.255	0.439	0.445	0.562	1.000				
y_5	0.240	0.238	0.415	0.418	0.496	0.512	1.000			
y_6	0.256	0.252	0.399	0.383	0.456	0.469	0.551	1.000		
y_7	0.240	0.219	0.387	0.364	0.445	0.442	0.500	0.544	1.000	
y_8	0.222	0.173	0.342	0.339	0.354	0.416	0.453	0.482	0.541	1.000

The variables include eight semesters of grade-point averages, $y_1, y_2, ..., y_8$, high school rank y_0 and a composite score on the American College Testing test y_0' for approximately 1600 undergraduate students at the University of Illinois.

We shall first use the variables $y_1, y_2, ..., y_8$ and illustrate what happens when one runs a model which is not identified. We have made three runs with the same data and model. In Run 1 we specified the model as if all the parameters were identified. The command file for this run is as follows (**EX66A.LIS** in the **LISREL Examples** folder):

```
Simplex Model for Academic Performance Run 1
DA NI=10 NO=1600
LA
(10A3)
 Y0 Y0' Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8
KM FI=EX66.COR
SE
3 4 5 6 7 8 9 10 /
MO NY=8 NE=8 LY=ID BE=FU PS=DI
FR BE 2 1 BE 3 2 BE 4 3 BE 5 4 BE 6 5 BE 7 6 BE 8 7
PD
OU SS SE AD=OFF
```

In run 2 we imposed the condition that $\theta_1 = \theta_2$ to eliminate the first indeterminacy and in Run 3 we imposed the condition $\theta_8 = \theta_7$, in addition, to eliminate the second indeterminacy also. The results are shown in the table below.

Table: Results for simplex model

Parameter Number	Parameter	Run 1	Run 2	Run 3
1	β_2	0.53	0.98	0.98
2	β_3	0.84	0.84	0.84
3	β_4	0.96	0.96	0.96
4	β_5	0.91	0.91	0.91
5	β_6	0.93	0.93	0.93
6	β_7	0.94	0.94	0.94
7	β_8	0.89	0.89	0.89
8	ψ_1	2.05	0.57	0.57
9	ψ_2	0.28	0.03	0.03
10	ψ_3	0.17	0.17	0.17
11	ψ_4	0.03	0.03	0.03
12	ψ_5	0.12	0.12	0.12
13	ψ_6	0.07	0.07	0.07
14	ψ_7	0.10	0.10	0.10
15	ψ_8	0.61	0.61	0.13
16	θ_1	-0.05	0.43	0.43
17	θ_2	0.43	0.43	0.43
18	θ_3	0.43	0.43	0.43
19	θ_4	0.44	0.44	0.44
20	θ_5	0.42	0.42	0.42
21	θ_6	0.42	0.42	0.42
22	θ_7	0.39	0.39	0.39
23	θ_8	-0.09	-0.09	0.39
	$\psi_1 + \theta_1$	1.00	1.00	1.00
	$\beta_2\psi_1$	0.56	0.56	0.56
	$\psi_8 + \theta_8$	0.52	0.52	0.52

Run 1 gave the message that the parameter TE(1) may not be identified. TE(1) is θ_1, the last of the four parameters involved in the first indeterminacy. In Run 2 the corresponding message was that the parameters TE(8) may not be identified. TE(8) is θ_8, the last parameter involved in the second indeterminacy. In Run 3 no such message was given indicating that the model is identified. All three solutions have the same χ^2 = 23.91 and it is seen that all parameters that are identified come out with the same parameter estimate in all three runs. Only the non-identified parameters vary over the three solutions. The values given for the non-identified parameters are of course arbitrary to some extent. However, these values are such that the following three quantities are invariant over all solutions:

$$\hat{\psi}_1 + \hat{\theta}_1 \quad \hat{\beta}_2 \hat{\psi}_1 \quad \hat{\psi}_8 + \hat{\theta}_8$$

These runs illustrate that LISREL behaves reasonably for model which are non-identified and that the program correctly identifies the last parameter involved in an indeterminacy as a non-identified parameter.

The option SS on the OU command gives the standardized solution, i.e., the correlation matrix of $\boldsymbol{\eta}$. The intercorrelations among η_2, η_3, …, η_7 are the same for all three solutions shown in the table below.

Table: Intercorrelations of a perfect simplex

	η_2	η_3	η_4	η_5	η_6	η_7
η_2	1.000					
η_3	0.838	1.000				
η_4	0.812	0.969	1.000			
η_5	0.724	0.865	0.892	1.000		
η_6	0.677	0.809	0.834	0.935	1.000	
η_7	0.619	0.739	0.763	0.855	0.914	1.000

Here every correlation ρ_{ij} with $|i-j|>1$ is the product of the correlations just below the diagonal. For example, $\rho(\eta_5,\eta_2)$ = 0.838 × 0.969 × 0.892 = 0.724. These correlations form a perfect simplex. The reliabilities of the semester grades y_2, y_3, …, y_7 can be obtained directly from the solution in which the η's are standardized. The reliabilities are:

y_2	y_3	y_4	y_5	y_6	y_7
0.569	0.575	0.563	0.584	0.581	0.608

A test of the hypothesis that all reliabilities are equal gives χ^2 = 2.17 with five degrees of freedom, so that this hypothesis is not rejected by the data despite the large sample size.

Without identification conditions imposed, as in Run 1, the correlations $\rho(\eta_1,\eta_j),\ j \neq 1$ and $\rho(\eta_i,\eta_8),\ i \neq 8$, and the reliabilities of y_1 and y_8 are not identified. However, in view of the above test of equality of reliabilities it seems reasonable to assume that all reliabilities or equivalently all error variances in the standardized solution are equal for y_1 through y_8. This assumption makes it possible to estimate the intercorrelations among all the η's.

Assuming that y_0 and $y_0^{'}$ are indicators of pre-college academic achievement η_0 which is assumed to influence the true academic achievement in the first semester η_1, one can estimate again the quasi-Markov simplex and show how this use of y_0 and $y_0^{'}$ helps identify the parameters of the model. The only parameters which are now not identified are ψ_8 and θ_8. This gives a χ^2 = 36.92 with 28 degrees of freedom. If we assume that the reliabilities of all the semester grades are equal, all parameters are identified and the goodness of fit becomes 45.22 with 34 degrees of freedom. The difference 8.30 with 6 degrees of freedom provides another test of equality of the reliabilities. Given that all error variances are equal, a test of the hypothesis that

$$\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_8$$

gives χ^2 = 6.48 with seven degrees of freedom so that this hypothesis cannot be rejected. The command file for the last run is (**EX66D.LIS**):

```
Simplex Model for Academic Performance Last model
DA NI=10 NO=1600
LA
(10A3)
 Y0Y0' Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8
KM FI=EX66.COR SY
MO NY=10 NE=9 BE=FU
FR LY 2 1 BE 2 1 BE 3 2 BE 4 3 BE 5 4 BE 6 5 BE 7 6 BE 8 7 BE 9 8
VA 1 LY 1 1 LY 3 2 LY 4 3 LY 5 4 LY 6 5 LY 7 6 LY 8 7 LY 9 8 LY 10 9
EQ TE 3 - TE 10
EQ BE 2 1 BE 3 2 BE 4 3 BE 5 4 BE 6 5 BE 7 6 BE 8 7 BE 9 8
ST .5 ALL
OU SS NS
```

Further analysis shows that the variances ψ_1, ψ_2, …, ψ_8 of the random disturbance terms are not equal, so the whole autoregressive process is not completely stationary.