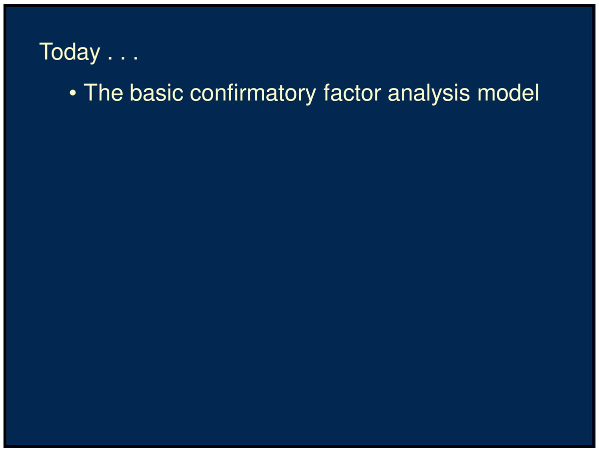
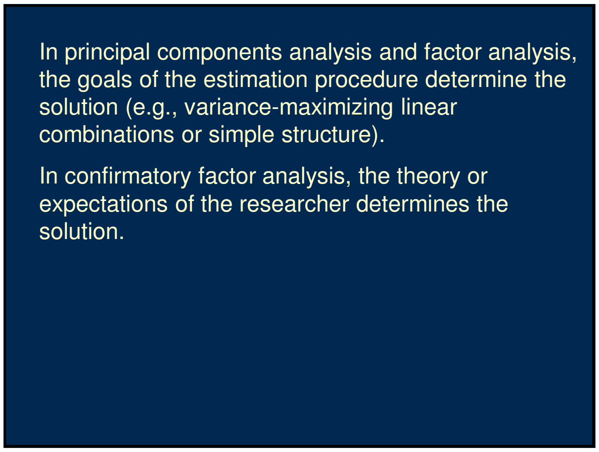
Confirmatory Factor Analysis



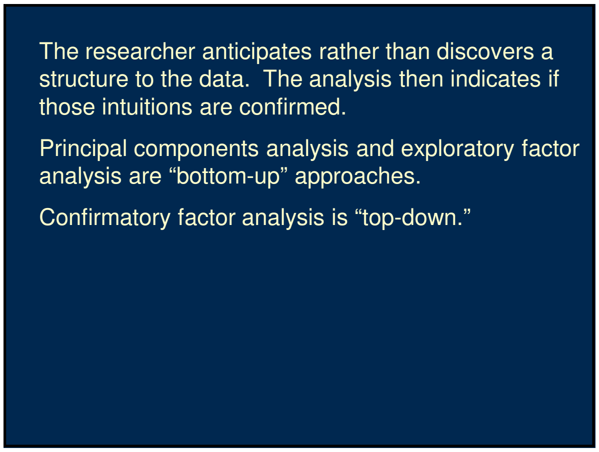
Today . . .

- The basic confirmatory factor analysis model



In principal components analysis and factor analysis, the goals of the estimation procedure determine the solution (e.g., variance-maximizing linear combinations or simple structure).

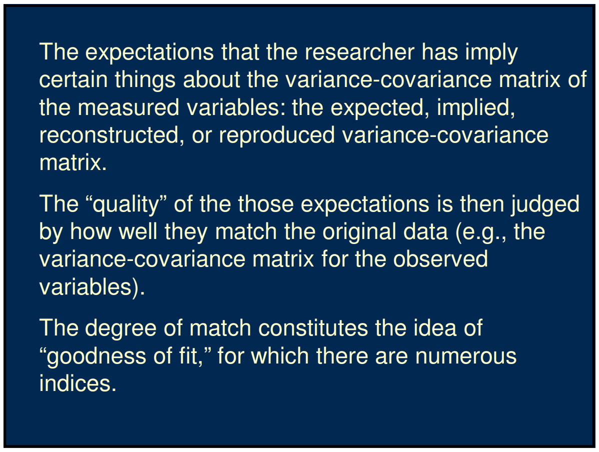
In confirmatory factor analysis, the theory or expectations of the researcher determines the solution.



The researcher anticipates rather than discovers a structure to the data. The analysis then indicates if those intuitions are confirmed.

Principal components analysis and exploratory factor analysis are “bottom-up” approaches.

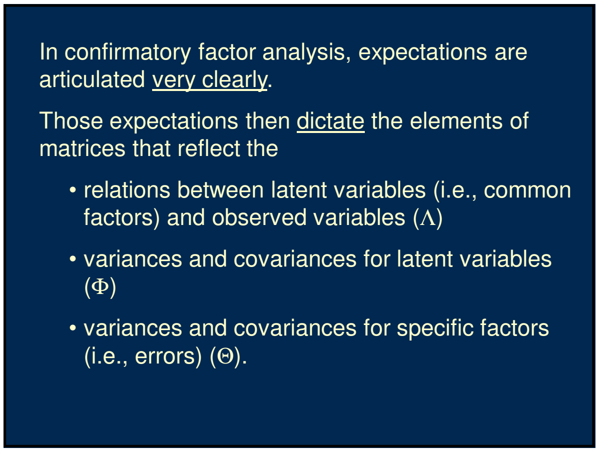
Confirmatory factor analysis is “top-down.”



The expectations that the researcher has imply certain things about the variance-covariance matrix of the measured variables: the expected, implied, reconstructed, or reproduced variance-covariance matrix.

The “quality” of the those expectations is then judged by how well they match the original data (e.g., the variance-covariance matrix for the observed variables).

The degree of match constitutes the idea of “goodness of fit,” for which there are numerous indices.



In confirmatory factor analysis, expectations are articulated <u>very clearly</u>.

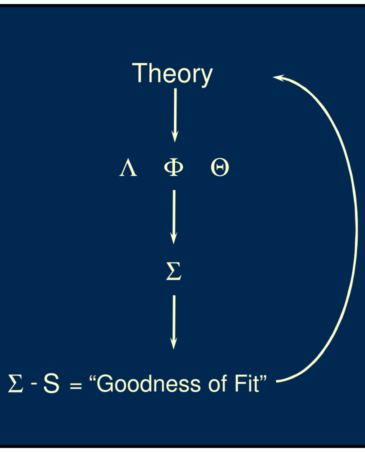
Those expectations then <u>dictate</u> the elements of matrices that reflect the

- relations between latent variables (i.e., common factors) and observed variables (Λ)
- variances and covariances for latent variables (Φ)
- variances and covariances for specific factors (i.e., errors) (Θ).

Some elements of these matrices are dictated to have certain values (e.g., 0) and others are free to be estimated.

Once the matrices have been estimated, they can be used to "reconstruct" the variance-covariance matrix for the original variables.

The similarity of this reconstructed matrix (Σ) to the actual variance-covariance matrix (S) indicates the quality of the solution and by implication, the quality of the original guiding expectations.



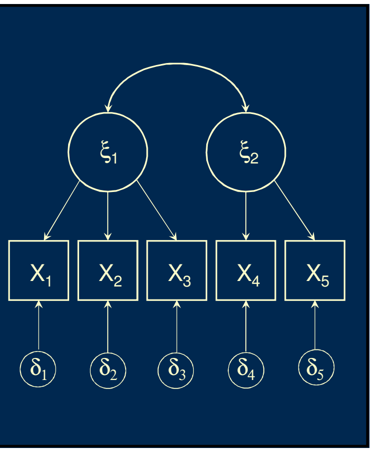
Theory

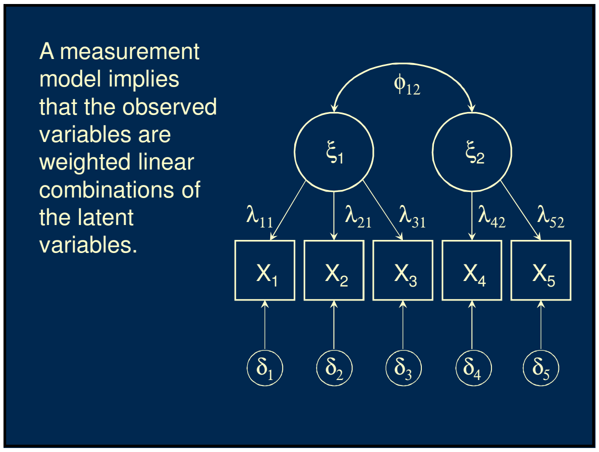
$\Lambda \quad \Phi \quad \Theta$

Σ

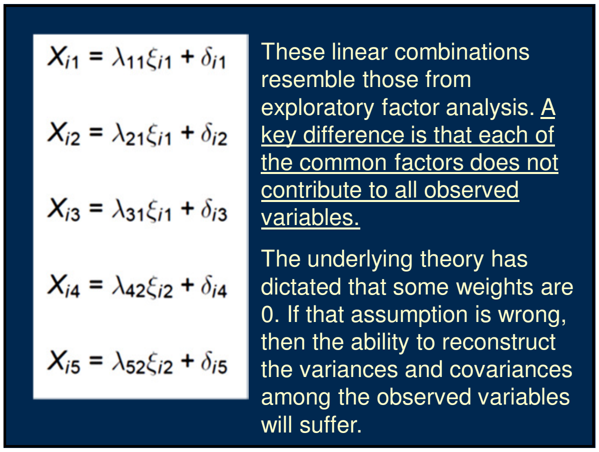
Σ - S = "Goodness of Fit"

Expectations can be represented in a measurement model that indicates how latent variables are related to observed variables and to other latent variables.





A measurement model implies that the observed variables are weighted linear combinations of the latent variables.



$$X_{i1} = \lambda_{11}\xi_{i1} + \delta_{i1}$$

$$X_{i2} = \lambda_{21}\xi_{i1} + \delta_{i2}$$

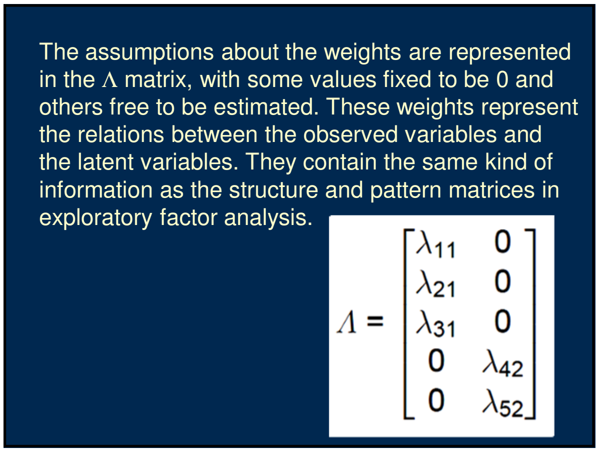
$$X_{i3} = \lambda_{31}\xi_{i1} + \delta_{i3}$$

$$X_{i4} = \lambda_{42}\xi_{i2} + \delta_{i4}$$

$$X_{i5} = \lambda_{52}\xi_{i2} + \delta_{i5}$$

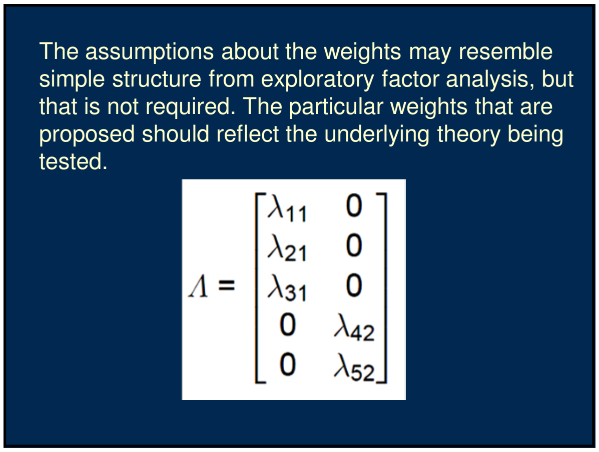
These linear combinations resemble those from exploratory factor analysis. A key difference is that each of the common factors does not contribute to all observed variables.

The underlying theory has dictated that some weights are 0. If that assumption is wrong, then the ability to reconstruct the variances and covariances among the observed variables will suffer.



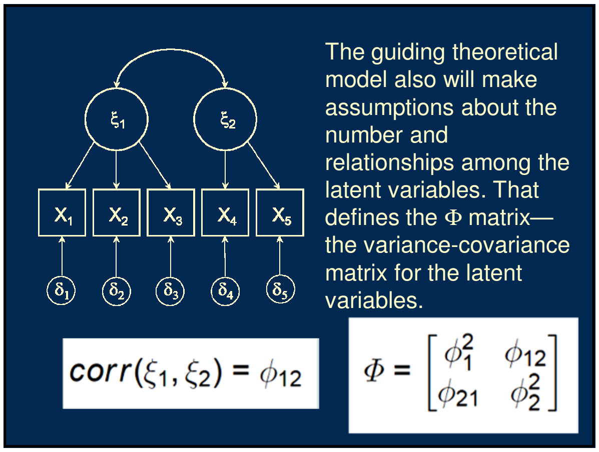
The assumptions about the weights are represented in the Λ matrix, with some values fixed to be 0 and others free to be estimated. These weights represent the relations between the observed variables and the latent variables. They contain the same kind of information as the structure and pattern matrices in exploratory factor analysis.

$$\Lambda = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \end{bmatrix}$$



The assumptions about the weights may resemble simple structure from exploratory factor analysis, but that is not required. The particular weights that are proposed should reflect the underlying theory being tested.

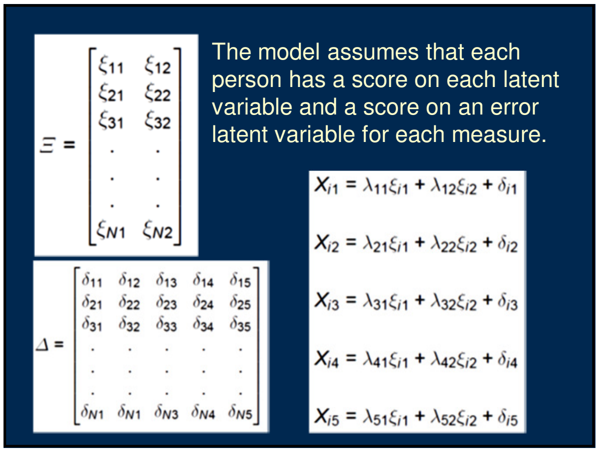
$$\Lambda = \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \end{bmatrix}$$



The guiding theoretical model also will make assumptions about the number and relationships among the latent variables. That defines the Φ matrix—the variance-covariance matrix for the latent variables.

$$corr(\xi_1, \xi_2) = \phi_{12}$$

$$\Phi = \begin{bmatrix} \phi_1^2 & \phi_{12} \\ \phi_{21} & \phi_2^2 \end{bmatrix}$$



The model assumes that each person has a score on each latent variable and a score on an error latent variable for each measure.

$$\Xi = \begin{bmatrix} \xi_{11} & \xi_{12} \\ \xi_{21} & \xi_{22} \\ \xi_{31} & \xi_{32} \\ . & . \\ . & . \\ . & . \\ \xi_{N1} & \xi_{N2} \end{bmatrix}$$

$$\Delta = \begin{bmatrix} \delta_{11} & \delta_{12} & \delta_{13} & \delta_{14} & \delta_{15} \\ \delta_{21} & \delta_{22} & \delta_{23} & \delta_{24} & \delta_{25} \\ \delta_{31} & \delta_{32} & \delta_{33} & \delta_{34} & \delta_{35} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \delta_{N1} & \delta_{N1} & \delta_{N3} & \delta_{N4} & \delta_{N5} \end{bmatrix}$$

$$X_{i1} = \lambda_{11}\xi_{i1} + \lambda_{12}\xi_{i2} + \delta_{i1}$$

$$X_{i2} = \lambda_{21}\xi_{i1} + \lambda_{22}\xi_{i2} + \delta_{i2}$$

$$X_{i3} = \lambda_{31}\xi_{i1} + \lambda_{32}\xi_{i2} + \delta_{i3}$$

$$X_{i4} = \lambda_{41}\xi_{i1} + \lambda_{42}\xi_{i2} + \delta_{i4}$$

$$X_{i5} = \lambda_{51}\xi_{i1} + \lambda_{52}\xi_{i2} + \delta_{i5}$$

Because of expectations about the weights, the linear combinations can be simplified.

$$X_{i1} = \lambda_{11}\xi_{i1} + 0\xi_{i2} + \delta_{i1}$$

$$X_{i2} = \lambda_{21}\xi_{i1} + 0\xi_{i2} + \delta_{i2}$$

$$X_{i3} = \lambda_{31}\xi_{i1} + 0\xi_{i2} + \delta_{i3}$$

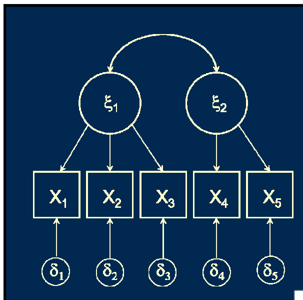
$$X_{i4} = 0\xi_{i1} + \lambda_{42}\xi_{i2} + \delta_{i4}$$

$$X_{i5} = 0\xi_{i1} + \lambda_{52}\xi_{i2} + \delta_{i5}$$

$$\Theta_\delta = \begin{bmatrix} \theta_{11}^2 & \theta_{12} & \theta_{13} & \theta_{14} & \theta_{15} \\ \theta_{21} & \theta_{22}^2 & \theta_{23} & \theta_{24} & \theta_{25} \\ \theta_{31} & \theta_{32} & \theta_{33}^2 & \theta_{34} & \theta_{35} \\ \theta_{41} & \theta_{42} & \theta_{43} & \theta_{44}^2 & \theta_{45} \\ \theta_{51} & \theta_{52} & \theta_{53} & \theta_{54} & \theta_{55}^2 \end{bmatrix}$$

The model also makes assumptions about the variances and covariances of the error latent variables. These are contained in the Θ_δ matrix.

When the errors are assumed to be uncorrelated, the off-diagonal elements of this matrix will be zero. This is an often made, but often unrealistic assumption. Confirmatory factor models allow correlated errors to be explicitly modeled.



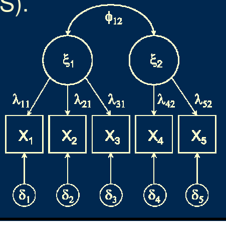
This model implies uncorrelated errors.

Why might correlated errors occur?

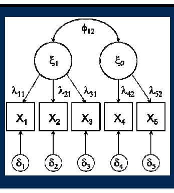
$$\Theta_\delta = \begin{bmatrix} \theta_{11}^2 & 0 & 0 & 0 & 0 \\ 0 & \theta_{22}^2 & 0 & 0 & 0 \\ 0 & 0 & \theta_{33}^2 & 0 & 0 \\ 0 & 0 & 0 & \theta_{44}^2 & 0 \\ 0 & 0 & 0 & 0 & \theta_{55}^2 \end{bmatrix}$$

The covariances (and variances) among the observed variables can be estimated from the latent variable parameters.

If the model-implied parameters are “correct”, then the reproduced variances and covariances (Σ) should be close to the observed variances and covariances (S).



The model implies that the observed variables are linear combinations.



$$\sigma^2_{X_1X_1} = \frac{\sum_{i=1}^{N} X_{i1}X_{i1}}{N-1}$$

$$\sigma^2_{X_1X_1} = \frac{\sum_{i=1}^{N}(\lambda_{11}\xi_{i1} + \delta_{i1})(\lambda_{11}\xi_{i1} + \delta_{i1})}{N-1}$$

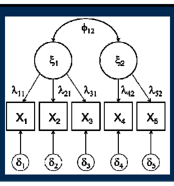
$$\sigma^2_{X_1X_1} = \frac{\sum_{i=1}^{N}(\lambda_{11}\xi_{i1}\lambda_{11}\xi_{i1})}{N-1} + \frac{\sum_{i=1}^{N}(\lambda_{11}\xi_{i1}\delta_{i1})}{N-1} + \frac{\sum_{i=1}^{N}(\delta_{i1}\lambda_{11}\xi_{i1})}{N-1} + \frac{\sum_{i=1}^{N}(\delta_{i1}\delta_{i1})}{N-1}$$

$$\sigma^2_{X_1X_1} = \frac{\lambda_{11}\lambda_{11}\sum_{i=1}^{N}(\xi_{i1}\xi_{i1})}{N-1} + \frac{\lambda_{11}\sum_{i=1}^{N}(\xi_{i1}\delta_{i1})}{N-1} + \frac{\lambda_{11}\sum_{i=1}^{N}(\delta_{i1}\xi_{i1})}{N-1} + \frac{\sum_{i=1}^{N}(\delta_{i1}\delta_{i1})}{N-1}$$

$$\sigma^2_{X_1X_1} = \lambda_{11}\lambda_{11}\sigma_{\xi_1\xi_1} + \lambda_{11}\sigma_{\xi_1\delta_1} + \lambda_{11}\sigma_{\delta_1\xi_1} + \sigma_{\delta_1\delta_1}$$

$$\sigma^2_{X_1X_1} = \lambda_{11}\lambda_{21}\sigma^2_{\xi_1} + 0 + 0 + \sigma^2_{\delta_1}$$

$$\sigma^2_{X_1X_1} = \lambda^2_{11} + \sigma^2_{\delta_1}$$



$$\sigma_{X_1X_2} = \frac{\sum_{i=1}^{N} X_{i1}X_{i2}}{N-1}$$

$$\sigma_{X_1X_2} = \frac{\sum_{i=1}^{N}(\lambda_{11}\xi_{i1} + \delta_{i1})(\lambda_{21}\xi_{i1} + \delta_{i2})}{N-1}$$

$$\sigma_{X_1X_2} = \frac{\sum_{i=1}^{N}(\lambda_{11}\xi_{i1}\lambda_{21}\xi_{i1})}{N-1} + \frac{\sum_{i=1}^{N}(\lambda_{11}\xi_{i1}\delta_{i2})}{N-1} + \frac{\sum_{i=1}^{N}(\delta_{i1}\lambda_{21}\xi_{i1})}{N-1} + \frac{\sum_{i=1}^{N}(\delta_{i1}\delta_{i2})}{N-1}$$

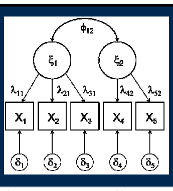
$$\sigma_{X_1X_2} = \frac{\lambda_{11}\lambda_{21}\sum_{i=1}^{N}(\xi_{i1}\xi_{i1})}{N-1} + \frac{\lambda_{11}\sum_{i=1}^{N}(\xi_{i1}\delta_{i2})}{N-1} + \frac{\lambda_{21}\sum_{i=1}^{N}(\delta_{i1}\xi_{i1})}{N-1} + \frac{\sum_{i=1}^{N}(\delta_{i1}\delta_{i2})}{N-1}$$

$$\sigma_{X_1X_2} = \lambda_{11}\lambda_{21}\sigma_{\xi_1\xi_1} + \lambda_{11}\sigma_{\xi_1\delta_2} + \lambda_{21}\sigma_{\delta_1\xi_1} + \sigma_{\delta_1\delta_2}$$

$$\sigma_{X_1X_2} = \lambda_{11}\lambda_{21}\sigma_{\xi_1\xi_1} + 0 + 0 + 0$$

$$\sigma_{X_1X_2} = \lambda_{11}\lambda_{21}$$

$$\sigma_{X_1X_4} = \frac{\sum_{i=1}^{N} X_{i1}X_{i4}}{N-1}$$



$$\sigma_{X_1X_2} = \frac{\sum_{i=1}^{N}(\lambda_{11}\xi_{i1} + \delta_{i1})(\lambda_{42}\xi_{i2} + \delta_{i4})}{N-1}$$

$$\sigma_{X_1X_2} = \frac{\sum_{i=1}^{N}(\lambda_{11}\xi_{i1}\lambda_{42}\xi_{i2})}{N-1} + \frac{\sum_{i=1}^{N}(\lambda_{11}\xi_{i1}\delta_{i4})}{N-1} + \frac{\sum_{i=1}^{N}(\delta_{i1}\lambda_{42}\xi_{i2})}{N-1} + \frac{\sum_{i=1}^{N}(\delta_{i1}\delta_{i4})}{N-1}$$

$$\sigma_{X_1X_2} = \frac{\lambda_{11}\lambda_{42}\sum_{i=1}^{N}(\xi_{i1}\xi_{i2})}{N-1} + \frac{\lambda_{11}\sum_{i=1}^{N}(\xi_{i1}\delta_{i4})}{N-1} + \frac{\lambda_{42}\sum_{i=1}^{N}(\delta_{i1}\xi_{i2})}{N-1} + \frac{\sum_{i=1}^{N}(\delta_{i1}\delta_{i4})}{N-1}$$

$$\sigma_{X_1X_2} = \lambda_{11}\lambda_{42}\sigma_{\xi_1\xi_2} + \lambda_{11}\sigma_{\xi_1\delta_4} + \lambda_{42}\sigma_{\delta_1\xi_2} + \sigma_{\delta_1\delta_4}$$

$$\sigma_{X_1X_2} = \lambda_{11}\lambda_{42}\sigma_{\xi_1\xi_2} + 0 + 0 + 0$$

$$\sigma_{X_1X_4} = \lambda_{11}\lambda_{42}\phi_{12}$$

More generally:

$$X = \Xi\Lambda' + \Delta$$

$$\Sigma = \Lambda\Phi\Lambda' + \Theta$$

Reminder: X is a matrix of linear combinations of latent variables and so its variance-covariance matrix can be obtainable from the variance-covariance matrix for the latent variables along with the weights for creating the linear combinations.

The estimation procedure most commonly used in confirmatory factor analysis is ***maximum likelihood estimation***.

This approach seeks the parameter estimates that maximize the probability of the data that were actually obtained.

In the context of confirmatory factor analysis, maximum likelihood estimates represent model parameters that make the obtained data most likely, within the constraints imposed by the model.

All of the model assumptions are contained in the reproduced variance-covariance matrix, Σ. The probability density function, assuming multivariate normality, can be used to estimate the relative likelihood of a score profile given Σ:

$$X_i' = (X_{i1}, X_{i2}, X_{i3}, \ldots, X_{ik})$$

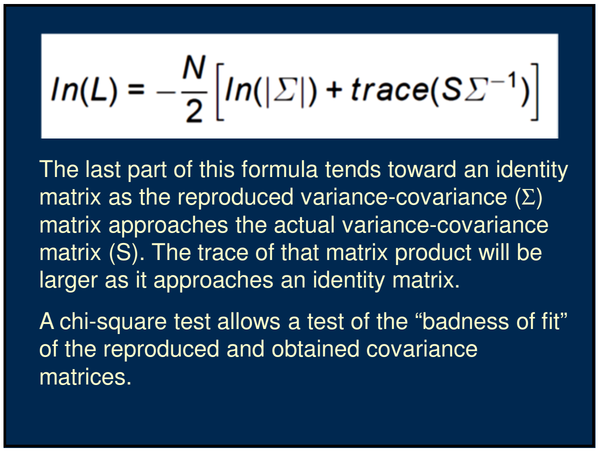
$$p(X_i', \Sigma) \propto \frac{1}{\sqrt{2\pi|\sum|}} exp\left(-\frac{1}{2}X_i'\Sigma^{-1}X_i\right)$$

The likelihood function that is maximized is thus:

$$Likelihood(L) = \prod_{i=1}^{N}\left[\frac{1}{\sqrt{2\pi|\sum|}} exp\left(-\frac{1}{2}X_i'\Sigma^{-1}X_i\right)\right]$$

It is more convenient to work with the log of the likelihood function, and, the function can be simplified:

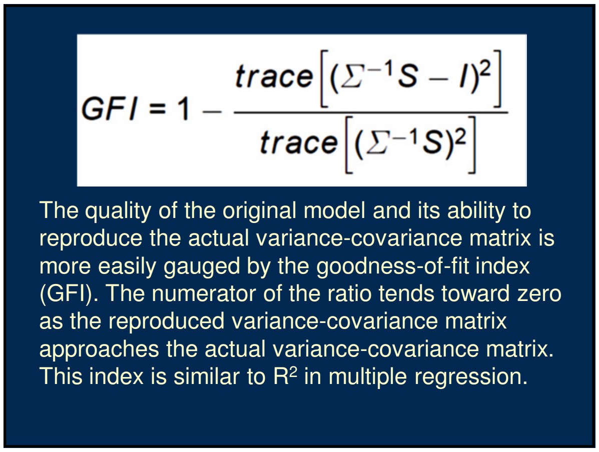
$$ln(L) = -\frac{N}{2}\left[ln(|\Sigma|) + trace(S\Sigma^{-1})\right]$$



$$In(L) = -\frac{N}{2}\left[In(|\Sigma|) + trace(S\Sigma^{-1})\right]$$

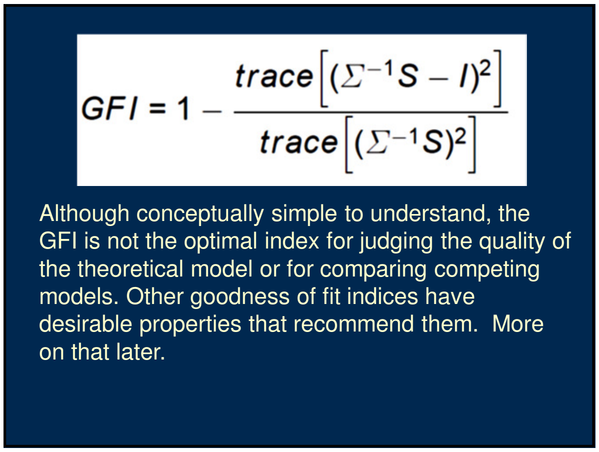
The last part of this formula tends toward an identity matrix as the reproduced variance-covariance (Σ) matrix approaches the actual variance-covariance matrix (S). The trace of that matrix product will be larger as it approaches an identity matrix.

A chi-square test allows a test of the “badness of fit” of the reproduced and obtained covariance matrices.



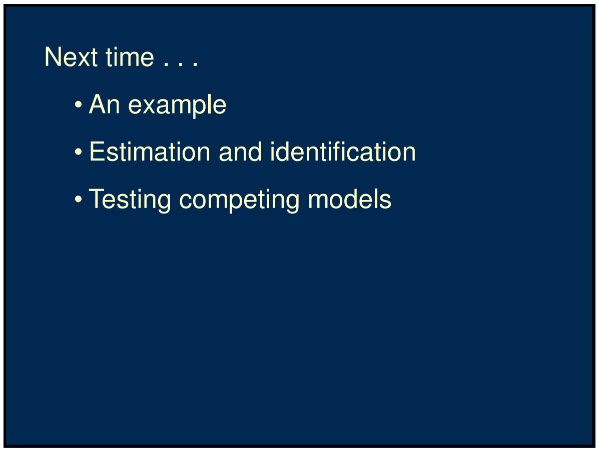
$$GFI = 1 - \frac{trace\left[(\Sigma^{-1}S - I)^2\right]}{trace\left[(\Sigma^{-1}S)^2\right]}$$

The quality of the original model and its ability to reproduce the actual variance-covariance matrix is more easily gauged by the goodness-of-fit index (GFI). The numerator of the ratio tends toward zero as the reproduced variance-covariance matrix approaches the actual variance-covariance matrix. This index is similar to R^2 in multiple regression.



$$GFI = 1 - \frac{trace\left[(\Sigma^{-1}S - I)^2\right]}{trace\left[(\Sigma^{-1}S)^2\right]}$$

Although conceptually simple to understand, the GFI is not the optimal index for judging the quality of the theoretical model or for comparing competing models. Other goodness of fit indices have desirable properties that recommend them. More on that later.



Next time . . .

- An example
- Estimation and identification
- Testing competing models