Maximum Likelihood Estimation

Overidentified models have more known values (or equations) than unknown parameters, and thus, no one correct solution, but the best estimate of the values can be obtained through an iterative process.[1] Even though I used correlations to illustrate path analysis, in practice, correlations are not typically used to derive path coefficients or factor loadings. The primary reason we do not typically use correlations to derive path estimates in practice is that correlations use standardized variables and important information about the variances of the variables is lost (i.e., the variances of each of the variables is assumed to be 1). It can also be shown that for many different factor models, use of correlations can lead to erroneous results (Cudeck, 1989). Thus, covariances, the unstandardized version of correlations, are used.

The logic of deriving estimates of the loadings remains the same. There are a set of equations that describe the model (i.e., *structural equations*) and some known values about the relations among all of the variables (i.e., a matrix of covariances). For complicated models, the most common approach to solving the set of structural equations is through a calculus-based method called *maximum likelihood* (ML). Maximum likelihood solves for the loadings by minimizing the discrepancy between the equations implied by the model and the obtained covariances.[2] This discrepancy is mathematically described as:

$$\mathbf{S} - \hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$$

Where $\mathbf{S}$ is the covariance matrix obtained from the data, and $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$ is matrix notation for a covariance matrix implied by the hypothesized model. For the measurement model, the implied covariance matrix is:

$$\boldsymbol{\Lambda}_x \boldsymbol{\Phi} \boldsymbol{\Lambda}_x' + \boldsymbol{\Theta}_\delta$$

where $\boldsymbol{\Lambda}_x$ is the matrix of factor loadings linking observed variables x to the factor, $\boldsymbol{\Phi}$ is the matrix of the variances and covariances of the factors, and $\boldsymbol{\Theta}_\delta$ is the matrix of measurement residuals.

Certain values for the relations among the variables are implied by certain specified models. For example, a typical confirmatory factor model assumes the measurement residuals are uncorrelated (all off diagonal elements of Θ_δ are zero) and, if there is more than one factor, items load on only one factor (all cross loadings in the Λ_x matrix are zero). We can examine the fit of the hypothesized model to the data, by comparing the implied (or "reproduced") covariances to those obtained. The ML solution is obtained by minimizing the following (somewhat frightening) *fit function*:

$$F_{\text{ML}} = \log\left|\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})\right| + tr\left(\mathbf{S}\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\theta})\right) - \log|\mathbf{S}| - (p+q)$$

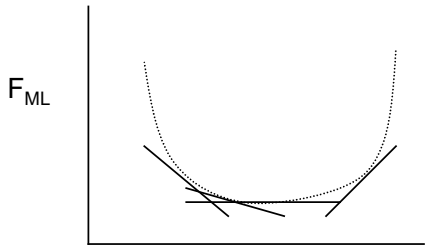
$\log$ is the natural logarithm function (base e), $\hat{\boldsymbol{\Sigma}}(\boldsymbol{\theta})$ is the covariance matrix implied by the model, $\mathbf{S}$ is the observed covariance matrix, tr is the trace matrix algebra function which sums diagonal elements, and $(p + q)$ is equal to the number of coefficients that need to be estimated in the model. The superscript in the middle, $^{-1}$, is the inverse matrix function, so the inverse of the implied matrix must be possible.[3]
The maximum likelihood estimator (Fisher, 1950) is *asymptotically unbiased*, meaning that in larger samples it is an unbiased estimation of the population value. It is an efficient estimator, so it provides a variance estimate that is smaller than other consistent estimation methods. ML is asymptotically normal, so it allows for convenient statistical tests like the Wald ratio test. With simple structural equation models, such as a multiple regression, and normal distribution assumptions, the maximum likelihood estimate will provide

[1] In mathematics, the system said to be *overdetermined*.
[2] The word "maximum" in "maximum likelihood" is used because the process maximizes the joint probability density function for the function of the parameters being estimated. The more general explanation of the maximum likelihood process is to find largest value among many natural log values (which are negative), and thus obtaining the parameter value(s) that are the most likely (i.e., has/have the maximum likelihood) for the observed values (see Enders, 2022, Chapter 2, for example). In SEM, this occurs as a minimization of the fit function instead (see Bollen, 1989, Appendices 4a and 4c; Ferron & Hess, 2007).
[3] Computer packages sometimes print an error message stating that the "inverse of sigma is not positive definite." The message is in reference to the assumption that the variance-covariance matrix must be positive definite—the inversion must be possible. This indicates a severe problem with the model because one or more of the implied variances from the variance/covariance matrix is negative.

identical values to the OLS regression. ML, however, is a more general method that can be used with much more complicated models.

Fit, Chi-square, and *df*

ML is an iterative process, so initial starting values (i.e., guesses) are generated by the computer, the discrepancy between the implied and the obtained covariance matrix is computed, then new guesses are entered, and so on, until the minimum possible discrepancy value is obtained. Each step is called an *iteration*. The idea is similar to the idea of ordinary least squares (OLS) in regression in which the squared errors or residuals are minimized to obtain the best fit of the regression line to the data and the regression coefficients.



To find the minimum value of the F_{ML}, the maximum likelihood fit function, derivatives (i.e., first- and second-order partial derivatives with respect to variance-covariance values) are used to draw tangent lines that correspond a point on the curve and confirm the minimum point. When the tangent line has a slope of zero, the minimized value of the function has been found. The computer stops and generates values for the fit of the overall model and the parameter values.[4] The final value can be used in a chi-square test [$\chi^2 = (N\text{-}1)F_{\text{ML}}$]. If the fit is perfect, there will be no discrepancy between the implied and obtained covariances, and the chi-square will be zero. A chi-square nonsignificantly different from zero indicates a good fit. Significantly positive chi-squares indicate poor fit.

Positive degrees of freedom do not guarantee all aspects of the model are identified and a solution can be found, but $df > 0$ is required for identification of the model (see Kenny & Milan, 2012, for an overview of identification). It should be noted that most textbooks give the following formula (or a variation using $p + q$ to distinguish paths between exogenous and endogenous variables from those between endogenous paths).

$$df = \frac{v(v+1)}{2} - p$$

This formula is used because the number of unique variance/covariance elements (including the diagonal) is $v(v+1)$. If the number of variance/covariance elements are counted, however, then one must include the number of variances in the model when determining the number of estimated parameters, p. Both methods of counting degrees of freedom generally lead to the same result.[5]

References and Further Reading

Bollen, K.A. (1989). Structural Equations with Latent Variables. New York: John Wiley. ISBN: 0-4710-1171-1.
Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 2,* 317-327.
Eliason, S.R. (1993). *Maximum likelihood estimation: Logic and practice.* Newbury Park, CA: Sage. QASS #96.
Enders, C.K. (2022). *Applied missing data analysis, second edition.* Guilford.
Ferron, J. M., & Hess, M. R. (2007). Estimation in SEM: A concrete example. *Journal of Educational and Behavioral Statistics, 32*(1), 110-120.
Fisher, R. A. (1950). *Contributions to mathematical statistics*. New York: Wiley.
Kenny, D. A., & Milan, S. (2012). Identification: A non-technical discussion of a technical issue. *Handbook of structural equation modeling (pp*. 145-163). Guilford
Mulaik, S.A. (2009). *Linear causal modeling with structural relations.* Boca Raton, FL: Chapman & Hall/CRC.
Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology, 47*(1), 90-100.

[4] The actual process by which this minimization occurs may vary between SEM software packages and depends on whether a model with missing data is estimated and the variable types. The simplest algorithm is the Newton-Raphson. Mplus uses a quasi-Newton, but other packages use other algorithms (LISREL uses Fletcher-Powell and EQS uses a modified Gauss-Newton algorithm). Parameter estimates and standard errors tend to vary minimally across packages, however.

[5] It can get bit trickier in some situations. If for example, special constraints are placed on variances in a model, then the correlation, $v(v\text{-}1)$ method, could be incorrect, because variances are not included in the count of parameters.