MORE ON EM FOR ML FACTOR ANALYSIS

DONALD B. RUBIN
UNIVERSITY OF CHICAGO

DOROTHY T. THAYER
EDUCATIONAL TESTING SERVICE

We address several issues that are raised by Bentler and Tanaka's [1983] discussion of Rubin and Thayer [1982]. Our conclusions are: standard methods do not completely monitor the possible existence of multiple local maxima; summarizing inferential precision by the standard output based on second derivatives of the log likelihood at a maximum can be inappropriate, even if there exists a unique local maximum; EM and LISREL can be viewed as complementary, albeit not entirely adequate, tools for factor analysis.

Key words: EM, LISREL, maximum likelihood, factor analysis, algorithms, precision of estimation.

We feel that Bentler and Tanaka's [1983] discussion of Rubin and Thayer [1982] raises three important questions:

(i) Is there only one local maximum to the likelihood for the Jöreskog [1969] model and data?
(ii) Assuming there is only one local maximum, is it appropriate to use standard output based on second derivatives of the log likelihood to measure inferential precision?
(iii) Is EM a good algorithm to use for ML factor analysis?

We address these questions in order.

Is Jöreskog's Original Solution the Only Local Maximum and How Do We Know?

Label the parameter estimates in the three Rubin and Thayer solutions as \( \theta_1, \theta_2, \) and \( \theta_3 \), where \( \theta_3 \) is the original Jöreskog solution. These values were obtained by using three different starting values and letting EM run through 50 iterations. We stopped at 50 iterations because the 1975 version of Jöreskog's program then available to us (ACOVSF) also converged to \( \theta_1, \theta_2, \) and \( \theta_3 \), as we mentioned in our article on page 74, and thus EM and this other algorithm were in agreement.

Now let \( \theta_1, \theta_2, \) and \( \theta_3 \) represent the same solutions except rotated to have a zero factor loading to agree with the Bentler and Tanaka restriction. Using highly accurate computations of first and second derivatives at the solutions, or employing EM with very strict convergence criteria, it evidently can be shown that, although \( \theta_1, \theta_2, \) and \( \theta_3 \) are all points where the tangent plane to the likelihood is nearly horizontal, neither \( \theta_1 \) nor \( \theta_2 \) is a true local maximum. Our reporting of \( \theta_1 \) and \( \theta_2 \) as local maxima cannot be blamed on either EM or Jöreskog's original program, but rather our willingness to accept convergence criteria used in ASCOVF.

This work was partially supported by the Program Statistics Research Project at Educational Testing Service.

Reprint requests should be addressed to Donald B. Rubin, Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, IL. 60637.

0033-3123/83/06000-505$00.75/0
© 1983 The Psychometric Society
In any case, is \( \theta_3 \) the only local maximum? Bentler and Tanaka believe so. If \( \theta_3 \) is the only local maximum, then there exists a monotonely increasing path from every \( \theta \) in the parameter space to \( \theta_3 \), in particular from \( \theta_1 \) to \( \theta_3 \) and from \( \theta_2 \) to \( \theta_3 \). The fact that programs like LISREL-V and EQS move from starting points \( \theta_1 \) and \( \theta_2 \) to \( \theta_3 \) provides no direct evidence that there exists a monotone increasing path from \( \theta_1 \) to \( \theta_3 \) or from \( \theta_2 \) to \( \theta_3 \), since these algorithms can jump valleys in the likelihood. In contrast, EM moves only along monotone increasing paths in problems like this where the expectation of the complete-data log likelihood is convex. Consequently, Schoenberg's experiment with EM reported in Bentler and Tanaka's Footnote 3 shows that there exists an increasing path from \( \theta_2 \) to \( \theta_3 \) since EM moves from \( \theta_2 \) to \( \theta_3 \). Judging from the one slightly positive eigenvalue in the second derivative matrix at \( \theta_2 \), apparently this path begins as a one-dimensional nearly horizontal ridge in the 35-dimensional parameter space.

Does there exist a similar monotone path from \( \theta_1 \) to \( \theta_3 \)? If such a path cannot be found, this suggests that there may in fact be another local maximum besides \( \theta_3 \). Schoenberg's experiment using EM finds a monotonely increasing path from \( \theta_1 \) to another nearly horizontal place in the likelihood (with value .0170) but not to \( \theta_3 \). Thus, we do not yet know whether \( \theta_3 \) is the only local maximum.

Bentler and Tanaka claim that multiple local maxima in ML factor analyses are rare. But how do they know? It is difficult to examine 35-dimensional space, and maximization algorithms like LISREL-V and EQS are not designed to investigate the question of multiple local maxima. We are not aware of any real evidence addressing the frequency of multiple local maxima in ML factor analyses with actual data.

Is There A Simple Measure of Inferential Precision with These Data?

Now suppose that there is only one local maximum for this problem, the original Jöreskog solution, \( \theta_3 \). Even then, the existence of points \( \theta_1 \) and \( \theta_2 \) where the likelihood is nearly horizontal can be exceedingly important inferentially, because they indicate that the likelihood function may not be well approximated by a normal distribution (i.e., the log-likelihood may not be essentially quadratic). In such cases, the MLE may not be a good estimate because of asymmetries in the likelihood function, and even when the MLE is a good estimate, standard statistical inferential procedures can be misleading since they rely on the approximate normality of the likelihood function. In contrast to our position, Bentler and Tanaka seem to believe that the existence of a unique local maximum is all that is needed to insure the optimality of the MLE as an estimate and the relevance of the second derivative matrix of the log-likelihood evaluated at the MLE as a measure of inferential precision.

To see that the existence of the solutions reported in our paper can matter, suppose we wish to compare the fit obtained at \( \theta_2 \) with the fit obtained at the Jöreskog solution, \( \theta_3 \). Let \( C \) be the variance-covariance matrix based on the second derivatives at \( \theta_3 \) printed out by LISREL-V. Since \( \{(\theta_2 - \theta_3)C^{-1}(\theta_2 - \theta_3)^T\}^{1/2} = 4.7 \), the conclusion based on this LISREL-V output is that \( \theta_2 \) is nearly five standard errors from \( \theta_3 \), implying that \( \theta_2 \) is entirely implausible, i.e., that \( \theta_2 \) provides a dramatically worse fit to the data than does \( \theta_3 \). Of course, there is another standard way to compare the fits provided by \( \theta_2 \) and \( \theta_3 \), and that is by comparing the values of the likelihood function at \( \theta_2 \) and \( \theta_3 \). The standard \( \chi^2 \) statistic is \(-2 \ln \text{likelihood ratio}\). The value of the likelihood ratio of \( \theta_2 \) to \( \theta_3 \) is given by

\[
\exp \left[ -\frac{n}{2} \{f(\theta_3) - f(\theta_2)\} \right] = .1143
\]

where \( f(\ ) \) is the function given in Rubin and Thayer's Table 3, and \( n = 710 \). In contrast