Standardized coefficients in DSEM/RDSEM

Tihomir Asparouhov & Bengt Muthén

February 5, 2020

This note discusses the following error message which appears in some DSEM/RDSEM Mplus estimations

WARNING: PROBLEMS OCCURRED IN SEVERAL ITERATIONS IN THE COMPUTATION OF THE STANDARDIZED ESTIMATES FOR SEVERAL CLUSTERS

This problem occurs because the standardized coefficients require the computation of the estimated variances for all dependent variables. For DSEM and RDSEM model, the estimation of the variance for an auto-regressive variable or model component is based on the assumption that the auto-regressive model is stationary, i.e., $Var(Y_t)$ is the same across all time points. An AR(1) model is stationary if the autoregressive coefficient is less than 1 by absolute value. For a VAR model, the condition of stationarity is somewhat more complex. If the VAR model is described by the following equation

$$Y_t = \alpha + R_1 Y_{t-1} + ... + R_L Y_{t-L} + \varepsilon_t,$$

the model is stationary if all the roots of the following equation

$$|I - xR_1 - ... - x^L R_L| = 0$$

are greater than 1 by absolute value, where I denotes the identity matrix and the absolute value in the above equation denotes the determinant of the matrix. These roots are currently not computed in Mplus, but typically non-stationarity occurs when the auto-regressive coefficients are large. A non-stationary model typically implies that the variance of Y_t increases with t and therefore the concept of standardized coefficients in not available. If a model is non-stationary and we attempt to compute the estimated variance

based on the stationarity assumption, the variance of some of the dependent variables becomes negative.

Because the estimation is Bayesian, the auto-regressive coefficients are estimated not just as point estimates but as entire posterior distributions. The most typical situation is the case where the point estimates of the auto-regressive parameters indicate a stationary model, but in a portion of the posterior distribution the non-stationarity assumption is violated. That portion of the posterior distribution is removed from the computation of the standardized coefficients, i.e., the standardized coefficients are based only on that part of the posterior distribution where the model is stationary.

Here we list some considerations that might be useful when this issue occurs.

1. In principle, the above message can be ignored. The standardized estimates that are printed in the Mplus output are the best possible that can be obtained with these data. Typically, not many of the MCMC iterations produce non-stationary model (less than 10%). If the process is non-stationary in large portion of the iterations a more sever problem would occur and the model would likely not converge.

2. One possible reasons for this problem is small sample size which leads to wide posterior distribution that extends far enough to reach the non-stationary part of the parameter space. This would also be the case when the auto-regressive coefficients are cluster specific and the cluster sample sizes are small. In such situations, the total sample size would not be relevant and the problem can occur even if the total sample is large. The size of the posterior distribution of the random auto-regressive coefficients is primarily driven by the size of the clusters and not by the total sample size. Possible model modifications might be useful to reduce the size of the posterior distribution. For example, auto-regressive coefficients with small variance can be converted from random to non-random. Any model simplification that improves the parsimony of the model can also lead to posterior distribution reduction that can lead to a solution of the above problem.

3. Switching from DSEM model to RDSEM model can also resolve the problem as typically RDSEM auto-regressive coefficients are smaller.

4. Another possible cause of the problem is trend in the data, i.e., the data is indeed non-stationary. If trend in the data exists, it should be modeled separately from the auto-regressive part of the model. One way to do that is to switch to an RDSEM model and to include a regression of Y on T and possibly other functions of T, such as T^2, $log(T)$, and $Exp(T)$. Mplus plots can be useful in that regard to evaluate the stationarity of the variables. For two-level models, it may be necessary for the trend model to be cluster specific, i.e., the regression of Y on T may need to be random. A different approach to dealing with trends in the data is to use a different scale for the dependent variable. For example, instead of modeling Y_t, one can model the change of Y_t, i.e., $Y_t - Y_{t-1}$. Instead of modeling Y_t, one can model $log(Y_t)$ which may be much more stationary than Y_t. If Y_t represents the "total of a population quantity" the variable can be replaced by "the total of the population quantity per 1000 people". This way the underlying increase in the population size can be removed from the model.

5. Weakly informative priors may be helpful in reducing the size of the posterior distribution of the auto-regressive parameters and eliminate some undesirable tail portions of the posterior distribution. This approach is much more effective when the auto-regressive coefficients are non-random. Random auto-regressive coefficients can not be given weakly informative priors. The prior for such a coefficient is essentially the two-level model. Weakly informative priors can be given for the between level parameters of the random auto-regressive coefficients but those will not necessarily shrink the posterior distribution of the random auto-regressive coefficients.

6. An alternative way to obtain the standardized coefficients is to use the observed variance instead of the estimated variance and by performing the standardization on the data before the model is estimated. In single level models a variable can be standardized with the STANDARDIZE option of the DEFINE command. In two-level models the standardization must be done for each cluster separately. This can be done by computing the variance of a dependent variable within each cluster with a separate Mplus run using the CLUSTER_MEAN option in the DEFINE command applied to Y and Y*Y. This approach is less reliable than the method used in Mplus but is a useful comparative alternative. The method is not recommenced in the presence of missing data as the sample variance would be estimated via listwise deletion.

7. The standardized results might be considered untrustworthy if there is a large discrepancy between the observed and the estimated within level variances. The estimated variances can be obtained using the RESIDUAL or the RESIDUAL(CLUSTER) options of the OUTPUT command. These are the same variances that are used for the computation of the standardized coefficients. The corresponding observed cluster specific variances can be obtained as in point 6 above. Alternatively, these quantities can be found and compared in the Mplus plot utility which can be obtained using the TYPE IS PLOT3 option of the PLOT command. Discrepancies between the observed and the estimated cluster specific variances could be mitigated by estimating random variances using a model like this
% within %
v | y;
Note, however, that if there are missing data, discrepancy between the observed and the estimated variances may actually be expected and the estimated quantities may be substantially more accurate than the observed values. In the presence of missing data, time series models may actually provide unbiased variance estimates for the within level variance while sample quantities (based on listwise deletion) could be biased.

8. Note also that the above error message concerns only the standardized coefficients, and not the model results. The model results section is not concerned with the issue discussed here.

9. In two-level models the above message will include the list of the particular clusters where the problem occurs. It may be useful to examine the data in those clusters. One possibility to further enlighten the issue is to run a particular cluster by itself using a single level model.

10. Further discussion on the computation of the standardized coefficients and the estimated variances can be found in

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. Structural Equation Modeling, 25, 359–388.
http://www.statmodel.com/download/DSEM.pdf

and in Chapter 3 of

Schuurman, N. (2016) Multilevel autoregressive modeling in psychology:
Snags and solutions. Utrecht University
https://dspace.library.uu.nl/bitstream/handle/1874/337475/Schuurman.pdf?sequence=1