# First-Order Derivative Check and Condition Number

*Tihomir Asparouhov & Bengt Muthén*

May 13, 2020

This note discusses the following error message that sometimes appears in the Mplus output

THE STANDARD ERRORS OF THE MODEL PARAMETER ESTIMATES MAY NOT BE TRUSTWORTHY FOR SOME PARAMETERS DUE TO A NON-POSITIVE DEFINITE FIRST-ORDER DERIVATIVE PRODUCT MATRIX. THIS MAY BE DUE TO THE STARTING VALUES BUT MAY ALSO BE AN INDICATION OF MODEL NONIDENTIFICATION. THE CONDITION NUMBER IS 0.889D-10. PROBLEM INVOLVING THE FOLLOWING PARAMETER:

A technical background is given, followed by several examples where the message appears and how to handle these.

## 1 Technical Background

To compute the maximum likelihood estimates the log-likelihood function $L$ is maximized with respect to all model parameters. If proper maximization has been achieved the negative of the matrix of the second derivatives, also known as the Fisher information matrix, $-L''$ should be a positive definite matrix. The information matrix is inverted $(-L'')^{-1}$ to obtain the standard errors. If the model is not identified, the information matrix is singular, i.e., it is not invertible and the determinant is 0. Mplus performs a check to determine if the information matrix is invertible, i.e., it is not singular. In principle, to do that one can simply compute the determinant of the matrix and check that it is 0. Numerically, however, the matrix will not be precisely zero. Mplus uses 15 decimal digits of precision for every number. In

addition, round off error can accumulate during the computation and thus it is not uncommon for a number that is theoretically zero to become as large as $10^{-10}$. In addition, computing the determinant of the information matrix to check the singularity of the matrix, is not a good idea as it exposes the computation to the scales of the observed variables. Instead of using the determinant to check for singularity, Mplus uses the condition number of the matrix. This is defined as the ratio of the smallest eigenvalue to the largest eigenvalue. If the matrix is singular the condition number is 0. Mplus uses as a cutoff value $10^{-10}$, i.e., if the condition number of the matrix is less than $10^{-10}$ Mplus concludes that the matrix is close to being singular which indicates that the model might not be identified. The cutoff value can be changed using the CONDITION option in the ANALYSIS command. The condition number can be found in the Mplus output for every model. Small condition numbers should be viewed as a problem, that may or may not need to be addressed. The smaller the condition number the flatter the likelihood of the model is, i.e., the weaker the identifiability of the model, i.e., the data set contains little or no information for some of the model parameters. The smaller the condition number the bigger the round off error in the matrix inversion, and as a result the bigger the round off error in the standard errors.

In addition to using the information matrix to check model identifiability, Mplus also uses the first-order derivative product matrix $\overline{L'(L')^T}$ which provides an approximation to the information matrix and is available in Mplus as the MLF estimator. This is the product matrix referred to in the above error message. Experience has shown that the MLF method is the best method to catch an unidentified model. It is the most accurate in terms of false negatives, i.e., least likely to not catch an unidentified model. Thus, Mplus performs the MLF singularity check even when the estimator that is used for the estimation is not the MLF estimator but is the MLR or the ML estimator. Unfortunately the MLF estimator also has a larger false positive error. That is, the MLF check may report potential non-identifiability in situations when the model is identified.

When the MLF check fails, Mplus will produce an error message described above. The parameter number that is listed may be the parameter that is not identified or may be closely connected to that parameter. The unidentified parameter can also be a parameter that is close to the listed parameter in terms of its TECH1 order. The way Mplus identifies the problematic parameter is by sequentially analyzing the information matrix. First we analyze the first parameter only, then the first and the second, then the first

three, etc. We conclude that the problematic parameter is the one that when added, the condition number drops below the cutoff value. Sometimes that drop will not occur at the exact unidentified parameter but at the next one (or the next few) as the condition number continues to decreases as we add more and more parameters.

An alternative method for identifying the problematic parameters is as follows. The analysis can be rerun using the settings: ANALYSIS: ESTIMATOR=MLF; CONDITION=0;. The parameters that are not identified will get very large standard errors and will be easy to spot.

When the above error message appears, Mplus can not guarantee that the model is identified. The identifiability of the model must be verified separately. For example, if the model is a well known model that has been used successfully with other data sets, the error message can be ignored.

In most situations, condition number below $10^{-12}$ indicates a true non-identification. Condition number between $10^{-10}$ and $10^{-12}$ is usually an indication of some other problem that could be addressed by model or data modification.

# 2 Examples

Here are some common causes of the message and how they can be resolved.

## 2.1 Binary variables treated as continuous

To avoid listwise deletion due to missing data on covariates, the covariates are often brought into the model by mentioning their means, variances, or covariances. With categorical variables, this means that they are treated as continuous variables. With binary variables, the MLF message gets triggered in this case because the mean of the variable $p$ is directly related to the variance $p(1-p)$ (one parameter $p$ exist while two are estimated). In almost all situations this should be addressed by model modification. Treating the variable as continuous is not ideal. Missing data estimation for continuous variables assumes normal distribution and this misspecification will likely (depending on the amount of missing data) bias the estimation. There are two options in this case. Option one is to model the variable as categorical. Option two is to use multiple imputations where the variable is specified

as categorical so that the imputed values are also binary (see User's Guide example 11.5 for how to impute missing values in Mplus).

Another situation where a binary covariate is treated as continuous is the situation where the variable is used in a WITH statement with all other covariate to ensure that correlations between the covariates are taken into account. However, if a variable is independent, it is automatically correlated with all other covariates in the model and such WITH statements are unnecessary and should be removed.

If a dependent binary variable is treated as continuous, causing the MLF message to appear, the variable should simply be declared as categorical.

## 2.2 Variables in the model are on very different scales

Sometimes variable have very different variances due to being measured on different scales and this may trigger the MLF message due to a low condition number. The scales of the variables can be changed so that they are more similar and that will improve the condition number. The scale of the variables can be changed either by DEFINE: Y=Y/10; or by DEFINE: STANDARDIZE Y.

## 2.3 There are more parameters in the model than observations

In this case the MLF check is triggered 100% of the time. It is possible to verify that the model is identified with a bigger data set. One way to do that is to use a data set that consists of multiple copies of the original data set. If the model is a standard well known model the MLF warning should simply be ignored.

## 2.4 There are more parameters than clusters for type=complex, type=twolevel or type=threelevel

In this case the MLF check is triggered 100% of the time. It is possible to verify that the model is identified with a bigger data set. In three-level models, the number of parameters is compared against the highest level number of clusters.

## 2.5 Parameter appears to be converging to a large value

Mplus will often identify parameters that are causing identifiability problems, in terms of causing low condition numbers, and fix those parameters. For example, when a binary indicator for a particular class is estimated to be a perfect indicator for that class, the threshold value will become +- large values. Such thresholds are equivalent to +- infinity and are essentially unidentified but are safe to fix. By fixing these parameters Mplus preserves the quality of the model and avoids the low condition number. Another example is the situation when a categorical variable is regressed on another categorical variable (latent or observed) and there are empty cells in the joint distribution of the two variables. Some of the regression parameters may be estimated to +- infinity and are safe to fix as well. Mplus automatically fixes threshold parameters if they reach the values of +-15 (this is controlled by the LOGHIGH and LOGLOW options of the ANALYSIS command). If a threshold is estimated at 14, it will not be fixed by Mplus and may produce low condition number. You can change the LOGHIGH/LOGLOW options or ignore the MLF warning message. In this situation it is useful to look at and report the model parameters in probability scale as those are well identified and do not have values converging to +- infinity.

## 2.6 Cluster invariant variable in two-level models

Suppose that a within-only variable in a two-level model takes the exact same values in each cluster. An example of such a variable is TIME in a two-level longitudinal model where the cluster is the person and the observations within cluster are the observations at particular time points. If TIME takes the exact same values in each cluster and is modeled as a dependent variable it will result in an MLF warning message due to a lack of variation in the score of TIME's parameters. In this case the message can be ignored or the TIME variable can be treated as a covariate instead of a dependent variable.

## 2.7 Group-mean centered within-only variable with estimated intercept or mean

Suppose that a within-only variable $Y$ in a two-level model has been group-mean centered using the command

DEFINE: CENTER Y(GROUPMEAN);
This command essentially removes the between part of the variable. The cluster sample mean is subtracted from each observed value in the cluster. The centered variable $Y$ has a cluster sample mean of 0 in every cluster. This kind of centering is necessarily followed by a within-only specification
VARIABLE: WITHIN=Y;
(otherwise the between part will be estimated to the constant 0). The issue that arises here is that not only the random part of the intercept is removed, but also the fixed part of the intercept is removed. Therefore, the mean parameter of $Y$ is necessarily 0. If that parameter is not fixed to 0, but is estimated as a free parameter, it is very likely that the MLF warning will appear. There are instances where it won't appear. For example, if the variable $Y$ is regressed on another variable $X$ that is not group-mean centered the warning will not appear (even though the underlying modeling problem is still there and should be fixed).

The MLF warning should not be ignored in this situation . The proper model setup is to have the intercept of $Y$ be fixed to zero and to also group-mean center all predictors of $Y$.