Bootstrap in Two-level models

Tihomir Asparouhov & Bengt Muthén

June 13, 2025

1 Introduction

In this note, we describe the bootstrap methodology implemented in Mplus for two-level models. We also provide details on this implementation for multiple group two-level models, two-level models with sampling weights, and two-level models with complex survey samples (i.e., stratified and cluster sampling). The methodology is illustrated with several simulation studies, where we compare alternative methods for bootstrap resampling.

2 Comparison of different bootstrap resampling methods

Van der Leeden et al. (2008) describe the following two alternative resampling methods in two-level models among other methods:

- A. Cluster resampling: A new bootstrap sample is constructed by sampling with replacement entire clusters. The data within each cluster is exactly the same as it appeared in the original sample. Each bootstrap sample has exactly the same number of clusters as the original sample.
- B. Cluster resampling followed by resampling of individual observations within each sampled cluster: A bootstrap sample also draws with replacement the same number of clusters as the original sample, but the drawn clusters are not identical to the original clusters. Instead, within each cluster, the observations are also bootstrapped (i.e., sampled with replacement) so that we obtain a cluster of the same size as the original cluster.

Method A is akin to the method implemented in Mplus for bootstrap sampling in the presence of complex survey data, specifically cluster sampling. In those situations, entire PSU (primary sampling unit)/clusters are sampled without alteration, see Asparouhov and Muthén (2010).

With both methods, the bootstrap samples may not have the same total size as the original sample when the clusters are not of the same size. However, this is somewhat of an artificial observation. In single-level complex sample bootstrapping, where the same observation applies, the bootstrap samples saved by Mplus are of the same size as the original sample. What changes is the sampling weight assigned to each observation. For example, two draws of the same observation is exactly the same as doubling the sampling weight of that observation. The bootstrap sample can thus be represented as a sample with the same data as the original sample but with different sampling weights.

To put this differently, if a bootstrap draw sample size increases dramatically due to unbalanced cluster designs, when we analyze that sample, the point estimates would generally be similar to those of the original sample, but the standard errors would be much smaller due to the fact that the bootstrap draw is much larger. These standard errors, however, are not used in the bootstrap procedure, and only the point estimates are used. Thus, for most practical purposes, the fact that the bootstrap draws are of different sizes is rather unimportant.

We compare methods A and B with a simulation study using a simple two-level regression model with a random intercept and a single covariate. Such a systematic comparison has not been reported previously and we consider this to be the main contribution of this paper. Figure 1 contains the Mplus input file for the simulation study. Figures 2 and 3 contain the results using methods A and B respectively.

The results show that method B substantially overestimates the standard errors on the within level. Typically, the quality of the standard errors for an estimator is monitored using the coverage column in Mplus, and values close to 95% indicate proper estimation. However, when the standard errors are substantially overestimated, the coverage becomes 100%, and the coverage is not as useful in determining the quality of the standard error estimates.

In such situations, a better approach is to consider the ratio between the columns "SE average" and "Std. Dev." The "SE average" is simply the average of the standard errors across the 100 replications in the simulation study. The column labeled "Std. Dev." contains the standard deviation

of the point estimates across the 100 replications. Asymptotically, the ratio between the columns should converge to 1. When the ratio is greater than 1, the standard errors are overestimated, and when it is less than 1, the standard errors are underestimated.

Because we usually work with moderate sample size and moderate number of replications, the ratio usually deviates from one but typically by not more than ± 0.1, i.e., standard error underestimation or overestimation in the single-digit percentages. What we see in our simulation study is that method B overestimates the standard errors for the two within-level parameters by 31% and 28%. For method A, those numbers are 0% and -8%. We clearly see that method A is superior. The overestimation of the standard error for method B is consistent across a variety of estimation settings and does not improve with more bootstrap draws, larger number of Monte Carlo replications, larger number of clusters, or larger cluster sizes.

Next, we conduct a simulation study using the two-level factor analysis model given in Figure 4. The results of the simulation study are given in Figures 5 and 6 for methods A and B respectively. Here again, we see the same problem with method B. Coverage for all within-level parameters is nearly 100%, and the average standard error overestimation across the 10 within-level parameters is 37%. For method A, the corresponding number is 1%.

Based on the results of the above simulations, we conclude that method B is not appropriate for two-level models. Only method A is implemented in Mplus, and only this method is used and discussed for the remaining portion of this article.

We should note here that resampling within clusters has its place when the modeling is multiple group instead of multi-level. In some situations when the number of clusters is small, such as less than 10, we often prefer single-level multiple group analysis instead of multilevel. This means that the cluster variable is treated as a grouping variable in Mplus, and random effects are replaced by non-random group-specific model parameters.

Two-level analysis is based on the asymptotic assumption that the number of clusters is sufficiently large. When the number of clusters is small, it is difficult to rely on the asymptotic theory, and often the ML estimation will exhibit larger biases in the point estimates. This, of course, has nothing to do with bootstrap, as the bootstrap method is responsible only for the standard error estimation.

When Mplus estimates a multiple group single-level model, the resam-

Figure 1: Two-level regression simulation study

```
montecarlo:
        names = y x;
        nobs = 2000;
        nreps = 100;
        ncsizes = 1;
        csizes = 100(20);
        within=x;

analysis: estimator=ml;
type=twolevel; bootstrap=100;

model population:
%WITHIN%
 y*1; x*1; y on x*1;
%BETWEEN%
 y*1; [y*1];

model:
%WITHIN%
 y*1; y on x*1;
%BETWEEN%
 y*1; [y*1];
```

Figure 2: Bootstrap results for two-level regression using method A

		ESTIMATES		S. E.	M. S. E.	95%	% Sig
	Population	Average	Std. Dev.	Average		Cover	Coeff
Within Level							
Y ON							
X	1.000	0.9981	0.0247	0.0227	0.0006	0.940	1.000
Residual Variances							
Y	1.000	1.0055	0.0326	0.0326	0.0011	0.940	1.000
Between Level							
Means							
Y	1.000	1.0106	0.0977	0.1013	0.0096	0.980	1.000
Variances							
Y	1.000	0.9846	0.1509	0.1453	0.0228	0.920	1.000

Figure 3: Bootstrap results for two-level regression using method B

	Population	ESTIMATES Average	Std. Dev.	S. E. Average	M. S. E.	95% Cover	% Sig Coeff
Within Level							
Y ON							
X	1.000	0.9981	0.0247	0.0317	0.0006	0.990	1.000
Residual Variances							
Y	1.000	1.0055	0.0326	0.0427	0.0011	1.000	1.000
Between Level							
Means							
Y	1.000	1.0106	0.0977	0.1029	0.0096	0.980	1.000
Variances							
Y	1.000	0.9846	0.1509	0.1492	0.0228	0.920	1.000

pling is done within each group, and each bootstrap sample preserves the sizes of the groups. This is akin to method B. With multiple group single-level bootstrap estimation, however, Mplus does not resample the groups, i.e., the resampling is only on the lower level.

3 Sample size requirements for two-level bootstrap

In this section, we illustrate the quality of the bootstrap estimation for samples with extreme sample designs. The three features we explore here are: highly unbalanced designs (large variation on the cluster size), small number of clusters, and small cluster sizes. For all three conditions, we use the regression model we used in the previous section.

Highly unbalanced designs have been discussed in the context of two-level bootstrap due to the fact that the bootstrap samples may have substantially different sizes than the sample size of the original data set. As we discussed in the previous section, however, this is somewhat irrelevant as the bootstrap sample can be presented for the purpose of obtaining the bootstrap point estimates as having the same sample size as the original sample but having different between-level weights.

In this simulation study, we use a sample design that has clusters with size 10, 20, 30, ..., 100. There are 10 clusters for each of the 10 different cluster sizes for a total number of clusters of 100 and a total sample size of 5500. The results of this simulation study are given in Figure 7. The results show that the bootstrap estimates the standard errors well.

Next, we consider the situation of a small number of clusters. This is important because we draw clusters as a whole, and if there are only a few clusters, the amount of variation in the bootstrap draws would be limited. For example, if there are only 2 clusters, there will be only 3 different bootstrap draws, and that could potentially cause issues with the estimation.

As we mentioned in the previous section, however, sample designs with very few clusters should be estimated as multiple group single-level models—not because of the bootstrap but to improve the point estimates. In this simulation study, we use 20 clusters of size 100. The results are given in Figure 8.

The bootstrap procedure works well here too. One of the between-level

Figure 4: Two-level factor analysis simulation study

```
montecarlo:
        names = y1-y5;
        nobs = 2000;
        nreps = 100;
        ncsizes = 1;
        csizes = 100(20);

analysis: estimator=ml;
type=twolevel;boot=100;

model population:
%WITHIN%
 y1-y5*1;
 f by y1-y5*1; f@1;
%BETWEEN%
 y1-y5*0.5;
 fb by y1-y5*0.8; fb@1;

model:
%WITHIN%
 y1-y5*1;
 f by y1-y5*1; f@1;
%BETWEEN%
 y1-y5*0.5;
 fb by y1-y5*0.8; fb@1;
```

Figure 5: Bootstrap results for two-level factor analysis using method A

		ESTIMATES		S. E.	M. S. E.	95%	% Sig
	Population	Average	Std. Dev.	Average		Cover	Coeff
Within Level							
F BY							
Y1	1.000	0.9976	0.0300	0.0294	0.0009	0.950	1.000
Y2	1.000	0.9958	0.0292	0.0305	0.0009	0.940	1.000
Y3	1.000	1.0018	0.0287	0.0301	0.0008	0.980	1.000
Y4	1.000	0.9996	0.0277	0.0299	0.0008	0.980	1.000
Y5	1.000	1.0041	0.0339	0.0306	0.0012	0.940	1.000
Variances							
F	1.000	1.0000	0.0000	0.0000	0.0000	1.000	0.000
Residual Variances							
Y1	1.000	0.9980	0.0353	0.0399	0.0012	0.980	1.000
Y2	1.000	1.0055	0.0431	0.0406	0.0019	0.900	1.000
Y3	1.000	1.0004	0.0408	0.0406	0.0016	0.940	1.000
Y4	1.000	1.0015	0.0419	0.0412	0.0017	0.930	1.000
Y5	1.000	0.9983	0.0410	0.0404	0.0017	0.940	1.000
Between Level							
FB BY							
Y1	0.800	0.7834	0.1094	0.1024	0.0121	0.920	1.000
Y2	0.800	0.7879	0.1075	0.1021	0.0116	0.940	1.000
Y3	0.800	0.7813	0.0970	0.1038	0.0097	0.950	1.000
Y4	0.800	0.7954	0.1096	0.1039	0.0119	0.950	1.000
Y5	0.800	0.7865	0.1038	0.1054	0.0108	0.950	1.000
Intercepts							
Y1	0.000	-0.0058	0.1139	0.1100	0.0129	0.890	0.110
Y2	0.000	-0.0098	0.1133	0.1113	0.0128	0.940	0.060
Y3	0.000	0.0111	0.1108	0.1106	0.0123	0.940	0.060
Y4	0.000	-0.0073	0.1116	0.1115	0.0124	0.950	0.050
Y5	0.000	-0.0006	0.1025	0.1119	0.0104	0.940	0.060
Variances							
FB	1.000	1.0000	0.0000	0.0000	0.0000	1.000	0.000
Residual Variances							
Y1	0.500	0.4707	0.1041	0.0931	0.0116	0.890	1.000
Y2	0.500	0.4894	0.1024	0.0952	0.0105	0.890	1.000
Y3	0.500	0.4857	0.0968	0.0933	0.0095	0.890	1.000
Y4	0.500	0.4765	0.0939	0.0929	0.0093	0.920	1.000
Y5	0.500	0.5002	0.0942	0.0969	0.0088	0.960	1.000

Figure 6: Bootstrap results for two-level factor analysis using method B

		ESTIMATES		S. E.	M. S. E.	95%	% Sig
	Population	Average	Std. Dev.	Average		Cover	Coeff
Within Level							
F BY							
Y1	1.000	0.9976	0.0300	0.0414	0.0009	0.990	1.000
Y2	1.000	0.9958	0.0292	0.0420	0.0009	0.990	1.000
Y3	1.000	1.0018	0.0287	0.0415	0.0008	1.000	1.000
Y4	1.000	0.9996	0.0277	0.0422	0.0008	1.000	1.000
Y5	1.000	1.0041	0.0339	0.0417	0.0012	0.980	1.000
Variances							
F	1.000	1.0000	0.0000	0.0000	0.0000	1.000	0.000
Residual Variances							
Y1	1.000	0.9980	0.0353	0.0535	0.0012	1.000	1.000
Y2	1.000	1.0055	0.0431	0.0538	0.0019	0.990	1.000
Y3	1.000	1.0004	0.0408	0.0541	0.0016	1.000	1.000
Y4	1.000	1.0015	0.0419	0.0542	0.0017	0.990	1.000
Y5	1.000	0.9983	0.0410	0.0543	0.0017	0.990	1.000
Between Level							
FB BY							
Y1	0.800	0.7834	0.1094	0.1079	0.0121	0.920	1.000
Y2	0.800	0.7879	0.1075	0.1073	0.0116	0.950	1.000
Y3	0.800	0.7813	0.0970	0.1086	0.0097	0.960	1.000
Y4	0.800	0.7954	0.1096	0.1077	0.0119	0.940	1.000
Y5	0.800	0.7865	0.1038	0.1094	0.0108	0.980	1.000
Intercepts							
Y1	0.000	-0.0058	0.1139	0.1123	0.0129	0.930	0.070
Y2	0.000	-0.0098	0.1133	0.1151	0.0128	0.940	0.060
Y3	0.000	0.0111	0.1108	0.1139	0.0123	0.950	0.050
Y4	0.000	-0.0073	0.1116	0.1144	0.0124	0.950	0.050
Y5	0.000	-0.0006	0.1025	0.1146	0.0104	0.960	0.040
Variances							
FB	1.000	1.0000	0.0000	0.0000	0.0000	1.000	0.000
Residual Variances							
Y1	0.500	0.4707	0.1041	0.1016	0.0116	0.920	1.000
Y2	0.500	0.4894	0.1024	0.1036	0.0105	0.950	1.000
Y3	0.500	0.4857	0.0968	0.1040	0.0095	0.930	1.000
Y4	0.500	0.4765	0.0939	0.1012	0.0093	0.940	1.000
Y5	0.500	0.5002	0.0942	0.1065	0.0088	0.970	1.000

Figure 7: Bootstrap results for two-level regression with unbalanced design

	Population	ESTIMATES Average	Std. Dev.	S. E. Average	M. S. E.	95% Cover	% Sig Coeff
Within Level							
Y ON							
X	1.000	0.9980	0.0137	0.0134	0.0002	0.960	1.000
Residual Variances							
Y	1.000	1.0026	0.0192	0.0191	0.0004	0.930	1.000
Between Level							
Means							
Y	1.000	1.0057	0.1017	0.1000	0.0103	0.940	1.000
Variances							
Y	1.000	0.9791	0.1550	0.1352	0.0242	0.890	1.000

parameters has a drop in coverage to 78%; however, this is mostly due to the bias in the point estimates (due to small number of clusters). The coverage for the MLR sandwich estimator in that case is slightly better at 84%. We conclude that if the number of clusters is small, some underestimation may exist in the bootstrap standard error on the between level.

We used 100 bootstrap draws for this estimation, and the coverage does not improve with a larger number of draws. In most cases, we expect that 100 bootstrap samples is sufficient to estimate the standard errors well, and a larger number of draws is likely unnecessary. The estimates will not change much by increasing the number of bootstrap draws beyond 100.

Next, we consider the situation of small cluster sizes. This is somewhat unnecessary, but we include this for completeness. The reason it is not necessary is that if the cluster sizes are 1, then the two-level bootstrap is essentially the same as the standard single-level bootstrap, which we know works well. Nevertheless, we conduct a simulation study with 400 clusters of size 5. The results are reported in Figure 9. The bootstrap estimation works well in this case.

4 Bootstrapping for two-level models with complex samples

In this section, we discuss the effects of complex sampling features on the bootstrap procedure. More specifically, we discuss the effects of stratification, cluster sampling, and sampling weights.

The most comprehensive sampling structure used in Mplus for two-level models is as follows. The sample is obtained from a population that consists of several distinct groups. Within each group, the population is stratified, i.e., the population is divided into several strata, and each stratum is sampled separately. Within each stratum, primary sampling units (PSUs) are drawn. These primary sampling units consist of multiple clusters from which the sample clusters are drawn. Furthermore, the clusters are sampled for individual observations.

Overall, this gives a 5-stage nested data structure: group / strata / psu / cluster / individual. In addition, two weight variables can be assigned to reflect unequal probability of selection: one that corresponds to the cluster, specified as BWEIGHT, and one that corresponds to the individual, specified

Figure 8: Bootstrap results for two-level regression with small number of clusters

	Population	ESTIMATES Average	Std. Dev.	S. E. Average	M. S. E.	95% Cover	% Sig Coeff
Within Level							
Y ON							
X	1.000	0.9979	0.0243	0.0218	0.0006	0.940	1.000
Residual Variances							
Y	1.000	1.0043	0.0328	0.0310	0.0011	0.900	1.000
Between Level							
Means							
Y	1.000	0.9980	0.2278	0.2127	0.0514	0.930	1.000
Variances							
Y	1.000	0.9130	0.2890	0.2580	0.0903	0.780	1.000

Figure 9: Bootstrap results for two-level regression with small cluster sizes

	Population	ESTIMATES Average	Std. Dev.	S. E. Average	M. S. E.	95% Cover	% Sig Coeff
Within Level							
Y ON							
X	1.000	0.9948	0.0252	0.0243	0.0007	0.930	1.000
Residual Variances							
Y	1.000	1.0043	0.0383	0.0356	0.0015	0.950	1.000
Between Level							
Means							
Y	1.000	1.0024	0.0581	0.0548	0.0033	0.910	1.000
Variances							
Y	1.000	1.0078	0.0804	0.0849	0.0065	0.950	1.000

as WEIGHT. The sampling weights are internally scaled (see Asparouhov, 2006; Asparouhov and Muthén, 2008).

Not all of the above features must be present for the methodology to apply, but every sampling design can be formulated as if it is the full design. If grouping and stratification are not present, we simply assume that there is just one group and one stratum. If PSU is not present, the PSU is the cluster. If sampling weights are not present, we can assume that the sampling weights are all 1.

The bootstrapping construction is as in Asparouhov and Muthén (2010). Bootstrap sampling is done separately and independently for each group and each stratum within each group. Within each stratum, PSUs are bootstrap sampled with replacement. The number of PSUs sampled in each stratum is the same as the number of PSUs in the original sample. The sampled PSUs are identical to the PSUs in the original sample, including the sampling weights, i.e., no further resampling is done at the cluster or individual levels.

It should be noted here that in Mplus, to analyze complex two-level sampling data, in the ANALYSIS command, the option TYPE=COMPLEX TWOLEVEL must be specified. If, however, only sampling weights are used for the complex sampling, i.e., no stratification or PSU, then the usual setting of TYPE=TWOLEVEL is used. The same methodology applies to both situations.

Note also that in two-level complex analysis with highly structured data, it's fairly likely that the number of PSUs in each stratum is small. As discussed earlier, this limits the variability in bootstrap draws, which will inevitably lead to standard error underestimation. For example, in the extreme case where each stratum contains precisely one PSU, the bootstrap standard errors will be zero since each bootstrap draw will be identical to the original sample. Similarly, if most strata have only one PSU, we can expect underestimation of the bootstrap standard errors. In such cases, the bootstrap methodology would not be appropriate, and the sandwich MLR standard errors should be used instead. However, if only a few strata have one PSU while most strata have many PSUs, there will be sufficient variability in the bootstrap draws, and we can expect the bootstrap standard errors to be consistent.

5 Saving the bootstrap draws

It is possible to save all bootstrap samples in separate files and all the point estimates obtained for each bootstrap sample using the following command:

```
SAVEDATA: SAVE=BOOTSTRAP; FILE IS breps*.dat; RESULTS = r.dat;
```

The bootstrap draws will be saved in the files breps1.dat, breps2.dat, etc. The parameter estimates obtained with each bootstrap draw are all saved in the file r.dat. As discussed earlier, there are two ways to save the bootstrap samples. One way is to directly save the sampled observations. With this method the bootstrap draws can vary in total sample size under unbalanced design. The second way is to save the modeled data as it is in the original sample but to change the between-level weight to reflect how many times a cluster has been sampled in the bootstrap draw. With this method, the total sample size remains the same as in the original data, albeit some clusters will have zero weight and some will have weight greater than 1. For complex sampling, Mplus uses the second method, while without complex sampling, the former method is used.

References

[1] Asparouhov, T. (2006). General multi-level modeling with sampling weights. Communications in Statistics - Theory and Methods, 35, 439-460.

[2] Asparouhov, T., & Muthén, B. M. (2008). Scaling of Sampling Weights For Two Level Models in Mplus
http://statmodel.com/download/Scaling3.pdf

[3] Asparouhov, T., & Muthén, B. M. (2010). Resampling methods in Mplus for complex survey data.
http://statmodel.com/download/Resampling_Methods5.pdf

[4] Van der Leeden, R., Meijer, E., & Busing, F. M. (2008). Resampling multilevel models. In Handbook of multilevel analysis (pp. 401–433). New York, NY: Springer.