Fitting Empirically Under-Identified Models: A Two-Factor Example

Dylan Boczar & Eric Loken, Department of Educational Psychology, UCONN UNIVERSITY OF CONNECTICUT

Introduction

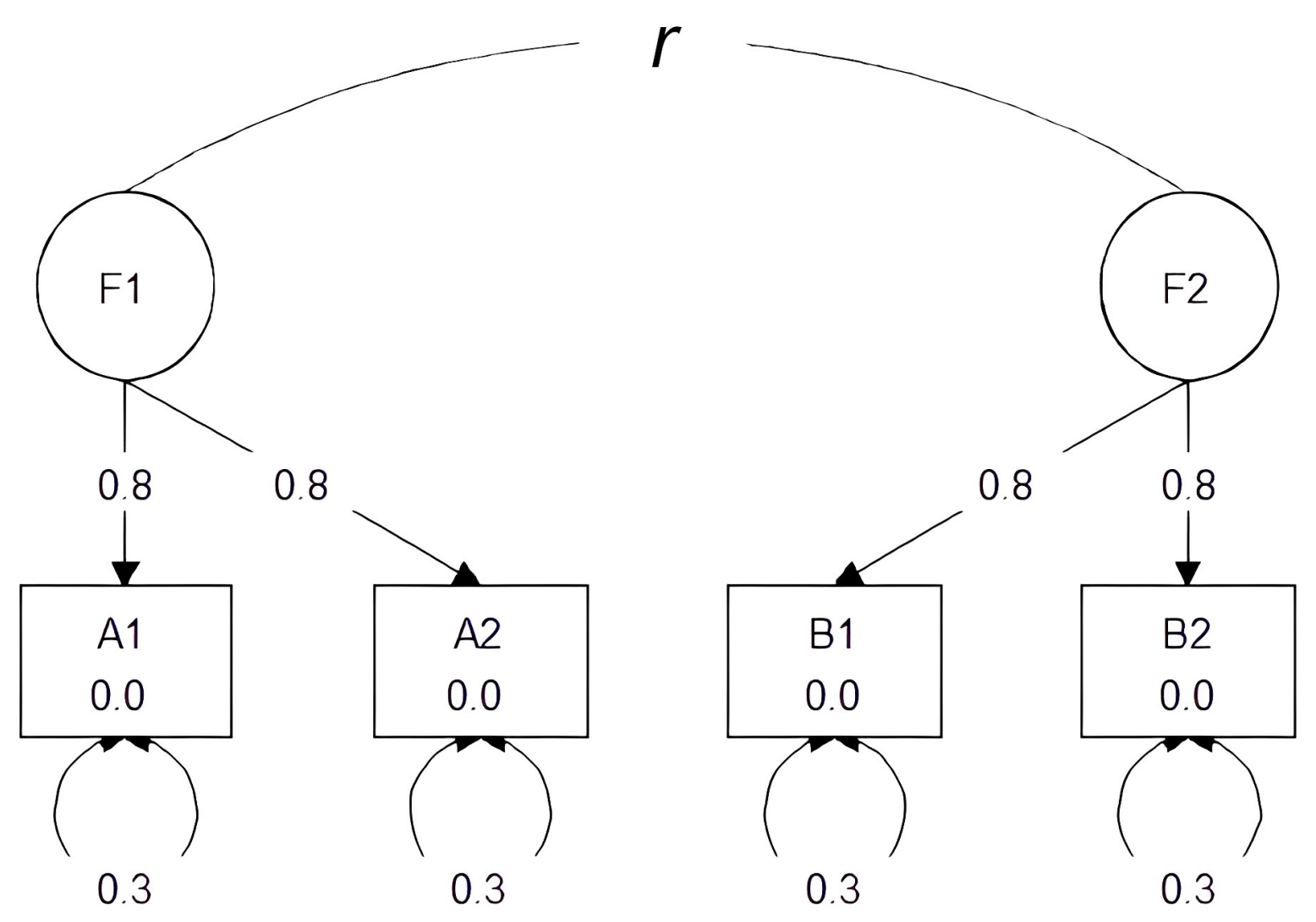
A CFA with two factors, and two indicators for each factor, is identified as long as the factor correlation (*r*) is not zero. According to Allman et. al. (2009), the model is not ***strictly*** identified, but is ***generically* identified** because the exact point *r*=0 has measure zero. However, if the **data-generating mechanism** is close to r=0, we run the risk of ***empirical under-identification*** (Kenny, Kashy, & Bolger. 1998). In this case, model fitting may be extremely difficult. Our example is only **illustrative;** more complex generically identified models can also suffer from empirical under-identification (Loken & Teitelbaum, 2023).

Model diagnostics for simulated data with decreasing factor covariances demonstrate the impact of empirical under-identification. We show issues with model convergence, parameter estimation, and standard errors when the population covariance nears zero. **Even for models that converged, several issues emerged**: extreme factor loadings, Heywood cases, and failure to estimate SEs. We also explore the role of sample size, noting that large sample sizes are more desirable when the factor covariance is far from zero, but are more problematic for *r* closer to zero.

Materials

Muthén, L.K. and Muthén, B.O. (1998-2017). **Mplus** User's Guide. Eighth Edition. Los Angeles, CA: Muthén & Muthén

Two-Factor Example



r ≠ 0, → model is **identified.**

r = 0 → **two disjoint, unidentified** two item factors.

Example solutions for n=1,000, F1 (A1, A2) ML

	Sample r=0.3		Sample r=0.1		Sample r=0.0	
Loading	0.862	0.802	0.863	0.741	0.060	10.50
SE	*0.047*	*0.052*	*0.096*	*0.084*	NA	NA
Residual Variance	0. 182	0.424	0.170	0.422	*0.980*	*-108.7*
SE	*0.072*	*0.057*	*0.161*	*0.120*	*NA*	*NA*
Factor Correlation	0.277		0.154		-0.001	
SE	*0.036*		*0.038*		*NA (can't estimate)*	

Typical modelling results by factor correlation and sample size (ML)

N	*r*=0.3	*r*=0.1	*r*=0.0
10,000	Normal estimation	Normal estimation	Error calculating SE
1,000	Normal estimation	*Some* Heywood cases	Heywood cases *Some* warnings/errors
300	*Some* Heywood cases	Heywood cases	*Some* Heywood cases *Some* warnings/errors *Some* converge fails
100	*Some* Heywood cases	Heywood cases	Heywood cases *Some* warnings/errors *Some* converge fails

Results

Factor Correlation: *r*=0.1 and 0 samples estimated using ML resulted in numerous Heywood cases, as well as information matrix errors, failure to compute SEs, and for some *r*=0 samples, failure to converge at all.

Warnings/errors: Warnings for *r*=0 did not flag the syntax of the model; instead, "problem parameters" were often flagged, with different problems and errors across similarly-generated samples and within the same sample across equivalent model constraints (fixed factor variances vs fixed first-indicators) and estimation methods (maximum likelihood vs Bayesian).

Sample size: n=10,000 showed consistency (*r*=0.3 and 0.1 no errors, *r*=0 consistent trouble with SE). **Lower sample sizes increased error rate, with lower *r*'s showing more errors.** Interestingly, low sample size for *r*=0 saw a dip in Heywood cases; n=300 models showed fewer Heywood cases than n=1000.

Standard errors: Even with model warnings/errors, all models estimate factor correlation with low SEs. However, both **smaller sample size** and **smaller factor correlation** resulted in larger SEs for loadings and residual variances, including in models without overt warnings or errors.

Bayesian estimation: Models always ran without error (prior constrains residuals to be positive and provides modest information). However, Bayes diagnostic pD (effective number of parameters) indicates problematic estimation at r=0.1 and 0.

Conclusions

When *r* is "far from" zero, models are easily estimated, software & methods agree. When *r* "close to" zero, **methods (i.e. Bayes – ML) differ considerably**. What counts as "close" depends on available information – sample size, factor loadings.

Navigating fit problems requires understanding the relevant mathematical issues affecting model identification. For instance, the estimate for *r* (the problematic parameter) has the *best* SEs. Standard software output indicates model fitting issues, but **does not diagnose the exact problem.**

While this small CFA is mathematically simple and produces easy to understand errors, **larger and more complex models (e.g. LCA) can be generically identified, but have many subtle, difficult to anticipate, parameter configurations that impact model estimation.** Understanding reasons why strict identifiability may fail helps navigate situations with empirical under-identification.

Literature cited

Allman, E. S., Matias, C., & Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics, 37* (6A), 3099–3132. doi: 10.1214/09-AOS689

Kenny, D. A., Kashy, D., & Bolger, N (1998). Data analysis in social psychology. In Gilbert, D. T., Fiske, S. T., & Lindzey, G. (Eds.). *The handbook of social psychology* (4th ed., pp. 233-265). New York: McGraw-Hill.

Loken, E., Teitelbaum, J., (2023). *Mathematical issues impacting the fitting of latent variable models.* Paper presented at the *National Council on Measurement in Education Annual Meeting*, Chicago, IL. (April 13, 2023).