Methods & Results

Valentin Kriegmair

Methods

The following description of the simulation studies is based on the established structure for simulation studies to coalign with the other conducted studies to facilitate a potential collaboration.

Aims, objectives and research questions

Both studies aimed to evaluate the performance of vanilla SEM compared to global SAM (gSAM), local SAM with maximum likelihood (lSAM-ML), and local SAM with unweighted least squares (lSAM-ULS) under various conditions. The two research questions we established prior to conducting the studies served as general basis for both studies

Population Models and Data Generation Mechanisms

Study 1:

Data were generated based on a 5-factor population structural model with 3 indicators for each factor. Four different models were simulated (see figure 1-4)

- Model 1.1: Correctly specified model.
- Model 1.2: Misspecified with cross-loadings ignored in the estimation model.
- Model 1.3: Misspecified with wrong direction of path and omitted correlated residuals.
- Model 1.4: Misspecified with multiple omitted structural paths.

For all models, the population-level values of the structural parameters were set to 0.1. Indicator reliability levels were manipulated with factor loadings set at low (= 0.3), moderate (= 0.5), or high (= 0.7).

Study 2:

Data were generated based on a 5-factor population structural model with 3 indicators for each factor, similar to Study 1, but with the additional condition of varying variance explained (R^2) by the endogenous factors was set at low (R^2 = 0.1) or medium (R^2 = 0.4), so that the regression weights were either between 0.183 and 0.224 (low) or between 0.365 and 0.447 (medium). Note however that the computation of this was a simplification and does not acurately result in said R^2 values. The aim here was only generally to modulate between lower and higher regression weights. Misspecifications included omitting a residual covariance and a factor cross-loading in either the exogenous or endogenous part of the model, or both (see figure 5-6).

Experimental Design of simulation procedures

Study 1

The study varied three main conditions: - Sample sizes: small (N = 100), moderate (N = 400), and large (N = 6400). - Indicator reliability: low ($\lambda = 0.3$), moderate ($\lambda = 0.5$), high ($\lambda = 0.7$). - Model specifications: correctly specified model and misspecified models (1.2, 1.3, and 1.4).

Study 2

The study varied five main conditions: - Sample sizes: small ($N = 100$), medium ($N = 400$), and large ($N = 6400$). - Variance explained by endogenous factors: low ($R^2 = 0.1$) and medium ($R^2 = 0.4$). - Indicator reliability: low ($\lambda = 0.3$), moderate ($\lambda = 0.5$), and high ($\lambda = 0.7$). - Model misspecifications: varying the population model by omitting a residual covariance and a factor cross-loading in different parts of the model. - Number of measurement blocks: separate measurement model per latent variable ($b = 5$) and joint measurement model for all exogenous variables ($b = 3$).

Both studies were conducted with 10000 replications for each condition.

Method Selection

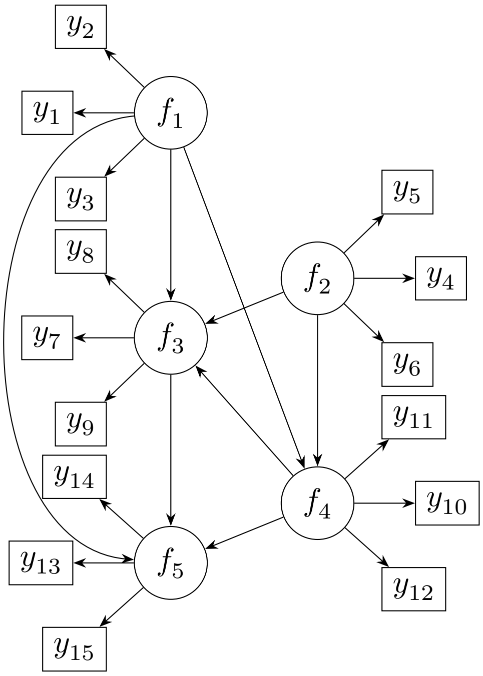

Figure 1: Note. Model 1.1: Error terms are not explicitly shown in the figure.

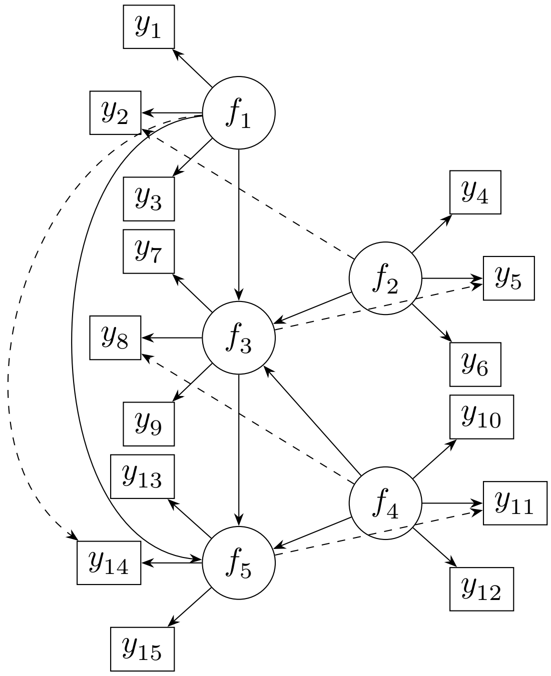

Figure 2: Note. Model 1.2: Error terms are not explicitly shown in the figure.

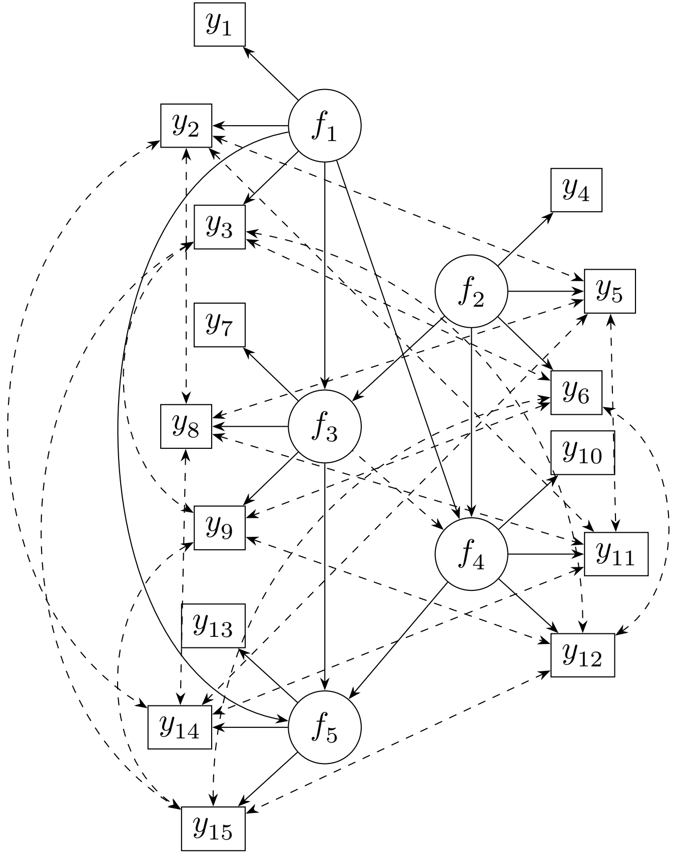

Figure 3: Note. Model 1.3: Error terms are not explicitly shown in the figure.

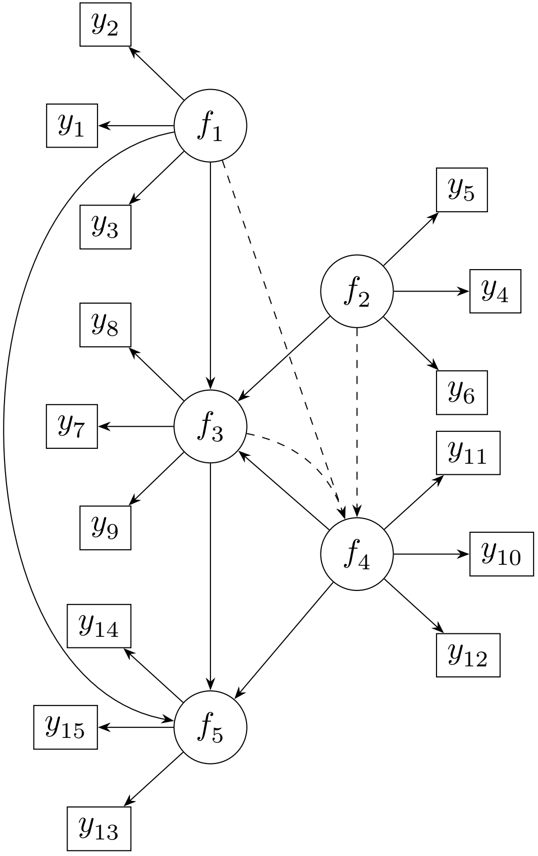

Figure 4: Note. Model 1.4: Error terms are not explicitly shown in the figure.

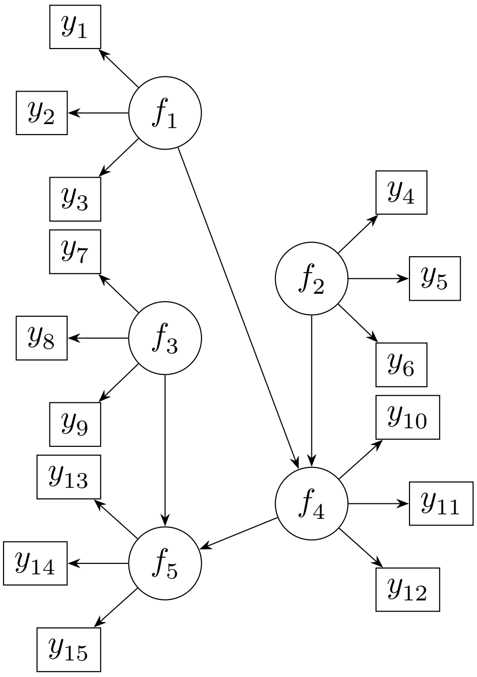

Figure 5: Note. Model 2.1: Error terms are not explicitly shown in the figure.

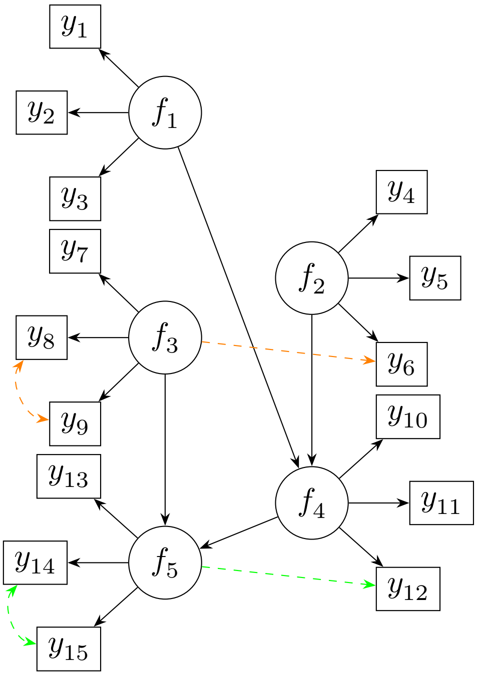

Figure 6: Note. Model 2.2: Error terms are not explicitly shown in the figure Orange lines represent misspecifications in the endogenous part of the model. Green lines represent misspecifications in the exogenous part of the model. These types of misspecifications result in different realisations of model 2.2 when they are modulated as factors in study 2 but are subsumed under one model here.

Study 1 and Study 2: The performance of four estimation methods was compared: - Vanilla SEM. - Global SAM (gSAM). - Local SAM with maximum likelihood (lSAM-ML). - Local SAM with unweighted least squares (lSAM-ULS).

Performance Measures

The following performance measures were captured: - Convergence rates. - Empirical relative biases. - Empirical coverage levels of 95% confidence intervals (CIs). - Root Mean Squared Errors (RMSE).

Software

All analyses were conducted in R Core Team (2023). See https://github.com/valentinkm/AdversarialSin for more details.

Analysis and Interpretation plan

The results were interpreted by descriptively comparing the performance measures (bias, MSE, convergence rates) of the different estimation methods under varying sample sizes, indicator reliability levels, and model misspecifications.

Results

In the following the main patterns and trends of the results are reported and only examplary of select conditions are shown for each study to illustrate the main findings. The full results can be found in the supplementary materials.

Study 1

Convergence Rate

For moderate and large sample sizes (N = 400, 6400), all methods achieved 100% convergence. At a small sample size (N = 100), SEM shows lower convergence rates, especially at lower reliability (0.3), while GSAM, SAM-ML, and SAM-ULS maintain high convergence rates. SEM faces significant convergence drops at the lowest reliability (0.3), particularly with misspecifications present.

Average Relative Biases

Here this study diverged from the original of Rosseel and Loh (2022) by focusing exclusively on the average relative bias under a single type of correctly specified model (model 1), which does not include any cross loadings or correlated residuals. This simplification, allowed to concentrate on the core advantage of SAM over traditional SEM: its more robust estimation of structural parameters, especially when measurement models are misspecified. For the correct model (1.1), SEM showed higher biases at smaller sample sizes and lower reliability, while GSAM, SAM-ML, and SAM-ULS exhibited lower biases. Under omitted cross loadings (model 1.2) and structural missepcifcation (model 1.4) , all methods demonstrated substantial biases, particularly SEM, at smaller sample sizes and lower reliability. With omitted correlated residuals and structural misspecification (1.3), GSAM, SAM-ML, and SAM-ULS showed more negative biases compared to SEM. Overall, GSAM, SAM-ML, and SAM-ULS were more robust with lower biases compared to SEM, especially under challenging conditions with smaller sample sizes and lower reliability.

Table 1 shows average relative bias percentages for different methods (SEM, GSAM, SAM-ML, SAM-ULS) under varying model specifications and sample sizes under under omitted correlated residuals (model 1.2) and Table 2 under correlated residuals and structural misspeifcation

Table 1: Average Relative Biases (in Percentages) under omitted Cross Loadings (Model 1.2)

N	Reliability	SEM	GSAM	SAM-ML	SAM-ULS
100	0.3	141.771	43.908	43.075	45.883
100	0.5	83.748	44.476	44.696	45.435
100	0.7	48.378	37.107	37.257	37.057
400	0.3	92.112	52.356	52.824	52.500
400	0.5	77.962	48.421	48.775	48.136
400	0.7	48.172	38.923	39.097	38.427
6400	0.3	81.625	54.110	54.676	53.948
6400	0.5	66.948	48.572	48.948	48.164
6400	0.7	46.667	39.142	39.316	38.585

Table 2: Average Relative Biases (in Percentages) under omitted Correlated Residuals (Model 1.3)

N	Reliability	SEM	GSAM	SAM-ML	SAM-ULS
100	0.3	-39.158	-56.090	-56.109	-55.530
100	0.5	-27.914	-32.240	-32.251	-30.610
100	0.7	-14.898	-16.446	-16.453	-15.897
400	0.3	-51.468	-53.692	-53.703	-53.543
400	0.5	-30.509	-31.317	-31.329	-31.132
400	0.7	-15.524	-15.859	-15.864	-15.711
6400	0.3	-52.334	-52.484	-52.500	-52.396
6400	0.5	-31.140	-31.225	-31.238	-31.117
6400	0.7	-15.893	-15.933	-15.938	-15.853

Coverage

For the correct model (1.1), SEM exhibited undercoverage at smaller sample sizes and lower reliability, indicating narrow CIs and less reliable estimates. GSAM, SAM-ML, and SAM-ULS showed slight overcoverage, reflecting more conservative estimates. Under the cross loadings model (1.2), all methods had substantial undercoverage, especially SEM. The correlated residuals model (1.3) consistently revealed undercoverage across all methods, with SEM showing the highest negative biases. For the structural model (1.4), SEM suffered from undercoverage at lower sample sizes and reliability, while GSAM, SAM-ML, and SAM-ULS performed better with relatively consistent biases. Overall, GSAM, SAM-ML, and SAM-ULS demonstrated better empirical coverage than SEM but the overcoverage indicates too wide confidence intervals.

Table 2 show the average difference between empirical coverage levels of the 95% confidence intervals (CIs) and their nominal level (95%) using each method for different reliability values and sample sizes under omitted cross loadings (model 1.2)

RMSE

For the correct model (1.1), SEM had higher RMSE values at smaller sample sizes and lower reliability, while GSAM, SAM-ML, and SAM-ULS demonstrated lower RMSE values. Under omitted cross loadings (1.2) and correlated residuals (1.3), SEM showed significantly higher RMSE values, although all methods exhibited increased RMSE under these

Table 3: Average Coverage Difference (in Percentages) under omitted Cross Loadings (Model 1.2)

N	Reliability	SEM	GSAM	SAM-ML	SAM-ULS
100	0.3	-1.008	3.671	3.650	3.451
100	0.5	-12.684	-2.277	-2.301	-3.112
100	0.7	-8.435	-4.447	-4.411	-5.164
400	0.3	-21.617	-6.320	-6.097	-6.914
400	0.5	-29.866	-16.801	-16.647	-17.457
400	0.7	-22.129	-15.936	-15.939	-16.359
6400	0.3	-65.231	-52.064	-51.036	-53.229
6400	0.5	-65.266	-54.583	-53.841	-55.944
6400	0.7	-55.541	-50.556	-50.237	-52.093

conditions. The structural model (1.4) revealed SEM had higher RMSE values at lower sample sizes and reliability, whereas GSAM, SAM-ML, and SAM-ULS showed better performance with lower RMSE. Overall, GSAM, SAM-ML, and SAM-ULS demonstrated more robustness with lower RMSE values compared to SEM, particularly in challenging conditions with smaller sample sizes and lower reliability.

Improper Solutions

Improper solutions occur when parameter estimates fall outside the boundary of the parameter space, such as when variances are estimated to be negative or correlations exceed an absolute value of one. These issues are more likely to arise in smaller sample sizes and can indicate problems with model estimation.

SEM exhibited a significantly higher count of improper solutions compared to GSAM, SAM-ML, and SAM-ULS, particularly in small sample sizes and lower reliability conditions. GSAM, SAM-ML, and SAM-ULS consistently showed minimal improper solutions across all scenarios, highlighting their robustness and reliability in parameter estimation.

Study 2

Convergence

In Study 2, all methods achieved a 100% convergence rate for moderate and large sample sizes (N = 400, 6400). For small sample sizes (N = 100), SEM showed lower convergence rates, especially at lower reliability levels (0.3), while GSAM, SAM-ML, and SAM-ULS maintained high convergence rates. R-squared values and measurement block size had minimal impact on convergence rates, indicating robustness of GSAM, SAM-ML, and SAM-ULS across various conditions, whereas SEM struggled particularly with lower reliability and smaller sample sizes.

The Influence of Number of Measurement Blocks on Bias and RMSE

For low reliability ($\lambda = 0.3$) and small sample size ($N = 100$), using 3 measurement blocks generally resulted in less negative bias compared to 5 blocks, although 5 blocks had lower RMSE. As reliability and sample sizes increased, the differences in bias and RMSE between 3 and 5 blocks diminished. Thus, for low reliability and small sample sizes, 3 measurement blocks were preferable for reducing bias, while 5 blocks performed better for RMSE. For subsequent comparisons, we considered lSAM with 3 measurement blocks as this difference was more pronounced especially in low sample size and reliability conditions.

Average Relative Bias

For the correctly specified model with low reliability ($\lambda = 0.3$) and small sample size ($N = 100$), traditional SEM showed a larger average relative bias compared to SAM methods. GSAM and lSAM-ML had small negative biases, while lSAM-ULS had a positive bias in challenging conditions. Higher R^2 had little impact on SEMs bias as absolute values, but SAM methods showed increased bias. For higher reliability and sample sizes, biases decreased for all methods.

Under misspecification (correlated residuals and crossloadings in the exogenous model), SEM had an even greater positive bias in challenging conditions, while SAM methods showed consistent negative biases, with SAM-ULS performing best.

In the endogenous misspecification model, SEM exhibited relatively small biases compared to SAM methods, which showed substantial negative biases. SEM's bias worsened with larger sample sizes and low reliability, while SAM methods' biases remained similarly large and negative.

For misspecifications in both endogenous and exogenous parts, SAM methods had consistent negative biases. SEM showed a positive bias with low reliability and sample size, turning negative with increasing sample size and higher R^2 values.

Table 4: Mean Relative Bias in Percentages for vanilla SEM and different SAM methods under Model Misspecifications (cross loadings and correlated residuals) in Exogenous and Endogenous Factors of the Model (Model 2.2-both)

N	Reliability	R^2	SEM	GSAM	SAM-ML	SAM-ULS
100	0.3	0.1	9.39	-24.32	-16.81	-15.15
100	0.3	0.4	6.44	-25.21	-22.61	-21.45
100	0.5	0.1	5.56	-9.54	2.57	3.01
100	0.5	0.4	-4.46	-12.24	-10.20	-10.51
100	0.7	0.1	3.73	-1.85	6.55	6.26
100	0.7	0.4	-2.76	-5.81	-4.33	-5.09
400	0.3	0.1	-12.30	-22.13	-16.74	-17.62
400	0.3	0.4	-17.65	-22.40	-22.73	-23.34
400	0.5	0.1	1.59	-8.36	1.20	-0.23
400	0.5	0.4	-8.48	-11.26	-10.92	-12.13
400	0.7	0.1	2.93	-0.95	3.85	2.93
400	0.7	0.4	-3.40	-5.10	-4.56	-5.83
6400	0.3	0.1	-16.73	-21.19	-17.64	-18.85
6400	0.3	0.4	-20.17	-21.73	-23.11	-24.04
6400	0.5	0.1	-1.78	-8.08	-1.13	-2.76
6400	0.5	0.4	-9.66	-11.15	-11.28	-12.74
6400	0.7	0.1	2.17	-0.79	2.13	1.27
6400	0.7	0.4	-3.49	-5.00	-5.21	-6.97

RMSE

For the correctly specified model with low reliability ($\lambda = 0.3$) and small sample size ($N = 100$), traditional SEM showed higher RMSE compared to SAM methods. GSAM and SAM-ML had lower RMSEs, while SAM-ULS had a slightly higher RMSE. Increasing regression weights (R^2) improved RMSE for both SEM and SAM methods. Under exogenous misspecifications, SEM had higher RMSE compared to SAM methods at low reliability and small sample size. This pattern persisted for endogenous misspecifications and misspecifications in both exogenous and endogenous factors, with GSAM and SAM-ML consistently showing lower RMSEs and SAM-ULS performing slightly worse. RMSE improved for all methods with higher R^2.

Improper Solutions

For small sample sizes ($N = 100$) and lower reliability ($\lambda = 0.3$), SEM exhibited a high percentage of improper solutions, while GSAM, SAM-ML, and SAM-ULS consistently showed very low or zero improper solutions across all conditions.

R Core Team. 2023. "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rosseel, Yves, and Wen Wei Loh. 2022. "A Structural After Measurement Approach to Structural Equation Modeling." *Psychological Methods*, No Pagination Specified–. https://doi.org/10.1037/met0000503.