

How robust are Structural Equation Models to model miss-specification? A simulation study

Lionel R. Hertzog

Terrestrial Ecology Lab, University of Ghent, K.L.
Ledeganckstraat 35, BE-9000 Gent.

Abstract

Structural Equation Models (SEMs) are routinely used in the analysis of empirical data by researchers spanning different scientific fields such as psychologists or econometricians. In some fields, such as in ecology, SEMs have only started recently to attract attention and thanks to dedicated software packages the use of SEMs have steadily increased. Yet, common analysis practices in such fields that might be transposed from other statistical techniques such as model acceptance or rejection based on p-value screening might be poorly fitted for SEMs. In this simulation study, SEMs were fitted via two commonly used R packages: lavaan and piecewiseSEM. Datasets were simulated under different modelling scenarios to test the impact of sample size and model complexity on various global and local model fitness indices. The results showed that not one single model indices should be used to decide on model fitness but rather a combination of different model fitness indices is needed. The global chi-square test for lavaan or the Fisher's C statistic are, in isolation, poor indicators of model fitness. Combining the different metrics explored here provided little safeguards against model overfitting, this emphasizes the need to cautiously interpret the inferred (causal) relations from fitted SEMs. I provide, based on these results, a tentative flowchart indicating how informations from different metrics may be combined to reveal model strength and weaknesses. Researchers in scientific fields with little experience in SEMs, such as in ecology, should consider and accept these limitations.

Keywords: R, lavaan, piecewiseSEM

Introduction

Structural equation models (SEMs) allow researchers to explicitly model measurement processes and complex inter-relations between variables [Kline and

Santor, 1999]. While classical regression analyses require one response variables and do not model correlations between explanatory variables, SEMs allow more flexibility where variables can be both depending on and affecting other variables. Other important features of SEMs include (but are not limited to): (i) possibility to model latent variables, theoretical and unobserved constructs, for instance motivation or stability, that affect a set of observed indicator variables, (ii) inclusion of mediation effects which allows the quantification of direct, indirect and total effects of one variable on another. Another appealing aspect to SEMs is the possibility to visually represent variables or constructs with boxes and draw complex relationships between them with arrows which potentially provide tighter links to theories [Grace, 2006]. Indeed, some have claimed that SEMs, if built correctly and carefully, have causal interpretation [Pearl, 2012] which would provide the possibility to test mechanisms and theories from observational data [Shipley, 2000b] something that is, at best, arduous in regression analysis.

While SEMs originated in the social sciences and have a long history both of development and of use by empirical researchers [i.e. Bentler and Weeks, 1980], in ecology SEMs are still perceived as a new method. Despite the presence of early ecological studies using SEMs [i.e. Power, 1972], there has been a recent surge of ecological studies using this approach driven by some influential papers [Grace et al., 2016], books written by and for ecologists [Shipley, 2000b, Grace, 2006] and the development of open-source, freely available R packages [Lefcheck, 2016, Rosseel, 2012]. It is interesting to note that while psychologists are wary in their use of causality statement preferring to leave out the exploration of mechanisms (Y. Rosseel pers. communication), ecologists are attracted to SEMs in part for its capacities to unravel mechanisms [Eisenhauer et al., 2015] a preeminent goal in current ecological research. The correctness of these assumptions have been discussed elsewhere [Petraitis et al., 1996, Pearl, 2012], and will not be considered further in the present article.

As in any other statistical framework, structural equation models require checking that the model fitted the data sufficiently well to proceed with interpretation of the fitted effects. While ample (psychological) literature exist on the issues of structural equation model testing [Bollen and Long, 1992], differences in terminology and publication in journals not read by researchers in other fields prevent the dissemination of methodological knowledge and best practices of SEMs testing across fields. In addition, SEMs being relatively new in the field of ecology there is little to none formal training on SEMs for (under)graduate students in ecology. As a result, techniques and traditions from other statistical framework such as ANOVAs or regressions might be applied to SEMs leading to sub-optimal or even biased inference. Books on SEMs written for ecologists do provide some guidance regarding model testing, for instance Shipley [2000b] discuss at length the chi-square test on fitted covariance matrices and alternative fitness tests, while Grace [2006] provide some guidance on different model evaluation strategies based on different approaches from strictly confirmatory to purely exploratory. Exploring the limits and the strengths of modelling framework can be performed with closed-form equations for simple models, such as for instance the power of ANOVA to detect differences between means. For more complex modelling framework such as SEM, simulation approaches can be handy to explore sensitivity of model checks or power to detect effects [Bolker, 2008].

Several software packages already exist to simulate datasets based on SEMs. The `simsem` package for R [Pornprasertmanit et al., 2016], for instance, allows great flexibility in fitting different types of SEMs with or without latent variables and structural equations but these packages usually assume that the correct model structure is known and so generate synthetic data from that model. In this paper I present a package to simulate simple SEMs with no latent variables, but with possibility to generate data following different scenarios in order to compare limits of model checking under different model miss-specification. In addition, two approaches to fit SEMs are compared, one with global estimation of parameters using the covariance matrix [Rosseel, 2012] and one with piecewise estimation of model parameters in individual linear models combined by checking independence claims [Lefcheck, 2016]. The aim is to compare: (i) model fitting strategy (global vs piecewise), (ii) model miss-specification type, (iii) sample size and (iv) model complexity on various common metrics used to check model fitness.

Methods

Structural equation models were created with varying numbers of covariates, ranging from 5 to 10. All covariates were observed, in other words there are no latent or composite variables in the models. The relations between the covariates were randomly generated to create a direct acyclic graph with a fixed connectance of 0.3, this resulted in models with 7 up to 30 relations.

Data were generated based on 5 scenarios: (i) random, (ii) exact, (iii) shuffled, (iv) complex and (v) simple, see Figure 1. In the random scenario, data were generated as random normal deviates with a mean of 0 and a standard deviation of 1 without any signal between covariates. In the exact scenario, covariates were sequentially generated following exactly the created model. Basically, the created models ensured that there was at least one exogenous covariate, the exact data generation first generated random uniform deviates for the exogenous covariates. Then for each subsequent covariates, regression coefficients were generated following a cauchy distribution (location of 0 and a scale of 2.5). Linear predictors were derived by combining the drawn regression coefficient and the effect of the other covariates. These were used as mean values for generating random normal data with a fixed standard deviation of 1. In the shuffled scenario, a fixed proportion of relations were reversed, in other words when the model set $A \rightarrow B$, a shuffled relation is $B \rightarrow A$. 25% of the relations were shuffled across model sizes, relations to be shuffled were randomly selected. Then the shuffled relation matrix was used to generate the data following the same steps as in the exact data generation. In the complex scenario, 25% of the relations from the model were randomly selected and dropped during the data generation process. For instance, the model might assume $A + B \rightarrow C$ but in the data generation we have $A \rightarrow C$. In other words in the complex scenario the assumed model is too complex compared to the data. The simple scenario is the opposite of the complex one, namely 25% additional relations are added during the data generation process which result in a model being too simple compared to the data. The models were then fitted to the data using `piecewiseSEM` v1.2 [Lefcheck, 2016] or `lavaan` v0.5 [Rosseel, 2012].

To further explore the impact of the signal / noise ratio on the simulation

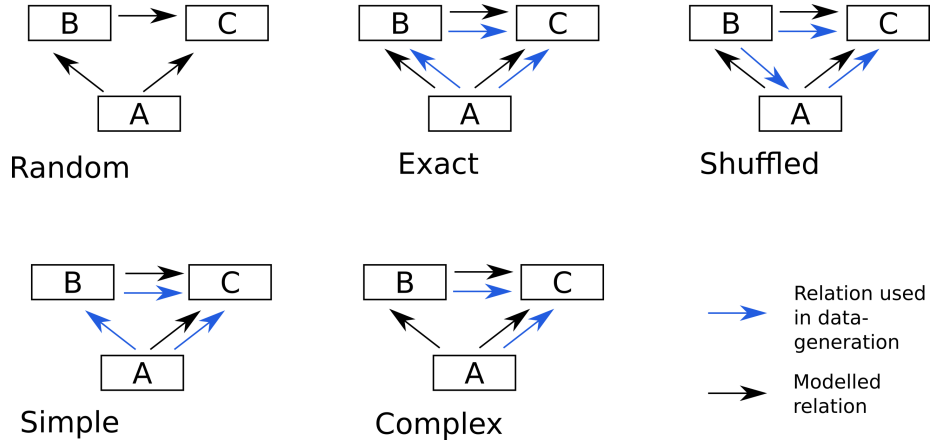


Figure 1: Schematic representation of the different data-generation scenarios. A, B and C represent hypothetical variables, the black arrows show the relationships assume in the fitted models and the blue arrow the relationship present during the data-generation process.

results additional simulations were ran. The residual standard deviation (noise) was varied taking values of 0.5, 1 and 2.5 crossed with variations in the scale of the cauchy distribution (signal) which took values of 1, 2.5 and 5. All results from these simulations are presented in the Appendix.

Four metrics were taken from the models: (i) the acceptance of the model based on the p-value of the Fischer’s C score in piecewiseSEM and on the p-value of the chi-square test in lavaan. Models with p-values higher than 0.05 were considered as accepted. In lavaan some of the models did not converge, these were considered as being rejected. (ii) The number of significant regression coefficients, so with p-values lower than 0.05. (iii) The average R-square values of the model derived from the adjusted R-square of the individual linear models composing the model. (iv) The proportion of conditional independence tests implied by the model structure that failed (p-value >0.05), this was computed based on the dagitty package v0.2 [Textor and van der Zander, 2016].

In the simulations, the sample size was varied from 20 to 100 with an increment of 10, the number of covariates in the models was varied from 5 to 10, the connectance in the resulting graph was kept constant at 0.3 and the models were fitted both using piecewiseSEM and lavaan. In total this generated 108 parameter sets, each were replicated 100 times to account for stochasticity in the data generation. Per parameter set the results were summarized as followed: (i) proportion of accepted models, (ii) proportion of significant paths, (iii) average of R-square values, (iv) average of proportion of failed conditional independence tests. All the code used here is available from: <https://github.com/lionel68/Blog/tree/master/SimSEM>

Results

Proportion of accepted models

Two groups of data generation scenarios showed up when looking at the proportion of accepted models (Figure 2), in the first one was: random, exact and complex scenarios, in the second one was: shuffled and simple scenarios. The first group always had larger acceptance than the second. Sample size and the number of covariates had little effect on models fitted via piecewiseSEM for the random, exact and complex scenarios, the acceptance proportion was constant at around 0.90. For the same data-generation scenarios with models fitted via lavaan, the proportion of accepted models tended to asymptotically increase with sample size, this effect was stronger for larger number of covariates. For the other two scenarios piecewiseSEM and lavaan led to almost identical acceptance rates. The shuffled scenario had higher model acceptance rate than the simple scenario for low number of covariates (5-7), with larger number of covariates (8-10) both scenarios had almost equivalent acceptance rates. With increasing number of covariates in the models the acceptance rate of these scenarios dropped, for the shuffled scenario acceptance rate was around 50% for 5 covariates but around 5% for 10 covariates. For the simple scenario acceptance rate dropped from around 20% to 5% as covariates number goes from 5 to 10.

Proportion of significant paths

Lavaan and piecewiseSEM led to an almost identical pattern for the proportion of significant paths (Figure 3). A clear and constant ordering of the data-generation scenarios was found across the number of covariates and the sample size: simple > shuffled > exact > complex > random. In the random data-generation scenario very low proportion of paths were significant, around 5%, close to the type I error rate. For the other scenarios the proportion of accepted paths increased slightly with the sample size. Interestingly, the exact scenario never reached the desired power in statistical analysis (80%), it was rather around 70%.

R-square

No consistent differences could be found between models fitted via lavaan or piecewiseSEM on the average R-square of the models (Figure 4). The random data-generation scenario had very low adjusted R-square, close to 0%. For the other scenarios a consistent ordering was again found across sample size and number of covariates, namely: exact > shuffled and complex > simple. Sample size had no effect on the R-square, but the number of covariates did. R-squares tended to be larger in models with higher number of covariates, for instance for the exact data-generation scenario, R-square was around 80% for 5 covariates, but it was around 90% for 10 covariates. So across data-generation scenarios, R-square increased by around 10% between 5 and 10 covariates.

Conditional independence

Two groups of scenarios appeared for the proportion of failed conditional independence, namely: (i) exact, random and complex, (ii) shuffled and simple



Figure 2: Effect of sample size, model complexity (number of covariates and parameters), data-generation process and model fitting type on the proportion of accepted structural equation models. Reported is the proportion of accepted models (p-values > 0.05) across 100 replications per parameter set (sample size, complexity, data-generation, model type). Models which failed to converge were considered as being rejected.

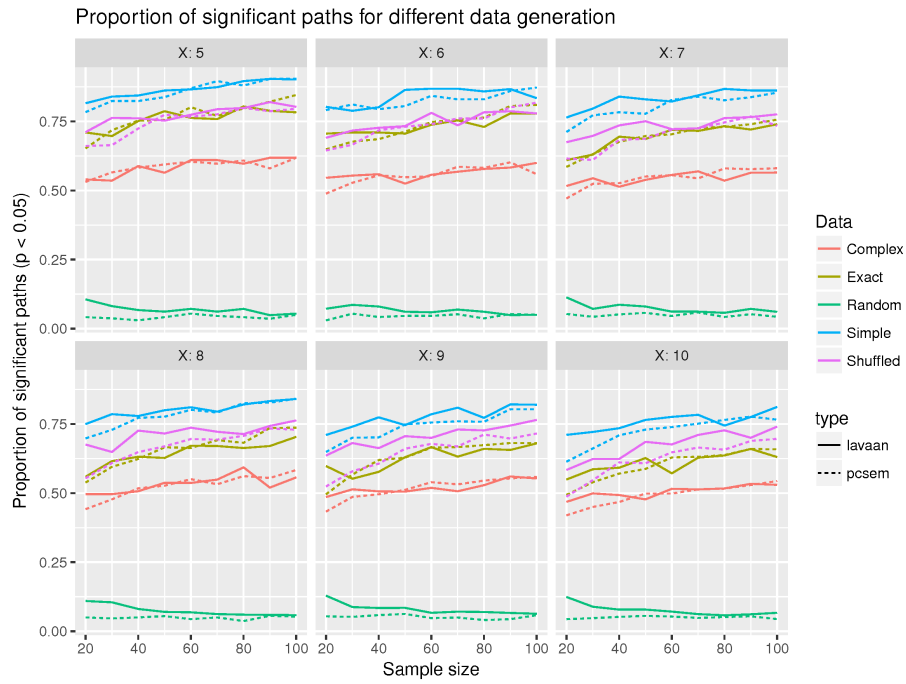


Figure 3: Effect of sample size, model complexity (number of covariates and parameters), data-generation process and model fitting type on the proportion of significant relations (p -value of individual paths < 0.05). Reported is the average proportion of significant relations across 100 replications per parameter set.

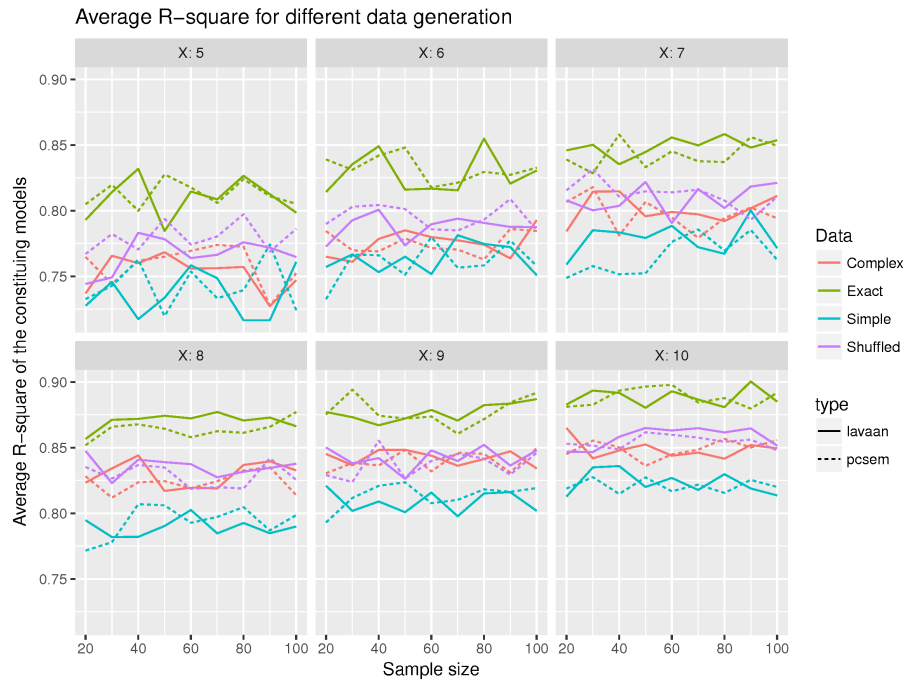


Figure 4: Effect of sample size, model complexity (number of covariates and parameters), data-generation process and model fitting type on the average R-square of the individual model regression. For each fitted model the adjusted R-square of the individual linear regression was extracted and averaged across 100 replications per parameter set.

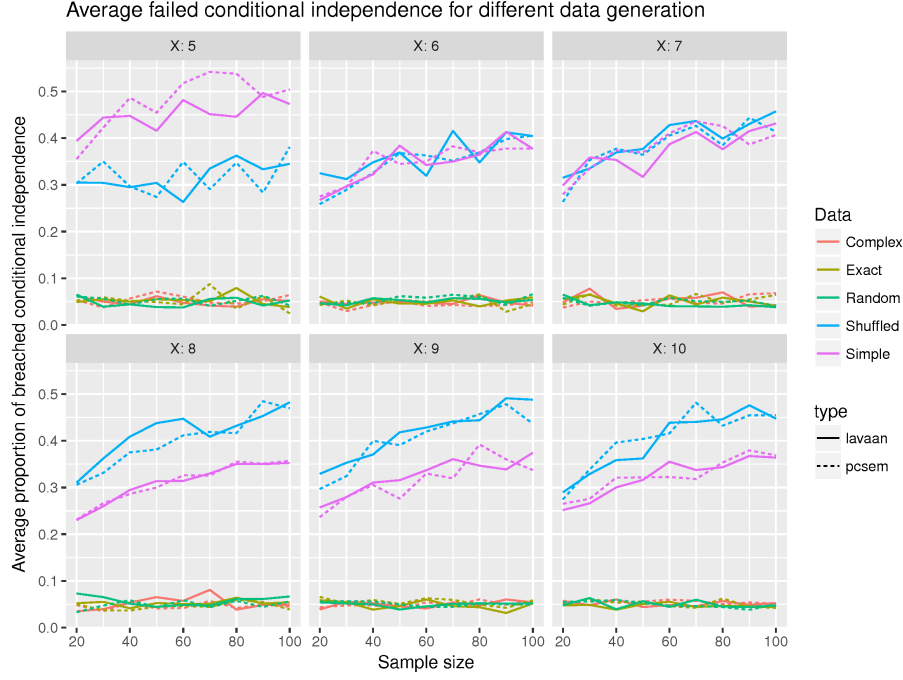


Figure 5: Effect of sample size, model complexity (number of covariates and parameters), data-generation process and model fitting type on the average proportion of conditional independence tests that failed (p-value < 0.05).

(Figure 5). In the first group the average proportion of failed conditional independence tests was consistently low at around 5% close to the type I error rate. For the other two scenarios, failure tended to increase with sample size. The effect of the number of covariates had opposite direction between the shuffled and simple scenario. While increasing the number of covariates increased the proportion of failures from around 30% to around 40% for the simple data-generation scenario, it reduced the proportion of failures from around 45% to around 30% for the shuffled data-generation one.

Signal / noise ratio effects

Independently varying the strength of the signal and the amount of noise did not substantially affect the results. All figures relating to the signal / ratio variation are shown in the Appendix.

Discussion

A first general pattern emerging from the simulations is the strong similarity between piecewiseSEM and lavaan. In other words, fitting SEM with piecewise assemblage of individual regressions or via the variance-covariance matrix leads to similar pattern in the explored metrics. Researchers fitting simple linear models with or without hierarchical structure or latent variables may use

both approaches indiscriminately. In more complex settings restrictions set by the two approaches may constrain the decision to use one or the other. For instance, latent variables structure are only available in lavaan, while piecewiseSEM allow non-normal regression and hierarchical structure. On a side note, departure from normality does not have strong impact on the chi-square p-value as reported by Shipley [2000b], in addition lavaan provide robust statistic that control for some of the non-normality (santorra-bentley stuff).

The first metric that researcher check on a fitted SEM is the p-value of the chi-square test on the covariance matrix for lavaan and the p-value for the Fisher's C statistic in piecewiseSEM. If this p-value is larger than 0.05 standard practice is to accept the model and explore the relationships further to unravel the relations between the covariates [Grace, 2006]. This approach is vulnerable to model miss-specification. The biggest concern being is that models may be readily accepted even if there is no signal in the data. Also relying solely on one p-value to accept model fit will tend to result in building overly complex models. Indeed the simulations, showed that models more complex than the processes that generated the data will be more often accepted than models being simpler than the data-generation process. Finally, reversing the direction of the effects lead to large decrease in acceptance rates of the models, so model rejection might also indicate some miss-specification in the directionality of the effects. In psychology, issues related with the global chi-square test and its associated p-values have been known for a long time [Bollen and Long, 1992]. This body of literature emphasize the fact that there cannot beis not one metric that can reliably tell if the SEM fit is acceptable or not [see also Kline and Santor, 1999]. These authors argued already at that time that exploration of the model component such as the magnitude of the path coefficient or the R-square should be inherent to any SEM model check. One of the most influential book on SEM in ecology [Shipley, 2000b] meticulously explores the issue of sample size and non-normality of the errors on the chi-square test and provide some solutions such as bootstrapping for complex cases. The use of such fixes is rare if not totally absent in the current ecological literature. piecewiseSEM use a different approach to model checking based on the independence claims derived from the models. I am not aware of any simulation or study exploring the impact of sample size on the Fisher's C statistic. This simulation showed that, for the range of sample size value used, model miss-specification had much stronger impacts on model checking than sample size. Researchers should therefore acknowledge that large sample size do not protect against error in model specification.

The next step in SEM inference is the exploration of the fitted relations between the covariates. The result of the simulation show little power for the detection of effects based on the p-values of the regression coefficients. Most striking is the fact that even when generating the data exactly as implied by the model only approximately 70% of the effects were detected. Interestingly, the direction of the effects had no impact on their significance. Indeed, the exact and shuffled data-generation had similar levels of significant paths. Deriving causality statements from fitted SEM should not be based on solely screening the significance of the regression coefficients. However, exploring coefficient significance is a strong test that some signal is present in the data and was captured in the models. The random data-generation revealed that the number of significant paths even under complete noise was close to the 5% level of the type I error. So researchers should be ready to modify or reject models that,

despite being accepted based on the chi-square statistics, have suspiciously low number of significant paths. To my limited knowledge, there is little literature about the power of SEM to detect individual effects. Wolf et al. [2013] reported power and bias to detect direct and indirect effects in a three latent-variables model each with three indicator variables. They reported that for detecting weak direct effects with a power of 80% sample size should be larger than 300, for detecting the indirect effect sample size should be higher than 400. Such sample sizes are rarely reached in ecological studies that usually consist of complex models [Scherber et al., 2010, Duffy et al., 2015].

The R square metric represent, in the case of linear regressions, the amount of variation explained by the data. Ecologists readily use R-square when reporting statistical analysis, this is also true for SEM [Duffy et al., 2015]. The simulations showed that the average R-square was a good indicator that some signal present in the data has been extracted by the model. There are, however, some limitations to what the average R-square can provide. Adding new covariates mathematically increases R-square, a pattern that is well-known, therefore R-square is most efficiently used to compare models with the same number of covariates. Also in SEMs, the R-square metric tend to favor more complex models over simpler ones.

The structure of SEM imply some conditional independence between covariates, these conditional independence can be tested and provides good indication if and where a SEM fails [Thoemmes et al., 2017]. The simulations showed that conditional independence tests are blind to random noise, data randomly generated create no conditional independence failures. Similarly conditional independence test will tend to favor more complex models, since models simpler than the process generating the data will be flagged by such tests, while models more complex than data-generation have very low levels of failures in conditional independence. More positively, such a metric can readily detect miss-specification of the directionality of the effects. In summary, local tests of conditional independence may provide good guidance to identify missing or miss-directed links, but these tests do not prevent from model overfitting. Shipley [2000a] already discussed techniques to derive all conditional independence implied by direct acyclic graphs and the advantages that such an approach provides. Most interesting is the possibility to identify parts of poor fit in the model to provide guidance for further model and/or theory improvements.

Combining all results lead to some recommendations regarding post-fitting checks of structural equation models, also summarized in Figure 6:

- The results re-emphasize the need to explore structural equation model fitness via several metrics
- Seemingly good fitting models with large p-values of the chi-square test or the C statistic with few or no significant paths and low R-squares imply that no real signal was extracted from the data
- Inversely poorly fitting models in regards to global p-values but having large number of significant paths, indication of un-modelled conditional dependence and/or large R-square may indicate missing relations between covariates (underfitting) or errors in the directionality of the effects
- The explored metrics provide little safeguards against fitting models with

larger numbers of relations than necessary (overfitting), this caution against validating implied causal structure from the tested fitness metrics alone.

Conclusion

Structural equation modelling is a powerful inferential framework but like any other approach it does not perform magic. SEMs have different histories and traditions in different scientific fields at different times. While SEMs have been widely used for a long time in psychology and are therefore an inherent part of the training of researcher in psychology. In other fields such as in ecology, SEMs gathered attraction more recently thanks in part to influential papers [Grace et al., 2016, Scherber et al., 2010], books [Shipley, 2000b, Grace, 2006] and R packages [Rosseel, 2012, Lefcheck, 2016]. In these fields SEMs might feel like a brave new world and little of the necessary caution is applied in deriving interpretation from the models [but see Grace et al., 2010]. The aim of this simulation paper was to reveal the limits of routinely used metrics from fitted SEMs to model miss-specification. The main conclusions are: (i) the failure of global fitness metric to separate signal from noise, (ii) the limited power of structural equation models to identify relation between variables for small datasets (20 - 100 sample units) and small models (7 to 30 parameters) and (iii) the need to explore local fitness of the constituting regressions in structural equation models.

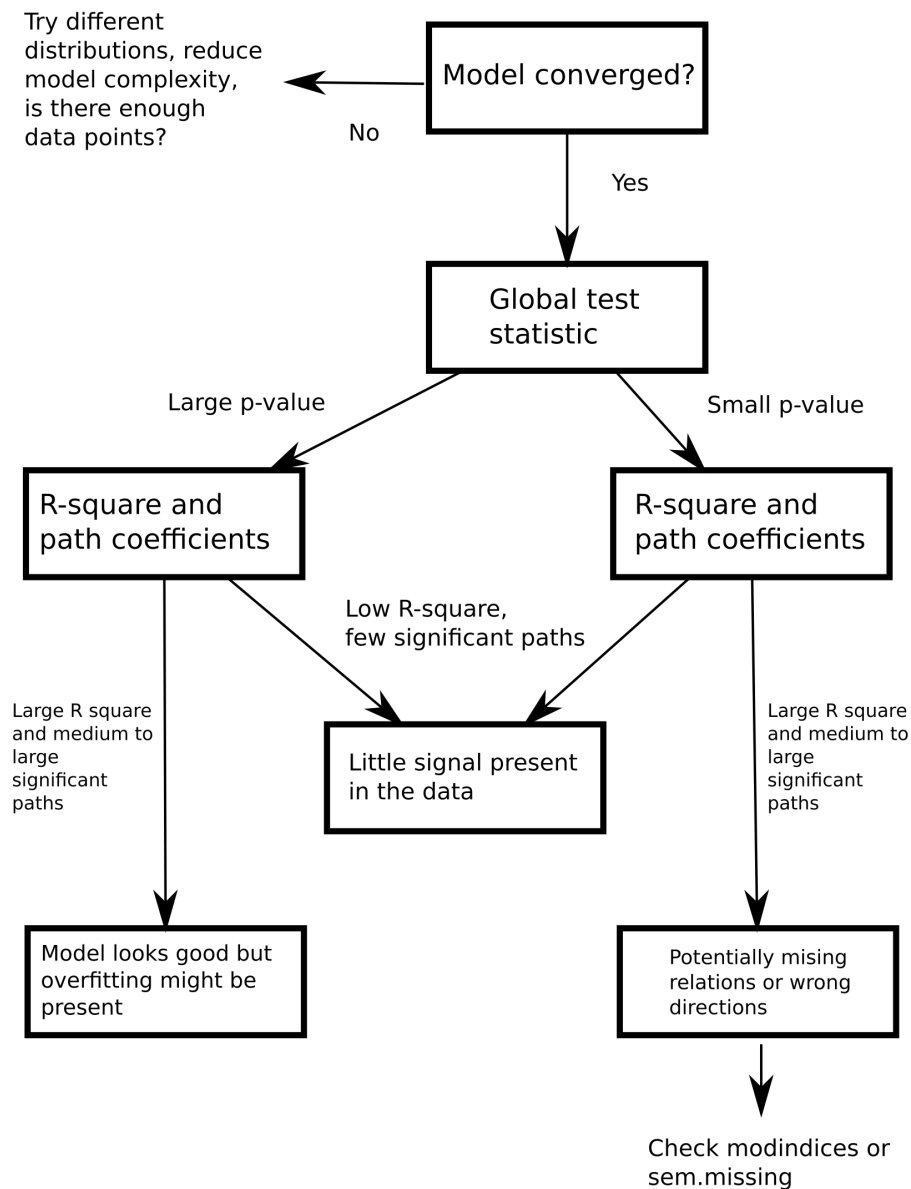


Figure 6: Flowchart describing how the information from the different metrics studied here maybe be combined to inform about model fitness. Issues of model convergence should lead to model re-specification as the estimates of unconverged models may not be trusted.

Acknowledgements

I thank Yves Rosseel and Johnathan Lefcheck for positive feedback and discussion of early simulation results. Bram Sercu, Femke Batsleer for providing helpful textual corrections.

References

- Peter M Bentler and David G Weeks. Linear structural equations with latent variables. *Psychometrika*, 45(3):289–308, 1980.
- Benjamin M Bolker. *Ecological models and data in R*. Princeton University Press, 2008.
- Kenneth A Bollen and J Scott Long. Tests for structural equation models: introduction. *Sociological Methods & Research*, 21(2):123–131, 1992.
- J Emmett Duffy, Pamela L Reynolds, Christoffer Boström, James A Coyer, Mathieu Cusson, Serena Donadi, James G Douglass, Johan S Eklöf, Aschwin H Engelen, Britas Klemens Eriksson, et al. Biodiversity mediates top-down control in eelgrass ecosystems: a global comparative-experimental approach. *Ecology letters*, 18(7):696–705, 2015.
- Nico Eisenhauer, Matthew A Bowker, James B Grace, and Jeff R Powell. From patterns to causal understanding: structural equation modeling (sem) in soil ecology. *Pedobiologia*, 58(2):65–72, 2015.
- James B Grace. *Structural equation modeling and natural systems*. Cambridge University Press, 2006.
- James B Grace, T Michael Anderson, Han Olff, and Samuel M Scheiner. On the specification of structural equation models for ecological systems. *Ecological Monographs*, 80(1):67–87, 2010.
- James B Grace, T Michael Anderson, Eric W Seabloom, Elizabeth T Borer, Peter B Adler, W Stanley Harpole, Yann Hautier, Helmut Hillebrand, Eric M Lind, Meelis Pärtel, et al. Integrative modelling reveals mechanisms linking productivity and plant species richness. *Nature*, 529(7586):390, 2016.
- Rex B Kline and Darcy A Santor. Principles & practice of structural equation modelling. *Canadian Psychology*, 40(4):381, 1999.
- Jonathan S Lefcheck. piecewissem: piecewise structural equation modelling in r for ecology, evolution, and systematics. *Methods in Ecology and Evolution*, 7(5):573–579, 2016.
- Judea Pearl. The causal foundations of structural equation modeling. Technical report, CALIFORNIA UNIV LOS ANGELES DEPT OF COMPUTER SCIENCE, 2012.
- PS Petraitis, AE Dunham, and PH Niewiarowski. Inferring multiple causality: the limitations of path analysis. *Functional ecology*, pages 421–431, 1996.

- Sunthud Pornprasertmanit, Patrick Miller, and Alexander Schoemann. *simsem: SIMulated Structural Equation Modeling*, 2016. URL <https://CRAN.R-project.org/package=simsem>. R package version 0.5-13.
- Dennis M Power. Numbers of bird species on the california islands. *Evolution*, 26(3):451–463, 1972.
- Yves Rosseel. Lavaan: An r package for structural equation modeling and more. version 0.5–12 (beta). *Journal of statistical software*, 48(2):1–36, 2012.
- Christoph Scherber, Nico Eisenhauer, Wolfgang W Weisser, Bernhard Schmid, Winfried Voigt, Markus Fischer, Ernst-Detlef Schulze, Christiane Roscher, Alexandra Weigelt, Eric Allan, et al. Bottom-up effects of plant diversity on multitrophic interactions in a biodiversity experiment. *Nature*, 468(7323): 553, 2010.
- Bill Shipley. A new inferential test for path models based on directed acyclic graphs. *Structural Equation Modeling*, 7(2):206–218, 2000a.
- Bill Shipley. *Cause and correlation in biology*. Cambridge University Press, 2000b.
- Johannes Textor and Benito van der Zander. *dagitty: Graphical Analysis of Structural Causal Models*, 2016. URL <https://CRAN.R-project.org/package=dagitty>. R package version 0.2-2.
- Felix Thoemmes, Yves Rosseel, and Johannes Textor. Local fit evaluation of structural equation models using graphical criteria. 2017.
- Erika J Wolf, Kelly M Harrington, Shaunna L Clark, and Mark W Miller. Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and psychological measurement*, 73(6):913–934, 2013.

Appendix

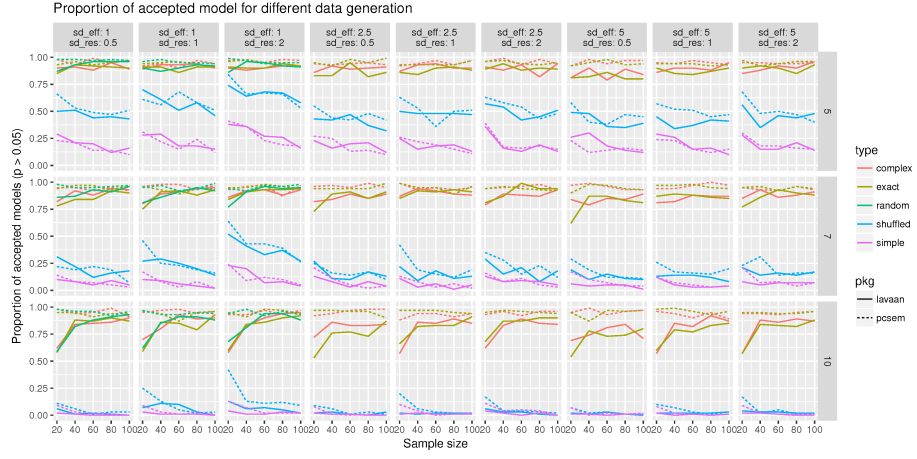


Figure A.1: Proportion of accepted models with varying sample size (x-axis), data-generation (colors), package type (linetype), number of covariates (rows) and signal / noise ratio (columns).

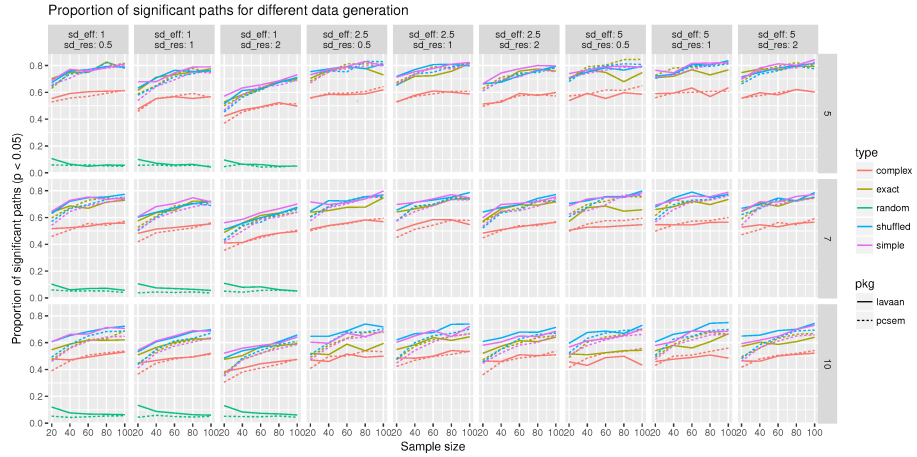


Figure A.2: Proportion of accepted models with varying sample size (x-axis), data-generation (colors), package type (linetype), number of covariates (rows) and signal / noise ratio (columns).

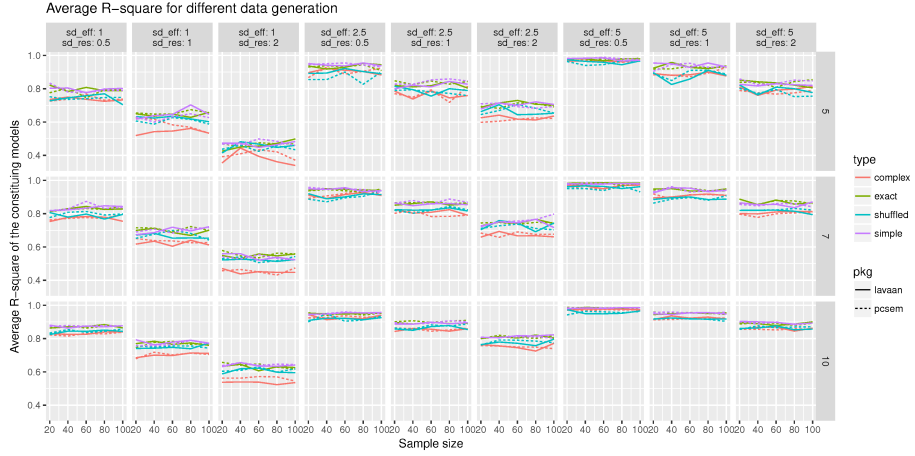


Figure A.3: Proportion of accepted models with varying sample size (x-axis), data-generation (colors), package type (linetype), number of covariates (rows) and signal / noise ratio (columns).

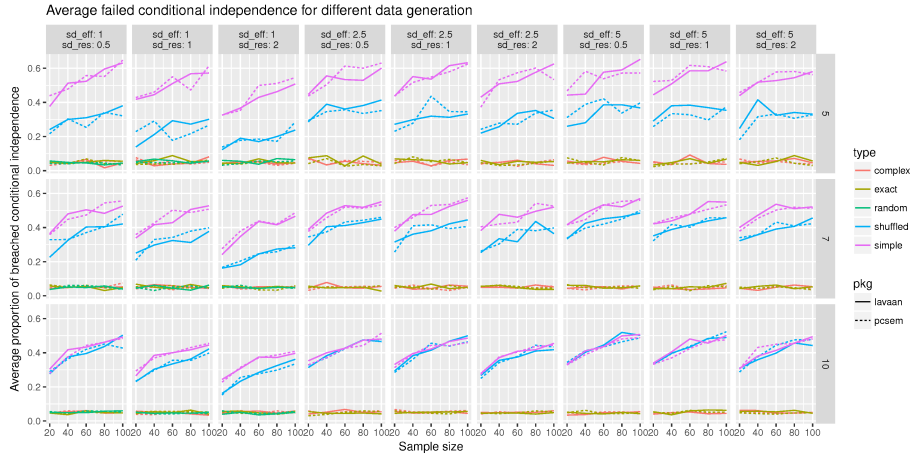


Figure A.4: Proportion of accepted models with varying sample size (x-axis), data-generation (colors), package type (linetype), number of covariates (rows) and signal / noise ratio (columns).