Getting a Step Ahead: Using the Regularized Horseshoe Prior to Select Cross-Loadings in Bayesian CFA

Research Report

Michael Koch (6412157)

Methodology and Statistics for the Behavioral, Biomedical, and Social Sciences

Supervisor: Dr. Sara van Erp

Email: j.m.b.koch@students.uu.nl

Word Count: 2477

Intented Journal of Publication: Structural Equation Modeling

The art of statistical modeling revolves around coming up with an appropriate simplification, a *model*, of a true *data-generating process*. Hereby, a fundamental trade-off between model simplicity and model complexity arises, that is mostly known as *bias-variance trade-off*. Simple models with few parameters have high bias, meaning that they deviate substantially from the true data-generating process, and low variance, such that they generalize well to other datasets from the same population. Moreover, simple models are easily identified and easy to interpret. Complex models with large numbers of parameters tend to have low bias and high variance. They are thus prone to over-fitting, i.e. picking up patterns that are only relevant in the dataset at hand, but do not generalize well to other datasets. Moreover, complex models can be cumbersome to interpret and often a large number of observations is required to estimate them (Cox, 2006; James, Witten, Hastie, & Tibshirani, 2021).

Regularization

A classic method of trying to find a balance in model complexity and model simplicity is *regularization* (Hastie, Tibshirani, & Wainwright, 2015). Regularization entails adding some bias to a model on purpose to reduce its variance. This helps to make models easier to interpret and more generalizable. In a frequentist context, regularization is achieved by adding a penality term to the cost function of a model. This ensures that model parameters that are irrelevant, e.g. small regression coefficients in a regression model with a large number of predictors, are shrunken to (or towards) zero. In a Bayesian context, the same is achieved by setting a so-called shrinkage-prior for the parameters (Van Erp, Oberski, & Mulder, 2019). The well-known ridge- (Hoerl & Kennard, 2000) and lasso-penalization (Tibshirani, 1996) in regression correspond to setting a ridge-prior (Hsiang, 1975) or a Laplace-prior (Park & Casella, 2008) for regression coefficients respectively.

Bayesian CFA: The Small Variance Normal Prior (SVNP)

In Confirmatory factor analysis (CFA, Bollen, 1989) it is common practice to deal with the bias-variance trade-off in a brute-force manner, by imposing a so-called simple structure. Here, cross-loadings, factor loadings that relate items to factors that they theoretically do not belong to, are fixed to zero to yield an identified and interpretable model. This often leads to poor model fit, which forces researchers to free some cross-loadings after the fact based on empirical grounds (modification indices) to improve fit. This procedure is flawed, as it risks capitalization on chance and thereby over-fitting (MacCallum, Roznowski, & Necowitz, 1992).

As solution to the issue Muthen and Asparouhov (2012) proposed *Bayesian CFA*, an alternative approach for identifying CFA models, which can be viewed as a form of regularized SEM (see Jacobucci, Grimm, & McArdle, 2016 for a frequentist perspective on regularized Structural Equation Modeling). Rather than identifying models by fixing *all* cross-loadings to zero, one should assume that *most* cross-loadings are zero. This is achieved by setting the so-called *Small Variance Normal Prior* (SVNP) for the cross-loadings, which is a normal distribution with mean zero and a very small variance (e.g. $\sigma^2 = 0.01$). This prior has a large peak at zero, and very thin tails (Figure 1). Hence, it attaches large prior mass to cross-loadings of or near zero, while attaching almost no prior mass to cross-loadings further from zero. Consequently, all cross-loadings in the model are shrunken. The larger the prior's variance, the more admissive the model is in the amount of deviation from zero it allows.

An issue with Muthen and Asparouhov (2012)'s Bayesian CFA is that not only the cross-loadings close to zero, which are considered irrelevant, are shrunken to zero, as desired. Also the ones further from zero are shrunken heavily towards zero, which introduces bias (Lu, Chow, & Loken, 2016). First, bias naturally occurs in the large cross-loadings itself. However, also in other parameters, such as factor-correlations or

main-loadings, substantial bias can arise, as they are estimated conditionally on the cross-loadings. Consequently, Bayesian CFA requires two steps in practice. First, the model is estimated with the SVNP set for the cross-loadings. In the original approach, cross-loadings are then selected as non-zero when their 95% credible intervals does not contain zero (Muthen & Asparouhov, 2012). The model is then re-estimated, where cross-loadings that have been selected to be non-zero are freely estimated without shrinkage, and the remaining cross-loadings are fixed to zero, avoiding the bias in the model of the previous step. Correctly selecting cross-loadings as non-zero can pose a challenge in practice, as the performance of different selection criteria depends on a broad set of conditions, making it difficult to formulate general recommendations for researchers (Zhang, Pan, & Ip, 2021). It is thus desirable to identify shrinkage-priors that can regularize CFA models without causing substantial bias, within a single step.

The Regularized Horseshoe Prior (RHSP)

A particularly promising candidate is the so-called *Regularized Horseshoe Prior* (RHSP, Piironen & Vehtari, 2017a, 2017b). This prior is an extension of the Horseshoe Prior (Carvalho, Polson, & Scott, 2010). The main idea of both priors is that there is a *global shrinkage parameter* τ, shrinking all cross-loadings to zero, and a *local shrinkage parameter* $\bar{\omega}_{jk}$[1] that allows truly large cross-loadings to escape the shrinkage. The issue with the original Horseshoe Prior is that not shrinking large parameters at all can lead to identification issues (see Ghosh, Li, & Mitra, 2018). The RHSP overcomes this by shrinking also large parameters a little bit, as it is designed such that for large parameters the prior approaches a normal (slab) prior with mean zero and variance c^2 (Piironen & Vehtari, 2017b).

[1] We deviate from the common notation of the local shrinkage parameter as $\bar{\lambda}$, as this letter is commonly used to denote factor loadings in CFA.

For every cross-loading of factor j on item k:

$$\lambda_{c,jk}|\bar{\omega}_{jk}, \tau, c \sim \mathcal{N}(0,\ \bar{\omega}_{jk}^2\tau^2),\ with\ \bar{\omega}_{jk}^2 = \frac{c^2\omega_{jk}^2}{c^2 + \tau^2\omega_{jk}^2},$$

$$\tau|df_{global}, s_{global} \sim half - t_{df_{global}}(0,\ s_{global}^2),\ with\ s_{global} = \frac{p_0}{p - p_0}\frac{\sigma}{\sqrt{N}},$$

$$\omega_{jk}|df_{local}, s_{local} \sim half - t_{df_{local}}(0,\ s_{local}^2),$$

$$c^2|df_{slab}, s_{slab} \sim \mathcal{IG}(\frac{df_{slab}}{2},\ df_{slab} \times \frac{s_{slab}^2}{2}),$$

where p_0 represents a prior guess of the number of relevant cross-loadings. It is, however, not necessary to use p_0. One can simply set s_{global} manually, whereby it is worth to consider that a s_{global} created based on a p_0 will typically be much lower than 1 (Piironen & Vehtari, 2017b).

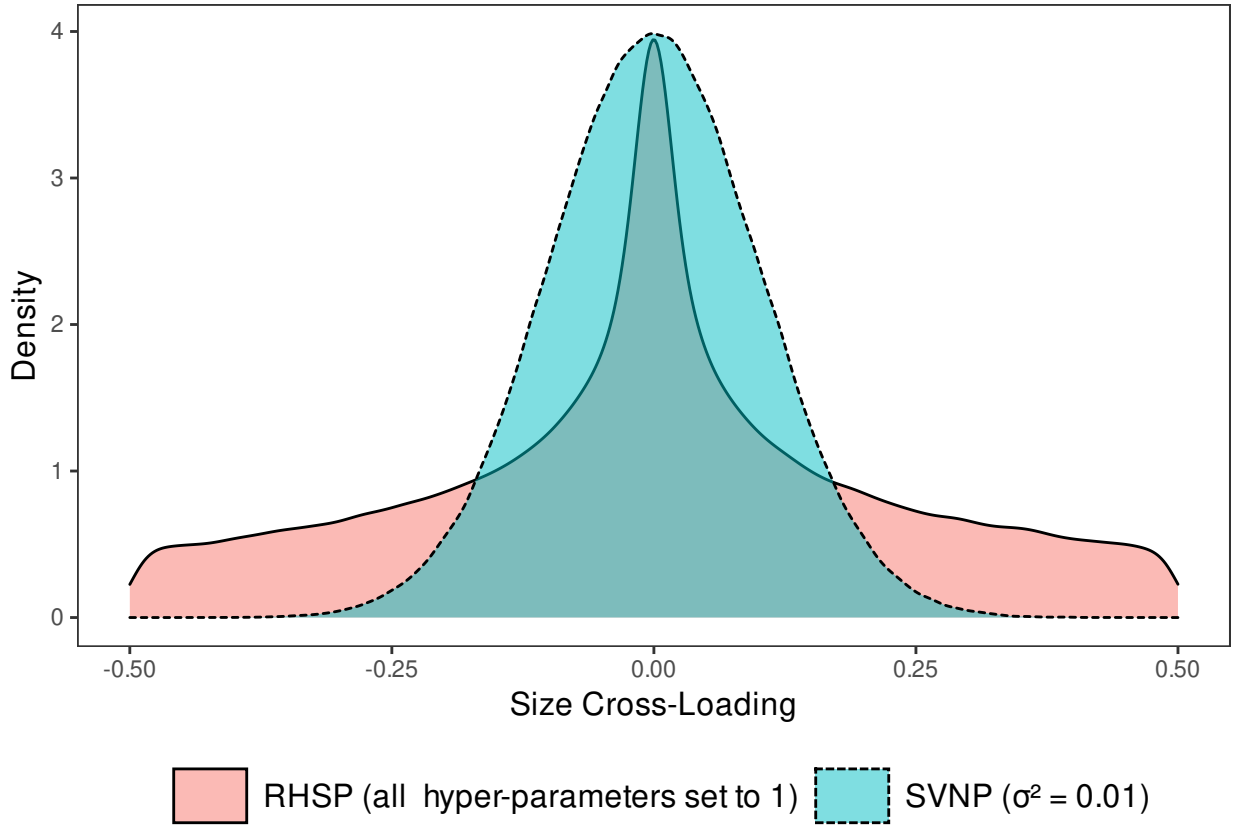

Figure 1. Density Plots of the Regularization Priors of Interest.

Figure 1 compares the two shrinkage-priors. Both priors share a large peak at zero, which ensures that cross-loadings are shrunken to(wards) zero. However, the RHSP has much fatter tails. Here, for larger cross-loadings, there is thus much more prior mass than with the SVNP. This is ought to ensure that large cross-loadings (and consequently other model parameters) can be estimated without bias within a single estimation step.

The current study

While the Regularized Horseshoe Prior has been shown to perform excellently in the selection of relevant predictors in regression (Piironen & Vehtari, 2017b; Van Erp et al., 2019), no previous research has validated its performance in selecting relevant cross-loadings in CFA. We therefore aim to compare the RHSP to the SVNP in their performance in regularizing cross-loadings in Bayesian CFA. Below we present our preliminary results regarding the performance of the SVNP.

Study Procedure and Parameters

A Monte Carlo simulation study was conducted using STAN (Stan Development Team, 2021). All code that was used to run the simulations can be openly accessed on the author's **github**[2]. The models were sampled using the No-U-Turn-Sampler (Homan & Gelman, 2014), with two chains, a burnin-period of 2000 and a chain-length of 4000. These sampling parameters were identified in pilot runs to be required for the RHSP to reach convergence, and were therefore also used for the SVNP in order to ensure a fair comparison.

[2] Specifically, the R-scripts needed to run the simulation can be found on https://github.com/JMBKoch/1vs2StepBayesianRegSEM/tree/main/R. `parameters.R` can be adjusted to adjust study parameters, and `main.R` is used to run the main simulation. Required packages are listed at the top of `parameters.R`.

True Model and Conditions

The datasets were simulated based on a true 2-factor model, with three items per factor, and a factor correlation of 0.5. The factors were scaled by fixing their means to zero and their variances to 1. The true model is summarized below, both in equations (Appendix A) and graphically (Figure 2).[3] All main-loadings were set to 0.75, and all residual variances to 0.3, to ensure that the largest proportion of variance in the items would be explained by their corresponding factor. We varied the size of the two truly non-zero cross-loadings λ_{c5} and λ_{c6} between 0.2, a negligible magnitude such that shrinkage to zero is desired, and 0.5, a size for which shrinkage towards zero should be avoided. We varied the sample sizes of the simulated datasets between 100 and 200. Larger sample sizes of for instance 500 were not included despite being common place in the literature, because adding them would have rendered the run-time of the simulations unfeasible. This is appropriate because for simple factor models researchers are unlikely to collect such larger sample sizes in practice. Finally, based on Muthen and Asparouhov (2012), we varied σ^2 between 0.001, 0.01 and 0.1. This left us with a total number of 2 x 2 x 3 = 12 individual sets of conditions. Per set of conditions, 200 replications were run, yielding a total of 2400 replications.

Outcomes

We focus[4] on the Mean (Absolute) Bias of the posterior mean estimates of all model parameters, per set of conditions ($\bar{\theta}|conditions$). For every model parameter θ and for

[3] The stan code of the model can be found on https://github.com/JMBKoch/1vs2StepBayesianRegSEM/blob/main/stan/SVNP.stan.

[4] We also computed the Mean Squared Error and Relative Bias of the posterior mean estimates, and the Power and Type-I Error-Rate in selecting truly non-zero cross-loadings as non-zero. Summaries of these outcomes can be found on https://github.com/JMBKoch/1vs2StepBayesianRegSEM/tree/main/Rmd/plots.

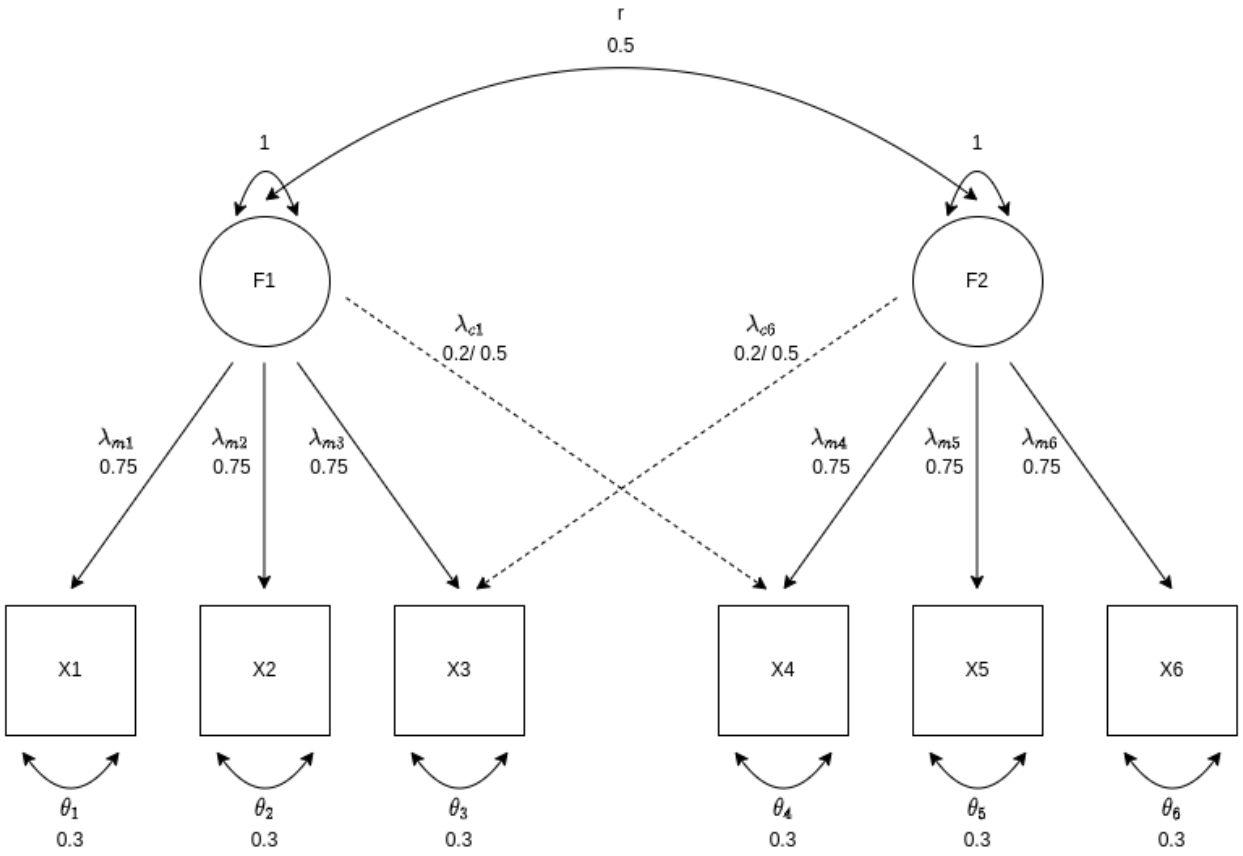

Figure 2. Graphical Representation of the True Model.

every set of conditions that has been sampled from for N_{rep} replications:

$$\bar{Bias}_{\bar{\theta}|conditions} = \frac{1}{N}\Sigma_{i=1}^{N_{rep}}|\bar{\theta}_i - \theta_{true}|.$$

Results

Convergence

In terms of convergence, the SVNP shows excellent performance. Across all 2400 replications there is no single parameter for which $\hat{R} > 1.05$. Across all parameters, the minimum value of the Effective Sample Size N_{eff} was 39.4% of the chain length. For the largest majority of runs N_{eff} even exceeded 50% of the chain length. Moreover, across all runs there was not a single divergent transition. All 2400 replications are therefore included in the results.

Main Results

The Mean Absolute Bias of all parameters is summarized in Figure 3. For parameter estimates that show an identical pattern ($\bar{\lambda}_{c2-5}$; $\bar{\lambda}_{c1}$ and $\bar{\lambda}_{c6}$; $\bar{\lambda}_{m1}$, $\bar{\lambda}_{m2}$, $\bar{\lambda}_{m5}$, and $\bar{\lambda}_{m6}$; $\bar{\lambda}_{m3-4}$; and $\bar{\theta}_{1-6}$), the first respecting estimate is presented representative for all, both in the plot and in the numbers presented below. As results are almost identical for the two sample sizes, we focus on presenting the findings for N = 100, to not distract from our main conclusions.[5]

Figure 3 shows that, as expected, there is substantial bias in some parameter estimates. While the bias in the posterior means of the truly zero cross-loadings $\bar{\lambda}_{c2-5}$ is relatively small, it is pronounced in the estimates of the truly non-zero cross-loadings $\bar{\lambda}_{c1}$ and $\bar{\lambda}_{c6}$. Particularly with a large true cross-loading of 0.5 and $\sigma^2 = 0.001$ the bias is very large, e.g. $\bar{Bias}_{\bar{\lambda}_{c1}} = 0.49$, since the estimates of the true cross-loadings of 0.5 were shrunken almost entirely to zero (e.g. $\bar{\lambda}_{c1} = 0.01$). The choice of σ^2 plays a crucial role here. Also with $\sigma^2 = 0.01$ (and true cross-loadings of 0.5) substantial bias occurs (e.g. $\bar{Bias}_{\bar{\lambda}_{c1}} = 0.35$), as the cross-loading are still under-estimated considerably ($\bar{\lambda}_{c1} = 0.15$), though not entirely shrunken to zero. With $\sigma^2 = 0.1$ the bias in the estimates of the cross-loadings is less pronounced (e.g. $\bar{Bias}_{\bar{\lambda}_{c1}} = 0.14$). Here σ^2 is large enough to estimate the cross-loadings closer to their true value, $\bar{\lambda}_{c1} = 0.37$.

Also the estimates of the main loadings of factor 1 on item 3 ($\bar{\lambda}_{m3}$) and of factor 2 on item 4 ($\bar{\lambda}_{m4}$) are substantially biased when the true cross-loadings are 0.5 and $\sigma^2 = 0.001$ (e.g. $\bar{Bias}_{\bar{\lambda}_{m3}} = 0.40$). These two loadings show much higher bias than the other four main-loadings as they load on the same two items as the two non-zero cross-loadings ($\bar{\lambda}_{c1}$ and $\bar{\lambda}_{c6}$, see Figure 2). As the cross-loadings are shrunken to zero, these main loadings now also account for the variance in the items that is truly explained by the cross-loadings.

[5] The Mean Absolute Bias visualized for the different sample sizes separately can be found on https://github.com/JMBKoch/1vs2StepBayesianRegSEM/blob/main/Rmd/plots/plotsBiasSVNP.html.

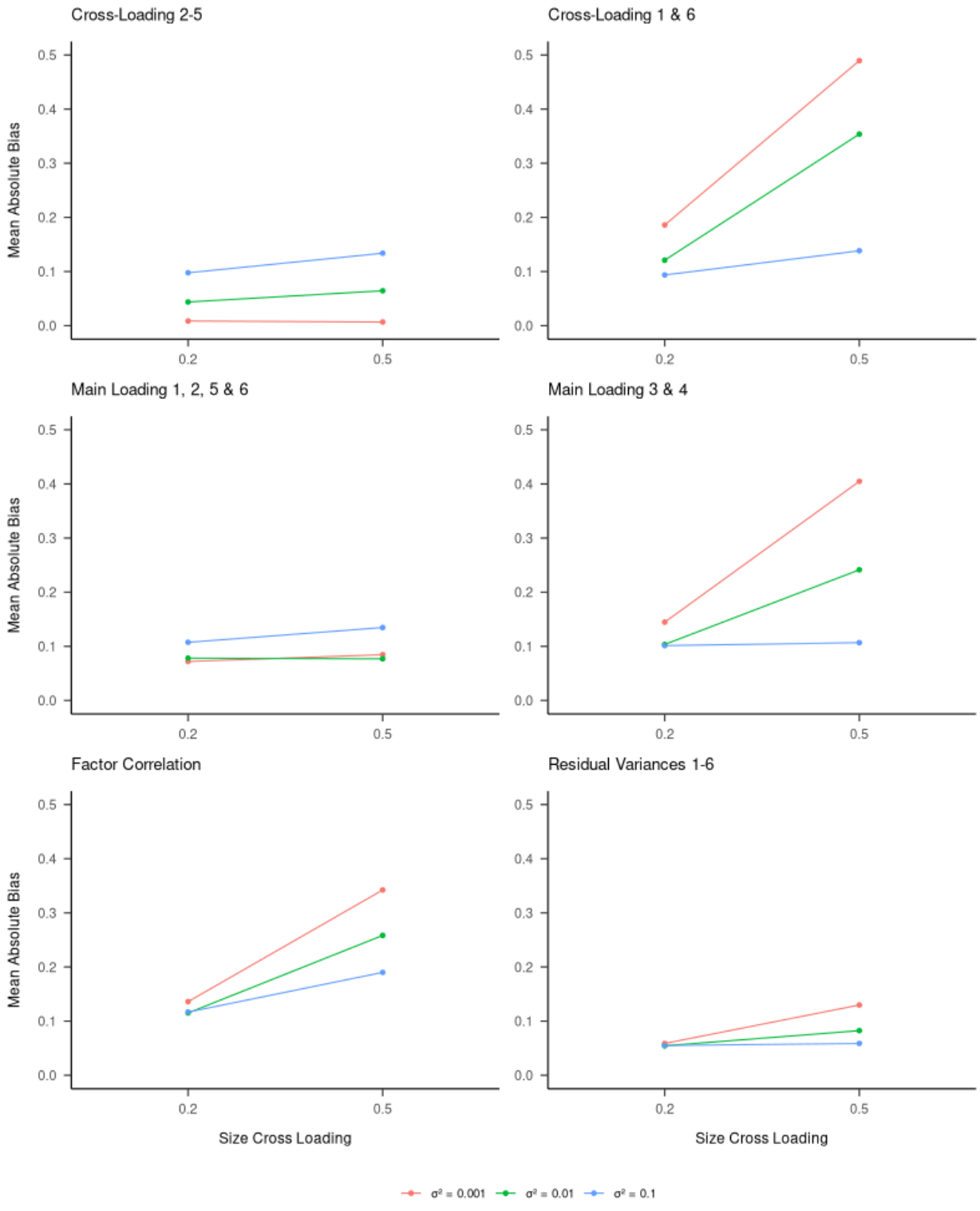

Figure 3. Main Results: Mean Absolute Bias in the Model Parameters (N = 100).

Consequently, the two main-loadings are over-estimated, e.g. $\bar{\lambda}_{m3} = 1.15$.

In the factor correlation the bias is also relatively small and approximately the same for the different values of σ^2 when the truly non-zero cross-loadings are 0.2. Again, bias becomes much more pronounced with true cross-loadings of 0.5, especially when $\sigma^2 = 0.001$ ($\bar{Bias}_{\bar{r}} = 0.34$). In this situation the factor correlation is heavily over-estimated ($\bar{r} = 0.84$). This is because the covariance between item 3 and 4 that arises from the two cross-loadings, is mis-attributed to the factor-correlation, as the cross-loadings are shrunken to zero.

The bias in the estimates of the residual variances $\bar{\theta}_{1-6}$ is not large across different conditions, although also here a noticeable increase occurs between true cross-loadings of 0.2 and 0.5 when $\sigma^2 = 0.001$.

Conclusions and Discussion

The results show a clear and consistent pattern. The SVNP performs well when the truly non-zero cross-loadings are small, in terms of estimating the model without substantial bias. This can be interpreted as a successful instance of regularization, where an acceptable amount of bias is added to the model by shrinking some parameters to zero, to reach a more sparse solution. However, with larger truly non-zero cross-loadings, the performance of the SVNP decreases. With smaller values of σ^2, particularly with $\sigma^2 = 0.001$, these cross-loadings are still shrunken to zero, even though they are much larger in practice. This causes substantial bias in some main-loadings, and in the factor correlation. In practice, bias in structural parameters is particularly concerning, as it may lead to wrong conclusions in research on structural relationships between latent constructs.

Bias occurs much less when $\sigma^2 = 0.1$. Such relatively large variance still allows for enough deviations from zero in the cross-loadings to yield relatively accurate estimates of the non-zero cross-loadings itself and consequently the other model parameters. However, simply using larger values of σ^2 is no general solution. In practice, models may include

more structural parameters, even more cross-loadings, or a number of residual co-variances. Under these circumstances, large values of σ^2 may lead to identification issues. Moreover, the larger σ^2, the more cross-loadings will be selected as non-zero, which may ultimately lead to over-fitting.

The high bias of the SVNP under large true cross-loadings and low values of σ^2 is not surprising, as it is clearly noted that the method requires a 2-step approach to avoid bias. However, this approach depends on a successful selection of non-zero cross-loadings. Muthen and Asparouhov (2012) advise a Power (true positive rate) in selecting non-zero cross-loadings of at least .80. However, only under a single set of conditions (N = 200, σ^2 = 0.01, cross-loadings = 0.5) this power was reached in our study (see Table B2), which suggests that even a 2-step approach is no robust solution. This serves to illustrate the need for more advanced priors such as the RHSP, although different selection rules (see Zhang et al., 2021) may show a better performance than the 95% credible intervals suggested by Muthen and Asparouhov (2012).[6]

The RHSP is expected to show less bias in the parameter estimates of the model within a single estimation step, even with true cross-loadings of 0.5. Estimates of these larger cross-loadings are expected to escape the shrinkage, which also prevents bias in other parameter estimates.

[6] We will assess differences between a broad variety of selection rules in the upcoming main study.

References

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480. https://doi.org/10.1093/biomet/asq017

Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press.

Ghosh, J., Li, Y., & Mitra, R. (2018). On the Use of Cauchy Prior Distributions for Bayesian Logistic Regression. *Bayesian Analysis*, *13*(2), 359–383. https://doi.org/10.1214/17-BA1051

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity. *Monographs on Statistics and Applied Probability*, *143*, 143.

Hoerl, A. E., & Kennard, R. W. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *42*(1), 80–86. https://doi.org/10.2307/1271436

Homan, M. D., & Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hsiang, T. C. (1975). A Bayesian View on Ridge Regression. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *24*(4), 267–268. https://doi.org/10.2307/2987923

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(4), 555–566. https://doi.org/10.1080/10705511.2016.1154793

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. New York, NY: Springer US.

https://doi.org/10.1007/978-1-0716-1418-1

Lu, Z.-H., Chow, S.-M., & Loken, E. (2016). Bayesian Factor Analysis as a Variable-Selection Problem: Alternative Priors and Consequences. *Multivariate Behavioral Research*, *51*(4), 519–539. https://doi.org/10.1080/00273171.2016.1168279

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Muthen, B., & Asparouhov, T. (2012). Bayesian SEM: A more flexible representation of substantive theory, 78. https://doi.org/10.1037/a0026802

Park, T., & Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Piironen, J., & Vehtari, A. (2017a). On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (pp. 905–913). PMLR. Retrieved from https://proceedings.mlr.press/v54/piironen17a.html

Piironen, J., & Vehtari, A. (2017b). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2), 5018–5051. https://doi.org/10.1214/17-EJS1337SI

Stan Development Team. (2021). Stan User Guide. Retrieved from https://mc-stan.org/docs/2_27/stan-users-guide-2_27.pdf

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian

penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50. https://doi.org/10.1016/j.jmp.2018.12.004

Zhang, L., Pan, J., & Ip, E. H. (2021). Criteria for Parameter Identification in Bayesian Lasso Methods for Covariance Analysis: Comparing Rules for Thresholding, p -value, and Credible Interval. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–10. https://doi.org/10.1080/10705511.2021.1945456

Appendix A

For every individual i in i = 1,..., N:

$$Y_i \sim \mathcal{N}(\mathbf{0}, \Sigma),$$

where

$$\Sigma = \Lambda \Psi \Lambda',$$

$$\Lambda = \begin{bmatrix} 0.75 & 0 \\ 0.75 & 0 \\ 0.75 & 0.2/0.5 \\ 0.2/0.5 & 0.75 \\ 0 & 0.75 \\ 0 & 0.75 \end{bmatrix},$$

$$\Psi = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix},$$

and

$$\Theta = diag[0.3, 0.3, 0.3, 0.3, 0.3, 0.3].$$

Appendix B

Table B1

N	σ^2	cross	Power λ_{c1}	Power λ_{c6}
100	0.00	0.20	0.00	0.00
100	0.01	0.20	0.01	0.01
100	0.10	0.20	0.00	0.00
100	0.00	0.50	0.00	0.00
100	0.01	0.50	0.22	0.18
100	0.10	0.50	0.30	0.28
200	0.00	0.20	0.00	0.00
200	0.01	0.20	0.12	0.12
200	0.10	0.20	0.00	0.00
200	0.00	0.50	0.00	0.00
200	0.01	0.50	0.92	0.91
200	0.10	0.50	0.60	0.52