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: 2500

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. On the one hand, models need to be simple enough to be (1) identified, i.e., estimate-able with the information available in the data; (2) interpretable; and (3) generalizable (not over-fitted). On the other hand, models need to be complex enough, i.e., have enough parameters, to accurately represent the data generating process, hence to not be too biased (Cox, 2006; James, Witten, Hastie, & Tibshirani, 2021).

In Confirmatory factor analysis (CFA, Bollen, 1989), an essential tool for modeling measurement structures, it is common practice to deal with this fundamental trade-off 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 straightforwardly interpretable model. This practice 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 (Lu, Chow, & Loken, 2016). This procedure is highly flawed, as it risks capitalization on chance and thereby over-fitting, hence ending up with a model that does not generalize well to other datasets from the same population (MacCallum, Roznowski, & Necowitz, 1992).

## Bayesian CFA: The Small Variance Normal Prior (SVNP)

As solution to the issue, Muthen and Asparouhov (2012) proposed a Bayesian alternative approach to CFA. Rather than identifying models by fixing *all* cross-loadings to zero, one should should assume that *most* cross-loadings are zero. Formally, 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.: $\sigma^2 = 0.1$, $\sigma^2 = 0.01$, $\sigma^2 = 0.001$). This prior has a large peak at zero, and very thin tails. 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. Lu et al. (2016) note that this approach is simply a form of regularization, where cross-loadings are regularized in an attempt to identify and select relevant cross-loadings as non-zero, such that one ends up with a sparse model.

An issue with Muthen and Asparouhov (2012)'s approach 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, which introduces bias (Lu et al., 2016). First, bias naturally occurs in the large cross-loadings itself. However, given that the parameters of a model are estimated conditionally on one another, also in other parameters, such as factor-correlations or main-loadings, substantial bias can arise. Consequently, the method requires a two-step approach. First, the model is estimated with the SVNP. Then the model is re-estimated, with cross-loadings that have been selected to be zero in the previous step are fixed to zero, and the remaining cross-loadings are estimated without shrinkage, avoiding the bias in the model of the previous step. This process is tedious, computationally expensive, and adds a number of undesired researchers degrees of freedom. Therefore, alternative priors need to be identified that can outperform the Small Variance Normal Prior in a single step. The literature on regularization in a regression context (see Van Erp, Oberski, & Mulder, 2019 for an overview) provides a variety of promising candidates for achieving this end.

## Spike-and-Slab Prior

One suitable regularization prior for the purpose of selecting cross-loadings in regularized Bayesian SEM is the so-called *Spike-and-Slab Prior* (George & McCulloch, 1993; Ishwaran & Rao, 2005; Mitchell & Beauchamp, 1988; Van Erp et al., 2019). This prior is a discrete mixture of an extremely peaked prior around zero (the spike), and a very flat prior for larger cross-loadings (the slab).

Formally, and applied to the cross-loadings in CFA, for every Cross-loading of factor j on item k, the Spike-and-Slab Prior can be specified as (Lu et al., 2016):

$$\lambda_{jk}|r_{jk} \sim (1 - r_{jk})\delta_0 + r_{jk}\mathcal{N}(0, c_{jk}^2),$$

with

$$r_{jk} \sim \mathcal{B}ernoulli(p_{jk}).$$

The basic intuition is as follows. When $r_{jk} = 1$ , $\lambda_{jk} \sim \mathcal{N}(0, c_{jk}^2)$, hence $\lambda_{jk}$ is assigned to the slab. When $r_{jk} = 0$, $\lambda_{jk} \sim \delta_0$, and is thus assigned to the spike.

Lu et al. (2016) found that this prior is performing well in shrinking truly zero cross-loadings to zero, while not shrinking (relevant) large cross-loadings to avoid bias. However, the big caveat of this prior is that it cannot be implemented in standard MCMC sampling software, such as STAN, due to being a discrete mixture prior (Betancourt, 2018; Stan Development Team, 2021).

**The Regularized Horseshoe Prior (RHSP)**

A promising alternative, that is a fully continuous mixture of distributions, and thus employable in STAN, 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* $\tau$, shrinking all cross-loadings to zero, and a *local shrinkage parameter* $\tilde{\omega}_{jk}^2$ that allows the relevant cross-loadings to escape the shrinkage. The issue with the original Horseshoe Prior is that not shrinking large cross-loadings at all can lead to identification issues (see Ghosh, Li, & Mitra, 2018). The RHSP solves this issue (Piironen & Vehtari, 2017b), by shrinking all cross-loadings at least a little bit. The prior is specified as follows.

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

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

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

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

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

where $p_0$ represents a prior guess of the number of relevant cross-loadings. Note that we deviate from the common notation of the local shrinkage parameter as $\lambda$, as this letter is commonly used to denote factor loadings in CFA.

Assuming that c is constant, with values of $\tau$ close to zero $\tau^2 \omega^2_{jk} << c^2$, and $\tilde{\omega}^2_{jk} \rightarrow \omega^2_{jk}$. Then, the RHSP approaches the Horseshoe Prior, ensuring that small cross-loadings are shrunken to zero as with the original Horseshoe. With large values of $\tau$ $\tau^2 \omega^2_{jk} >> c^2$ and $\tilde{\omega}^2_{jk} \rightarrow \frac{c^2}{\tau^2}$. Then, the prior approaches $\mathcal{N}(0, c^2)$. This slab then ensures that even large cross-loadings are shrunken a little.

## 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. To fill this gap, we aim to compare the RHSP to the SVNP in their performance in selecting the true factor structure in CFA. Below we present our preliminary results regarding the performance of the SVNP.

### Study Procedure and Parameters

In order to assess the performance of the SVNP in regularizing cross-loadings in Bayesian Regularized SEM, a Monte Carlo simulation study was conduced using STAN (Stan Development Team, 2021). All code that was used to run the simulation study can be openly accessed on the author's github. 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.

### 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. All main-loadings were set to 0.75, and all residual variances to 0.3. We included two truly non-zero cross-loadings, that of factor 1 on item 4, and that of factor 2 on item 3. The true model is summarized in the Appendix both in equations and graphically. We varied the magnitude of the two non-zero cross-loadings between 0.2 and 0.5. Next, we varied the sample sizes of the simulated datasets between 100 and 200. This choice was made because for simple factor models researchers would be unlikely to collected larger sample sizes in practice. Finally, based on the recommendations of Muthen and Asparouhov (2012), we included three levels of the hyper-parameter $\sigma^2$: 0.001, 0.01, 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 iterations were run, yielding a total of 2400 posterior samples.

### Outcomes

As main outcome, we considered the Mean (Absolute) Bias of all estimated model parameters. As secondary outcomes, we also computed the Relative Bias and Mean Squared Error (MSE, Morris, White, & Crowther, 2019). Next, we computed the power for the truly non-zero cross-loadings, i.e. the probability of correctly identifying them as non-zero. For the truly zero cross-loadings we computed the Type-I Error Rate, hence the probability of wrongly selecting these cross-loadings as non-zero. For these last two outcomes, in order to select cross-loadings as zero, several selection rules were used based on recommendations Zhang, Pan, and Ip (2021). First, a number of thresholds were considered, where a cross-loading is selected to be zero when the absolute value of its posterior estimates falls below a certain value. Specifically we considered three thresholds: 0, 0.1, 0.15. Moreover, we selected cross-loadings based on whether or not their 95%, 90%, 80%, and 50% credible interval contained zero. Note that for all outcomes we computed two versions, one based on mean and one based on median posterior estimates. The latter is only reported in case of relevant deviations from the former.

## Results

### Convergence

In terms of convergence, the SVNP showed excellent performance. For all parameters, across all iterations and configurations of conditions, there were none in which $\hat{R} < 1.05$. The lowest value of the Effective Sample Size $N_{eff}$ was still a good 39.4% of the chain length, with the largest majority of runs even exceeding an excellent 50% of the chain length (96% for the parameter with the largest percentage of $\frac{N_{eff}}{N_{chain}} < 0.5$). Moreover, across all runs there was not a single divergent transition. Therefore, none of the 2400 posterior samples had to be disregarded.

### Main Results

## Conclusions and Discussion

## References


Betancourt, M. (2018). A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv:1701.02434 [Stat]*. Retrieved from http://arxiv.org/abs/1701.02434

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.

George, E. I., & McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association*, *88*(423), 881–889. https://doi.org/10.2307/2290777

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

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.

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773. https://doi.org/10.1214/009053604000001147

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

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032. https://doi.org/10.2307/2290129

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

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

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

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

## Appendix A: True Model

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

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

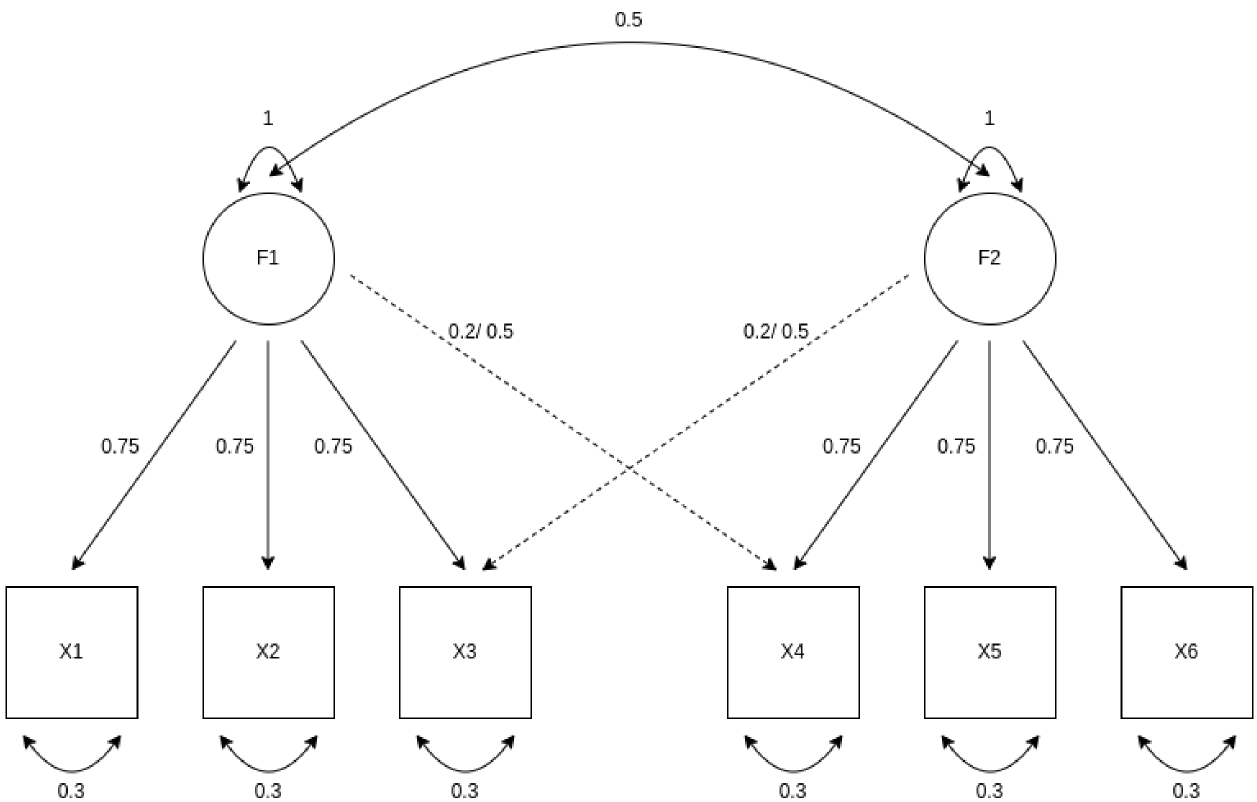


*Figure 1.* Graphical Representation of the True Model.