

# 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: 250 

Intended 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. However, these models have low variance, hence they generalize well to other datasets from the same population. Moreover, simple models are easily identified (estimatable with the information available in the data) and easy to interpret. Complex models with large numbers of parameters tend to have low bias and high variance. Consequently, complex models are 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 way of dealing with the bias-variance trade-off is *regularization* (Hastie, Tibshirani, & Wainwright, 2015). Here, during the estimation process of a model it is actively chosen to add some bias to the model to reduce its variance. This helps to ensure that the model becomes easier to interpret and more generalizable. In a frequentist context, regularization is achieved by adding a penalty term to the cost function of a model. Such penalty ensures that some model parameters that are deemed irrelevant, e.g. small regression coefficients in a regression model with a large number of predictors, are shrunk to (or towards) zero. In a Bayesian context, the same is achieved by setting a so-called shrinkage-prior for the parameter in question (Van Erp, Oberski, & Mulder, 2019). For every frequentist penalty term a bayesian counterpart exist. For instance, 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.

## Simple Structure in CFA


In Confirmatory factor analysis (CFA, Bollen, 1989), an essential tool for modeling measurement structures, 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. This is done to yield an identified and straightforwardly interpretable model. However, the 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. 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 an alternative to imposing simple structure to identify CFA models, Muthen and Asparouhov (2012) proposed *Bayesian CFA*, which can be viewed as a form of regularized SEM (see also Jacobucci, Grimm, & McArdle, 2016 for a summary of frequentist approaches to regularized Structural Equation Modeling). Rather than identifying models by fixing *all* cross-loadings to zero, one 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 (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 shrunk to zero, as desired. Also the ones further from zero are shrunk heavily towards zero, which introduces bias (Lu, Chow, & Loken, 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, Bayesian CFA requires two steps in practice. First, the model is estimated with the SVNP set for the cross-loadings. Cross-loadings are selected as non-zero when their 95% credible intervals does not contain zero (which can be viewed as a ‘Bayesian way to determine statistical significance’). The model is then 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. **It is desirable to identify** alternative priors that can outperform the Small Variance Normal Prior in a single step  The literature on regularization in a regression context (see Van Erp et al., 2019) provides a variety of promising candidates for achieving this end.

## 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*  $\tau$ , 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 parameters at all can lead to identification issues (see Ghosh, Li, & Mitra, 2018). The RHSP solves this issue (Piironen & Vehtari, 2017b), **by shrinking also large parameters a little bit, by setting a slab for them as prior** . The full prior is specified as follows.

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

$$\begin{aligned}\lambda_{jk}|\tilde{\omega}_{jk}, \tau, c &\sim \mathcal{N}(0, \tilde{\omega}_{jk}^2 \tau^2), \text{ with } \tilde{\omega}_{jk}^2 = \frac{c^2 \omega_{jk}^2}{c^2 + \tau^2 \omega_{jk}^2}, \\ \tau|s_{global}^2 &\sim \text{half} - t_{df_{global}}(0, s_{global}^2), \text{ with } s_{global} = \frac{p_0}{p - p_0} \frac{\sigma}{\sqrt{N}}, \\ \omega_{jk} &\sim \text{half} - t_{df_{local}}(0, s_{local}^2), \\ c^2|df_{slab}, s_{slab} &\sim \mathcal{IG}(\frac{df_{slab}}{2}, df_{slab} \times \frac{s_{slab}^2}{2}),\end{aligned}$$

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

Figure 1 compares the two shrinkage-priors. Both priors share a large peak at zero, which ensures that cross-loadings are shrunk 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 SVN. This ensures that large cross-loadings, that would have been shrunk heavily towards zero with the SVN, can escape the shrinkage.

### 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 SVN in their performance in selecting the true factor structure in CFA. Below we present our preliminary results regarding the performance of the SVN.

## Study Procedure and Parameters

In order to assess the performance of the SVNPs in regularizing cross-loadings in Bayesian Regularized SEM, a Monte Carlo simulation study was conducted 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 SVNPs 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 below, both in equations (Appendix A) and graphically (Figure 2). 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 collect larger sample sizes in practice. Finally, based on the recommendations of Muthén 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  $2 \times 2 \times 3 = 12$  individual sets of conditions. Per set of conditions, 200 iterations were run.

## Outcomes

As outcomes, we initially considered the Mean (Absolute) Bias ( $Bias_\theta = |\hat{\theta} - \theta_{true}|$ ), the Relative Bias ( $RelativeBias_\theta = \frac{|\hat{\theta} - \theta_{true}|}{\theta_{true}}$ ) and the Mean Squared Error

( $MSE = \frac{1}{N}(\hat{\theta} - \theta_{true})^2$ ) of all estimated model parameters. 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. Across all iterations and configurations of conditions, there were not a single parameter for which  $\hat{R} < 1.05$ . Across all parameters, the minimum value of the Effective Sample Size  $N_{eff}$  was still 39.4% of the chain length, which is still a very acceptable proportion. 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. Therefore, none of the 2400 posterior samples had to be disregarded.

### Main Results

The mean absolute bias of all model parameters is summarized below in Figure 3. We focus on presenting this outcome as very similar patterns emerged for the other outcomes. For all parameters that showed an identical pattern (cross-loadings 2-5; cross-loadings 1

and 6; main-loadings 1, 2, 5, and 6; main-loadings 3 and 4; and theta 1-6), the first respecting parameter is presented representative for all, both in the plot and in the numbers presented below. All findings presented are averaged over the two sample sizes considered, as there were no noticable differences between the two sample sizes.

Figure 3 shows that, as expected, substantial bias can arise in the model parameters when using the SVNP to regularize cross-loadings. While the bias in the truly zero cross-loadings 2-5 is relatively small, substantial bias can arise for the truly non-zero cross-loadings. Particular with a large cross-loading of 0.5 and  $\sigma^2 = 0.001$  the bias is substantial, e.g.  $Bias_{\hat{\lambda}_{23}} = 0.485$ ). Here, the cross-loading that truly is 0.5 was still shrunk almost entirely to zero ( $\hat{\lambda}_{23} = 0.015$ ), explaining the substantial bias. The choice of  $\sigma^2$  plays a crucial role here. Also with  $\sigma^2 = 0.01$  substantial bias still occurs ( $Bias_{\hat{\lambda}_{23}} = 0.296$ ). Here the cross-loading is still substantially under-estimated ( $\hat{\lambda}_{23} = 0.203$ ). With a  $\sigma^2 = 0.1$  the bias in the estimate of the cross-loading is less pronounced, but still present ( $Bias_{\hat{\lambda}_{23}} = 0.121$ ). Here the variance of the prior of the cross-loadings is large enough that the cross-loadings are estimated closer to their large population value, e.g.  $\hat{\lambda}_{23} = 0.383$ .

Next, looking at the main-loadings it is clear that also in the main loadings of factor 1 on item 3 and of factor 2 on item 4 substantial bias arises, again in particular under the most extreme combination of conditions. When the true cross-loadings are 0.5 and  $\sigma^2 = 0.001$  the bias becomes very pronounced (e.g.  $Bias_{\hat{\lambda}_{13}} = 0.397$ ). The two loadings have much higher bias than the other four main-loadings as these are the two main-loadings that load onto the same two items on which the truly non-zero cross-loadings load. When these cross-loadings are shrunk to zero, these main loadings now also have to account for the variance in the items that should be accounted for by the cross-loadings. Consequently, these main-loadings are over-estimated, e.g. under the above configuration  $\hat{\lambda}_{13} = 1.147$ .

Also in the structural parameter of the model, the factor correlation, a similar



pattern emerges. While the bias is relatively small and approximately the same for the different values of  $\sigma^2$  when the truly non-zero cross-loadings are 0.2, it becomes more pronounced when they are 0.5, particularly when  $\sigma^2 = 0.001$  ( $Bias_{fact\hat{Corr}} = 0.345$ ). The underlying pattern becomes clear when considering the estimates of the factor correlation. When  $\sigma^2 = 0.001$  and the non-zero cross-loadings are 0.5, the factor correlation is heavily over-estimated ( $fact\hat{Corr} = 0.845$ ). 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 shrunk to zero.

The bias in the residual variances  $\theta_{1-6}$  is not substantial across different conditions, although also here a noticable increase occurs between cross-loadings of 0.2 and 0.5, with  $\sigma^2 = 0.001$ .

## Conclusions and Discussion

In sum, a clear pattern arose. In general, the SVNPs appear to perform well in situations where the truly non-zero cross-loadings are small, in terms of not leading to extreme bias in the model parameters. However, with larger non-zero cross-loadings, the performance of the SVNPs decreases. With smaller values of  $\sigma^2$ , particularly with  $\sigma^2 = 0.001$ , these cross-loadings are still shrunk to zero, even though they are much larger in practice. This, consequently, causes also substantial bias in main-loadings, and in the factor correlation. In particular the bias in such structural parameters is concerning, as it may lead to highly misleading conclusion in research in which structural relationships between latent constructs are of interest.

Note that in our results, these problems occurred in a much less pronounced fashion with  $\sigma^2 = 0.1$ . A shrinkage prior with such a relatively large variance still allowed 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. This may

suggest at first sight that the SVNPs can be used safely even with large cross-loadings in the population, by simply specifying such relatively large value for  $\sigma^2$ . However, one must be cautious with this conclusion, in the light of the fact that the models employed in this study are rather simple and therefore easily identifiable. In practice, models may include more structural parameters, even more cross-loadings, or a number of residual co-variances. Under these circumstances, such a large value of  $\sigma^2$  may lead to identification issues, and is therefore no general solution.

The RHSP is expected to perform better with large non-zero cross-loadings of 0.5, with estimates of these cross-loadings being able to escape the shrinkage. While some hyper-parameter configurations of the RHSP are likely to show much worse performance in terms of identification, other configurations should allow for regularizing cross-loadings without risking substantial bias nor identification issues.

## 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](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

### Appendix A: True Model

For every individual  $i$  in  $i = 1, \dots, N$ :

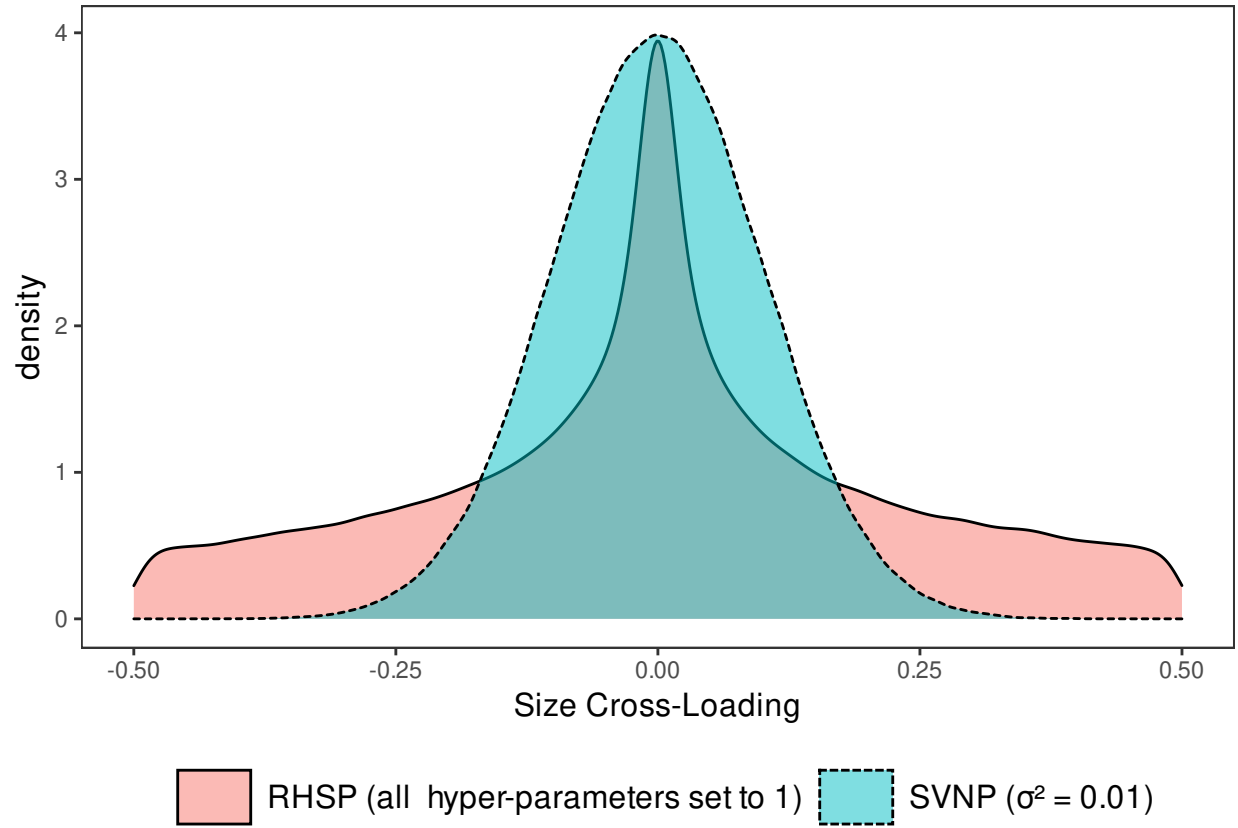
$$Y_i \sim \mathcal{N}(\mathbf{0}, \Sigma),$$

where

$$\begin{aligned} \Sigma &= \Lambda \Psi \Lambda', \\ \Lambda &= \begin{bmatrix} 0.75 & 0 \\ 0.75 & 0 \\ 0.75 & 0.2/0.5 \\ 0.2/0.5 & 0.75 \\ 0 & 0.75 \\ 0 & 0.75 \end{bmatrix}, \\ \Psi &= \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \end{aligned}$$

and

$$\Theta = \text{diag}[0.3, 0.3, 0.3, 0.3, 0.3, 0.3].$$



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

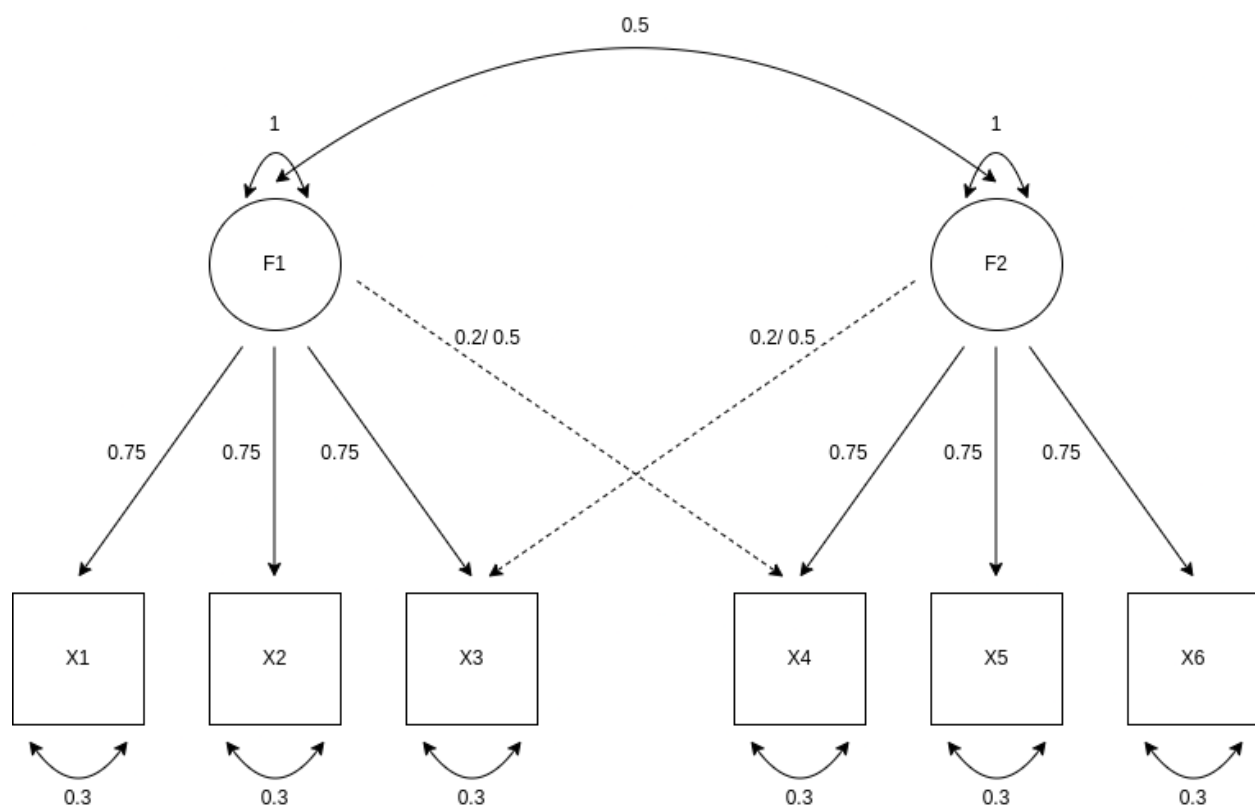


Figure 2. Graphical Representation of the True Model.



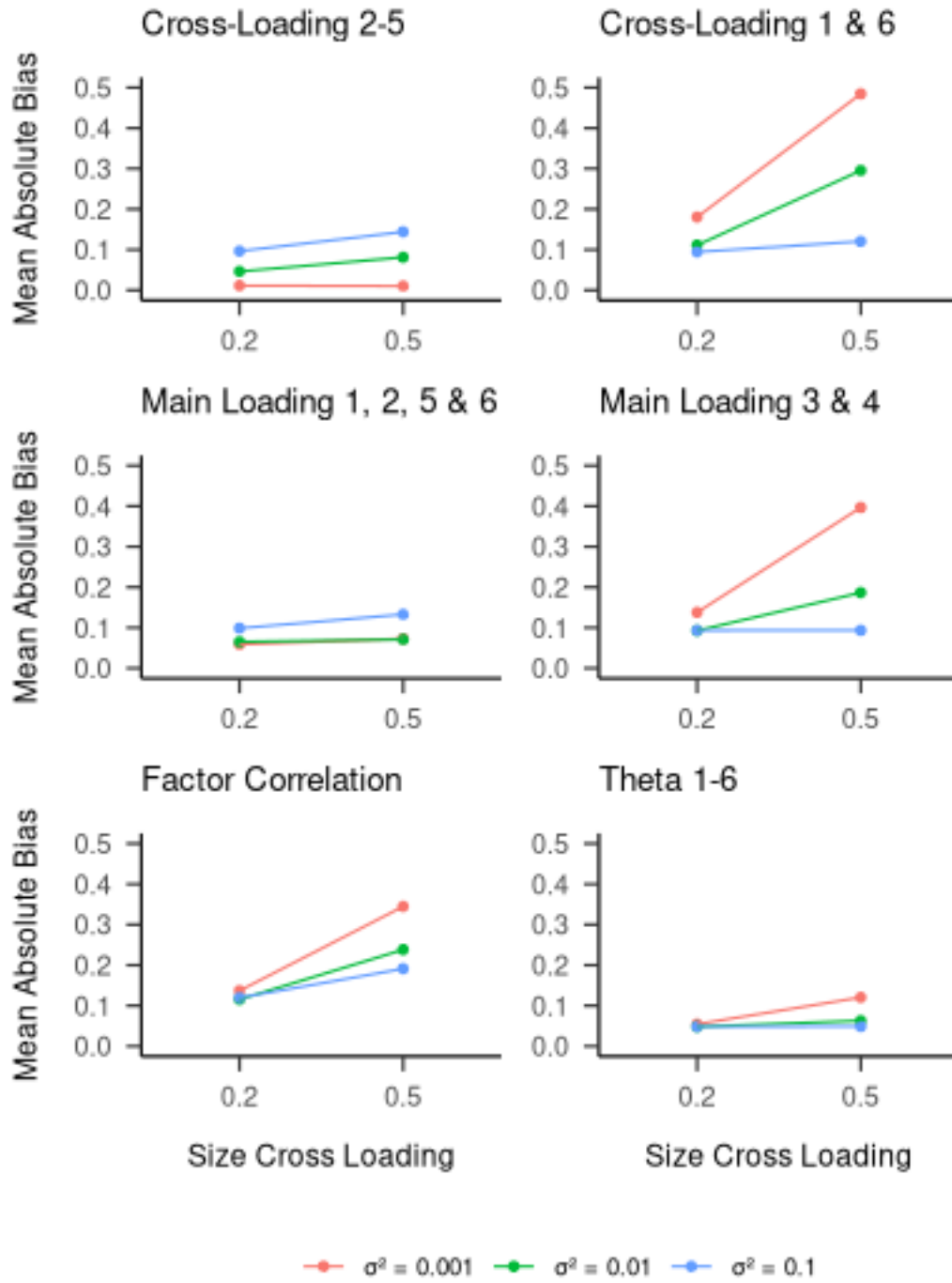


Figure 3. Main Results: Mean Absolute Bias in the Model Parameters.