Research Master's programme Methodology and Statistics for the
Behavioural, Biomedical and Social Sciences
Utrecht University, the Netherlands

MSc Thesis Johannes Michael Benjamin Koch (6412157)
TITLE: "Getting a Step Ahead: Using the Regularized Horseshoe Prior to Select Cross-Loadings in Bayesian CFA"
June 2022

Supervisor:
Dr. Sara van Erp

Second grader:
Dr. Beth Grandfield

Preferred journal of publication: Structural Equation Modeling
Word count: 9465

Introduction

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 between 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. For a regression model:

$$y_i = \beta \mathbf{x_i} + e_i, \ where$$

$$e_i \sim \mathcal{N}(0, \sigma^2),$$

the Ordinary Least Squared Residuals estimates of β are obtained by minimizing the sum of squared residuals:

$$\hat{\beta} = \underset{\beta}{argmin}\{\Sigma_{i=1}^{N}(y_i - \beta\mathbf{x_i})^2\}.$$

Penalized regression adds a a penalty term to this cost function, which is generally denoted as $||\beta||_L$:

$$\hat{\beta} = \underset{\beta}{argmin}\{\Sigma_{i=1}^{N}(y_i - \beta\mathbf{x_i})^2 + \lambda||\beta||_L\}.$$

When L = 1, the so-called L-1 norm, $||\beta||_1 = \Sigma_{j=1}^{p}|\beta_j|$, the well-known LASSO penality (Tibshirani, 1996, 2011). When, L = 2, the L2-norm, $||\beta||_2 = \sqrt{\Sigma_{j=1}^{p}\beta_j^2}$. This is the famous ridge penalty (Hoerl & Kennard, 2000). The so-called tuning-parameter λ is a hyper-parameter that is in practice often determined through cross-validation.

Regularization can also be applied outside of regression, for instance in Structural Equation Modeling (SEM, Jacobucci, Grimm, & McArdle, 2016). Regularized SEM entails adding penalties to the cost function of SEM models (typically a variant of the maximum likelihood cost function) to reach sparser models. Jacobucci et al. (2016) showed that applying the ridge or lasso penalty to model parameters of SEM models, for instance factor loadings in CFA or regression coefficients in MIMIC models, is a feasible and efficient way of yielding easier to interpret and more generalizable models, especially under large sample sizes. One disadvantage of the frequentist regularized SEM method is that SEM models can become hard to optimize in practice when more complex penalties are added to their cost function.

In a Bayesian context, instead of adding a penalty to the cost function of a model, so-called shrinkage-priors are set for parameters (see Van Erp, Oberski, & Mulder, 2019 for an overview). In Bayesian model estimation, the so-called Joint Posterior Distribution of the model-parameters given the data $P(\theta|data)$ is a combination of the data and the prior. Priors can not only be set to steer model estimates towards expected outcomes, for instance based on previous research. Also shrinking model parameters to(wards) zero can

be achieved by setting priors that, in general, attach a lot of prior mass to the parameter in question being zero. In the most simple case one can simply set a normal prior for regression coefficients that is centered around zero, which resembles the ridge-penalty (Hsiang, 1975). The lasso penalty can be mimicked by setting a Laplace- (double exponential) prior for the regression coefficients (Park & Casella, 2008; see Van Erp et al., 2019 for the Bayesian equivalents of other relevant penalties). In general, an advantage of Bayesian Regularization over the frequentist approach is that it does not rely on optimization, since the the model estimation is achieved through MCMC methods. This allows for more flexibility in regularization, as shrinkage priors are not limited by having to lead to an optimizable cost function together with the model, as is the case with frequentist penalties.

Bayesian CFA: The Small Variance Normal Prior (SVNP)

Confirmatory Factor Analysis (CFA, Bollen, 1989) is an essential tool for modeling measurement structures, falling under the class of Structural Equation Modeling (SEM). For every individual i, the scores on a vector of p observed indicators $\mathbf{y}_i$ (typically items of a psychological test):

$$y_i = \mu + \Lambda\eta_i + e_i,$$

where y_i is a $p \times 1$ vector of observed indicators, μ is a $p \times 1$ vector of intercepts, Λ is a $p \times q$ matrix of factor loadings, η_i is a $q \times 1$ vector of scores on the latent factors, and e_i is a $p \times 1$ is a random vector of random (measurement) error terms. Here, Λ is thus the part of the equation that relates the latent variables to the observed scores on the items. We can differentiate between so-called main-loadings, and cross-loadings. The former are factor loadings that relate factor and items to one another that are theoretically expected to have a relationship. Cross-loadings are factor loadings that relate factors to items between which, theoretically, no relationship should exist.

In confirmatory factor analysis it is common practice to deal with the bias-variance

trade-off in a brute-force manner, by imposing a so-called simple structure. While generally, the model allows for *some* cross-loadings to not be fixed to zero, this practice entails fixing all cross-loadings 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 Muthén and Asparouhov (2012) proposed *Bayesian CFA*, an alternative, more flexible approach for identifying CFA models, which can be viewed as a form of regularized SEM. 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 Muthén 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 (Muthén & 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.

- ADD little alinea on Bayesian LASSO in SEM here?

The 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). This prior is a discrete mixture of an extremely peaked prior around zero (the spike), and a very flat prior for larger parameters (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):

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

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

The basic intuition is as follows. When $r_{jk} = 1$, $\lambda_{c,jk} \sim \mathcal{N}(0, c^2)$, hence $\lambda_{c,jk}$ is assigned to the slab. When $r_{jk} = 0$, $\lambda_{c,jk} \sim \delta_0$, and is thus assigned to the spike. This ensures that large cross-loadings, that are are relevant are not shrunken while small, negligible cross-loadings are shrunken to zero.

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,

especially under favorable conditions with large sample sizes and cross-loadings. However, the Spike and Slab Prior cannot be implemented in STAN, as STAN does not allow for discrete mixture priors (Betancourt, 2018; Stan Development Team, 2021). Not only is STAN the most advisable MCMC-package for complex, highly-dimensional Bayesian models. STAN also forms the basis of the R-packages that allow to apply Bayesian SEM in practice (e.g., Blavaan, Merkle et al., 2022). It is thus crucial that a prior with the desired properties can be implemented in STAN, such that it can be added to existing software packages, making it accessible to applied researchers. This calls for a *non-discrete* alternative shrinkage-prior that also outperforms the SVNP within a single estimation step.

The Regularized Horseshoe Prior (RHSP)

A fully continuous alternative to the Spike and Slab prior that is implementable 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 the original Horseshoe Prior is that there is a *global shrinkage parameter* τ, shrinking all cross-loadings to zero. Next to this, there is a *local shrinkage parameter* $\bar{\omega}_{jk}$[1] that allows truly large cross-loadings to escape the shrinkage, by setting thick Cauchy tails for the local scales ω_{jk} (Polson & Scott, 2010). Formally, the Horseshoe prior for every cross-loading of factor j on item k is specified as follows:

$$\lambda_{c,jk}|\omega_{jk}, \tau, c \sim \mathcal{N}(0,\ \omega_{jk}^2\tau^2),\ where$$

$$\omega_{jk} \sim \mathcal{C}^{+}(0, 1).$$

The name-giving intuition behind the horseshoe prior becomes clear when considering the finding that (Carvalho et al., 2010; Piironen & Vehtari, 2017b):

$$\bar{\lambda}_{c,jk} = (1\ -\ k_{jk})\hat{\lambda}_{c,jk},\ where$$

[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.

$$k_{jk} = \frac{1}{1 + n\sigma^{-2}\tau^2 s_{jk}\omega_{jk}^2}.$$

Here k_{jk}, denotes the so-called *shrinkage factor* for cross-loading $\lambda_{c,jk}$. When plotting the density of k_{jk} there is a very high peak at at very low values and a very high peak of high values, resulting in a plot that resembles a horseshoe, illustrating that the Horseshoe Prior has the desired property of either shrinking parameters very little, or very much, with very few parameters that are shrunken in a non-extreme fashion.

The Horseshoe Prior was found consistently to possess the theoretical properties outlined above in practice (Carvalho et al., 2010; Datta & Ghosh, 2013; Polson & Scott, 2010; Van Der Pas, Kleijn, & Van Der Vaart, 2014). However, due to its Cauchy tails it suffers from the same issues as a Cauchy prior. Specifically, not shrinking large parameters at all can lead to estimation issues, especially when parameters are weakly identified. This happens for instance in logistic regression with separable data, where a flat likelihood and thereby a weakly identified model arises (Ghosh, Li, & Mitra, 2018). The RHSP prevents such issues by shrinking also large parameters a little bit. 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 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). Note that we specify the RHSP in its most general form. Setting the degrees of freedoms of the half-t-distributions to 1 results in half-Cauchy distributions. Strictly speaking, the prior is only a Regularized *Horseshoe* Prior when this is the case. In

the current study we vary the degrees of freedoms of all scale parameters to assess the extent to which the sparcifying properties as well as the convergence of the RHSP are influenced by these parameters.

The intuition of how the RHSP shrinks large parameters a little bit is best illustrated by assuming that c is a given constant. Now, when $\tau^2\omega_{jk}^2 < c^2$, which is the case under small cross-loadings, $\bar{\omega}_{jk}^2 \rightarrow \omega_{jk}^2$. Hence, in this case the RHSP approaches the original Horseshoe Prior, with equally pronounced shrinkage to zero. However, when τ is far from zero, hence with large cross-loadings, $\tau^2\omega_{jk}^2 > c^2$, and $\bar{\omega}_{jk}^2 \rightarrow \frac{c^2}{\tau^2}$. Then, the prior of $\lambda_{c,jk}$ approaches a slab $\mathcal{N}(0, c^2)$. Under the above specification, when c is no constant but a parameter for which an Inverse-Gamma hyper-prior is set, the slab becomes a t-distribution with df_{slab} degrees of freedom, a mean of zero and a scale of $scale_{slab}^2$ (Piironen & Vehtari, 2017b).

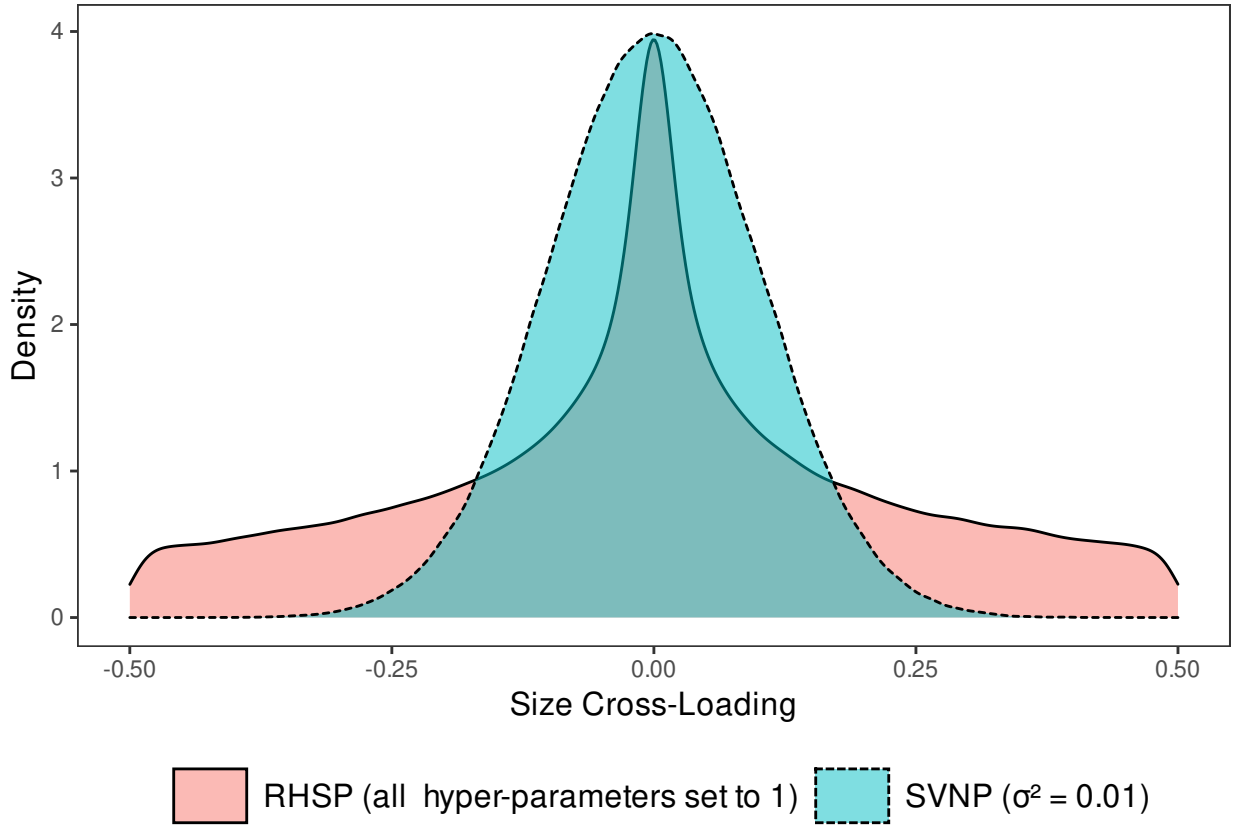

Figure 1. Density Plots of the Regularization Priors of Interest.

Figure 1 compares the two shrinkage-priors that are the focus of our study. Both priors share a large peak at zero, which ensures that cross-loadings are shrunken to(wards) zero. However, the RHSP has much thicker tails. Here, for larger cross-loadings, there is thus much more prior mass than with the SVNP. This ensures 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 regularizing cross-loadings in CFA. We therefore aim to compare the RHSP to the SVNP in their performance in regularizing cross-loadings in Bayesian CFA.

Study Procedure and Parameters

A Monte Carlo simulation study was conducted using STAN (Stan Development Team, 2021) and R (R Core 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`.

Conditions

Population 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 true model is summarized below, both in equations (Appendix A) and graphically (Figure 2).[3] 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, 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 for the RHSP unfeasible. This is appropriate because for simple factor models researchers are unlikely to collect such larger sample sizes in practice.

SVNP: Prior Conditions. We varied the hyper-parameter of the SVNP σ^2 between 0.001, 0.01 and 0.1, based on Muthén and Asparouhov (2012). For the SVNP 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 for this prior.

RHSP: Prior Conditions. The RHSP has six hyper-parameters in the specification that we apply. We varied the scales of the global shrinkage parameter τ, s_{global} between, 0.1 and 1. Here 1, is a natural maximum given that the scale would never be larger than 1 when applying a prior guess p_0, and 0.1 a logical minimum given the scale of the model. Also the scale of the local shrinkage parameter ω_{jk} was varied between, 0.1 and 1. The degrees of freedoms of these two parameters, df_{local} and df_{global} were varied between 1 and 3. For the local shrinkage parameter, larger degrees of freedoms may help to

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

overcome sampling issues that can arise when $df_{local} = 1$, i.e. when the prior reduces to a half-Cauchy prior. Finally, for the scale of the distribution of c^2, $scale_{slab}$ was varied between 0.1, 1 and 5, and df_{slab} between 1 and 3. This left a total of 96 individual hyper-parameter conditions for the RHSP. In combination with the 2x2 population conditions we were left with 384 individual sets of conditions for this prior. In total there were thus $384 \times 200 = 76800$ replications run for this prior.

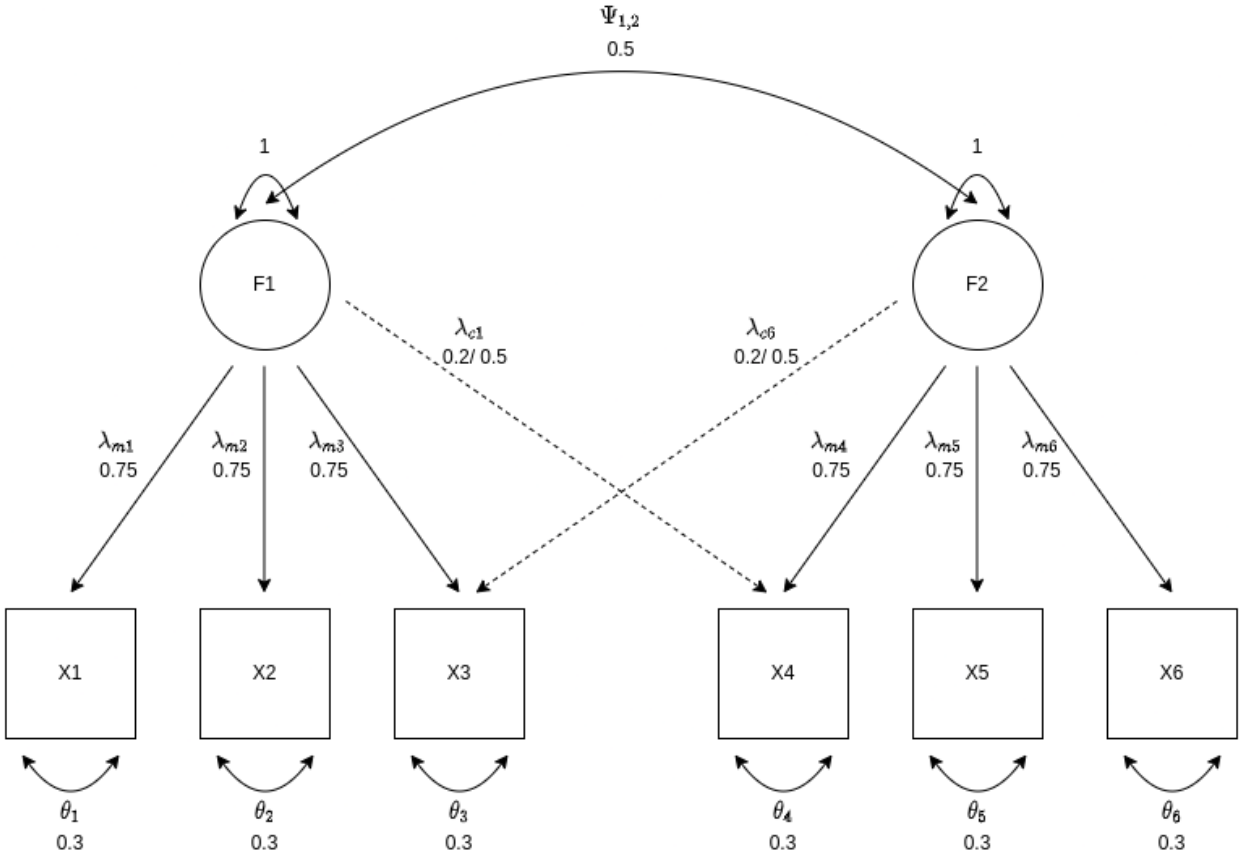

Figure 2. Graphical Representation of the True Model.

Outcomes

All outcomes[4] were computed based on both mean and median posterior estimates of the model parameters. We only present the results of the mean estimates, but those

[4] Summaries of all outcomes can be found on https://github.com/JMBKoch/1vs2StepBayesianRegSEM/tree/main/Rmd/plots.

concerning the median estimates (which do not differ relevantly from those of the mean estimates) can be accessed on github[5].

Mean Absolute Bias. For every model parameter θ and for every set of conditions that has been sampled from for N_{rep} replications, we computed the Mean Absolute Bias:

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

Given that the core issue of the SVNP is biased model estimates, this outcome naturally plays a central role in our study.

Relative Bias. The (Mean) Relative Bias was computed per model parameter estimate and set of conditions by dividing the estimates of the Mean Absolute Bias by the true value of the parameter:

$$\bar{Bias}_{rel,\ \bar{\theta}} = \frac{\bar{Bias}_{\bar{\theta}}}{\theta_{true}}.$$

This outcome gives an indication of the magnitude of the bias by expressing it relative to the parameter's true value. However, given the standardized scale of the true model, the Mean Absolute Bias is a quantity that can be interpreted rather intuitively in the context of this study. We therefore do not discuss these results in detail, and refer the interested reader to the study repository on github[6].

Mean Squared Error: The Mean Squared Error (MSE) was computed per model parameter and set of conditions as:

$$MSE_{\bar{\theta}} = \frac{1}{N_{rep}} \Sigma_{i=1}^{N_{rep}} (\bar{\theta}_i - \theta_{true})^2.$$

Another way to express the MSE is as the sum of the bias and the variance of a model parameter, which explains its added value over the Mean Absolute Bias alone. As with the Relative Bias we refrain from presenting results here as they do not add to the conclusions based on the Mean Absolute Bias[7].

[5] see TBA LINK for the SVNP and TBA LINK for the RHSP

[6] see TBA LINK for the relative bias of the SVNP and TBA LINK for the relative bias of the RHSP

[7] MSE estimates and plots can be found on TBA Link for the SVNP and TBA LINK for the RHSP

Power and Type-I-Error Rate. We computed the Mean Power (true positive rate) per set of conditions in selecting truly non-zero cross-loadings as non-zero by calculating the proportion of replications where the truly non-zero cross-loadings were selected as non-zero, and averaging this over the 2 truly non-zero cross-loadings.

The Mean Type-I-Error (false positive) rate in selecting truly zero cross-loadings as non-zero, was computed as the proportion of truly zero cross-loadings selected as non-zero, averaged over the four truly zero cross-loadings.

For both of these outcomes, we applied a variety of selection criteria for selecting cross-loadings as non-zero, based on Zhang et al. (2021). First, we used a variety of thresholding rules, where a cross-loading is selected as non-zero when the absolute value of its estimate exceeds a specific threshold: 0, 0.05, 0.1, 0.15. Next, we we considered four credible intervals (50%, 80%, 90%, 95%), where cross-loadings are selected as non-zero when the interval does not contain zero.

Results

Convergence

SVNP. In terms of convergence, the SVNP showed excellent performance. Across all 2400 replications there was no single parameter for which $\widehat{R} > 1.05$. 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 were therefore included in the results.

RHSP. The RHSP showed weaker performance in terms of convergence than the SVNP, although with most hyper-parameter configurations it was still acceptable, especially conisdering the very complex nature of the underlying model.

A total of 156 replications failed entirely. They all happend under one set of

conditions: N = 100, size $\lambda_{c1,6} = 0.2, N = 100, scale_{global} = scale_{local} = scale_{slab} = 0.1, df_{global} = df_{local} = df_{slab} = 1$. This likely happened due to identification issues. We removed the remaining 44 replications under this set of conditions, as they were too little to give a reliable picture.

Next, we removed all replications in which at least one model parameter had a value of $\hat{R} >= 1.05$, or a value for N_{eff} smaller than 10% of the chain-length. This removed a total of 542 replications. The maximum number of removed replications for a given set of conditions was 37, which corresponds to 18.5% of the of replications under these conditions. Below in Table 1 we present all combinations of conditions under which more than 5% of the replciations had to be removed.

Table 2 presents all sets of conditions under which there were, on average, at least 5% divergent transitions per chain. We decided not to remove such replications, as this would have remove a substantial number of 4474 replications. In general, it is advised not to included any divergent transitions, since they introduce bias. Given the complex nature of the RHSP it is hard to follow this advise in practice. However, it needs to be taken into account in the interpretation of the findings that the divergent transions may have added bias to the model estimates of the RHSP.

Main Results

SVNP: Mean Absolute Bias. The Mean Absolute Bias of the SVNP for all parameters is summarized in Figure 3. For parameter estimates that show an identical pattern ($\bar{\lambda}_{c2-5}$, $\bar{\lambda}_{c1,6}$, $\bar{\lambda}_{m1,2,5,6}$, $\bar{\lambda}_{m3-4}$, and $\bar{\theta}_{1-6}$), the first respecting estimate is presented representative for all, both in Figure 3 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,

Table 1

Conditions under which more than 5% of replications were removed due to not reaching convergence (N = 542).

$scale_{global}$	df_{global}	$scale_{local}$	df_{local}	$scale_{slab}$	df_{slab}	N	Size $\lambda_{c1,6}$	N removed Rep.
0.10	3	0.10	1	0.10	1	100	0.50	10
0.10	3	0.10	1	1.00	3	100	0.50	11
0.10	1	0.10	1	5.00	3	100	0.50	12
0.10	3	0.10	1	5.00	1	100	0.50	12
0.10	3	0.10	1	1.00	1	100	0.50	13
0.10	3	0.10	3	0.10	3	100	0.50	13
0.10	1	0.10	1	5.00	1	100	0.50	15
0.10	3	0.10	1	5.00	3	100	0.50	15
0.10	1	0.10	3	0.10	1	100	0.50	20
0.10	1	0.10	3	1.00	1	100	0.50	24
0.10	1	0.10	3	1.00	3	100	0.50	24
0.10	1	0.10	3	5.00	3	100	0.50	27
0.10	1	0.10	3	5.00	1	100	0.50	30
0.10	3	0.10	3	0.10	1	100	0.50	33
0.10	3	0.10	3	1.00	1	100	0.50	34
0.10	3	0.10	3	5.00	1	100	0.50	34
0.10	3	0.10	3	1.00	3	100	0.50	37
0.10	3	0.10	3	5.00	3	100	0.50	37

Note. Replications were removed for having an $\hat{R} >= 1.05$ or an N_{eff} smaller that 10% of the chain-length, for any of the model parameters.

Table 2

Conditions with on average more than 5% divergent transitions.

$scale_{global}$	df_{global}	$scale_{local}$	df_{local}	$scale_{slab}$	df_{slab}	N	Size $\lambda_{c1,6}$	Mean Prop. Div.
0.10	1	0.10	3	0.10	1	100	0.50	0.09
0.10	1	0.10	3	1.00	1	100	0.50	0.08
0.10	1	0.10	3	5.00	1	100	0.50	0.08
0.10	1	0.10	3	5.00	3	100	0.50	0.08
0.10	3	0.10	1	0.10	1	100	0.50	0.10
0.10	3	0.10	1	1.00	1	100	0.50	0.08
0.10	3	0.10	1	5.00	1	100	0.50	0.09
0.10	3	0.10	1	5.00	3	100	0.50	0.08
0.10	3	0.10	3	0.10	1	100	0.50	0.10
0.10	3	0.10	3	1.00	1	100	0.50	0.11
0.10	3	0.10	3	1.00	3	100	0.50	0.07
0.10	3	0.10	3	5.00	1	100	0.50	0.11
0.10	3	0.10	3	5.00	3	100	0.50	0.12

Note. There was a total of 4474 replications were the divergent transitions exceeded 5% of the chain-length. There were 19036 replications with more than 1% of divergent transitions. There were 1970 replications with more than 10% of divergent transitions. There were 186 replications with more than 50% of divergent transitions.

to not distract from our main conclusions.[8]

Figure 3 shows that, as expected, there was substantial bias in some parameter estimates. While the bias in the posterior means of the truly zero cross-loadings $\bar{\lambda}_{c2-5}$ was relatively small, it was 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 was 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 played a crucial role here. Also with $\sigma^2 = 0.01$ (and true cross-loadings of 0.5) substantial bias occured (e.g. $\bar{Bias}_{\bar{\lambda}_{c1}} = 0.35$), as the cross-loading were 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 was less pronounced (e.g. $\bar{Bias}_{\bar{\lambda}_{c1}} = 0.14$). Here σ^2 was 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 ($\bar{\lambda}_{m4}$) were substantially biased when the true cross-loadings were 0.5 and $\sigma^2 = 0.001$ (e.g. $\bar{Bias}_{\bar{\lambda}_{m3}} = 0.40$). These two loadings showed much higher bias than the other four main-loadings as they loaded 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 were shrunken to zero, these main loadings now also had to account for the variance in the items that was truly explained by the cross-loadings. Consequently, the two main-loadings were over-estimated, e.g. $\bar{\lambda}_{m3} = 1.15$.

In the factor correlation the bias was also relatively small and approximately the same for the different values of σ^2 when the truly non-zero cross-loadings were 0.2. Again, bias became much more pronounced with true cross-loadings of 0.5, especially when

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

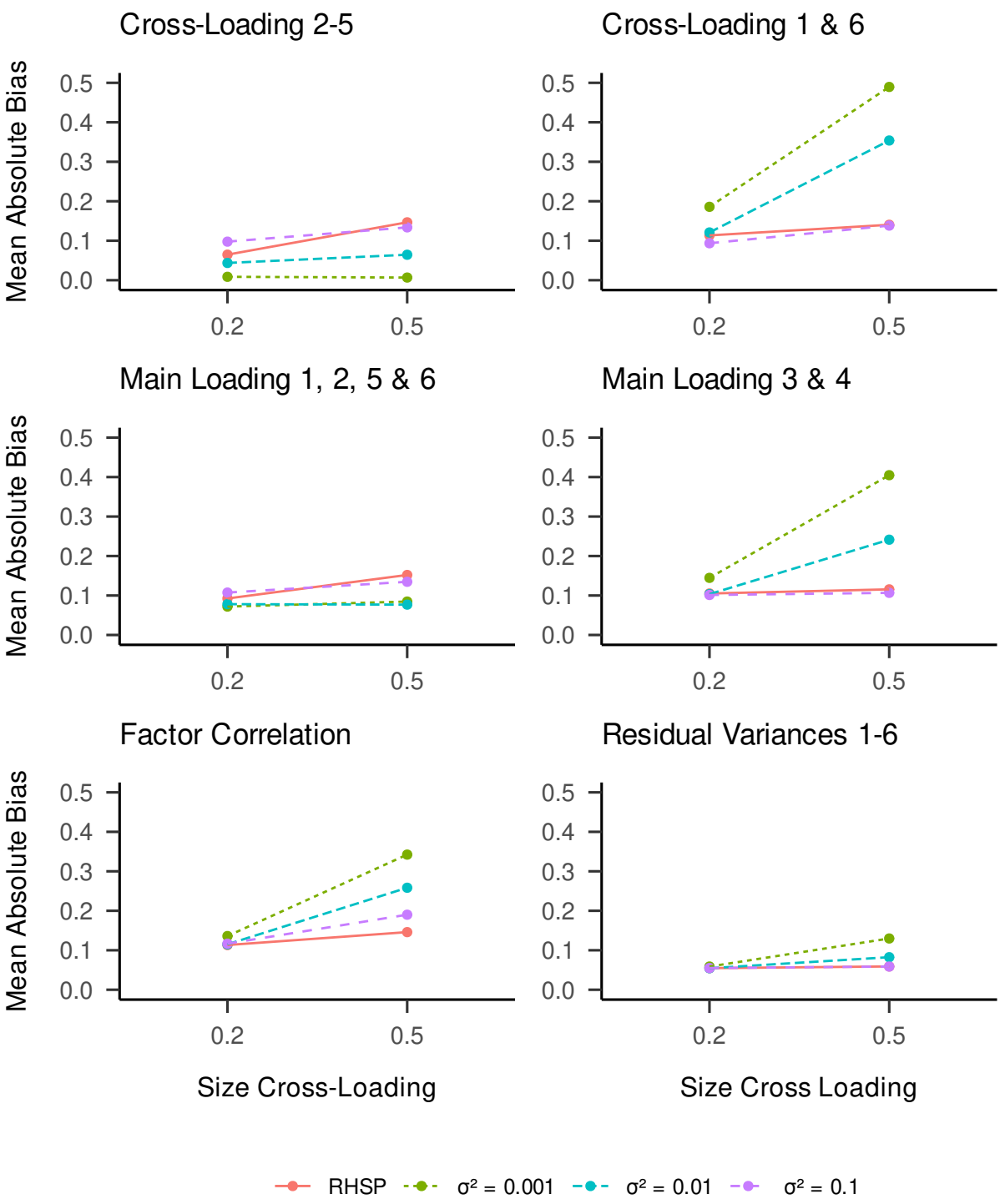

Figure 3. Mean Absolute Bias in the Model Parameters (N = 100). Per set of parameters that showed an identical pattern, the first parameter was used to represent all other parameters, e.g. cross-loading 1 was plottet representative for cross-loading 1 and 6. All hyperparameters of the RHSP are set to 1 in the results presented here.

$\sigma^2 = 0.001$ ($\bar{Bias}_{\bar{r}} = 0.34$). In this situation the factor correlation was heavily over-estimated ($\bar{\Psi}_{1,2} = 0.84$). This was because the covariance between item 3 and 4 that arose from the two cross-loadings, was mis-attributed to the factor-correlation, as the cross-loadings were shrunken to zero.

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

RHSP: Mean Absolute Bias.

SVNP: Power and Type-I-Error Rate. The top left panel Figure 4 summarizes the Mean[9] Power (true-postive rate) in selecting the truly two non-zero cross-loadings as non-zero of the SVNP, per set of conditions and selection criterion. The horizontal red dash line indicates the recommended minimum power of .80 suggested by First, we can see that with a treshold of 0.00 there is a perfect power of 1 in selecting. This is logical, since in posterior means will never be entirely zero. Therefore this results mostly servers this porperty of Bayesian inference and thereby the need for more complex selection rules in Bayesian regularization. Next, we can see that across most conditions.

RHSP: Power and Type-I-Error Rate.

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

[9] We present the Mean Power, averaged over the two truly non-zero cross-loadings, as seperating cross-loadings leads to identical conclusions.

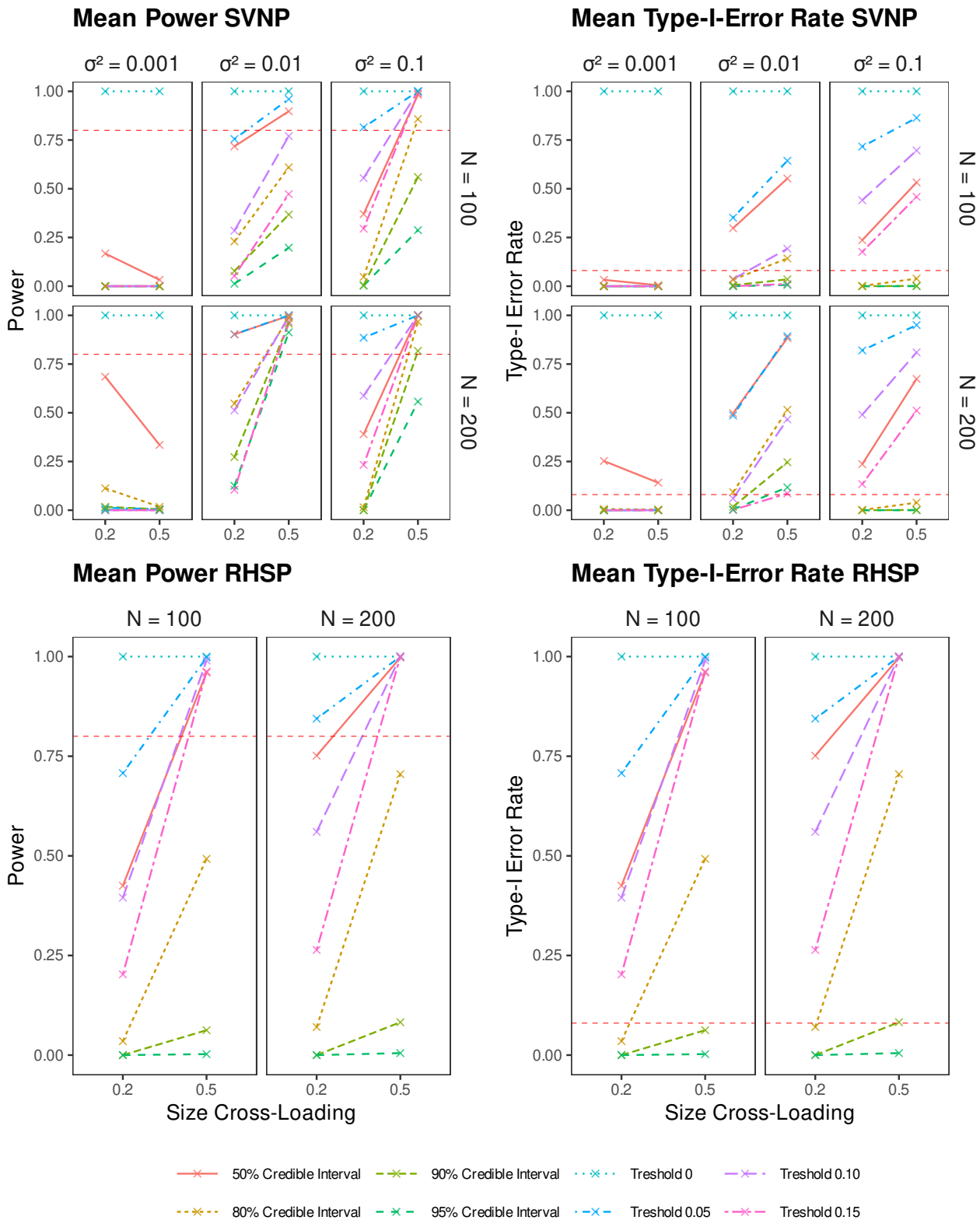

Figure 4. Mean Power and Type-I-Error Rates in Selecting non-zero Crossloadings. All hyper-parameters of the RHSP are set to 1 in the results presented here.

$\sigma^2 = 0.001$, 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. Muthén 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, size cross-loadings = 0.5) this power was reached in our study (see Table B1), which suggests that also the 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 Muthén and Asparouhov (2012).

Future Research:

- More factors
- Residual Co-variances

- Binary, ordinal, nominal outcomes
- Larger sample sizes
- Invite reader to do this based on my code.

Other important steps: - Implementation in Practice!

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.

Datta, J., & Ghosh, J. K. (2013). Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis*, *8*(1), 111–132.

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

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

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

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

Merkle, E., Yves Rosseel, Ben Goodrich, Mauricio Garnier-Villarreal, Terrence D. Jorgensen, Huub Hoofs, … Matthew Emery. (2022). Blavaan: Bayesian Latent Variable Analysis. Retrieved from https://cran.r-project.org/web/packages/blavaan/blavaan.pdf

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

Muthén, 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

Polson, N. G., & Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, *9*(501-538), 105.

R Core Team. (2021). R: A Language and Environment for Statistical Computing. Retrieved from https://www.R-project.org/

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

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *73*(3), 273–282. Retrieved from https://www.jstor.org/stable/41262671

Van Der Pas, S. L., Kleijn, B. J., & Van Der Vaart, A. W. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, *8*(2), 2585–2618.

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

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

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