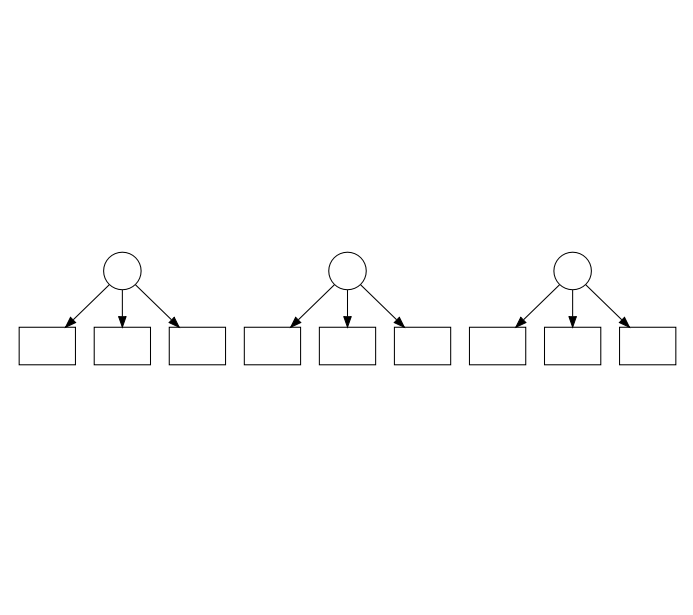
When specifying the measurement component of a structural equation model (SEM) careful thought has to be given to the specific pattern of factor loadings that will be estimated. A number of concerns will motivate this decision. The loading pattern should reflect substantive theory, there should be sufficient flexibility to allow for unpredicted findings to arise from observed data, the resulting model fit must be adequate, and the practical concerns of model identifiability must be met. Unfortunately these concerns are often in conflict with one another, for example a more flexible specification may complicate the relation between model and theory (Stromeyer, et al., 2015). As a result it is necessary that some compromise is reached between such concerns. How this compromise can best be reached is an open question, with an increasing recognition that traditional approaches may not provide an optimal balance.

The independent clusters (IC) factor loading pattern, where only a single factor loading is estimated for each observed variable, represents the most common compromise between the above concerns in specifying a loading matrix. Figure 1. provides an example diagram and loading matrix for an IC structure with three latent and nine observed variables. The majority of loading parameters are fixed at zero, with only a small proportion freely estimated.

FIGURE 1 GOES HERE



By only estimating a single loading for each observed variable the IC structure ensures model identifiability under reasonable conditions, allows for straightforward mapping of theory on to the model, and provides easily interpretable results. However, it achieves this at the cost of model flexibility. Better fitting loading patterns may exist but the IC structure is unable to provide evidence for, or against, these due to its enforced rigidity. As such, the IC structure is increasingly being recognised as an unsatisfactory solution to the problem of specifying an effective loading pattern (Asparouhov & Muthén, 2009; Muthén & Asparouhov, 2012).

Muthén and Asparouhov (2012) propose Bayesian Structural Equation Modelling (BSEM) as a method for simultaneously estimating all possible factor loadings in SEM. This method overcomes the limitations of IC structure by allowing for any loading estimate to differ from zero if the observed data suggest it, greatly increasing flexibility in SEM loading specification. This is achieved by setting informative Bayesian priors, specifically small variance Gaussian distributions with mean zero, on all cross-loading parameters, rather than fixing them to zero as in IC structure. This allows for cross-loading estimates to differ from zero whilst maintaining model identifiability.

However BSEM, as proposed by Muthén and Asparouhov, achieves this at the expense of the clear relationship between model and theory that IC models promote. As Stromeyer, et al. (2015) discuss in their critique of the method, the use of Gaussian priors in this manner can result in a complex pattern of cross-loadings, with sampling error alone resulting in a large number of apparently non-zero cross loading parameters. This not only complicates the interpretation of the latent variables but also makes one goal of BSEM, to systematically identify reliably non-zero cross-loadings, more difficult as some decision criteria must then be chosen to separate ‘significant’ from ‘non-significant’ cross loadings.

BSEM would therefore seem to free SEM from a restrictive approach to cross-loading estimation only at the expense of losing the direct and easy to interpret relationship between theory and model that makes SEM such a productive tool. However, the literature on Bayesian regularisation, concerned with modelling sparse parameter sets where most population values are equal to zero, may provide a means to achieve systematic cross-loading estimation whilst maintaining a sparse, easy to interpret cross-loading pattern similar to IC structure.

This paper investigates the potential for using informative priors common in the Bayesian regularisation literature, instead of Gaussian priors, to estimate SEM cross-loadings. It was hoped that the use of sparsity promoting priors would provide a better compromise between model flexibility and interpretability than either BSEM with Gaussian priors or IC models.

A brief overview of the principles of Bayesian regularisation is first provided, and candidate prior distributions are introduced. Simulations comparing performance of small variance Gaussian priors against two priors from the regularisation literature, the horseshoe and horseshoe plus, are then reported and discussed. Finally, the impact of prior distribution selection on cross loading estimation is explored using the Holzinger and Swineford (1939) ability testing dataset.

**Bayesian Regularisation and Sparse Estimation**

Regularisation describes a range of approaches to model estimation which broadly seek to simplify the estimated parameter solution, subject to some penalty on model complexity (Hastie, Tibshirani, & Friedman, 2001). This almost always entails certain parameter estimates being ‘shrunk’ towards zero, relative to a non-regularised solution, in order to better meet the conditions of the penalty. For this reason regularisation is also often referred to as shrinkage modelling.

Regularisation models have gained popularity in disciplines where complex models are at risk of overfitting sample data, or where a large number of candidate predictor variables make traditional means of variable selection (e.g. null hypothesis significance testing) untenable. Certain regularisation techniques can also result in improved performance in poorly posed modelling situations, such as in the presence of high multicollinearity (Vinod, 1978). They can also enable the estimation of models which would have been unidentified without the application of regularisation, such as regression models where candidate predictor variables outnumber data observations (Zou & Hastie, 2005).

Within a frequentist framework regularisation is achieved by adding some penalty on model complexity to the relevant loss/likelihood function. For example, a linear ridge regression model is estimated via ordinary least squares (OLS) plus an additional penalty that is proportional to the sum of squared regression coefficients (Hoerl & Kennard, 1970), as below.

The parameter λ is a tuning parameter which controls the strength of regularisation to be applied. When λ = 0 the model reduces to basic OLS regression, whilst as it increases all relevant coefficients will be asymptotically shrunk to zero. The value of λ in frequentist regularisation is typically selected using cross-validation (Tibshirani, 1996). Different specifications of the penalty term will result in different shrinkage behaviour. For example, the extremely popular LASSO regularisation method uses a linear, rather than quadratic, penalty term (Tibshirani, 1996) whilst Elastic Net regularisation uses a weighted mixture of linear and quadratic penalties (Zou & Hastie, 2005).

Comparable results are achieved in a Bayesian context by setting tight prior distributions, centred on zero, on the relevant model parameters (typically regression coefficients). By specifying prior distributions with a high probability density near zero, equivalent shrinkage of posterior estimates towards zero can be achieved, bringing with it the benefits associated with frequentist regularisation described above (Park & Casella, 2008). So long as they have a high probability density near zero, a variety of different distributions can be used to achieve different regularisation properties, similar to the use of different complexity penalties in frequentist regularisation.

Specific prior distributions have been shown to produce regularised estimates closely related to common frequentist regularisation methods. For example, the frequentist LASSO is broadly equivalent to a Bayesian model in which Laplace distributions are set as priors on the relevant parameters (Park & Casella, 2008). Importantly for the present discussion, Bayesian ridge regression is achieved by setting small variance Gaussian priors on the regression coefficients (Pasanen, Holmström, & Sillanpää, 2015). The BSEM approach to cross loadings advocated by Muthén and Asparouhov (2012) can therefore be interpreted as a use of ridge regression, applied only to cross loading parameters, to overcome the problem of identifiability when all SEM loadings are freely estimated.

This perspective casts the critique of Stromeyer, et al. (2015) in a productive light. Their major concerns with BSEM, using small variance Gaussian priors, relate to the fact that models estimated in this way will commonly display a relatively complex pattern of cross loading estimates. They raise the issue that having such a large number of, theoretically secondary, loadings estimated at non-zero values harms the interpretability of the estimated model. They also raise the concern that the observed tendency of estimated cross loading posteriors to follow the data away from zero, could complicate model falsifiability with even poorly stated models achieving reasonable fit through an overfitting of cross-loading parameters.

These issues arise due to the behaviour of ridge regression as a regularisation method, specifically that it does not promote sparse parameter solutions. This is best seen by comparison with LASSO regression, which does promote sparsity. **Figure 2** shows, on a simulated dataset, how LASSO and ridge regression coefficient estimates change as the regularisation penalty parameter is increased. Even for very large values of λ ridge regression produces non-zero estimates for almost all variables. The LASSO by comparison quickly shrinks most estimates to zero, retaining only a sparse set of non-zero estimates for most values of λ.

FIGURE 2 GOES HERE

A potential development of BSEM, in response to the concerns of Stromeyer, et al. (2015), is therefore to explore the use of regularisation priors that promote a sparse parameter solution, rather than the noisy solution provided by the use of Gaussian priors (comparable to ridge regression). Such a sparse solution could not only aid interpretability of the estimated model, but could also help avoid the overfitting of cross-loadings near zero that the use of Gaussian priors may invite. The literature on Bayesian regularisation provides several such candidate priors.

**Sparsity Promoting Prior Distributions**

A number of prior distributions can be used to promote sparsity in Bayesian regularisation models. The Bayesian LASSO is achieved by setting a Laplace prior distribution on the parameters to undergo regularisation (Park & Casella, 2008). The Laplace distribution can be parameterised in a number of equivalent forms and is sometimes referred to as the double exponential distribution (Andrews & Mallows, 1974) as it can be expressed as two exponential distributions placed back to back, centred over zero.

In addition to the Laplace, the Horseshoe prior (Carvalho, Polson, & Scott, 2009) is a recent development, which promotes sparsity more powerfully than the LASSO but is also better placed to handle substantive non-zero estimates due to its fatter tails. Following the notation of Carvalho, Polson and Scott (2009), the Horseshoe distribution can be described as a Gaussian distribution whose variance is the product of two other random variables. If represents a vector of values, distributed according to a horseshoe distribution, then its conditional distribution can be written as follows:

Here represents the Cauchy plus distribution, a Cauchy distribution limited to non-negative values (Gelman, 2006). By associating both ‘global’ and ‘local’ parameters with each observation, the horseshoe distribution allows for flexibility in the estimation of sparse parameter sets that is lacking from the Laplace distribution. Comparisons tend to show the use of horseshoe priors outperforming the Bayesian LASSO (Peltola, Havulinna, Salomaa, & Vehtari, 2014).

A further development on the horseshoe prior is the horseshoe+ prior, which was explicitly designed to provide more optimal estimation of sparse parameter sets than competitors, including the horseshoe and Laplace distributions (Bhadra, Datta, Polson, & Willard, 2015). The horseshoe+ is defined similarly to the horseshoe distribution, but with an additional ‘local’ half Cauchy mixture component in the variance term.

The Laplace, horseshoe, and horseshoe+ distributions were all considered as alternatives to the Gaussian priors of Muthén and Asparouhov (2012) in the below simulations. **Figure 3** shows how these distributions differ from one another, both at the centre of the distribution and in the tails[[1]](#footnote-1). The key properties they share are large probability volume around the centre of the distribution, and fat tails by comparison with the Gaussian distribution. The high density near zero promotes sparsity by strongly shrinking small estimates towards zero. The fat tails help reduce shrinkage of values that are further from the central value.

FIGURE 3 GOES HERE

**Simulation Study of BSEM Using Sparsity Promoting Priors**

In order to compare different cross-loading priors, data were simulated from two ‘true’ latent population models. One where there were a small number of non-zero cross loadings, and one where all population cross-loadings were zero (that is, the IC structure was accurate). As researchers will likely not know whether there are any non-zero cross-loadings in advance of modelling, including this second condition provides a fuller evaluation of the performance of BSEM in realistic situations.

Both datasets were simulated using the Simsem R package (Pornprasertmanit, Miller, & Schoemann, 2015) from a population model with 15 observed variables, each with 5 main loadings on to 3 latent variables with values between 0.4 and 0.8. Latent variables were uncorrelated and observed variables were all conditionally independent on the latent variables. All latent and observed variables had population variances of 1.

The only difference between the IC population model and the sparse cross-loadings (SCL) model were that in the former all cross-loadings were equal to 0, whilst in the SCL model 3 cross-loadings were given non-zero population values. The two loading matrices are shown in figure 4. 100 datasets, each with n = 200, were simulated for both the IC and SCL models.

FIGURE 4 GOES HERE

**IC Factor Matrix**  **SCL Factor Matrix**

The simulation study aimed to evaluate model behaviour on four points:

1. **The ability to estimate population parameters with minimal bias.** It was expected for cross-loadings with a population value of zero that all priors would perform equivalently in returning un-biased estimates. However it was expected that for non-zero values the Gaussian prior would show larger bias towards zero than the alternative distributions due to its thinner tails.
2. **The variation of estimates from sample to sample.** As it does not promote sparsity, the Gaussian prior was expected to show considerably greater variation in its estimates from sample to sample, particularly in its estimates of cross-loadings with population values of zero.
3. **Effect of prior choice on model fit.** It was expected that the Gaussian priors would provide poorer out of sample model fit than the alternatives, due to the larger bias of estimates for non-zero parameters, and to the overfitting of near zero values.
4. **Model Complexity**. It was expected that the Gaussian prior would result in a more complex model than the alternatives (as measured by the effective number of parameters) due to its non-sparse solutions.

A direct comparison of model interpretability is provided in a later section using a single non-simulated dataset.

**Modelling Information**

All BSEM models were estimated using the Stan probabilistic programming language (Carpenter, et al., in press) from within R using the RStan package. Stan was chosen as its use of Hamiltonian Monte Carlo (implemented using its No-U-turn sampler) tends to provide better Markov-chain Monte-Carlo (MCMC) sampling behaviour than alternative sampling procedures such as Metropolis-Hastings or Gibbs sampling (Hoffman & Gelman, 2014).

Gaussian priors with a mean of 0 and standard deviation of 3 were set on the item intercepts. Primary factor-loadings were set with vague priors so as not to influence estimation; these were Gaussians with mean zero and standard deviations which were in themselves estimated from the data (priors on the standard deviation parameters were uniform(0, 100)). Primary loading parameters were also restricted to positive values to achieve model identification.

Whilst their population values equalled zero, latent variable correlations were estimated to allow for sampling variation. Observed variables were modelled as conditionally independent without correlated residuals. All fitted models were identical except for the prior distributions set on the cross-loading parameters. The Stan code for each model can be found at (GITHUB GOES HERE).

All models were explored individually to ensure that the estimation process converged reliably in each case. This meant that the number of MCMC iterations ran, the thinning parameters used, and the Metropolis acceptance step parameters, differ somewhat between models, depending on how well the MCMC process sampled from the posterior in each case. In all cases between 1000 and 2000 MCMC iterations were ran with a burn in of 50% of samples. Thinning was between three and five cases. Reliable convergence was identified in all cases before simulations were run. Full details are recorded in the simulation code which can be found at (GITHUB GOES HERE).

Approximate leave-one-out cross-validation (LOO) was used a measure of model fit for comparison (Vehtari, Gelman, & Gabry, 2016). LOO is a recently proposed measure of Bayesian model fit which asymptotically estimates full leave-one-out cross validation but without the need to carry out the computationally expensive process of cross validation. This is achieved through log-likelihood evaluations using samples from the fitted posterior distribution (full details are reported in Vehtari, Gelman and Gabry, 2016). The strength of LOO is it’s ability to produce out of sample fit estimates (helping to identify and avoid overfitting)whilst avoiding identified issues with alternative such as the deviance information criterion (DIC; Plummer, 2008). LOO was calculated for all models after estimation using the loo package for R. This package was also used to estimate the effective number of parameters for each model in order to compare model complexity.

**Laplace prior estimation**

During both the initial tests of model convergence, and later when the full simulation study was ran, issues were identified with the convergence of models using the Laplace prior distribution on it’s cross loading parameters. Even when using a large set of MCMC samples, and taking steps such as increasing the thinning parameters, convergence of these models was found to be poor, with non-convergence more common than not.

As reliable convergence could not be achieved the Laplace prior is not considered further below. It was considered that this failure was worth reporting as the LASSO method which the Laplace prior achieves is the most popular method of sparsity promoting regularisation. As such it’s exclusion from consideration entirely would have been peculiar.

**Sparse Cross Loading Comparison**

Simulations were first ran on the SCL datasets to compare the performance of different prior values in the presence of a small number of non-zero cross loadings. Figures 5 - 7 display the median and 95% coverage interval of all cross-loading parameter estimates over the 100 simulated datasets (the parameter estimate here is the posterior median) for each prior distribution.

FIGURES 5 – 7 GO HERE

The horseshoe and horseshoe+ priors both reliably produced near exact estimates of zero where the population value was zero. Across almost all simulated models they resulted in an appropriately sparse solution to help mitigate potential difficulties of model interpretation. The Gaussian prior, whilst providing a median estimate across all models very near to zero, displayed a high level of variation from sample to sample, highlighting it’s susceptibility to over fit estimates based on sampling variation.

The horseshoe and horseshoe+ priors also produced a less biased estimate of the cross-loadings with population values of .6 and .4 than the Gaussian. Due to their relatively fat tails the horseshoe and horseshoe+ estimates were shrunk more minimally than with the thin tailed Gaussian. However, it’s clear that the non-Gaussian priors struggle with the cross-loading with population value of .2, with the median estimate across samples lying very close to zero. This results from the relatively small sample sizes (n=200) and the strong shrinkage behaviour of these priors near zero. Due to its weaker shrinkage behaviour the Gaussian prior tended to provide a less biased estimate of the true population value. This weakness of the non-Gaussian priors could be overcome with larger sample sizes.

In order to study the effects of prior choice on model fit a multilevel model was used, at this stage models were also fit to the 100 simulated datasets which had cross-loading priors fixed at 0, producing strict independent clusters structure. This was to act as a baseline for model fit. All 400 estimated LOO values were regressed firstly on to an indicator variable for which dataset each model was fit to, thus accounting for variance between simulated datasets (these parameters were modelled as random effects). Against this baseline model was compared a model that, in addition to using a dataset indicator variable, added a fixed indicator of which prior distribution was set on the cross-loading parameters. A likelihood ratio test of the two models showed that the choice of priors had a significant impact on model fit, χ2 (3) = 268.19, p < .001.

Figure 8 shows the coefficients for each prior distribution from the above model, with the independent clusters model acting as a baseline. Consistent with the results of Muthén & Asparouhov (2012) the use of Gaussian priors on cross-loadings provides notably better fit than independent clusters structure when there are cross-loadings not equal to zero. However, both the horseshoe and horseshoe+ priors can be seen to provide just as notable an increase in model fit over the Gaussian prior models as they do over the independent clusters model. This is consistent with the hypothesis that larger bias for non-zero values, and a tendency to overfit based on sampling variation, would harm the LOO of models using Gaussian priors relative to the alternatives.

FIGURE 8 GOES HERE

The final aspect on which models were compared was their complexity, as measured by their LOO effective number of parameters (see Vehtari, Gelman, & Gabry, 2016, for full details on the calculation of this value). As with LOO fit a multilevel model was employed which regressed the effective number of paramaters on to a dataset indicator to control for variation between simulated datasets. This model was then compared against one in which indicators for cross-loading priors were included (no intercept term was included in either model to allow direct comparison between prior distributions) , a likelihood ratio test showed that inclusion of prior distribution indicators significantly improved model fit, χ2 (4) = 863.1, p < .001.

Figure 9 shows the coefficients for each prior distribution when modelling the effective number of parameters (with no intercept term). The independent clusters model is clearly the simplest model with the free estimation of all cross-loadings significantly increasing the number of effective parameters. As expected the use of Gaussian priors results in a more complex fitted model than either the horseshoe or horseshoe+ models. This is a result of weaker shrinkage towards zero and, to some extent, overfitting of estimates that had population values of zero. The effective number of parameters shown here are linked directly to the LOO fit values shown in figure 8, so the difference between the Gaussian and alternative priors not only highlights a difference in fitted model complexity, but also contributes to the difference in LOO fit above.

FIGURE 9 GOES HERE

**Independent Clusters Loading Comparison**

The above process was repeated on the datasets simulated under the assumption that all population cross-loading values were equal to zero. Figures 10 – 12 again display the median and 95% coverage interval of cross-loading estimates over the simulated datasets.

FIGURES 10 – 12 GO HERE

Both methods provide reliable, un-biased median estimates close to zero over all 100 datasets. However it’s clear immediately that this estimate is only a long run property when using Gaussian priors, with individual estimates varying widely between -.2 and .2, following random variation from simulated dataset to dataset. By comparison the strong shrinkage of the horseshoe and horseshoe+ priors result in extremely consistent estimates, returning values indistinguishable from zero in almost all cases.

The analysis of model fit was carried out identically to the above analysis using the SCL datasets. A multilevel model with random dataset indicator variables was estimated, along with a further model which added indicators for cross-loading priors used (including models in which cross-loadings were fixed at zero). A likelihood ratio test of these two models showed that choice of prior distribution had a significant effect on LOO model fit, χ2 (3) = 517.55, p < .001. The fitted coefficients for prior distribution (with priors fixed at zero acting as a baseline) are plotted in figure 13.

FIGURE 13 GOES HERE

Compared to a model in which all cross-loadings are a priori fixed at zero, the use of Gaussian priors results in notably worse model fit in cases where all cross-loadings have population values equal to zero. Interestingly the use of the horseshoe and horseshoe+ priors does not result in a notably worse out of sample model fit than restricting these values at zero. Whilst it might be expected that running a larger number of simulations may allow for a statistically significant increase in model fit to be detected, this finding does suggest that any such effect will be minor next to the worsening of model fit associated with the use of Gaussian priors.

The effect of prior distribution choice, with an independent clusters population structure, on the effective number of model parameters was again approached in the same manner as the above analysis of this question on the SCL datasets. Again, a likelihood ratio test comparing the base model against one which included indicators for cross-loading prior distribution showed a significant improvement in explanatory power with the variables inclusion, χ2 (4) = 1109.7, p < .001. Figure 14 plots the estimated coefficients for the four approaches to cross-loading estimation (again no intercept term was modelled to allow for direct interpretation of effective parameter numbers).

FIGURE 14 GOES HERE

Fixing cross-loadings at zero produced a simpler model than any of the freely estimated alternatives. However, both the horseshoe and horseshoe+ models were had a significantly smaller number of parameters than the Gaussian models. The non-Gaussian priors estimate a model a much simpler model than the Gaussian approach, which assigns greater probability density at values further from zero.

**Discussion of simulation study results**

The effects of cross-loading prior distributions on model fit shown above are considered the most important point of comparison. As expected, in the presence of non-zero cross-loadings there was a significant improvement in fit when these parameters were freely estimated rather than fixed at zero. However, in these cases the use of Gaussian priors significantly underperformed both of the alternative regularisation priors. A combination of greater bias towards zero for large estimates, and overfitting of small estimates near zero, drove this difference in out of sample fit.

Not only did the horseshoe and horseshoe+ priors outperform the Gaussian when there were a small number of non-zero cross-loadings, but they also outperformed it when all cross-loadings had population values of zero. Most notably, when the independent clusters structure was true in the population, no significant difference in model fit was identified between fixing cross-loadings at zero and freely estimating them using the horseshoe or horseshoe+. The use of these regularisation priors at least matches the model fit of independent clusters models when the assumption holds true, and significantly outperforms when it is false.

Both the Gaussian and alternative distributions showed that they were able to fairly reliably identify moderate to large cross-loadings (.4 and .6 respectively). Due to the regularisation being carried out in all cases, through the use of priors with high density around zero, these estimates all showed bias towards zero. However, due to their fatter tails the horseshoe and horseshoe+ priors resulted in smaller overall bias than the use of Gaussians. The one area where the Gaussian priors were seen to outperform the alternatives was in the identification of small (.2) cross loading values. With the sample size used in each simulated dataset (n=200) the large majority of estimates produced using the non-Gaussian priors were shrunk to zero, whilst the use of Gaussian priors did not show this same strong shrinkage.

All three studied priors provided unbiased estimates of near zero cross-loadings over the 100 simulations. However, the Gaussian showed a high degree of variation from sample to sample whereas the horseshoe and horseshoe+ priors reliably returned estimates of exactly zero across almost all samples. When the true parameter structure was simple, with no major cross-loadings, the Gaussian prior resulted in noisy estimates driven by sampling error. By comparison the alternatives reliably identified the correct, simple structure from sample to sample.

These results indicate that the horseshoe and horseshoe+ outperform the Gaussian distribution as a prior on cross-loading parameters, in both individual samples and aggregated over many simulations. On the basis of the above, Stromeyer, et al’s (2015) concern that BSEM would lead to overfitting and consistently complex parameter solutions are considered to be valid when small variance Gaussian priors are used. However, the above simulations show that the strong regularisation of the horseshoe and horseshoe+ distributions avoid these pitfalls. Likewise, concerns over falsifiability are mitigated as the use of the non-Gaussian priors retains a simple structure where appropriate, and avoids the overfitting from sample to sample observed using Gaussian priors.

However, the above simulations can’t speak to one aspect of Stromeyer, et al’s critique, that BSEM models, with cross-loadings freely estimated, will often result in difficult to interpret results due to the large number of loading parameters that must then be considered when interpreting the overall model. Based on the above results it was expected that the use of strong Bayesian regularisation priors would also help minimise this concern by returning relatively simple patterns of cross-loading estimates. This is explored below in a comparison of BSEM prior distributions using a real dataset.

**Comparison of cross-loading priors using the Holzinger and Swineford dataset**

In order to compare the interpretability of BSEM models in a single sample situation the widely studied Holzinger and Swineford (1939) mental abilities dataset was modelled. The models fitted had a loading pattern mirroring that reported in Muthén & Asparouhov (2012), with 19 observed variables regressed on to 4 latent variables. Table 1 lays out this loading structure, observed and latent variable names are those used by Muthén & Asparouhov.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Observed Variable | Spatial | Verbal | Speed | Memory |
| Visual perception | X | 0 | 0 | 0 |
| Cubes | X | 0 | 0 | 0 |
| Paper form board | X | 0 | 0 | 0 |
| Flags | X | 0 | 0 | 0 |
| General information | 0 | X | 0 | 0 |
| Paragraph comprehension | 0 | X | 0 | 0 |
| Sentence completion | 0 | X | 0 | 0 |
| Word classification | 0 | X | 0 | 0 |
| Word meaning | 0 | X | 0 | 0 |
| Addition | 0 | 0 | X | 0 |
| Code | 0 | 0 | X | 0 |
| Counting groups of dots | 0 | 0 | X | 0 |
| Straight and curved capitals | 0 | 0 | X | 0 |
| Word recognition | 0 | 0 | 0 | X |
| Number recognition | 0 | 0 | 0 | X |
| Figure recognition | 0 | 0 | 0 | X |
| Object-number | 0 | 0 | 0 | X |
| Number-figure | 0 | 0 | 0 | X |
| Figure-word | 0 | 0 | 0 | X |

Apart from the changes in observed and latent variable numbers, and the associated new loading structure, the BSEM models fit are exactly as described in the above simulation studies, with uncorrelated observed variable errors, Gaussian main loading priors with mean of 0 and standard deviation of 3, and only cross-loading priors differing between models. In both this study, and the above simulations, the horseshoe and horseshoe+ distributions showed no substantive differences in estimation. For this reason only the models using Gaussian and horseshoe priors are reported in the below comparison. Full Stan code for all models can be found at (GITHUB GOES HERE). In this comparison only the data from the Grant-White school (N=145) was used.

Table 2 lists for each model all factor loadings greater than .01 in absolute value, factor correlations, and their LOO information criterion statistics. As in the simulations, the use of horseshoe priors results in a marginally better fitting model than the use of Gaussian priors. The differences in cross-loading estimate sparsity are also clear, with Gaussian priors only very rarely resulting in absolute estimates less than .01, whilst absolute values greater than .01 are the exception using horseshoe priors. The interpretation of the latent factors in the horseshoe model is relatively straightforward without having to employ further decision criterion (e.g. ignoring values whose 95% credible interval crosses zero).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Gaussian priors | | | |  | Horseshoe priors | | | |
| Observed Variable | Spatial | Verbal | Speed | Memory |  | Spatial | Verbal | Speed | Memory |
|  | Loadings  (posterior median) | | | | | | | | |
| Visual perception | **.67\*** | .02 | .06 | .07 |  | **.71\*** |  |  |  |
| Cubes | **.54\*** |  | -.02 |  |  | **.50\*** |  |  |  |
| Paper form board | **.44\*** | .07 | .06 | .08 |  | **.52\*** |  |  |  |
| Flags | **.66\*** | .10 | -.03 | .02 |  | **.67\*** |  |  |  |
| General information | .03 | **.86\*** | .07 | -.10 |  |  | **.81\*** |  |  |
| Paragraph comprehension | -.02 | **.92\*** | -.12 | .04 |  |  | **.85\*** |  |  |
| Sentence completion | -.07 | **.98\*** | .02 | -.12 |  |  | **.88\*** |  |  |
| Word classification | .06 | **.64\*** | .14 |  |  |  | **.65\*** | .05 |  |
| Word meaning | -.03 | **.97\*** | -.14 | .01 |  |  | **.89\*** | -.02 |  |
| Addition | -.30\* | .04 | **.91\*** |  |  | -.23 |  | **.86\*** |  |
| Code | -.03 | .04 | **.54\*** | .21 |  |  |  | **.56\*** | .14 |
| Counting groups of dots | .02 | -.16 | **.94\*** | -.09 |  |  | -.03 | **.80\*** |  |
| Straight and curved capitals | .30\* | .03 | **.65\*** | -.10 |  | .34 |  | **.57\*** |  |
| Word recognition | -.09 | .06 | -.08 | **.63\*** |  |  |  |  | **.54\*** |
| Number recognition |  | -.03 | -.08 | **.62\*** |  |  |  |  | **.54\*** |
| Figure recognition | .23 | -.08 | -.12 | **.63\*** |  | .06 |  |  | **.56\*** |
| Object-number | -.25 |  | .05 | **.81\*** |  | -.29 |  |  | **.86\*** |
| Number-figure | .15 | -.15 | .16 | **.56\*** |  |  | -.01 | .02 | **.62\*** |
| Figure-word | .02 | .06 |  | **.45\*** |  |  |  |  | **.47\*** |
|  |  | Latent Variable Correlations | | | | | | | |
| Spatial | - |  |  |  |  | - |  |  |  |
| Verbal | .54\* | - |  |  |  | .54\* | - |  |  |
| Speed | .50\* | .52\* | - |  |  | .47\* | .43\* | - |  |
| Memory | .55\* | .60\* | .63\* | - |  | .60\* | .52\* | .54\* | - |
| LOO information criterion | 6,560.9 | | | |  | 6,548.5 | | | |

*Note*. Values in bold indicate hypothesized major loadings. Statistically significant cross-loadings (marked with asterisks) have a 95% credibility interval that does not cover zero. Absolute values less than .01 not shown.

Whilst main loading estimates differ between models, a paired t-test indicated no significant differences, t(18) = 2.03, p = .06. Factor correlations are slightly smaller in general for the horseshoe model than the Gaussian. The one final point of difference to note is that none of the large cross-loadings in the horseshoe prior model were significantly different from zero, whilst 2 were when Gaussian priors were used. The horseshoe prior seems to produce a more conservative measure of significance than the Gaussian.

**Discussion**

*The use of strong regularisation priors for cross-loadings*

The concerns of Stromeyer, et al. (2015), that the use of Gaussian priors to estimate cross-loadings could lead to overfitting, a lack of interpretability, and potentially wekeaned falsifiability of measurement theories, are considered to be valid in light of the above simulations. However, the above findings are considered to demonstrate that these risks are largely avoid through the use of regularisation priors such as the horseshoe. They outperform the use of Gaussian priors in both situations studied above. Even more significantly, they provide practically equivalent results to IC models when the IC assumptions are true, and outperform these models when they are not. Both the horseshoe and horseshoe+ priors seem to provide a well performing solution to the systematic estimation of cross-loadings, whilst avoiding the risks Stromeyer, et al. raise for the use of small variance Gaussian priors.

The two models fitted to the Holzinger and Swineford (1939) dataset illustrate the clear benefits of using prior values that enforce a sparse structure in producing models which are easily interpretable. Most cross-loading estimates from the model using horseshoe priors were negligible (< .01) whereas the use of Gaussian priors returned a solution which required close analysis to interpret, particularly in deciding which cross-loadings should be considered relevant to the latent variables.

The simulation results above show that the use of horseshoe or horseshoe+ priors on cross-loadings produces models that consistently outperform ones in which Gaussian priors are used. Not only do these models display better out of sample fit, but have a lower number of effective parameters, provide more consistent estimates near zero when the population parameter value is zero, and show smaller shrinkage for substantive, non-zero parameters than when Gaussian priors are used. The difference in out of sample fit is likely driven in part by all three of the above factors. This was the case regardless of whether any non-zero cross loadings were present or not.

The one area where the Gaussian priors were seen to outperform the alternatives was in the identification of small cross-loading values (.2 in this case). With the sample size used in each simulated dataset (n=200) the large majority of estimates produced using the non-Gaussian priors were shrunk to zero. The use of Gaussian priors did not show this same strong shrinkage. In this case, the shrinkage behaviour of the horseshoe models can be thought of similarly to statistical power, where the failure to consistently identify the true value is a function of the sample size, and the magnitude and variance of the true effect. With a larger sample size the horseshoe and horseshoe+ estimates of small values would likely avoid being shrunk near zero, and would more closely resemble the estimates for moderate or large cross-loadings seen in the simulations.

Also of note is how models using different cross-loading priors compared to the traditional independent clusters model, in which all cross loadings are fixed at zero. Muthén and Asparouhov’s (2012) finding that the use of Gaussian priors on cross-loadings improved model fit over the IC model when there were substantive cross-loadins was replicated. However, in cases where there were no non-zero cross-loadings the use of Gaussian priors was shown to harm model fit when compared to IC models. This is a result of the overfitting that Stromeyer, et al. (2015) raised as a potential concern, with the Gaussian model essentially chasing noise around zero, with the improvement in model ft not outweighing the increase in model complexity. This also highlights the importance of using measures of model fit which properly account for model comlexity when highly flexible models are being studied.

The horseshoe and horseshoe+ models outperformed the IC model when there were non-zero cross-loadings present. However, no significant difference in model fit was identified when the assumptios of the IC model were correct and all cross-loadings equalled zero in the population. This is due to the strong shrinkage behaviour of the horseshoe and horseshoe+ priors. When cross-loading estimates were near zero these models reduced to a structure nearly indestiguashable from that of the IC model.

That no significant difference in model fit was seen is considered an artifact of the study design, rather than reflecting a true difference of zero between fitted models. The horseshoe and horseshoe+ priors *will* assign positive density away from zero in cases with realistic sample sizes and so it can be expected that under a much larger simulation study significant differences in fit could be identified for n = 200. However, the fact that any such difference is statistically undetected with 100 simulations ran provide a good indicator that such differences in model fit are unlikely to be of practical significance or concern.

*Specification of Bayesian priors*

It is possible that the informative Bayesian aproach proposed, where informative priors are leveraged in order to promote desirable model behaviour, may be seen to be providing the analyst too much influence over the estimated model. The above Holzinger and Swineford analysis highlights two routes by which the priors selected by analysts can influence results; either different prior distributions can be used (as in the Gaussian and Horseshoe example above) or different variance parameters can be set on the same family of prior distribution. This latter case can be seen when comparing the Holzinger and Swineford Gaussian priors model above with the Gaussian priors model reported by Muthén & Asparouhov (2012) on the same dataset. Due to minor differences in prior variances, the above model used Gaussians with variance of .04 whereas Muthén & Asparouhov report results with variance of .01, parameter estimates differ noticably[[2]](#footnote-2).

In the first case, the flexibility afforded by the family of distribution chosen for the prior is not considered an undue introduction of influence, as similar control is readily available and commonly used in frequentist modelling. Whilst different prior families can provide control over certain aspects of Bayesian model behaviour, differing estimation procuedres provide similar control in the frequentist context. Maximum likelihood (Jöreskog, 1967), partial least squares (Chin, 1998), weighted least squares (Muthén & Satorra, 1995), and recent regularisation techniques (Jacobucci, Grimm, & McArdle, 2016) are only a small sample of the published frequentist methods which can be used to produce different SEM estimates under different conditions. How analysts should best approach and use this flexibility of approaches is a proper area for discussion, however it is not unique to the use of informative Bayesian priors.

By comparison, the proper selection of the prior’s variance parameter is largely unique to the Bayesian approach to SEM. Indeed, Muthén & Asparouhov (2012) devote a significant section of their paper to describing procedures for the selection of optimal cross-loading prior variances. However, it is not neccesary to directly specify the variance of any given prior dsistribution in a Bayesian analysis. Rather a hierarchical distribution, in which one or more distribution parameters are themselves modelled as random variables, allows for the prior variance itself to be estimated from the observed data (Gelman et al., 2013). This approach is used for both the horseshoe and horseshoe+ distributions, rather than setting a fixed variance on the prior, it is modelled as the product of a set of cauchy+ distributions. When the model is fit in this way the full posterior distribution of the prior variance will be estimated, allowing the data to guide variance selection. This approach still requires users to specify fixed hyperparameters (parameters on the distributions used to define the variance term) but so long as these are set with an appropriate distribution the variance parameters are unlikely to have any undue influence on the analysis.

*Bayesian and Frequentist approaches to regularisation*

As discused at the beginning of this paper, the use of informative, small variance priors in order to overcome a modelling difficulty (non-identifiability in this case) is the Bayesian approach to regularised modelling. This raises the possibility that frequentist regularisation could be used to achieve similar results. A recent treatment of frequentist regularisation can be found in Jacobucci, Grimm and McArdle (2016), who also provide a means of implementing these models through the regsem R package. Given the poor convergence when using Laplace priors above, it is unclear how effective the frequentist LASSO method would be in estimating cross loading values. However, ridge regression regularisation applied through the regsem package could be expected to provide similar performance to the use of small variance Gaussian priors, due to their equivalence (Pasanen, Holmström, & Sillanpää, 2015).

Whilst the differences between Bayesian and frequentist uses of regularisation to estimate sparse cross-loading sets may turn out to be minor, the Bayesian approach taken above is considered to have two advantages. Firstly, as in all fuly Bayesian modelling, complete density distributions are estimated for all parameters, allowing for a natural treatment of uncertainty. By comparison, frequentist methods provide only a point estimate, and further assumptions (typically regarding asymptotic behaviour) are required to produce estimates of uncertainty such as confidence intervals. Secondly, where direct comparisons are available, recently developed Bayesian regularisation priors have been shown to outperform priors corresponding to commonly the commonly used LASSO method in frequentist regularisation (see for example Peltola, et al., 2014).

It is also possible that certain developments in regularised SEM may be easier to approach using a Bayesian MCMC framework. As probablistic programming langauges such as stan (Carpenter, et al., in press) and PyMC (Salvatier, Wiecki, & Fonnesbeck, 2016) continue to develop provide greater general capabilities for model estimation. Stan in particular was developed explicitly to overcome difficulties in estimating complex multi-level models (CITE). They can therefore enable development of novel statistical models without also having to develop new estimation procedures in parallel. Regularised SEM with a multi-level parameter structure (such as those used for multiple measurement sampling designs; CITE) is one such development that may likely be more easily implemented in a Bayesian than frequentist framework. Such structures are easily estimated using Bayesian estimation, whereas frequentist developments would likely need to consider the estimation techniques employed (EXAMPLE HERE, LME4?).

*Future Work*

This paper provides solid evidence that the use of horseshoe or horseshoe+ priors on cross-loading parameters is to be prefered to small variance Gaussians as proposed by Muthén and Asparouhov (2012). They also are considered to largely mitigate the concerns raised by Stromeyer, et al. (2015) that free estiation of cross-loadings can harm model interpretability, falsifiability and lead to overfitting.

1. The horseshoe and horseshoe+ distributions lack closed form representations but are tightly bounded by definite upper and lower bounds. The densities plotted here represent the upper bound on the distributions as derived in Carvalho, Polson and Scott (2010) and Bhadra, Datta, Polson and Willard (2015). [↑](#footnote-ref-1)
2. When the same variance parameter was used results equivalent to Muthén & Asparouhov (2012) were obtained, identifying the cross-loading priors as the source of variation, rather than software used or some other aspect of model specification. [↑](#footnote-ref-2)