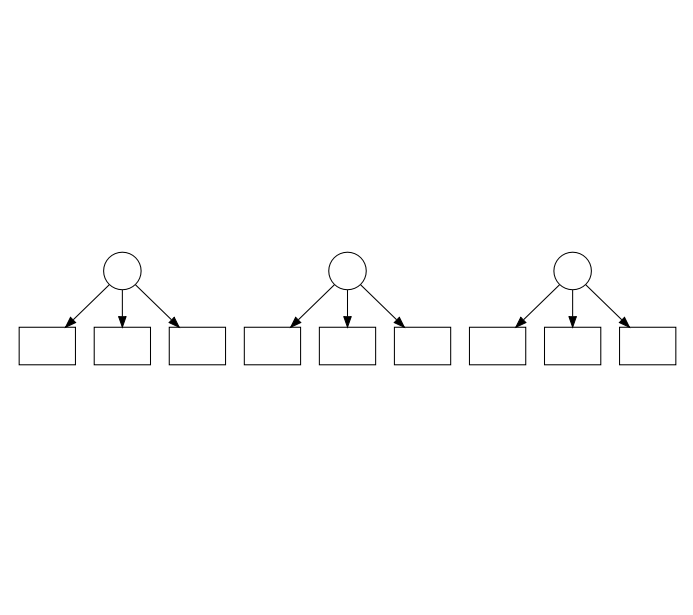
In Bayesian structural equation modelling (BSEM) placing small variance Gaussian priors on cross-loading parameters can be used to obtain full estimation of these parameter’s posterior probabilities (Muthen and Asparouhov, 201232). However, these posteriors are sensitive to sampling error and can result in models with many small, non-zero cross loadings which complicate theory driven modelling (Stromeyer, et al, 23211). They can also result in notable shrinkage of non-negligible cross-loadings towards zero. The literature on Bayesian regularisation reports non-Gaussian priors which are intended for sparse modelling situations. This paper compares the performance of Gaussian priors with the Laplace, horseshoe, and horseshoe plus priors from said literature. Simulations show that all three alternatives provide better out of sample model fit, smaller shrinkage of major cross-loadings, and more easily interpretable loading patterns than Gaussian priors. Non-Gaussian shrinkage priors are recommended for future modelling of cross-loading parameters in BSEM.

**Introduction**

In cases of structural equation modelling (SEM) where observed variables are regressed on to more than a single latent variable, care has to be taken in specifying which regression coefficients (or loadings) are directly estimated. The estimated coefficients must be chosen not only to reflect substantive theory, but must also collectively meet the requirements of model identifiability. These joint concerns often motivate an independent cluster structure, in which every observed variable is regressed on to only a single latent variable, with all other potential loadings fixed at zero. This is sometimes referred to as a *simple structure* pattern of loadings, after Thurstone’s (1947) criteria for simple structure in factor analytic models. However as Browne (2001) discusses such a strict pattern of latent variable loadings is only the most restrictive form of Thurstone’s simple structure.

Figure 1. provides an example diagram and loading matrix for an independent clusters structure with three latent and nine observed variables. The majority of factor loading parameters are fixed at zero, with only a relatively small number freely estimated.

FIGURE 1 GOES HERE



Whilst this form of measurement model is highly restrictive, unable to capture more complex patterns of latent-observed variable relationships, there are several benefits to its sparse form. From a theoretical point of view, it posits neat and immediately interpretable relationships between latent and observed variables. If every observed variable loads on only a single latent variable it helps researchers to clearly state their theories through the model. More complex patterns of latent-observed loadings, such as those produced by exploratory factor analysis, may allow researchers to claim evidence for poorly stated or unsupported theories (Stromeyer, et al., 2015). It also enforces the greatest degree of model simplicity in the loading parameters; all else equal strict independent clusters are to be preferred to more complex loading patterns on the principal of parsimony.

Even in cases where reality is more complex than an independent clusters model implies such structure is useful as the most straightforward solution to the issue of model identifiability in a frequentist modelling context. In order for a model to be identified for estimation purposes a number of parameters need to be restricted to fixed values. In most cases it only makes sense to restrict parameters to zero. Theory may well suppose that certain variables have no direct relation to one another; it’s much rarer for an exact non-zero relationship to be stated. Unfortunately there’s currently no widely agreed empirical approach for deciding which loading parameters should be restricted to zero based on observed data. In the absence of such an approach, theory is often stated in the form of an independent clusters model which, due to the number of loadings restricted to zero, can neatly handle the issue of identifiability.

Muthén & Asparouhov (2012) advocate Bayesian Structural Equation Modelling (BSEM) as a method for simultaneously estimating all possible latent-observed variable loadings. Thus relaxing the restrictions of independent cluster models and allowing for an observed variable to load on any latent variable where data suggests an association. They achieve this by setting informative Bayesian priors, specifically small variance Gaussian distributions with mean zero, on all cross-loading parameters. That is, all loading parameters not specifically hypothesised to be non-zero (those which would be fixed to zero in an independent clusters model).

Using these priors, alongside broad variance Gaussian priors on primary loadings, allows researchers to systematically test their hypotheses that cross-loading parameters are equal to zero. The tight prior distributions achieve model identifiability, but allow for posterior estimates to move away from zero if suggested by the data. BSEM as proposed by Muthén & Asparouhov thus overcomes the restrictive form required in most likelihood-based approaches whilst meeting the requirements of identifiability.

However BSEM, as proposed by Muthén & Asparouhov does this at the expense of the clear relationship between model and theory that independent cluster 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 the desired aim 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 seem to free SEM from a restrictive approach to cross-loading estimation only at the expense of losing the direct and clear 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.

This paper investigates the potential of using informative priors common in the Bayesian regularisation literature, instead of Gaussian priors, to estimate SEM cross-loadings. A brief overview of the principles of regularisation from a Bayesian perspective is provided, along with a comparison of the problem of identifying non-zero cross loadings with the problems that typically motivate regularisation. Simulations comparing performance of small variance Gaussian priors against several priors from the regularisation literature are then reported and discussed. Finally an example of applying such a regularised BSEM on a real data set is provided.

**Bayesian Regularisation**

Regularisation captures a range of approaches to model estimation which broadly seek to simplify the estimated parameter solution, subject to some penalty on model complexity. 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 (Hastie, Tibshirani, & Friedman, 2001). Regularisation models have attained wide spread usage 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. Regularisation penalties can also provide a means of estimating model parameters which may have been unidentified in equivalent non-regularised models, such as when the number of predictor variables is greater than the number of observations (Zou & Hastie, 2005).

Regularisation was originally developed within a frequentist statistical framework, where parameters are shrunk by adding some penalty on model complexity to the relevant loss/likelihood function. For example, an ordinary least squares linear regression model is fit by minimising the sum of the squared model residuals.

The LASSO, a common regularisation method, minimises this same term, but subject to the condition that the sum of the absolute coefficients (excluding the intercept term) is less than some constant value (Tibshirani, 1996).

In practice the value of is typically selected using cross-validation to maximise the out of sample model fit (Tibshirani, 1996). The LASSO is particularly illustrative in this context as a regularisation method with a direct Bayesian equivalent; however different complexity penalties will result in regularisation techniques with different behaviours. Ridge regression (Hoerl & Kennard, 1970) and elastic net (Zou & Hastie, 2005) are two examples of commonly used regularisation techniques which employ different complexity penalties to the LASSO.

By contrast, regularisation in the Bayesian context is achieved by setting tight prior distributions on the relevant model parameters (typically regression coefficients) that place a very high prior probability on the central value of the distribution, which rapidly decreases out to relatively flat tails. The steep slope near the centre results in shrinkage of near central posterior estimates towards the prior’s central value, usually set equal to zero. This results in a similar effect to the complexity penalties of frequentist regularisation. Some values are shrunk towards a pre-specified value, reducing model complexity and allowing for the fitting of otherwise unidentified models. Muthén & Asparouhov’s (2012) use of small variance Gaussian priors to estimate cross-loadings in BSEM can therefore be seen as a case of Bayesian regularisation. The use of a small-variance Gaussian prior on regression coefficients can specifically be interpreted as a bayesian equivalent to ridge regression (Pasanen, Holmström, & Sillanpää, 2015).

However, Gaussian distributions are relatively uncommon regularisation priors in the literature on Bayesian regularisation. Rather, Gaussian scale-mixture distributions are more commonly used. This family of distributions is similar to the Gaussian in form, but allow for steeper slopes near the prior’s central value and flatter distributions in the tail (Polson & Scott, 2010). As the name implies these distributions can be expressed as Gaussian distributions where the variance parameter is itself a product of two or more prior distributions. This hierarchical structure on the variance parameter allows for non-standard distributions which display the features described above.

The Bayesian equivalent to frequentist Lasso regularisation is achieved by setting a special case of such a scale-mixture prior on the relevant model parameters, specifically the Laplace, or double-exponential, distribution (Park & Casella, 2008). The Laplace distribution is equivalent to a Gaussian distribution with variance equal to the product of a fixed Gaussian and Exponential distribution. **Figure 1** compares the Laplace and Gaussian distributions, centred on zero with a scale parameter of one, around zero and then in the tail of the distributions.

**Figure 1 goes here**

The Laplace distribution (and other commonly used regularisation priors) has a much sharper peak close to zero than the Gaussian. This results in much stronger shrinkage of parameters near zero, resulting in a sparser solution (where the posterior distributions are very tightly distributed around zero). It also has a flatter distribution in the tail than the Gaussian, this results in smaller shrinkage of parameters whose unbiased median estimate is far from zero. The Laplace prior effectively draws a distinction between zero and non-zero posterior estimates (as does the frequentist LASSO), whereas the Gaussian prior results in global shrinkage, but will often not shrink even small posterior estimates to zero. The Gaussian prior is a poor enforcer of posterior sparsity.

**Bayesian structural equation modelling as a regularisation problem**

Muthén and Asparouhov (2012) demonstrated how placing small variance priors on all cross-loadings in a structural equation model can produce identifiable models which provide systematic estimates for all cross-loadings. However, the use of Gaussian priors leaves the posterior cross-loading estimates open to several risks alluded to above. Specifically, parameters far from zero may be subject to significant shrinkage towards zero, an undesirable result for values we’re confident are non-zero in our sample. Shrinkage of small values with Gaussian priors will also stop short of total shrinkage to near zero. Whilst Gaussian priors do carry out shrinkage they don’t encourage true sparsity of the final solution.

This latter risk is discussed at length by Stromeyer, et al. (2015) in their comment on Muthén and Asparouhov’s method. They raise the issue that, by not enforcing any kind of sparsity on cross-loadings, the Gaussian prior method could result in models that were difficult to interpret and which may weaken the capability for analysts to clearly support or reject specific measurement hypotheses. With many non-zero cross loadings present it would prove a much harder task for researchers to directly relate models to tractable theory.

Both Muthén and Asparouhov and Stromeyer, et al. raise important issues in the treatment of cross-loadings in structural equation modelling. Muthén and Asparouhov are right to draw attention to the, likely often, over restrictive approach to cross-loadings found in the independent clusters approach. Likewise Stromeyer, et al. are correct that, by not enforcing the relative sparsity of parameter estimates, the Gaussian prior approach risks limiting the interpretability of estimated models and increasing the disparity between clearly stated theory and difficult to interpret, non-sparse models.

The problem of estimating cross-loadings can thus viewed as similar to the problems which have motivated regularisation in other areas of statistical modelling. Specifically, the need to identify a small number of non-zero signals against a large background of noise. This paper explores the performance of non-Gaussian regularisation priors in producing a sparse set of estimates for cross-loading values. These priors will be compared against the Gaussian prior approach suggested by Muthén and Asparouhov using simulated data.

The hypothesised outcomes of the below simulation studies were:

1. That the Gaussian prior models would display poorer out of sample model fit, measured using leave one out cross validation (Vehtari, Gelman, & Gabry, 2016), than those using the comparison regularisation priors
2. That the Gaussian prior models would have a higher average number of effective parameters (Vehtari, Gelman, & Gabry, 2016) than their regularised equivalents, representing less parsimonious solutions.
3. As a result of the above two hypotheses it was expected that Gaussian prior methods would show higher within sample fit than alternatives, representing overfitting of parameter values near zero.
4. That shrinkage towards zero of non-zero cross-loadings would be significantly greater for the Gaussian prior method than the alternatives. This would result in greater bias in estimated non-zero cross loading values produced using Gaussian priors.
5. That Non-Gaussian priors would result in significantly sparser and more interpretable solutions than the Gaussian prior method.

**Simulation Studies**

Simulation studies were used to evaluate the first four of the above hypotheses. The fifth was investigated by exploring individual fitted models, rather than performance over multiple solutions. Two different population models were compared, to investigate performance under different circumstances. These Two conditions were independent clusters, with no non-zero population cross-loadings, and sparse cross-loadings, with a small number of non-zero cross loadings to identify.

The independent clusters model provides a basis on which the others are built, simply by switching out zero valued cross-loadings for non-zero values. The basic population model regressed fifteen observed variables on to three latent variables, each latent variable underlying five observed. These primary loadings ranged from 0.4 to 0.8 in value.

The sparse model added three cross loadings to the basic model, one on each latent variable. These varied in size from small (0.2), through medium (0.4), to large (0.6). The factor loading matrices for each case are shown in figure 2.

**Figure 2 goes here**

The residual item variance in the population model was set to 1 for all measured variables. The latent variable covariance matrix was a three by three identity matrix. Using the R package Simsem (Pornprasertmanit, Miller, & Schoemann) 100 datasets, each with 200 observations across the 15 measured variables, were simulated from each of these population models.

The Bayesian structural equation models were fit using the probabilistic programming language Stan (Carpenter, et al., in press) managed from R using the RStan package. The models fitted were basic common factor models (Brown, 2006). Gaussian priors with a mean of zero and variance of three were set on the primary factor loadings (these were presumed known); these parameters were restricted to positive values to achieve model identification. Whilst their population values equal zero, latent variable correlations were estimated to allow for sample variance, measured items were modelled as independent without correlated residuals. All models fitted were identical except for the prior distributions set on 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. The number of MCMC iterations ran, and the thinning parameters used to minimise chain autocorrelation varied 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. Details are recorded in the simulation code which can be found at (github goes here).

**Choice of Cross-Loading Priors**

Five prior distributions for cross-loadings were compared; these are laid out in Table 1. The point mass prior on zero is equivalent to estimating an independent cluster model and is reported simply to provide a baseline comparison for model fit. Figure 3. shows the probability density distributions for each investigated prior except the point mass on zero.

**Table 1. goes here**

**Figure 3. goes here**

The baseline for performance is the small variance Gaussian centred on zero, proposed by Muthén and Asparouhov (2012), a standard deviation of 0.2 was set on these priors representing a prior belief that approximately 95% of cross loading values are less than 0.4 in absolute value.

The Laplace distribution corresponds to the commonly used LASSO method of sparse regularisation (ref). This was parametrised as a Gaussian distribution with mean = 0 and a scale parameter following a Rayleigh distribution with a uniform prior on its scale parameter[[1]](#footnote-1).

Two other distributions from the Bayesian regularisation literature, the horseshoe and horseshoe plus distributions were also explored. The horseshoe prior can be expressed as a normal distribution with mean = 0 and the variance is the product of two terms, a local and global term. The global term is a single value, equal for all parameters (here it’s set a prior distribution of half-Cauchy(0, 1)), whilst the local term is a vector with an entry for each estimate parameter. The horseshoe prior is achieved by also setting half-Cauchy(0, 1) distributions on each value for the local term. Previous research indicates that the horseshoe may well outperform the Laplace distribution in sparse regularisation cases (refernce), resulting in small parameters being shrunk closer to 0, and minimal comparative shrinkage of larger parameters.

The horseshoe plus distribution is a direct extension of the horseshoe distribution. In the horseshoe plus, the local scale terms are not distributed according to a Cauchy (0, 1) distribution, but by a Cauchy(0, tau) distribution where tau is a vector of local scale values distributed according to a Cauchy(0, 1,) distribution. This additional Cauchy term results in a distribution which has been found to marginally outperform the original horseshoe distribution in identifying very sparse signals (reference).

1. There are a number of methods to parameterise the Laplace distribution, this form proved adequate for the purposes of this paper but more efficient sampling may be achieved with other parametrisations. [↑](#footnote-ref-1)