Published in final edited form as:

Struct Equ Modeling. 2018 ; 25(4): 639–649. doi:10.1080/10705511.2017.1410822.

Comparison of frequentist and Bayesian regularization in structural equation modeling

Ross Jacobucci¹ and Kevin J. Grimm²

¹University of Notre Dame

²Arizona State University

Abstract

Research in regularization, as applied to structural equation modeling (SEM), remains in its infancy. Specifically, very little work has compared regularization approaches across both frequentist and Bayesian estimation. The purpose of this study was to address just that, demonstrating both similarity and distinction across estimation frameworks, while specifically highlighting more recent developments in Bayesian regularization. This is accomplished through the use of two empirical examples that demonstrate both ridge and lasso approaches across both frequentist and Bayesian estimation, along with detail regarding software implementation. We conclude with a discussion of future research, advocating for increased evaluation and synthesis across both Bayesian and frequentist frameworks.

Keywords

regularization; structural equation modeling; Bayesian; lasso; ridge; factor analysis; shrinkage

Introduction

In structural equation modeling (SEM), the amount of estimation flexibility has significantly increased in recent years. Broadly speaking, this has occurred in two overlapping domains. The first is in Bayesian structural equation modeling (BSEM; Lee, 2007), particularly given the implementation of small variance priors in Mplus (Muthén and Asparouhov, 2011). The second area is regularization, where most work has been focused in factor analysis (Choi et al., 2010; Hirose and Yamamoto, 2014; Jung and Takane, 2008; Ning and Georgiou, 2011; Lu et al., 2016), with more recent work in general SEM (Jacobucci et al., 2016; Huang et al., 2017; Serang et al., 2017; Feng et al., 2017). The focus of this paper is to provide a detailed overview and demonstration of the overlap between these two areas, highlighting how each method can be utilized. Furthermore, there are a dizzying array of contemporary regularization methods available to users. Many of these methods can only be used in specific software environments, further increasing the difficulty involved in testing and comparing methods. Therefore, scripts used to run all models in this paper are provided on both author's websites (footnote added at later time).

As a way to compare and contrast these methods, we provide an empirical example where a confirmatory factor model is fit to six items from the Holzinger and Swineford (1939) dataset. Regularization methods are particularly adapted for high-dimensional settings (large numbers of variables in comparison to sample size), however, we use this simplistic example to highlight how each method is unique and can be implemented. We first start with the use of small variance priors in Mplus, and how this is equivalent to the use of ridge regularization. In the second step, we demonstrate difference across both frequentist and Bayesian estimation in the case of the lasso. Finally, we extend this demonstration to more complex methods, specifically a number of more recent hierarchical Bayesian methods: the Bayesian adaptive lasso (Feng et al., 2017) and the horseshoe prior (and extension; Carvalho et al., 2010; Bhadra et al., 2017).

Bayesian SEM

Although SEMs have traditionally been estimated in the frequentist framework with maximum likelihood estimation (MLE; Lawley, 1940; Jöreskog, 1969), Bayesian estimation is becoming increasingly popular (van de Schoot et al., 2017). In this, general formulations have been proposed (Lee, 2007; Scheines et al., 1999), along with more specific applications in growth curve models (Liu et al., 2016; Zhang, 2016; Zhang et al., 2007), confirmatory and exploratory factor models (Lu et al., 2016, 2017; Moore et al., 2015; Muthén and Asparouhov, 2013), and measurement invariance models (Asparouhov and Muthén, 2014; Muthén and Asparouhov, 2013), to name just a few. Comparing both Bayesian and frequentist estimation has occurred (e.g., Lee and Song, 2004), particularly through the use of informative priors in the context of small sample research. This is due to the fact that in Bayesian estimation, as the sample size increases, the likelihood distribution becomes more and more influential in the estimation of the posterior distribution, to the point that the priors have little to no influence.

With the increasing availability of easy to use software, researchers are likely to fit certain models with MLE and other models with Bayesian estimation. In some settings, BSEM has a number of advantages in comparison to using a frequentist estimation technique, such as MLE. One advantage is that instead of point estimates for parameters, marginal distributions are generated for each parameter. This better allows models to account for sampling variability. In addition, BSEM does not rely on large-sample theory, which leads to the potential for less bias with small sample sizes (Muthén and Asparouhov, 2011; McNeish, 2016). Finally, with more complex models, estimation may be difficult with MLE. This has become increasingly evident with item factor models, where models with more than five latent variables are intractable with MLE (Wirth and Edwards, 2007).

We do not present a technical overview of Bayesian estimation in SEM (see Lee, 2007; Muthén and Asparouhov, 2011) and instead focus on the specification of prior distributions for model parameters and how these specifications influence posterior distributions. In BSEM, researchers can provide a diffuse prior distribution or an informative prior distribution for a given parameter. A diffuse prior leads to the posterior distribution of the parameter being determined almost entirely from the data, where an informative prior leads to the posterior distribution of the parameter being a combination of the prior information

and the data. The degree of informativeness and sample size determines the relative weighting of the two.

In BSEM, when a diffuse prior (large variance) is used, the resulting mean estimates of the posterior distribution are often close to estimates obtained from MLE (e.g., Muthén and Asparouhov, 2011; Hamagami et al., 2009). The use of diffuse priors, along with large sample sizes, allows the likelihood to contribute a large portion of the information to the formation of the posterior distribution. In contrast, using priors that limit the influence of the likelihood on the marginal posterior distribution, leads us to both the formulation of small variance priors, and more generally, Bayesian regularization.

Specific to our purpose is the recent implementation of Mplus' (Muthén and Muthén, 2012) BSEM (Muthén and Asparouhov, 2011) estimation, which focuses on the application of small-variance priors. Similar in a sense to using previous research to inform the choice of the prior distribution, the use of small variance priors that are centered around zero restricts the influence of non-target (e.g. cross-loadings) parameters. This may allow for less biased estimation in situations where small parameter estimates, not of empirical interest, are inappropriately constrained to zero. For example, it is common in confirmatory factor analysis (CFA) for researchers to specify few, if any, cross-loadings and opt for simple structure (e.g., Thurstone, 1935). True simple structure in CFA is rare and forcing cross-loadings to zero can bias estimated factor loadings and factor correlations. Instead of imposing a zero parameter value for a cross-loading, a small variance prior can be used to restrict the range of the estimate, without forcing the estimate to zero.

As an example, we present results from fitting a one factor CFA model to six items from the Holzinger and Swineford (1939) dataset. Of these six indicators, only three of the loadings were strong (Paragraph Comprehension, Sentence Completion, and Word Meaning) and the other three indicators (Visual Perception, Cubes, and Lozenges) are generally thought to be the result of an additional factor (e.g., Visualization). For identification purposes, the factor variance was constrained to one. Thus, the model is fairly misspecified, containing unnecessary indicators of the latent variable. In many situations, researchers may want a more pure formulation of the latent variable, thus performing variable selection by removing the three weak indicators from the model. Although our initial model did not fit well, with a RMSEA of 0.187, removing the three indicators would reduce our degrees of freedom from 9 to zero, effectively removing any form of model testing.

We used Mplus (Muthén and Muthén, 2011) to run two BSEM models. For BSEM, two different small variances priors were used for non-target loadings, $\sim N(0, 0.01)$ so that 95% of the loading variation is between -0.2 and 0.2 , and $\sim N(0, 0.005)$ with 95% of the loading variation between approximately -0.14 and 0.14 . As displayed in Table 1, although our fit does not improve with the use of small variance priors, this is mainly due to the number of parameters remaining consistent in the calculation of the Bayesian information criterion (BIC; Schwarz, 1978). For alternative Bayesian fit criteria that allow the number of parameters to vary, thus exhibiting a preference for models with fewer estimated parameters, see the use of the deviance information criteria (Spiegelhalter et al., 2002; Asparouhov et al.,

2015), the leave-one-out cross-validation (Gelfand and Sahu, 1996), or the widely applicable information criterion (Watanabe, 2010).

Effectively, this act of putting a constraint on the parameter estimate is a form of regularization (i.e. penalized likelihood estimation, shrinkage). Across estimation frameworks, using small variance Normal priors has been shown to be equivalent to that of ridge regularization, both in the context of regression (Kyung et al., 2010; Park and Casella, 2008; Tibshirani, 1996), and factor analysis (Lu et al., 2016). To discuss this connection further, we present a brief overview of frequentist regularization in SEM, termed regularized SEM (RegSEM; Jacobucci et al., 2016).

Regularized Structural Equation Modeling

Although there are multiple formulations of adding regularization into SEM (see Huang et al., 2017, for an alternative), we discuss the formulation from Jacobucci et al. (2016). Specifically, RegSEM builds in an additional element to the traditional maximum likelihood fit function to penalize chosen model parameters. This is formulated as

$$F_{regsem} = F_{ML} + \lambda P(\cdot), \quad (1)$$

where λ is the regularization parameter and takes on a value between zero and infinity, and $P(\cdot)$ is a general function for summing the values of one or more of the model's parameter matrices. When λ is zero, MLE is performed, and when λ is infinity, all penalized parameters are shrunk to zero. Although there are many forms of regularization (see Jacobucci, 2017, for extensions), the two most common forms of $P(\cdot)$ include both the lasso ($\|\cdot\|_1$), which penalizes the sum of the absolute values of the parameters, and ridge ($\|\cdot\|_2$), which penalizes the sum of the squared values of the parameters. The lasso penalty shrinks parameters by a constant amount, driving certain parameter estimates to zero, whereas the ridge penalty shrinks estimates proportionally. The ridge penalty helps stabilize parameter estimates, particularly if multi-collinearity is present, but does not drive estimates to zero. This makes lasso regularization more attractive when model simplification is desired.

Parameters from either the *asymmetric* (A; e.g. regressions or factor loadings) or *symmetric* (S; e.g. variances or covariances) matrices in *Reticular Action Model* (RAM; McArdle and McDonald, 1984; McArdle, 2005) notation, can be penalized. To translate CFA model using six items from the Holzinger Swineford dataset to RAM notation involves first separating directed and undirected paths. In this model, the only directed paths are that of the factor loadings, one for each indicator. These parameters are included in the A matrix. As there are no regression parameters in the model, the rest of the matrix is filled with zeroes. For the S matrix, there are residual variance parameters for each indicator, as well as a variance for the single latent variable, which was fixed to one. The rest of the matrix is filled with zeros outside of these seven parameters. The final matrix is the *filter* matrix which is used to separate manifest from latent variables. For more detail on the RAM notation and its extension to regularization, see Jacobucci et al. (2016).

RegSEM is implemented as a package in R (R Core Team, 2017), termed *regsem* (Jacobucci, 2017). The *regsem* package makes it easy for the researcher to specify a model with the *Javaan* package (Rosseel, 2012). As in regularized regression, a final model is chosen by selecting a large number of λ values (e.g. 50) and running the model for each value of the penalty. Among these models, the final model is the one that achieves the best performance on a chosen fit index. Previous research has shown that the BIC works well for choosing a final model (Jacobucci et al., 2016).

To demonstrate the connection between the RegSEM ridge and the use of small variance Normal distribution priors in Mplus, we ran two RegSEM ridge models on the Holzinger-Swineford dataset. As the exact conversion between prior variance and shrinkage parameter is not known in the context of SEM, RegSEM penalties were modified until correspondence of parameter estimates was achieved. To show correspondence, asserting a conversion from prior variance to value of λ is not as important as a demonstration that the parameter estimates are shrunken at parallel rates as the prior variance is constrained (λ is increased).

In addition to the previously detailed small variance prior models, two RegSEM ridge models with λ equal to 0.17 and 0.34 on the same three penalized factor loadings were run. In this example, as the prior variance was halved (precision doubled), the equivalent penalty for RegSEM Ridge was doubled. Although a seemingly simple conversion between prior variance for BSEM and RegSEM Ridge exists for this example, further testing is needed to assess invariance across model and sample size. The parameter estimates and fit statistics are displayed in Table 2. The parameter estimates were almost identical across BSEM and RegSEM, which confirms that the use of small variance normal distribution priors can be characterized as a form of BSEM ridge regularization, or *BRidge*. We use this term for the remainder of the paper.

The Mplus formulation of small variance priors deviates from other forms of regularization in two ways: 1. it is typical to test a range of prior variances, not just one or two, and 2. variable selection is often desired, and both the Bayesian and frequentist versions of Ridge regularization do not push estimates all the way to zero. As a result, we demonstrate extensions to address these limitations, testing a sequence of penalties, and showing the resultant parameter trajectories (how the parameter estimates change across the sequence of penalties), while also demonstrating the use of the lasso, and compare implementations as a means to perform variable selection.

Extensions

Lasso estimates can be extended to Bayesian estimation by taking the mode of the posterior distribution under independent Laplace distribution priors (Park and Casella, 2008; Tibshirani, 1996). To give an idea of why the use of the Laplace distribution, in comparison to the Normal distribution, is more likely to push posterior estimates towards zero, we plot both the Normal and Laplace distributions in Figure 1. Specifically notice how the Laplace distribution has more density around zero and much less density at values further away from zero. This translates to pushing smaller estimates (in absolute values) more towards zero than when using a Normal prior, along with keeping larger estimates larger than when

Normal priors are used. This is more in line with the aims of variable selection – pare away small estimates while simultaneously not biasing large estimates.

To further test the comparison between ridge procedures, while also examining lasso regularization, a set of 30 penalties or small variance priors were tested on the same one factor model with all six factor loadings penalized. Mplus does not currently allow the use of Laplace distribution priors on factor loadings. As a result, Bayesian models were estimated using the *blavaan* package (Merkle and Rosseel, 2015), which interfaces JAGS (Plummer et al., 2003) in R. Normal distribution priors were used for BRidge, and Laplace distribution priors were used for BSEM lasso (BLasso). We note that JAGS uses precision instead of variances for priors. Across each of the 30 prior precisions, inference for each parameter was made based on the mean for BRidge and the mode for BLasso. Parameter trajectories are displayed in Figure 2A and 2B for ridge. For ridge, the two parameter trajectories are nearly coincident indicating the strong alignment between the two regularization approaches. The parameter trajectories from lasso regularization are displayed in Figure 2C and 2D, exhibiting clearly discrepant trajectories. In RegSEM lasso, estimates approach zero at a similar rate to the BLasso estimates. However in RegSEM lasso the parameter estimates are driven to zero, whereas in BLasso estimates are driven to near zero, without estimates actually reaching zero. This finding agrees with Park and Casella (2008) in that the BLasso acts as a form of hybrid between the frequentist ridge and lasso.

In this example, testing 30 different models can be computationally expensive, particularly using the Bayesian approach to estimation. Additionally, neither form of Bayesian regularization drove the parameters to zero, thus not performing variable selection. However, alternative forms of Bayesian regularization exist, most notably using what is termed hierarchical priors, resulting in the need to test only one model. These newer methods simplify the process of applying Bayesian regularization. Although there are numerous formulations, we detail and demonstrate the use of three: the Bayesian adaptive lasso (BaLasso; Feng et al., 2017) and the horseshoe and horseshoe+ priors (Carvalho et al., 2010; Bhadra et al., 2017). Both variants of the horseshoe have not been applied in the context of SEM before, thus we give particular attention to this form of regularization. However, to first understand both of these formulations, we discuss hierarchical Bayesian models.

Hierarchical Models

As an alternative formulation of the BLasso, Park and Casella (2008) used a scale mixture of normals with an exponential mixing density to represent the Laplace prior. In this model, there is no longer a need to test various values of the prior variance. Instead, the parameter we are penalizing is itself modeled, known as a hierarchical model (aka hyper-prior or hyperparameter; Gelman, 2006). Most recent developments in Bayesian regularization use a hierarchical representation of the model, as this conducts parameter estimation and variable selection simultaneously. Using small variance priors as an example, instead of specifying a fixed variance for a factor loading as

$$\lambda_i \sim N(0, .001), \quad (2)$$

we place a prior on the variance of the factor loading prior, such as

$$\lambda_i \sim N(0, \tau), \quad (3)$$

$$\tau \sim \text{Uniform}(0, 1). \quad (4)$$

This allows the data to be the primary influence of the estimated variance for the factor loading.

Adaptive Lasso Prior.—The hierarchical formulation of the Bayesian Lasso has been generalized to the context of SEM (Feng et al., 2016, 2017; Guo et al., 2012). In these models, the same specification for the hyperprior is set for each of the penalized parameters. However, analogous to findings in frequentist regularization, using the same penalty for each of the parameters can result in bias. Instead, each parameter can be given its own penalty, known as the *adaptive* lasso (alasso Zou, 2006). Extending this to Bayesian estimation involves specifying a unique hyper-prior for each parameter subjected to textbfregularization, known as the BaLasso (Feng et al., 2017), which has been shown to outperform the hierarchical BLasso. In this, the prior on the parameters of interest (θ_j) is specified as

$$\theta_j \sim N(0, \psi_j \tau_j^2), \quad (5)$$

$$\tau_j^2 \sim \text{Gamma}\left(1, \frac{\gamma_j^2}{2}\right), \quad (6)$$

$$\psi_j^{-1} \sim \text{Gamma}(\alpha, \beta), \quad (7)$$

$$\gamma_j^2 \sim \text{Gamma}(\alpha, \beta), \quad (8)$$

with α set to 1 and β as .05, following the recommendations by Feng, Wu, and Song (2017). This prior on a prior, resulting in a gamma mixture of normals, capitalizes on the fact that the Laplace distribution can be expressed as a scale mixture of normal distributions with independent exponentially distributed variances (i.e. Gamma with $\alpha=1$; Andrews and Mallows, 1974).

Horseshoe and Horseshoe+ Prior.—Similar to the BLasso in that it is also a member of the family of multivariate scale mixtures of normals, the horseshoe prior (Carvalho et al., 2010) was developed to deal with scenarios where variable selection is the goal. The horseshoe prior will leave obvious signals unshrunk, meaning robustness when the model is not sparse, while exhibiting efficiency at shrinking noise parameters (Carvalho et al., 2010). This is specified as

$$\theta_j \sim N(0, \rho_j^2 \tau^2), \quad (9)$$

$$\rho_j \sim C^+(0, 1), \quad (10)$$

$$\tau \sim C^+(0, 1), \quad (11)$$

where $C^+(0,1)$ is a half-Cauchy distribution. In this, ρ_j is a local shrinkage parameter, while τ is the global shrinkage parameter.

An extension of the horseshoe, derived for more sparse models is the horseshoe+ of Bhadra et al. (2017). This is specified as

$$\theta_j \sim N(0, \rho_j^2), \quad (12)$$

$$\rho_j \sim C^+(0, \tau \eta_j), \quad (13)$$

$$\eta_j \sim C^+(0, 1), \quad (14)$$

$$\tau \sim C^+(0, 1). \quad (15)$$

In comparison to the horseshoe, the horseshoe+ involves the specification of an additional prior, adding a hierarchical representation of ρ_j . In a comparison, the horseshoe+ exhibits lower bias in estimating *ultra-sparse* signals (Bhadra et al., 2017). Given that both of these methods were developed for very sparse data settings, and that this is rare in the case of SEM, it was our aim to examine the performance of both methods, particularly in comparison to the BaLasso.

Second Example

To provide a further example of the application of regularization methods in SEM, we examine the effect of covariates on changes in reading achievement over time using a latent growth curve model (McArdle and Epstein, 1987; Meredith and Tisak, 1990). Data come from the Early Childhood Longitudinal Study - Kindergarten Cohort (Tourangeau et al., 2009, ECLS-K). The ECLS-K includes a nationally representative sample of 21,260 children attending kindergarten in 1998–1999, with measurements spanning the fall of kindergarten to the end of eighth grade. For these analyses, we used a measure of reading achievement collected during the fall and spring of kindergarten, fall and spring of 1st, and the spring of 3rd, 5th, and 8th grades. For the analyses, we took a random subset of 500 participants. These are the same data that were illustrated in Jacobucci et al. (2017). In comparison to using the covariates to identify important subgroups with structural equation model trees (Brandmaier et al., 2013), we used both frequentist and Bayesian regularization to perform variable selection of covariates in the prediction of latent means.

For this model, eight covariates were included to evaluate their influence on the reading trajectories over time. These variables included fine motor skills (fine), gross motor skills (gross), approaches to learning (learn), self-control (control), interpersonal skills (interp), externalizing behaviors (ext), internalizing behaviors (int), and general knowledge (gk), which were all measured in the fall of kindergarten. Particularly when the sample size is small, including this number of covariates may be problematic (see Jacobucci, 2017, for further detail). Additionally, researchers may be interested in performing variable selection to simplify the depiction of mean changes in both the intercept and slope attributable to covariates.

To model reading trajectory, we used a latent basis growth model to capture non-linearity in the trajectories. To accomplish this, we specified a factor loading matrix of

$$\begin{bmatrix} 1 & 0 \\ 1 & \lambda_{22} \\ 1 & \lambda_{32} \\ 1 & \lambda_{42} \\ 1 & \lambda_{52} \\ 1 & \lambda_{62} \\ 1 & 1 \end{bmatrix}$$

where the first column defines the latent variable intercept and the second column defines the latent shape or slope variable, representing change over time.

This latent basis growth model, including the eight predictors of both the latent intercept and slope, is displayed in Figure 3. To select which predictors are important, without relying on the basis of p-values, we used both the RegSEM Lasso, along with BaLasso. For pedagogical purposes we did not include additional forms of frequentist or Bayesian regularization. We used the same specification for each model as detailed previously while

comparing these estimates to that of MLE. The parameter estimates from these models are displayed in Table 3.

In this, we first note that the parameter estimates (mean of the posterior distribution for BaLasso) are almost identical for each parameter in the measurement model across the three forms of estimation. For the covariate estimates, the BaLasso model resulted in sparser estimates in comparison to the RegSEM Lasso. Counting parameters estimated at ± 0.1 , following the recommendation of Feng et al. (2017), the BaLasso selected six parameters, while the RegSEM Lasso estimated eleven parameters as non-zero. In contrast, although MLE had many more non-zero parameter estimates, only six parameters were estimated as significant at $p < 0.05$. Of the parameters estimated as significant, two did not have mean BaLasso estimates as ± 0.1 . The point to keep in mind when comparing methods is that by performing variable selection, one is also changing the conditional distributions between variables, which can affect which parameters are counted as important.

In a setting such as this, where the number of covariates is small (purposely done for ease of understanding), researchers may choose to perform variable selection according to a number of different methods. The use of p-values relies on asymptotic performance, whereas the use of regularization relies on the assumption of sparsity (e.g. Hastie et al., 2015). However, in higher dimensional settings, relying on the assumption of sparsity becomes increasingly important, as either the sample size or amount of signal in the data limits what can be extracted.

Software

Due to the creation of general purpose software for Bayesian estimation, it is no longer necessary for researchers to program their own sampler, such as Gibbs (Geman and Geman, 1984), in order to estimate the model of interest. Instead, general purpose software such the more traditional JAGS (Plummer et al., 2003) or BUGS (Lunn et al., 2009; Zhang, 2014), or the more recently developed Stan (Carpenter et al., 2016) and PyMC3 (Salvatier et al., 2016) can be used, as well as specific SEM software that contain Bayesian samplers, such as Mplus, AMOS (Arbuckle, 2010), or *blavaan* (Merkle and Rosseel, 2015). For each of the general purpose Bayesian software, every form of regularization can currently be implemented. One alternative form of regularization, the spike-and-slab prior (Ishwaran and Rao, 2005), which was formulated in the context of factor analysis (Lu et al., 2016), is not available for testing with these software packages, and was therefore not detailed here.

The first frequentist regularization software packages for latent variable modeling were created for use in factor analysis. These include both the *fanc* package (Hirose and Yamamoto, 2015) and *FANet* (Blum et al., 2014) in the R statistical environment (R Core Team, 2017). In the context of SEM frequentist regularization, two different general SEM packages exist in the R statistical environment: *regsem* (Jacobucci, 2017) and *IsI* (Huang et al., 2017). These packages use different SEM matrix formulations, as well as different structures for specifying the models in R. Outside of work evaluating the application of the *fanc* package for use in small samples (Finch and Finch, 2016), very little research has gone into evaluating the application of any frequentist regularization latent variable software.

For the purposes of this paper, frequentist analyses were conducted using the *regsem* package, while two different general purpose Bayesian samplers were used. For the empirical example, the original models were specified using the *blavaan* package to easily translate latent growth curve model, as specified using *lavaan*, to JAGS. For the simulation, each of the models were specified and run using Stan because there is active community discussing the different ways for specifying the horseshoe and horseshoe+ priors (see Piironen and Vehtari, 2015, for Stan code in specifying it using regression), as well as for the specification of the frequentist lasso and ridge.

Discussion

The main purpose of comparing forms of regularization across frequentist and Bayesian estimation was to demonstrate the options available to researchers, while highlighting the generalization of Bayesian regularization, from small variance priors to hierarchical forms of regularization. Our demonstration of similarities began with showing how identical estimates can be achieved with ridge penalties in RegSEM and small variance Normal distribution priors in BSEM. While both Bayesian and frequentist ridge methods produced comparable parameter estimate trajectories, this was not the case for the lasso methods. In line with previous research, the BLasso acted as a form of hybrid between ridge penalties and the frequentist lasso. The RegSEM lasso drove parameter estimates all the way to zero, whereas the BLasso drove parameter estimates at almost an identical rate; however, parameter estimates hovered above zero. This was extended to three hierarchical forms of Bayesian regularization: the Bayesian adaptive lasso, the horseshoe, and horseshoe+. Across both demonstrations, our goal was to provide a framework for understanding the options available to researchers for regularization in Bayesian estimation, and how this related to their frequentist counterparts.

Framework Comparison

The purpose of this paper was not to highlight the strengths of one approach while discussing the corresponding weaknesses of the other estimation method. Instead, our aim was to detail both estimation frameworks and in particular, highlight Bayesian regularization, as this has received comparatively less coverage in the literature. For instance, we began our overview of the methods with the use of small-variance Normal distribution priors, demonstrating its correspondence with the use of Ridge regularization in frequentist estimation. In extension, with more sparse forms of regularization, there is less similarity across estimation frameworks.

In our own research, we have utilized both regularization frameworks depending on the context. With complex longitudinal SEM models, hierarchical Bayesian regularization worked better in comparison to frequentist estimation, as the Bayesian forms had less difficulty reaching convergence in the more complex models (Jacobucci and Grimm, 2018). If researchers have simpler models, or desire for the model parameters to be estimated as exactly zero, then the frequentist framework may be preferable. Both frameworks of regularized estimation are undergoing rapid research and development, which is likely to lead to fewer distinctions between frameworks will be come less and less.

Future Research

We strongly believe that one of the areas of greatest potential for future research is in studying the various form of SEM regularization for choosing models that are more likely to generalize. The natural integration of various cross-validation methods with SEM regularization makes this process increasingly simple for researchers. Particularly in small samples, future research should build off of work highlighting the limitations of default BSEM procedures (McNeish, 2016). Regularization in both estimation frameworks is particularly tailored to problems specific to small samples by limiting the dimensionality of the model. Specifically, integrating various methods tailored for small sample problems with both ridge and lasso regularization opens up a variety additional options for producing less biased results.

As the size of SEMs grow, less emphasis should be placed on the actual size of the sample, and more on the ratio of estimated parameters to the number of observations. With the integration of regularization, this entails accurately measuring the *effective* number of parameters (or degrees of freedom), as placing constraints, either through a prior or penalty, limits the influence of a parameter, and thus the size of the model. In particular, this affects the calculation of fit indices that incorporate degrees of freedom. As an example, in Mplus, the BIC does not adjust the number of estimated parameter when small variance priors are used. If a parameter is given a small variance prior, which heavily constrains the estimate to be near zero, the BIC counts the estimate as a full parameter when determining the number of estimated parameters. As an alternative, Asparouhov et al. (2015) argued for the use of the deviance information criterion (DIC; Spiegelhalter et al., 2002), as the DIC takes into account constraints placed on estimated parameters by adjusting the effective number of parameters. The same difficulties exist in RegSEM, where RegSEM with lasso penalties results in increasing degrees of freedom as the penalty is increased (as some parameters are estimated as zero). However, in RegSEM with ridge penalties (along with the adaptive lasso and elastic net), an appropriate method for calculating the changes in degrees of freedom has not been determined. In contrast to both of these, the use of hierarchical Bayesian models is in and of itself a form of model selection. More research is warranted on both the calculation of the effective number of estimated parameters, and how this relates to accurately choosing the final model.

Currently, there has been an influx in the creation of new hierarchical Bayesian forms of research. For instance, an extension to both the horseshoe and horseshoe+ was recently proposed (Piironen and Vehtari, 2017). Despite this, very little work has been done on how these methods fair in typical SEM research, with the exception of research by Feng et al. (2017) and Lu et al. (2016). Going forward, more research is needed on setting recommendations for which method may be best and in what settings. Our simulation was limited in the number of conditions tested, and did not find any particular differences in performance across the three hierarchical methods. Future research should consider larger models, both in terms of the number of variables and estimated parameters.

Concluding Remarks

In social and behavioral research, it is common to have complex hypotheses (models) and limited datasets. In these, our datasets may not contain enough information, either in the form of small sample size or missing data, to test complex models. SEM regularization, both frequentist and Bayesian forms, is one way to overcome this mismatch. By reducing the influence of parameters that are not of central interest in order to overcome problems with identification and to limit bias, researchers can maximize the information gleaned from the data.

The development of methods for SEM regularization in both frequentist and Bayesian estimation gives researchers a host of options for accomplishing a number of goals. As SEMs grow with the increasing availability of large datasets, regularization should and will become a necessary tool in a researcher's statistical toolbox. Our focus was on comparing multiple forms of Bayesian and frequentist regularization methods, as very little research has touched on the over-arching commonalities and discrepancies across methods. Although this research is in its infancy, we believe that its use and number of applications will continue to grow rapidly.

Acknowledgments

The first author was supported by funding through the National Institute on Aging Grant number T32AG0037.

References

- Andrews DF and Mallows CL (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–102.
- Arbuckle JL (2010). *Ibm spss amos 19 user's guide*. Crawfordville, FL: Amos Development Corporation, 635.
- Asparouhov T and Muthén B (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(4):495–508.
- Asparouhov T, Muthén B, and Morin AJ (2015). Bayesian structural equation modeling with cross-loadings and residual covariances comments on stromeyer et al. *Journal of Management*, 41(6): 1561–1577.
- Bhadra A, Datta J, Polson NG, Willard B, et al. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*.
- Blum A, Houee M, Lagarrigue S, and Causeur D (2014). The r package fanet: sparse factor analysis model for high dimensional gene co-expression networks. In *The International R Users Conference*, page np.
- Brandmaier AM, von Oertzen T, McArdle JJ, and Lindenberger U (2013). Structural equation model trees. *Psychological Methods*, 18(1):71–86. [PubMed: 22984789]
- Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker MA, Guo J, Li P, and Riddell A (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, 20:1–37.
- Carvalho CM, Polson NG, and Scott JG (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.
- Choi J, Zou H, and Oehlert G (2010). A penalized maximum likelihood approach to sparse factor analysis. *Statistics and Its Interface*, 3(4):429–436.
- Feng X-N, Wu H-T, and Song X-Y (2016). Bayesian adaptive lasso for ordinal regression with latent variables. *Sociological Methods & Research*, page 0049124115610349.

- Feng X-N, Wu H-T, and Song X-Y (2017). Bayesian regularized multivariate generalized latent variable models. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(3):341–358.
- Finch WH and Finch MEH (2016). Fitting exploratory factor analysis models with high dimensional psychological data. *Journal of Data Science*, 14(3).
- Gelfand AE and Sahu SK (1996). Identifiability, propriety, and parametrization with regard to simulation-based fitting of generalized linear mixed models,”. In *Technical Report 96–36*. Citeseer.
- Gelman A (2006). Multilevel (hierarchical) modeling: what it can and cannot do. *Technometrics*, 48(3):432–435.
- Geman S and Geman D (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741. [PubMed: 22499653]
- Guo R, Zhu H, Chow S-M, and Ibrahim JG (2012). Bayesian lasso for semiparametric structural equation models. *Biometrics*, 68(2):567–577. [PubMed: 22376150]
- Hamagami F, Zhang Z, and McArdle J (2009). Modeling latent difference score models using bayesian algorithms In Chow SM, Ferrer E., F. H., editor, *Statistical methods for modeling human dynamics: An interdisciplinary dialogue*, pages 319–348. Erlbaum Mahwah, NJ.
- Hastie T, Tibshirani R, and Wainwright M (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press.
- Hirose K and Yamamoto M (2014). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing*, pages 1–13.
- Hirose K and Yamamoto M (2015). Sparse estimation via nonconcave penalized likelihood in factor analysis model. *Statistics and Computing*, 25(5):863–875.
- Holzinger KJ and Swineford F (1939). A study in factor analysis: the stability of a bi-factor solution. *Supplementary Educational Monographs*.
- Huang P-H, Chen H, Weng L-J, et al. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika*, 82(2):329–354. [PubMed: 28417228]
- Ishwaran H and Rao JS (2005). Spike and slab variable selection: frequentist and bayesian strategies. *Annals of statistics*, pages 730–773.
- Jacobucci R (2017). regsem: Regularized structural equation modeling. arXiv preprint arXiv: 1703.08489.
- Jacobucci R and Grimm KJ (2018). Regularized estimation of multivariate latent change score models In *Advances in Longitudinal Models for Multivariate Psychology: A Festschrift for Jack McArdle*.
- Jacobucci R, Grimm KJ, and McArdle JJ (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(4):555–566. [PubMed: 27398019]
- Jacobucci R, Grimm KJ, and McArdle JJ (2017). A comparison of methods for uncovering sample heterogeneity: Structural equation model trees and finite mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, 24(2):270–282. [PubMed: 29225453]
- Jöreskog KG (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, 34(2):183–202.
- Jung S and Takane Y (2008). Regularized common factor analysis In Shigemasu K, editor, *New trends in psychometrics*, pages 141–149. Universal Academy Press.
- Kyung M, Gill J, Ghosh M, and Casella G (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5(2):369–411.
- Lawley DN (1940). Vi.—the estimation of factor loadings by the method of maximum likelihood. *Proceedings of the Royal Society of Edinburgh*, 60(01):64–82.
- Lee SY (2007). *Structural equation modeling: A Bayesian approach*, volume 711 John Wiley & Sons.
- Lee S-Y and Song X-Y (2004). Evaluation of the bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, 39(4):653–686. [PubMed: 26745462]
- Liu H, Zhang Z, and Grimm KJ (2016). Comparison of inverse wishart and separation-strategy priors for bayesian estimation of covariance parameter matrix in growth curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(3):354–367.

- Lu ZH, Chow SM, and Loken E (2016). Bayesian factor analysis as a variable-selection problem: Alternative priors and consequences. *Multivariate Behavioral Research*, pages 1–21. [PubMed: 26881950]
- Lu Z-H, Chow S-M, and Loken E (2017). A comparison of bayesian and frequentist model selection methods for factor analysis models. *Psychological Methods*, 22(2):361. [PubMed: 28594228]
- Lunn D, Spiegelhalter D, Thomas A, and Best N (2009). The bugs project: evolution, critique and future directions. *Statistics in medicine*, 28(25):3049–3067. [PubMed: 19630097]
- McArdle JJ (2005). The development of the ram rules for latent variable structural equation modeling. *Contemporary psychometrics: A festschrift for Roderick P. McDonald*, pages 225–273.
- McArdle JJ and Epstein D (1987). Latent growth curves within developmental structural equation models. *Child development*, pages 110–133. [PubMed: 3816341]
- McArdle JJ and McDonald RP (1984). Some algebraic properties of the reticular action model for moment structures. *British Journal of Mathematical and Statistical Psychology*, 37(2):234–251. [PubMed: 6509005]
- McNeish DM (2016). On using bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(5):750–773.
- Meredith W and Tisak J (1990). Latent curve analysis. *Psychometrika*, 55(1):107–122.
- Merkle EC and Rosseel Y (2015). blavaan: Bayesian structural equation models via parameter expansion. arXiv preprint arXiv:1511.05604.
- Moore TM, Reise SP, Depaoli S, and Haviland MG (2015). Iteration of partially specified target matrices: Applications in exploratory and bayesian confirmatory factor analysis. *Multivariate behavioral research*, 50(2):149–161. [PubMed: 26609875]
- Muthén B and Asparouhov T (2011). Bayesian sem: A more flexible representation of substantive theory. *Psychological Methods*.
- Muthén B and Asparouhov T (2013). Bsem measurement invariance analysis. *Mplus Web Notes*, 17:1–48.
- Muthén B and Muthén L (2012). *Mplus Version 7: User's guide*. Muthén and Muthén.
- Muthén LK and Muthén BO (2011). *Mplus: Statistical analysis with latent variables: User's guide*. Muthén & Muthén Los Angeles.
- Ning L and Georgiou TT (2011). Sparse factor analysis via likelihood and l1-regularization. In *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, pages 5188–5192. IEEE.
- Park T and Casella G (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.
- Piironen J and Vehtari A (2015). Projection predictive variable selection using stan+ r. arXiv preprint arXiv:1508.02502.
- Piironen J and Vehtari A (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. arXiv preprint arXiv:1707.01694.
- Plummer M. Jags: A program for analysis of bayesian graphical models using gibbs sampling; *Proceedings of the 3rd international workshop on distributed statistical computing*; Vienna, Austria. 2003. 125volume 124
- R Core Team (2017). *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria.
- Rosseel Y (2012). lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, 48(2):1–36.
- Salvatier J, Wiecki TV, and Fonnesbeck C (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55.
- Scheines R, Hoijtink H, and Boomsma A (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, 64(1):37–52.
- Schwarz G (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.
- Serang S, Jacobucci R, Brimhall KC, and Grimm KJ (2017). Exploratory mediation analysis via regularization. *Structural Equation Modeling: A Multidisciplinary Journal*, pages 1–12.

- Spiegelhalter DJ, Best NG, Carlin BP, and Van Der Linde A (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.
- Thurstone LL (1935). *The vectors of mind*. University of Chicago Press, Chicago, IL.
- Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.
- Tourangeau K, Nord C, Lê T, Sorongon A, Najarian M, and Hausken E (2009). Combined user's manual for the ecls-k eighth-grade and k–8 full sample data files and electronic codebooks. National Center for Education Statistics, Institute of Education Sciences, US Department of Education Washington, DC.
- van de Schoot R, Winter SD, Ryan O, Zondervan-Zwijnenburg M, and Depaoli S (2017). A systematic review of bayesian articles in psychology: The last 25 years. *Psychological Methods*, 22(2):217. [PubMed: 28594224]
- Watanabe S (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec): 3571–3594.
- Wirth R and Edwards MC (2007). Item factor analysis: current approaches and future directions. *Psychological methods*, 12(1):58. [PubMed: 17402812]
- Zhang Z (2014). Webbugs: Conducting bayesian statistical analysis online. *Journal of Statistical Software*, 61(1):1–30.
- Zhang Z (2016). Modeling error distributions of growth curve models through bayesian methods. *Behavior research methods*, 48(2).
- Zhang Z, Hamagami F, Lijuan Wang L, Nesselroade JR, and Grimm KJ (2007). Bayesian analysis of longitudinal data using growth curve models. *International Journal of Behavioral Development*, 31(4):374–383.
- Zou H (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.

Figure 1.
Comparison of Normal and Laplace Distributions

(a) RegSEM Ridge Parameter Estimates

(b) BRidge Parameter Estimates

(c) RegSEM Lasso Parameter Estimates

(d) BLasso Parameter Estimates

Figure 2.

Parameter estimate trajectories for all four regularization methods across a wide range of penalties or prior variances. Note that the parameter estimates were scaled as relative to their un-penalized estimates.

Figure 3. Latent Basis Growth Model for the ECLS-K Data. Note that some regression parameters (β) are omitted.

Table 1.

Parameter estimates and BIC fit for ML and both BSEM small variance priors. Small variance priors were placed on the bolded scales.

Scales	Factor Loadings		
	ML	BSEM N(0,0.01)	BSEM N(0,0.005)
Visual Perception	0.418	0.309	0.244
Cubes	0.212	0.154	0.120
Lozenges	0.202	0.140	0.107
Paragraph Comprehension	0.851	0.835	0.830
Sentence Completion	0.845	0.835	0.832
Word Meaning	0.839	0.819	0.810
Fit			
BIC	4656.5	4662.3	4669.9

Table 2.

Parameter estimates and BIC fit for both ML, BSEM ridge, and RegSEM ridge. Small variance priors were placed on the bolded scales.

Scales	Factor Loadings				
	ML	BSEM $N(0,0.01)$	RegSEM $\lambda = 0.17$	BSEM $N(0,0.005)$	RegSEM $\lambda = 0.34$
Visual Perception	0.418	0.309	0.308	0.244	0.246
Cubes	0.212	0.154	0.149	0.120	0.116
Lozenges	0.202	0.140	0.139	0.107	0.107
Paragraph Comprehension	0.851	0.835	0.832	0.830	0.833
Sentence Completion	0.845	0.835	0.832	0.832	0.836
Word Meaning	0.839	0.819	0.819	0.810	0.818
Fit					
BIC	4656.5	4662.3	4661.8	4669.9	4669.2

Table 3.

Parameter estimates for the latent-basis growth model using the ECLS-K dataset. MLE refers to maximum likelihood estimation, Lasso is the RegSEM Lasso model, and BaLasso is the Bayesian adaptive lasso. Note that * refers to covariate parameters that were significant at $p < .05$.

Par	MLE	Lasso	BaLasso
β_{11}	1.73*	1.03	1.81
β_{21}	0.53	0.00	0.00
β_{31}	3.83*	3.12	2.90
β_{41}	-0.52	0.00	-0.11
β_{51}	-1.58	-0.88	-0.24
β_{61}	-0.25	0.00	0.01
β_{71}	-0.63	0.00	-0.01
β_{81}	3.32*	2.65	3.34
β_{12}	1.15	0.50	0.05
β_{22}	-1.74	-1.04	-0.02
β_{32}	0.40	0.00	0.00
β_{42}	-3.55*	-2.80	-0.04
β_{52}	1.70	1.00	-0.01
β_{62}	-2.36*	-1.54	-0.03
β_{72}	1.42	0.71	0.01
β_{82}	12.41*	11.69	12.15
α_1	37.52	37.50	37.53
α_2	137.92	137.90	137.84
λ_{22}	0.09	0.09	0.09
λ_{32}	0.14	0.14	0.14
λ_{42}	0.35	0.35	0.35
λ_{52}	0.71	0.71	0.71
λ_{62}	0.87	0.87	0.87
θ_{11}	47.44	46.09	46.98
θ_{22}	22.97	23.76	24.09
θ_{33}	65.09	66.12	66.82
θ_{44}	184.01	184.72	184.62
θ_{55}	164.51	164.41	164.84
θ_{66}	77.58	77.70	78.34
θ_{77}	199.94	199.73	200.48
ψ_{11}	106.63	102.02	107.86
ψ_{22}	301.02	297.19	309.97
ψ_{21}	-53.35	-52.60	-53.65