Handbook of Structural Equation Modeling, 2nd ed, edited by Rick Hoyle

Foundations and Extensions of Bayesian Structural Equation Modeling

Online Supplementary Material

Sarah Depaoli, David Kaplan, and Sonja D. Winter

Six Examples of Bayesian SEM

This section provides five examples of Bayesian SEM. Example 1 presents a simple two-factor Bayesian confirmatory factor analysis (CFA). This model is extended in Example 2 to include near-zero priors on cross-loadings. Example 3 presents a Bayesian latent growth curve model (LGCM) with diffuse and informative priors. Example 4 presents an application of Bayesian Model Averaging (BMA). Example 5 presents a structural equation model (SEM) estimated with and without residual covariances. Finally, example 6 presents two forms of Bayesian regularization to avoid overfitting a model with many predictors. All data, code, and results can be found in the online supplementary material at: https://osf.io/24xyv/.

Example 1: Bayesian CFA with Diffuse Priors

CFA is a restricted factor analysis model that can be used to help define latent subgroups of items. The current example illustrates Bayesian CFA without cross-loadings (i.e., a simple structure CFA) and implements diffuse prior settings. Data were obtained from a sample of 2000 Canadian 4th graders who participated in the IEA sponsored Progress in International Reading and Literacy Study (PIRLS). Questions focused on the extent to which students felt that they were confident in reading. An example item was "I usually do well in reading," measured on a four-point scale from "agree a lot" to "disagree a lot." Items were reverse coded. For the sake of this example, we randomly selected 400 participants and tested a two-factor model.

Factor 1 reflected a positive reading experience ("PosRead") and was comprised of three items: "I usually do well in reading" (*dowell*), "Reading is easy for me" (*easy*), and "If a book is interesting, I don't care how hard it is to read" (*interesting*). Factor 2 was defined as a negative reading experience ("NegRead") and was also made up of three items: "Reading is harder for me

than for many of my classmates" (*harder*), "I have trouble reading stories with difficult words" (*difficultwords*), and "Reading is harder for me than any other subject" (*otherthings*). Figure 1 includes a picture of the model with relevant notation. In this example, only primary loadings (solid lines) were estimated, and all cross-loadings (dashed lines) were fixed to zero. In addition, the loadings for the first item on each factor (Item 1 for Factor 1, and Item 4 for Factor 2) were fixed to 1.0 to set the scale of the latent factor.

Priors and Estimation Settings. For the first analysis, the `blavaan` package (Merkle & Rosseel, 2018) in R was used with default prior settings. It is important to be mindful that some software languages scale the model and priors to variances, while others use precisions (i.e., the inverse of the variance) or standard deviations. For this example, we used the JAGS implementation of the Gibbs sampler, which presents some priors in terms of precisions. However, in order to keep a consistent framing throughout the examples, when we discuss the settings for a precision hyperparameter, we will convert it to the variance scale in the text.

The relevant default prior settings for this model are: $N(0, \text{precision} = 0.01)$ or $N(0, \text{variance} = 100)$ for factor loadings and latent factor means, Gamma(1, 0.5) placed on the residual precision, and Wishart($\mathbf{I}$, 3) placed on the latent covariance matrix. Two MCMC chains were generated with 200,000 total iterations, and the first 100,000 were discarded as burn-in. Convergence was monitored using the $\widehat{R}$ convergence criterion (Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2019). The highest $\widehat{R}$ obtained in the analysis was 1.00, which indicates chain convergence for all parameters. The lowest effective sample size (ESS; for the latent factor mean for Factor 2, "NegRead") was 18626.07, which was deemed adequate for this model (Kruschke, 2015; Zitzmann & Hecht, 2019).

Model Interpretation. Table 1 presents estimates for the six items. This table includes the following information: posterior median, mean, and standard deviation; upper and lower 95% credible interval values (equal tails), upper and lower 95% HDI (unequal tails), and ESS for each model parameter. Likewise, various plots have been included as an example of the type of results that can be examined for each model parameter in a Bayesian analysis. Figure 2 presents the

following plots for Item 2 loading onto Factor 1: trace-plot, autocorrelation plot, posterior histogram and density, and the HDI histogram and density plots. Plots for all model parameters can be found in the online supplementary material.

Taking a closer look at Table 1, the posterior mean and median, and similarly the 95% credible interval and HDI, were nearly identical for all parameters. This implies that the posterior distributions were all relatively symmetric, which is confirmed for Item 2 in Figure 2. Focusing on the specific parameters, we can see that Item 5 (*difficultwords*) had the strongest association with the underlying "NegRead" latent factor ($M_{posterior}$ = 1.23, 95% HDI = [1.16; 1.30]). The 95% HDI of all factor loadings excluded zero, indicating that the factor loadings were of a meaningful magnitude. Furthermore, the covariance between the "PosRead" and "NegRead" factors was negative ($M_{posterior}$ = -0.19) and its 95% HDI did not include zero. The posterior mean level of "PosRead" was 3.50 (95% HDI = [3.43; 3.56]) whereas the posterior mean level of "NegRead" was 1.78 (95% HDI = [1.69; 1.87]).

Model Fit. The quality of fit in this CFA was examined using the posterior predictive checking procedure based on the chi-square discrepancy function. The procedure produced a posterior predictive *p*-value that was rounded to 0.00, indicating misfit. Figure 3 illustrates a visual depiction of this misfit (top plot), where adequate fit would have the points plotted along the 45-degree line. The results from this analysis indicate that there is a considerable lack of fit. Therefore, alternative model specifications should be examined.

Example 2: Bayesian CFA with Approximate-Zero Cross-Loadings

A common goal of scale construction is to create items that neatly load onto one latent factor. However, when factors are correlated, it may not be realistic to assume that items load solely onto a single factor (i.e., that all cross-loadings are exactly zero). Items may have some level of construct-relevant association with the other factors. Classic (simple structure) CFA models assume that all of the cross-loadings are exactly zero. With Bayesian SEM, it becomes possible to place small variance priors centered around zero on these cross-loadings. This replaces the assumption of *exactly* zero cross-loadings with *approximately* zero cross-loadings. This example

explores the implementation of approximate-zero priors on cross-loadings.

The same data and two-factor structure from Example 1 were used here, but we extended the model to allow for cross-loadings. In order to keep the basic two-factor structure hypothesized in Example 1, the cross-loadings were not allowed to be freely estimated. Instead, we used approximate-zero priors on these cross-loadings, centered at zero and with small variance hyperparameters (i.e., large precision hyperparameters). The following items received approximate-zero priors for Factor 1 loadings: *harder*, *difficultwords*, and *otherthings*. Likewise, the following items received an approximate-zero prior for Factor 2: *dowell*, *easy*, and *interesting*. The model depicted in Figure 1 shows the presence of the cross-loadings with dashed lines.

Prior Settings and Model Interpretation. Several different prior settings were examined for these cross-loadings, each with a different variance hyperparameter. Specifically, we tested prior settings with variance hyperparameters of 0.001, 0.005, 0.01, and 0.02 in order to determine the optimal hyperparameter setting for this model. In addition, we compared results to the findings from Example 1, where the cross-loadings were fixed to exactly zero. It is important to reiterate that `blavaan` code using JAGS implements precisions as opposed to variances. With a cross-loading prior of $N(0, 0.005)$, the variance hyperparameter is 0.005 and the precision is 200 (precision = 1/0.005 = 200).

Typically, several small variances are assessed to see at what point model fit no longer improves, or the small variance prior becomes too wide and causes convergence issues. At that point, cross-loadings can be inspected to see if any have reached a meaningful level of association. Here, we examine a two-factor CFA with exact zero cross-loadings, as well as four approximate-zero cross-loading specifications. In the models with approximate-zero priors, all possible cross-loadings were estimated. The results for these models are presented in Table 2; parameter estimates for all models are presented in the online supplementary material.

Results from Table 2 indicate that using a variance hyperparameter of 0.005 for the cross-loadings produces the best model fit with a posterior predictive p-value of 0.193 and the smallest BIC of 5703.334. Although the variance hyperparameter settings of 0.01 or 0.02 produce smaller

DICs, these prior settings resulted in convergence problems, even after doubling the number of iterations. Table 2 shows that the highest $\widehat{R}$ value for the 0.01 and 0.02 variance settings (bottom two rows) were indicative of non-convergence, while the other models appeared to have converged based on the $\widehat{R}$ values.

It is important to highlight the differences in the ESSs across the models. Notice that the ESSs are notably higher for the model with exact-zero cross-loadings. This is an important point to consider in practice because it may be that requesting approximate-zero cross-loadings will require a much longer chain to obtain adequate ESS levels. Kruschke (2015) indicates that ESSs should be near 10,000 to obtain adequate information in the chains, however, Zitzmann and Hecht (2019) shows that values of 1,000 can be adequate depending on the model type. In either case, the model with a cross-loading variance hyperparameter of 0.005 did not reach the ESS recommendations despite evidence of convergence (largest $\widehat{R} \leq 1.01$); this could be an indication additional models should be explored. ESS values were only less than 1,000 for two parameters in the final model selected: the factor mean of "NegRead" and the residual variances of Item 4. Whereas for the models with a variance hyperparameter of 0.01 and 0.02, ESS values were less than 1,000 for the majority of model parameters. These results led us to conclude that the small variance hyperparameter of 0.005 was a viable selection for the final model. All loadings and cross-loadings for the model using 0.005 as the variance hyperparameter produced stable trace-plots, showing visual support for convergence. The trace-plots for the cross-loadings can be found in Figure 4.

The posterior distributions of the cross-loadings can also be examined to see if any are noticeably different from 0 (i.e., if their 95% highest density interval (HDI) does not include 0). Figure 5 shows that three out of six cross-loadings are centered around zero. Items 2, 3, and 5 all have HDIs that do not include zero. Item 2 ("Reading is easy for me") has an HDI that is shifted negative, indicating there is a negative relationship with the factor. Item 3 ("If a book is interesting, I don't care how hard it is to read") and Item 5 ("I have trouble reading stories with difficult words") have HDIs that are shifted positive, indicating a positive relationship with the

corresponding factor.

Finally, we can compare the posterior distribution for the latent factor covariance estimate for the model with and without approximate-zero cross-loadings. If cross-loadings exist, not including them in the model will inflate the latent factor covariance estimate. Thus, we expect the covariance estimate to be larger for the model without cross-loadings. As we can see in Figure 6, the median and mean covariance is slightly more negative for the model without cross-loadings, although the difference is negligible.

Model Fit. The posterior predictive *p*-value results can be compared for the model with cross-loadings of exactly zero (from Example 1) to the model with approximate-zero cross-loadings with the prior $N(0, 0.005)$. Figure 3 presents this comparison, and it further confirms that including the approximate-zero cross-loadings improves model fit (the subtle difference in posterior predictive *p*-value as reported in the table versus this figure is due to the fact that the figure is created through a different function than the numeric output).

Example 3: Latent Growth Curve Model with Math Achievement Data: Comparing Diffuse and Informative Priors

Data for this example are comprised of an unweighted sample of children from the Early Childhood Longitudinal Study–Kindergarten (ECLS-K) class of 1998-1999 (National Center for Education Statistics [NCES], 2001). Math assessment data were extracted for students in kindergarten and first grade (fall and spring assessments each year). Item response theory was used to derive scale scores across these four time-points used for the LGCM. Estimation of growth reflects math skill development over the 18 months of the study.

Two samples were extracted for this example. The first dataset (Dataset 1) consisted of 600 children. An initial model was estimated through the R package `blavaan`. Two chains were generated using NUTS from the `rstan` package, with 10,000 iterations of which 5,000 were discarded as burn-in. Default priors were used for all parameters with the exception of the mean parameters of the growth factors (i.e., intercept, linear slope, quadratic slope). For this type of parameter, the default prior in `blavaan` is *N*(0, standard deviation = 10). Although this may

represent a diffuse prior in some circumstances, it is not appropriate for this example, as the math assessment scores ranged from 11 to 171. Thus, we specified a prior with the following hyperparameters: $N(0, \text{standard deviation} = 100)$. The remaining default priors used were: Gamma(1, 0.5) placed on the residual standard deviations and factor standard deviations, and an LKJ(1) (Lewandowski, Kurowicka and Joe) distribution assigned to the lower Cholesky factor of the correlation matrix of the latent factors for the latent factor correlations/covariances.[1] Information about the growth factor means was extracted from the results obtained here and used to help construct informative prior distributions for a subsequent analysis on a second sample of data consisting of 600 children (Dataset 2).

Observed trajectories from Dataset 1 were plotted to obtain a visual depiction of growth over time and determine an initial model form to estimate. The math scores appeared to follow a quadratic (positive) slope. As a result, a quadratic LGCM was estimated for this example. The model can be found in Figure 7.

Model Results for the Initial Analysis (Diffuse Priors). Convergence for the initial quadratic LGCM with Dataset 1 was monitored using the $\widehat{R}$ statistic. The lowest $\widehat{R}$ value obtained was 1.00, which indicated that convergence was obtained for this analysis. The lowest ESS was 1946.47 (for the residual variance at Time 4). According to the results obtained, the posterior chains were stable. Posteriors for the intercept, slope, and quadratic term can be found in Figure 8. The full set of results for all parameters can be found in the online supplementary material. Regarding fit, the posterior predictive p-value indicated model misfit with a p-value of 0.024.

The posterior mean estimate for the intercept mean was 26.226 (posterior standard deviation = 0.348). We used this information to form a prior that was $N(26.226, 0.348)$, in line with the prior parameterization in `rstan` for the normal distribution, which is in terms of a mean and standard deviation hyperparameter. This prior can be used as an informative prior in a subsequent analysis with a new sample. Similarly, the posterior mean estimate for the linear slope mean was 3.889 (posterior standard deviation = 0.338). We used this information to form a prior that was $N(3.889, 0.338)$. Finally, the posterior mean estimate for the quadratic slope was 6.789 (posterior

standard deviation = 0.140). We used this information to form a prior that was $N(6.789, 0.140)$.

The priors that were based on the posterior estimates for Dataset 1 were then used as informative priors for a second dataset (Dataset 2). The analysis conducted for Dataset 2 illustrates how to implement informative prior settings on growth factor means. In this case, the informative priors were constructed based on a data-splitting technique out of convenience, but priors can come from a variety of places.

Model Results for Subsequent Analysis (Informative Priors). A visual inspection of the observed trajectories for Dataset 2 indicated that a quadratic LGCM was a good choice for modeling the data patterns. The model priors for the intercept, slope, and quadratic means were derived from Dataset 1 as described above. The remaining priors followed software default settings. The lowest $\widehat{R}$ value obtained was 1.00, which indicated that convergence was obtained for this analysis. The lowest ESS was 1046.62 (for the residual variance at Time 1). According to the results obtained, the posterior chains were stable. In addition, a visual inspection of all plots indicated that convergence was obtained and that the chains appeared stable.

Regarding fit, the posterior predictive p-value indicated there was improved fit for this analysis as compared to the diffuse prior settings implemented for Dataset 1. The posterior predictive p-value was 0.294.

Comparing Results from Diffuse Priors to Informative Priors (Dataset 2). Next, we can compare findings for Dataset 2 when informative priors were implemented versus diffuse priors. In Figure 8, the left column shows the prior, likelihood, and posterior distribution for the three main parameters when diffuse priors ($N(0$, standard deviation = $100)$) were specified for these parameters. The prior distribution density is close to the x-axis across the range shown, indicating that it is flat and wide, allowing for a wide range of values to be plausible. With this specification, the likelihood and posterior distributions overlap almost perfectly.

The second column of Figure 8 shows the results when informative priors (based on a previous sample's posterior estimates) were specified. The prior distribution can be clearly seen in these plots (dotted lines). For parameters where the prior and likelihood are almost completely

aligned (e.g., the intercept mean), the posterior is centered above their center, looks relatively normally distributed, and is narrower than the likelihood distribution. The slope and quadratic means illustrate that the prior and likelihood did not entirely align. The resulting posteriors represent a compromise between the prior and the likelihood for each of these parameters. Even though the prior and likelihood do not completely agree, their combined information still results in a narrower posterior distribution compared to the model with diffuse priors.

The impact of diffuse versus informative priors can be further highlighted for Dataset 2. Figure 9 presents the 95% HDI histograms for the intercept, slope, and quadratic means. For each parameter, the 95% HDI is narrower when informative priors were used. Although the posterior median (an example of a posterior point estimate) does not change meaningfully between the two prior specifications, the advantage of using informative priors can be seen in this narrowing of the posterior. Conversely, if the (informative) prior and likelihood really diverge, then the posterior HDI will be wider compared to using diffuse priors, reflecting the conflict between the two sources of information.

This LGCM example has been further extended in the accompanying R Markdown file on the companion website to highlight the process for conducting a sensitivity analysis. In the example code, we manipulate the prior distribution settings (via the hyperparameters) for the intercept mean. This manipulation of the prior showcases how prior impact, or robustness of results, can be examined via a prior sensitivity analysis. The R Markdown file walks through modifications of the code and describes the influence of different prior settings in this modeling context.

Example 4: Implementation of Bayesian Model Averaging using PISA 2009 Reading Data

For this example, we use data from PISA 2009 (Organization for Economic Cooperation and Development (OECD), 2010), consisting of 2,489 PISA-eligible students in the United States. The data included several background and reading strategy variables, which we used to predict reading proficiency (see Figure 10). The background variables act as exogenous variables in the initial SEM and are Gender (male = 0, female = 1), immigrant status (Immigr), and economic, social, and cultural status of the student (ESCS). The reading strategy variables act as a

first level of mediating endogenous variables and include memorization strategies (MEMO), elaboration strategies (ELAB), and control strategies (CSTRAT). We used a variable measuring joy in reading (JoyR) as a second level mediating variable. Finally, we used the first plausible value of reading assessment (Reading) as the outcome variable. The model is an adaptation of the example presented in Kaplan and Lee (2015).

The BMA algorithm considered an initial set of 100 possible models (C = 100) and selected four models. An overview of the included regression paths in each model is shown in Table 3. Several regression paths that were not included in the original model (Figure 10) were included in multiple proposed models: e.g., direct effects of reading on gender and reading on ESCS, and MEMO on Immigr. Two regression paths were never included in a model: ELAB on Gender (which was included in the intended model in Figure 10) and Reading on Immigr. Model 1 included all remaining regression paths. The other three models each excluded one regression path: CSTR on Immigr (Model 2), ELAB on Immigr (Model 3), or MEMO on Immigr (Model 4). The best model (Model 1) accounted for only 69% of the total posterior model probability (PMP), indicating that there is a fair amount of uncertainty in model selection.

Table 4 presents the combined results of the BMA analysis. As most parameters were included across all models, they could be considered indispensable for the model. Only the three regression paths listed above have a lower probability of being included in the model, although their probability is still high. Overall, BMA can be useful for examining the inclusion of model parameters (e.g., paths) because the procedure provides an inclusion probability for model parameters. This result indicates which parameters are certain to be included and which are relatively dispensable to the prediction model.

Example 5: Model Selection with Leave-One-Out Cross Validation (LOO-CV) for a Structural Equation Model

For this example, we use data from the classic SEM example by Bollen (1989). The data consist 75 observations with 11 variables. The variables measure different aspects of political democracy (in 1960 and 1965) and industrialization (in 1960) in developing countries. The hy-

pothesized model is shown with LISREL notation in Figure 11. In this model, industrialization in 1960 functions as an exogenous latent factor (ξ_1), that is used to predict the endogenous latent factors of political democracy in 1960 (η_1) and in 1965 (η_2). The model also includes residual covariances between equivalent items measured in 1960 and 1965, as well as residual covariances between Y_2 and Y_4 and Y_6 and Y_8.

Priors and Estimation Settings. Two chains were generated using NUTS from the `rstan` package, with 3,000 iterations of which 1,500 were discarded as burn-in. Default priors were used for all parameters. Convergence was monitored using the $\widehat{R}$ convergence criterion (Vehtari et al., 2019). The highest $\widehat{R}$ obtained in the analysis was 1.00, which indicates chain convergence for all parameters. The lowest effective sample size (ESS; for the intercept of Y_8) was 1024.01, which was deemed adequate for this model (Kruschke, 2015; Zitzmann & Hecht, 2019). A full assessment of convergence (including trace plots, density plots, and autocorrelation plots) can be found in the online supplementary material.

Model Interpretation. Table 5 presents estimates for the model parameters. This table includes the following information: posterior median, mean, and standard deviation; upper and lower 95% credible interval values (equal tails), upper and lower 95% HDI (unequal tails), and ESS for each model parameter. Taking a closer look at Table 5, the posterior mean and median, and similarly the 95% credible interval and HDI, were nearly identical for all parameters. This implies that the posterior distributions were all relatively symmetric. The exceptions to this pattern are the residual variances of some of the observed variables. This makes sense as variance parameter distributions are often skewed. Focusing on the specific parameters, we can see that the 95% HDI of all factor loadings excluded zero, indicating that the factor loadings were of a meaningful magnitude. Furthermore, the 95% HDI of all regression paths also excluded zero and were positive. Thus, industrialization in 1960 had a meaningful, positive effect on democracy in 1960 and 1965. Similarly, democracy in 1960 had a meaningful, positive effect on democracy in 1965. Finally, the 95% HDI of three residual covariances (Y_2 with Y_4, Y_3 with Y_7, and Y_4 with Y_8) included zero, indicating that any residual relationship between these variables may not be mean-

ingful.

Model Fit. The quality of fit in this SEM was examined using the posterior predictive checking procedure based on the chi-square discrepancy function. The procedure produced a posterior predictive p-value of 0.537, indicating good fit. Figure 12 illustrates a visual depiction of the predictive accuracy of this model. We can see that the points are plotted in a cloud around the 45-degree line. However, given that some of the 95 % HDI of the residual covariances included zero, we might want to explore a model that excludes residual covariances to see if a simpler model may be equally adequate.

Model Comparison Through LOO-CV. One approach to model comparison is to examine leave-one-out cross-validation (LOO-CV) through the LOO information criterion (LOOIC) (Vehtari, Gelman, & Gabry, 2017). As this index is based on cross-validation, the LOOIC provides an assessment of the predictive accuracy of the model. Contrary to other information criteria such as the BIC and the DIC, the LOOIC is fully Bayesian in that it is estimated after each posterior sample is drawn and has its own posterior distribution. This distribution can be used to estimate a standard error for the difference between LOOIC for competing models. For this example, we estimated a second SEM that excluded all residual covariances between observed variables (more information about this second model can be found in the online supplementary material). The LOOIC for this simpler model was 3200.99. In contrast, the LOOIC for the original model was 3181.79.

To compare across models, we need to use the LOO estimate of the expected log predictive density (ELPD) instead of the LOOIC. These two quantities are related as follows: $LOOIC = -2\widehat{\text{elpd}}_{loo}$; multiplying the $\widehat{\text{elpd}}_{loo}$ by -2 places the estimate on the same scale as the DIC or BIC (hence: LOOIC). Comparing these two models, we find that the original model performs better in terms of predictive accuracy than the simpler model; the difference in the $\widehat{\text{elpd}}_{loo}$ is $-9.60(SE = 6.60)$. The absolute difference is larger than 4, which indicates that there is a meaningful difference in predictive accuracy of the two models. As the sample size of the data used for this example is quite small, the SE may not be accurate, and we may not be able to assume that

the difference in $\widehat{\text{elpd}}_{loo}$ approaches a normal distribution. With a larger sample size, we could have looked at $-9.60/6.60 = -1.45$ and interpreted this as an approximation of a z-score. In general, if the ratio of the difference in $\widehat{\text{elpd}}_{loo}$ to SE is larger, it implies that there is stronger evidence in favor of one model over a second. In this case, we can conclude that including the residual covariances improves the predictive accuracy of the model (Vehtari, 2020).

Example 6: Penalized SEM with Shrinkage Priors

For this example, we used a classic SEM example with data from Holzinger and Swineford (1939). The full sample consists of 301 seventh- and eight-grade children (from two schools) who completed mental ability tests. To illustrate the penalization methods, the full sample was divided into seven smaller samples of 43 participants each. Before estimating any model that uses a ridge or lasso penalty, it is important to standardize all variables (both predictors and outcomes). We will estimate a multiple indicators/multiple causes (MIMIC) model for each sample, where the three mental abilities (visual, textual, speed) are measured with nine items. Sex, school, and grade are used as dichotomous predictors (or causes) of the mental abilities (see Figure 13).

Initial Analysis. As an initial step, we estimated the MIMIC model without using penalization methods, using `blavaan`. To estimate the model, we implemented default priors and `rjags` in the background. For each of the seven samples of students, we estimated three MCMC chains with 10,000 burn-in iterations and 10,000 posterior samples.

Each plot in Figure 14 shows the posterior distributions of the nine regression slope estimates for a specific sample. The pattern of results is different for each sample. Different predictors seem relevant in different samples. For example, textual $\sim$ school appears to be a meaningful predictor in Sample 1, 2, 4, 6, and 7, but not in 3 and 5. Similarly, visual $\sim$ grade appears to be a meaningful predictor in Sample 1, 4, 5, 6, but not in 2, 3, and 7.

Ridge Penalization. Next, we estimated the models using the ridge penalty. To do so, we needed to specify normal priors on the regression slope parameters, with hyperparameters 0 and τ_β. Then, we needed to create a separate character object, in which we transformed τ_β from a precision (which is what `rjags` uses to specify a normal distribution) to a standard deviation

(which is what we want to specify the hyperprior for). Then, we specified the hyperprior for σ_β as a half Cauchy prior.

To estimate the model, we used default priors for all other parameters and `rjags` in the background. To include the ridge penalty, we use the `mcmcextra` argument to include the extra syntax and ensure that `blavaan` saves the posterior samples for the σ_β hyperparameter. For each sample, we estimated three MCMC chains with 10,000 burnin iterations and 10,000 posterior samples.

Each plot in Figure 15 shows the posterior distributions of the nine regression slope estimates for a specific sample. Some regression slope estimates have shrunk towards zero across all samples (e.g., speed $\sim$ school). However, some estimates were more likely to "escape" the penalty prior in some samples (e.g., speed $\sim$ grade in Sample 2 and 5). Figure 16 illustrates that the penalty parameter differed across samples, resulting in more or less shrinkage towards zero for the regression slope estimates.

Lasso Penalization. Next, we estimated the models using the lasso penalty. To do so, we needed to specify double exponential priors on the regression slope parameters, with hyperparameters 0 and τ_β. Then, we created a separate character object, in which we transformed τ_β from a precision (which is what `rjags` uses to specify a double exponential distribution) to a standard deviation (which is what the hyperprior is specified on). Just as with the ridge penalty, we specified the hyperprior for σ_β as a half Cauchy prior.

To estimate the model, default priors were implemented for all other parameters and `rjags` was used in the background. To include the lasso penalty, we used the `mcmcextra` argument to include the extra syntax and ensure that `blavaan` saves the posterior samples for the σ_β hyperparameter. For each sample, we estimated three MCMC chains with 10,000 burnin iterations and 10,000 posterior samples.

Each plot in Figure 17 shows the posterior distributions of the nine regression slope estimates for a specific sample. We can now see that the estimates for a number of regression slopes have shrunk towards zero across all samples. For Sample 2, 3, and 6 all regression estimates have

been shrunk to zero. In contrast, some larger effects in Sample 4 were not affected by the lasso penalty. Compared to the ridge penalty, the lasso penalty was more likely to shrink posterior distributions towards zero. Figure 18 illustrates that the penalty parameter differed across samples, resulting in more or less shrinkage towards zero for the regression slope estimates.

Comparing Ridge and Lasso Penalties. To further illustrate the similarity between ridge and lasso, we have taken the posterior mode of the penalty terms for Sample 1, and created this simple plot (Figure 19). This plot shows the prior distributions used for the regression slope parameters with the ridge (dashed line) and lasso (solid line) penalty methods. The lasso approach results in a distribution with heavier tails. The consequence of this is that when you use the lasso method, it results in more shrinkage for small estimates, but less shrinkage for large estimates (compared to the ridge method). The heavier tails allow larger estimates to "escape" the penalty.

To make it easier to compare across methods, we focus on one sample: Sample 1. The three columns of plots in Figure 20 show the posterior distribution across the regression slope parameters for the regular SEM (left column), the SEM using the ridge penalty (middle column), and the SEM using the lasso penalty (right column). To make comparisons more straightforward, the *x*-axis limits are identical across plots, and the shaded region within each density plot represents the 90% highest posterior density (HPD) interval.

When looking at the regular SEM posterior densities, we can see that the posterior distributions often cover 0, meaning that 0 is a plausible value. We can see that the ridge penalty is unlikely to shrink estimates to zero (as compared to the lasso penalty). Although estimates are closer to zero than when examining the regular SEM, most parameter estimates still place their mode or highest posterior probability at a value that diverges from zero.

We can also see that, for those parameters, the lasso penalty does a good job at shrinking the estimates towards 0 even further. As was discussed above, using the lasso penalty resulted in posterior distributions that were narrower and more peaked at 0 compared to the ridge penalty. This example illustrates the difference between the lasso and ridge penalty. Although the exam-

ple is beneficial in highlighting these methods, it should also be noted that the model in this example likely does not include enough predictors to truly showcase the benefits of penalization methods.

References

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Kaplan, D., & Lee, C. (2015). Bayesian model averaging over directed acyclic graphs with implications for the predictive performance of structural equation models. *Structural Equation Modeling*. doi:10.1080/10705511.2015.1092088

Kruschke, J. K. (2015). *Doing Bayesian analysis: A tutorial with R, Jags, and STAN*. San Diego, CA: Elsevier Inc.

Merkle, E. C., & Rosseel, Y. (2018). blavaan: Bayesian structural equation models via parameter expansion. *Journal of Statistical Software*, *85*, 1–30.

National Center for Education Statistics [NCES]. (2001). *Early childhood longitudinal study: Kindergarten class of 1998-99: Base year public-use datafiles user's manual (NCES 2001-029).* Washington, DC: U.S. Government Printing Office.

Organization for Economic Cooperation and Development (OECD). (2010). *PISA 2010 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2019). Rank-normalization, folding, and localization: An improved ”0562*R* for assessing convergence of MCMC. Retrieved from https://arxiv.org/pdf/1903.08008.pdf

Vehtari, A. (2020). More limitations of cross-validation and actionable recommendations. Retrieved from https://statmodeling.stat.columbia.edu/2020/08/27/more-limitations-of-cross-validation-and-actionable-recommendations/

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. doi:10.1007/s11222-016-9696-4

Zitzmann, S., & Hecht, M. (2019). Going beyond convergence in Bayesian estimation: Why precision matters too and how to assess it. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 646–661.

Table 1
Example 1: CFA Results with Diffuse Prior Settings

	Posterior Median	Posterior Mean	Posterior Standard Deviation	95% CI (Equal Tails) Lower	95% CI (Equal Tails) Upper	95% HDI (Unequal Tails) Lower	95% HDI (Unequal Tails) Upper	Effective Sample Size
Factor 1 (F1) Loadings								
Item 2: easy	0.99	0.99	0.01	0.98	1.01	0.98	1.01	41558.71
Item 3: interesting	0.97	0.97	0.01	0.94	0.99	0.94	0.99	68430.96
Factor 2 (F2) Loadings								
Item 5: difficultwords	1.23	1.23	0.03	1.17	1.30	1.16	1.30	23151.39
Item 6: otherthings	0.97	0.97	0.03	0.92	1.03	0.92	1.03	27670.30
Factor (Co)Variances								
F1	0.21	0.21	0.02	0.17	0.26	0.17	0.26	40817.68
F2	0.35	0.35	0.04	0.27	0.43	0.27	0.43	31121.31
F1 WITH F2	-0.19	-0.19	0.02	-0.24	-0.15	-0.24	-0.15	58908.67
Factor Means								
F1	3.50	3.50	0.03	3.43	3.56	3.43	3.56	26620.93
F2	1.78	1.78	0.05	1.69	1.87	1.69	1.87	18626.07
Residual Variances								
Item 1: dowell	0.21	0.21	0.02	0.17	0.25	0.17	0.25	43142.91
Item 2: easy	0.26	0.26	0.02	0.21	0.31	0.21	0.31	47895.30
Item 3: interesting	0.68	0.68	0.05	0.59	0.79	0.58	0.79	91712.76
Item 4: harder	0.52	0.53	0.05	0.44	0.63	0.44	0.62	56536.08
Item 5: difficultwords	0.69	0.69	0.06	0.58	0.83	0.57	0.82	54313.34
Item 6: otherthings	0.57	0.57	0.05	0.48	0.68	0.48	0.67	60257.71

Note. The factor loading for Item 1 on Factor 1 was fixed to 1.0 to set the scale of the latent factor, so an estimate is not provided here for the loading. The same holds true for Item 4 on Factor 2. Unstandardized loadings are presented here.

Table 2
Example 2: Comparing No Cross-Loadings to Four Different Approximate-Zero Priors for Cross-Loadings

Prior Setting	PPP	BIC	DIC	Low ESS	High $\widehat{R}$
Exact Zero	0.000	5709.593	5648.454	18626.073	1.000
$N(0, 0.001)$	0.025	5717.467	5623.684	2952.138	1.000
$N(0, 0.005)$	0.193	5703.334	5612.273	869.626	1.003
$N(0, 0.01)$	0.116	8327.925	3054.472	280.300	28.437
$N(0, 0.02)$	0.126	8331.177	3009.757	126.259	18.512

Note. PPP = Posterior predictive p-value. BIC = Bayesian information criterion. DIC = Deviance information criterion. Low ESS = The lowest effective sample size obtained in the model (i.e., one parameter had the lowest ESS). High $\widehat{R}$ = The highest $\widehat{R}$ value produced for the model (i.e., one parameter had the highest $\widehat{R}$ value). Exact Zero = The model presented in Example 1, where cross-loadings were fixed to zero in the model. $N(0, \sigma^2)$ = Prior settings placed on the cross-loadings. Rather than fixing the loadings to zero, they received approximate-zero priors centered at zero with a small variance hyperparameter.

Table 3
Example 4: Overview of Included Parameters of Selected Models for the PISA 2009 Example

Regression Effects	Model 1	Model 2	Model 3	Model 4
MEMO∼ESCS	x	x	x	x
MEMO∼Gender	x	x	x	x
MEMO∼Immigr	x	x	x	
ELAB∼ESCS	x	x	x	x
ELAB∼Gender				
ELAB∼Immigr	x	x		x
CSTR∼ESCS	x	x	x	x
CSTR∼Gender	x	x	x	x
CSTR∼Immigr	x		x	x
Reading∼ESCS	x	x	x	x
Reading∼Gender	x	x	x	x
Reading∼Immigr				
Reading∼MEMO	x	x	x	x
Reading∼ELAB	x	x	x	x
Reading∼CSTR	x	x	x	x
RCOM∼Reading	x	x	x	x
Posterior Model Prob.	0.69	0.12	0.10	0.09
BIC	35884.11	35887.60	35887.95	35888.23

Note. "x" denotes the parameters included for each model.

Table 4
Example 4: Results of Bayesian Model Averaging for the PISA 2009 Example

Parameter	Posterior Mean	Posterior SD	Inclusion Probability
Regression Effects			
MEMO∼ESCS	0.09	0.03	1.00
MEMO∼Gender	0.18	0.04	1.00
MEMO∼Immigr	-0.19	0.08	0.91
ELAB∼ESCS	0.16	0.03	1.00
ELAB∼Gender	0.00	0.00	0.00
ELAB∼Immigr	-0.18	0.08	0.90
CSTR∼ESCS	0.29	0.03	1.00
CSTR∼Gender	0.25	0.04	1.00
CSTR∼Immigr	-0.17	0.08	0.88
Reading∼ESCS	0.13	0.02	1.00
Reading∼Gender	0.64	0.04	1.00
Reading∼Immigr	0.00	0.00	0.00
Reading∼MEMO	-0.11	0.02	1.00
Reading∼ELAB	0.08	0.02	1.00
Reading∼CSTR	0.28	0.02	1.00
RCOM∼Reading	0.34	0.02	1.00
Intercepts			
MEMO	0.01	0.07	1.00
ELAB	0.02	0.07	1.00
CSTR	-0.09	0.07	1.00
Reading	-0.36	0.03	1.00
RCOM	5.00	0.02	1.00
Residual Variances			
MEMO	1.19	0.03	1.00
ELAB	1.23	0.03	1.00
CSTR	1.18	0.03	1.00
Reading	0.89	0.03	1.00
RCOM	0.76	0.02	1.00

Note. SD = Standard deviation.

Table 5
Example 5: SEM Results

	Post. Median	Post. Mean	Post. *SD*	95% CI (Equal Tails) Lower	95% CI (Equal Tails) Upper	95% HDI (Unequal Tails) Lower	95% HDI (Unequal Tails) Upper	Effective Sample Size
Ind60 Loadings								
X2	2.23	2.24	0.16	1.95	2.58	1.93	2.56	1551.67
X3	1.84	1.85	0.17	1.54	2.17	1.55	2.18	2154.26
Dem60 Loadings								
Y2	1.32	1.33	0.22	0.94	1.80	0.91	1.76	1464.40
Y3	1.10	1.10	0.17	0.80	1.46	0.79	1.44	2144.34
Y4	1.33	1.34	0.19	1.03	1.77	0.99	1.71	1283.34
Dem65 Loadings								
Y6	1.23	1.24	0.20	0.90	1.66	0.87	1.62	1626.39
Y7	1.32	1.33	0.18	1.01	1.74	0.99	1.70	1725.87
Y8	1.31	1.33	0.19	1.01	1.75	1.01	1.75	1581.02
Regression Paths								
Dem60∼Ind60	1.44	1.44	0.39	0.70	2.26	0.71	2.26	3373.78
Dem65∼Ind60	0.54	0.54	0.24	0.06	1.01	0.09	1.03	2883.13
Dem65∼Dem60	0.85	0.85	0.11	0.66	1.08	0.64	1.05	1375.82
Residual Variances								
X1	0.09	0.09	0.02	0.05	0.14	0.05	0.13	1940.57
X2	0.11	0.11	0.08	0.00	0.28	0.00	0.25	1529.50
X3	0.51	0.52	0.10	0.34	0.76	0.33	0.73	2945.24
Y1	2.12	2.17	0.53	1.26	3.35	1.16	3.22	1752.70
Y2	7.64	7.81	1.50	5.21	11.04	5.03	10.83	2771.57
Y3	5.35	5.48	1.08	3.74	7.95	3.59	7.71	2760.44
Y4	3.22	3.27	0.86	1.74	5.08	1.71	5.03	2212.32
Y5	2.55	2.61	0.54	1.69	3.85	1.61	3.66	2043.79
Y6	5.15	5.27	0.99	3.64	7.48	3.48	7.25	2638.91
Y7	3.65	3.72	0.80	2.40	5.52	2.28	5.32	2974.38
Y8	3.35	3.43	0.78	2.09	5.16	2.06	5.09	1811.66
Ind60	0.45	0.46	0.10	0.30	0.67	0.27	0.63	2331.28
Dem60	3.75	3.84	0.99	2.16	6.01	1.95	5.66	1806.65
Dem65	0.17	0.22	0.20	0.00	0.72	0.00	0.61	2259.09
Residual Covariances								
Y1 WITH Y5	0.71	0.74	0.40	0.04	1.59	0.03	1.55	1763.63
Y2 WITH Y4	1.21	1.28	0.76	-0.10	2.91	-0.04	2.94	2356.38
Y2 WITH Y6	2.13	2.19	0.78	0.82	3.88	0.66	3.65	2620.74
Y3 WITH Y7	0.88	0.91	0.68	-0.35	2.35	-0.46	2.22	2327.59
Y4 WITH Y8	0.30	0.33	0.50	-0.58	1.40	-0.68	1.28	2524.84
Y6 WITH Y8	1.32	1.35	0.61	0.25	2.63	0.25	2.62	1960.07

Note. The factor loading for X1 on Ind60 was fixed to 1.0 to set the scale of the latent factor, so an estimate is not provided here for the loading. The same holds true for Y1 on Dem60 and Y5 on Dem65. Unstandardized loadings are presented here. Post. = posterior; *SD* = standard deviation.

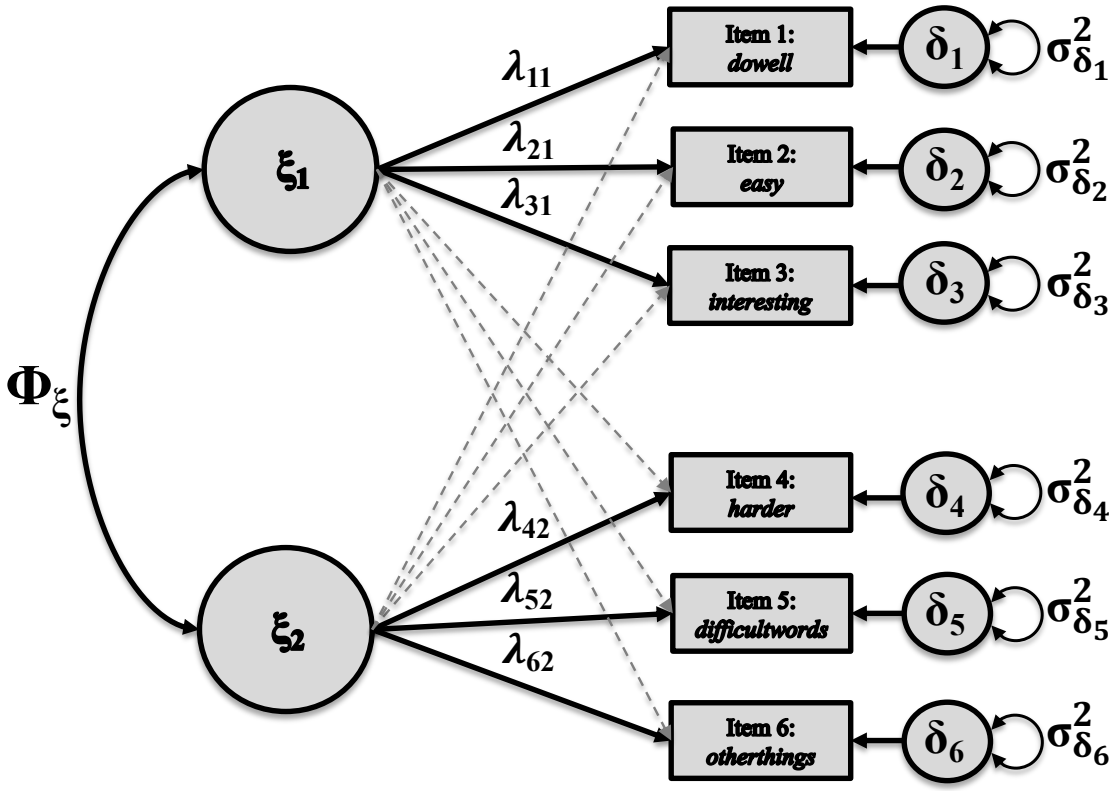


Figure 1: Confirmatory factor analysis model. Solid lines represent primary loadings (estimated in Examples 1 and 2). Dashed lines represent cross-loadings (fixed to zero in Example 1, and estimated with approximate-zero priors in Example 2). Notation follows basic LISREL notation as follows: ξ = latent factors, Φ = latent factor covariance matrix, λ = factor loading in the Λ matrix, δ = measurement errors, and σ^2_δ = error variances. The factor loading for the first item loading on each factor was fixed to 1.0 to set the scale of the latent factor (Item 1 on Factor 1, and Item 4 on Factor 2).

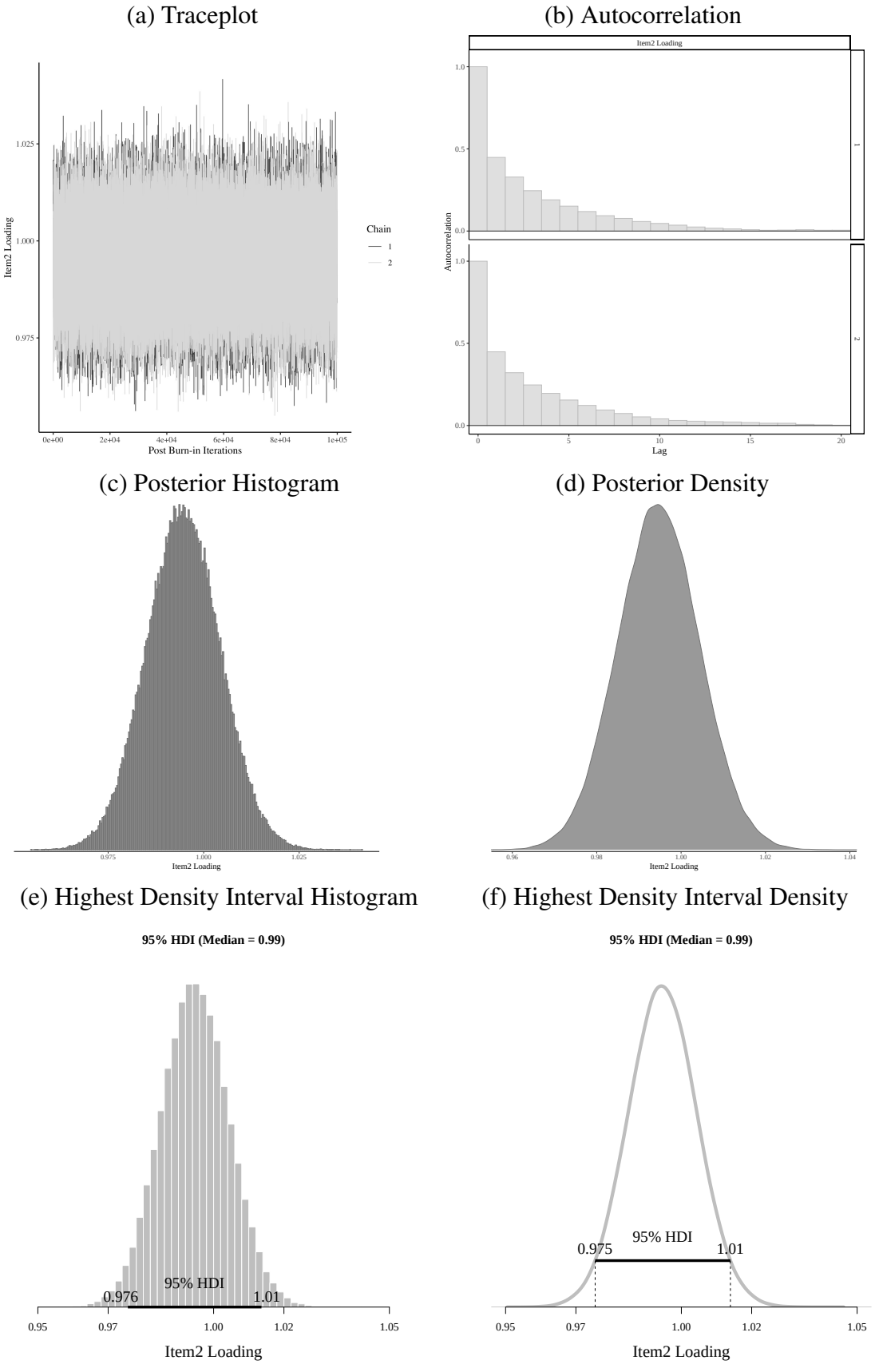
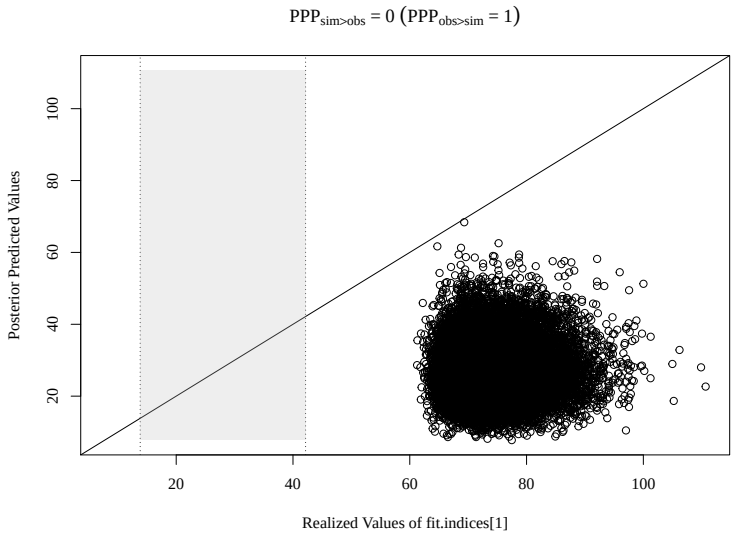


Figure 2: Plots for Item 2 (*easy*) Factor Loading for Factor 1 (“PosRead”).

No Cross-Loadings:



Approximate-Zero Cross-Loadings of $N(0,0.005)$:

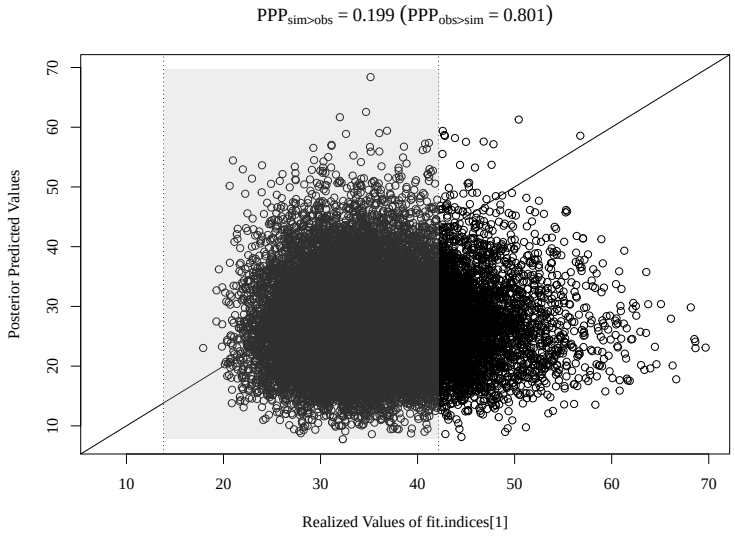


Figure 3: Posterior predictive *p*-value plots. These plots are for the model in Example 1 without cross loadings (top), and the model in Example 2 with approximate-zero priors on cross-loadings of $N(0,0.005)$ (bottom).

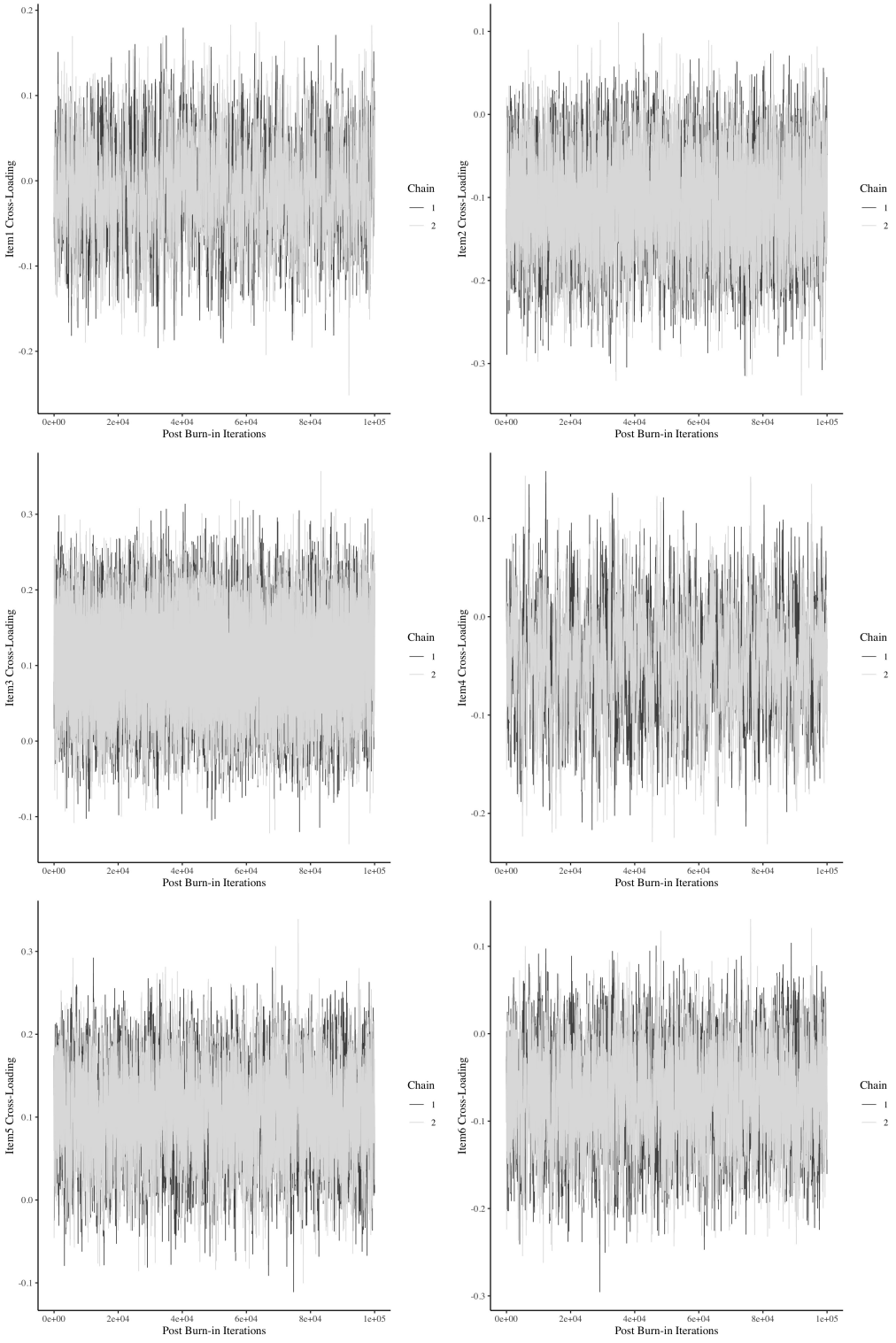


Figure 4: Trace-plots for cross-loadings in Example 2 with approximate-zero priors of $N(0,0.005)$.

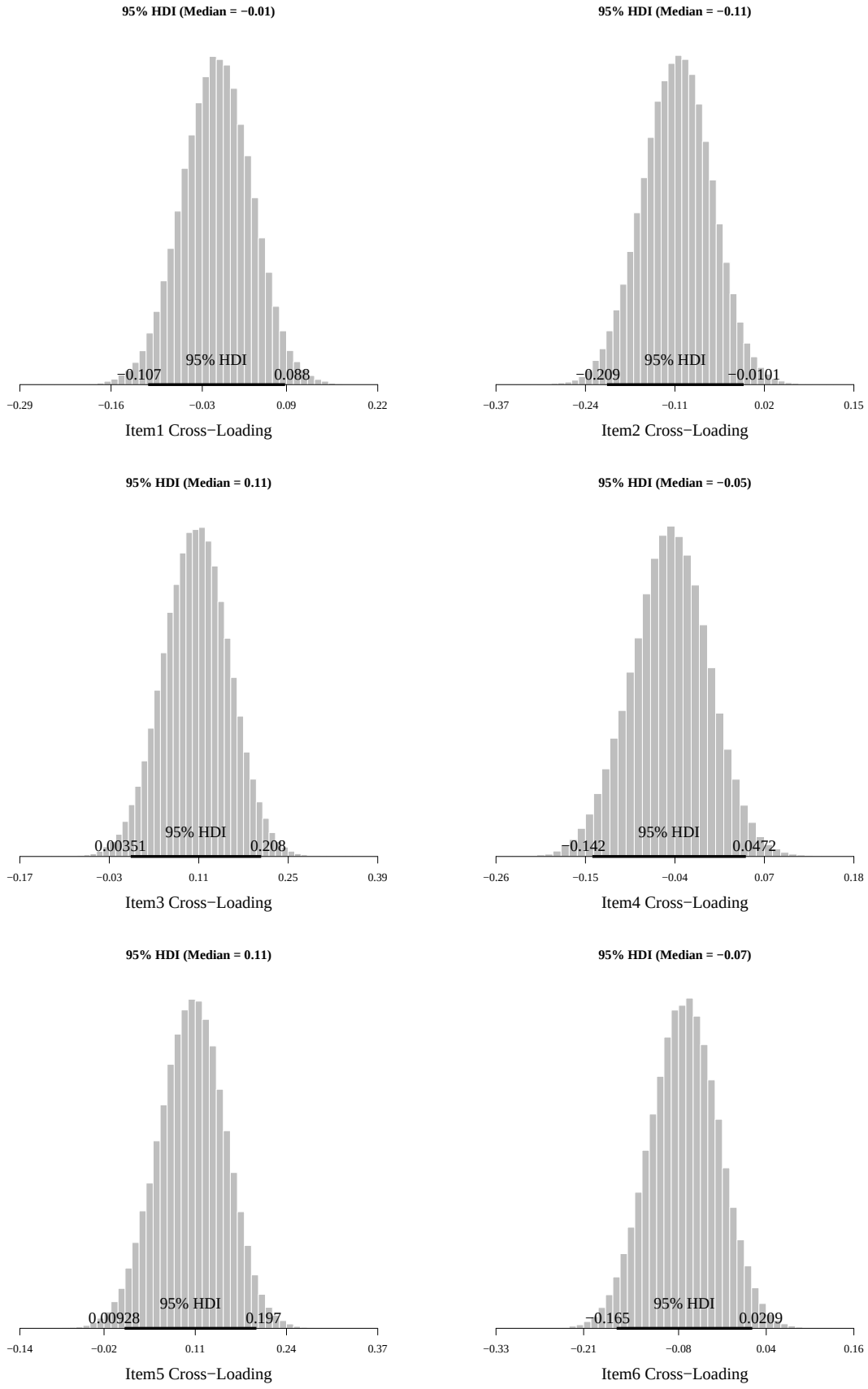


Figure 5: HDI histograms for cross-loadings in Example 2 with approximate-zero priors of $N(0,0.005)$.

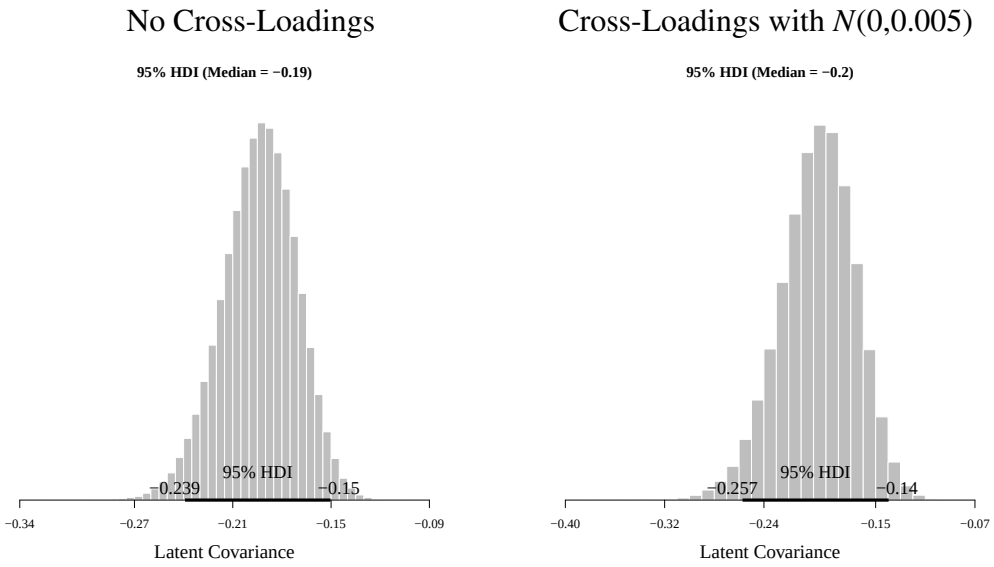


Figure 6: Latent factor covariance HDI histograms. The model in Example 1 without cross-loadings (left plot), and the model in Example 2 with approximate-zero priors of $N(0,0.005)$ (right plot).

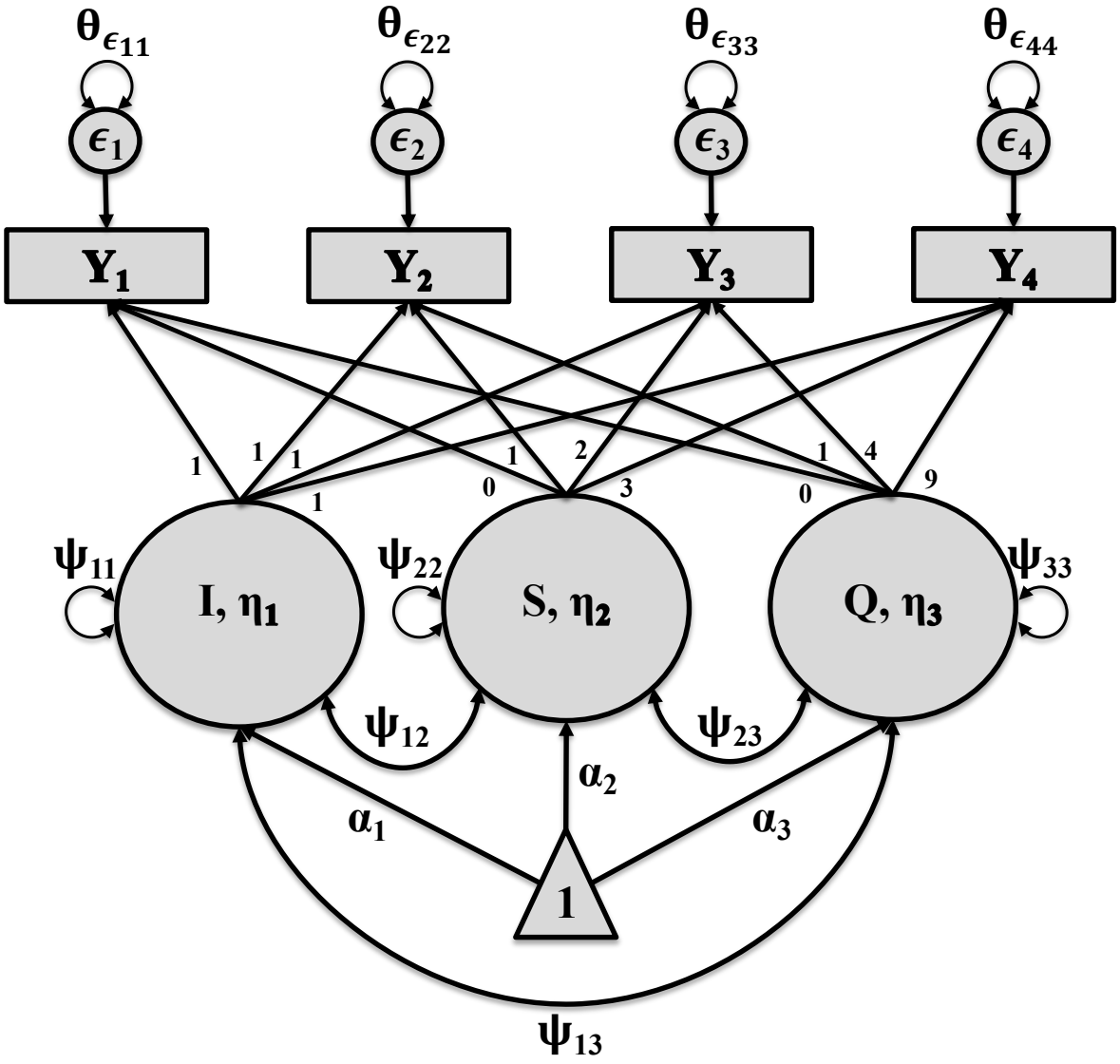


Figure 7: Latent growth curve model. Notation follows basic LISREL notation as follows: the triangle is notation that represents a constant or intercept, η = latent growth factors (intercept, linear slope, and quadratic term), Ψ = latent factor covariance matrix, α = growth factor means, ε = errors tied to observed repeated measure outcomes (Y), and θ = error variances. The factor loadings are fixed values to represent the specific growth model being estimated (in the case of this figure, a quadratic model with equal time-spacing and an intercept defined at the first time-point).

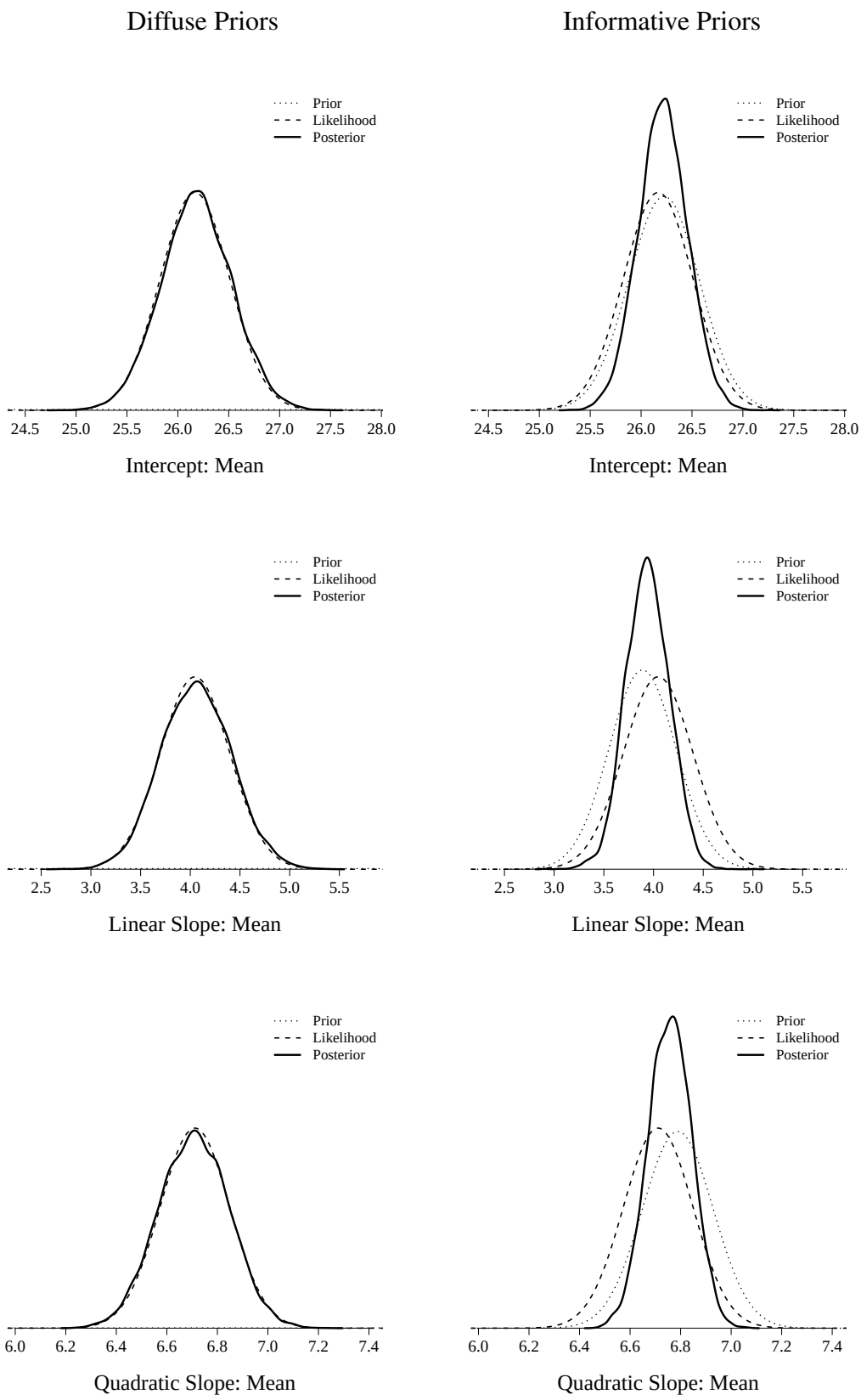


Figure 8: Prior, likelihood, and posterior for Dataset 2. Diffuse priors are in the left-hand column, and informative priors (based on the data-splitting technique) are in the right-hand column.

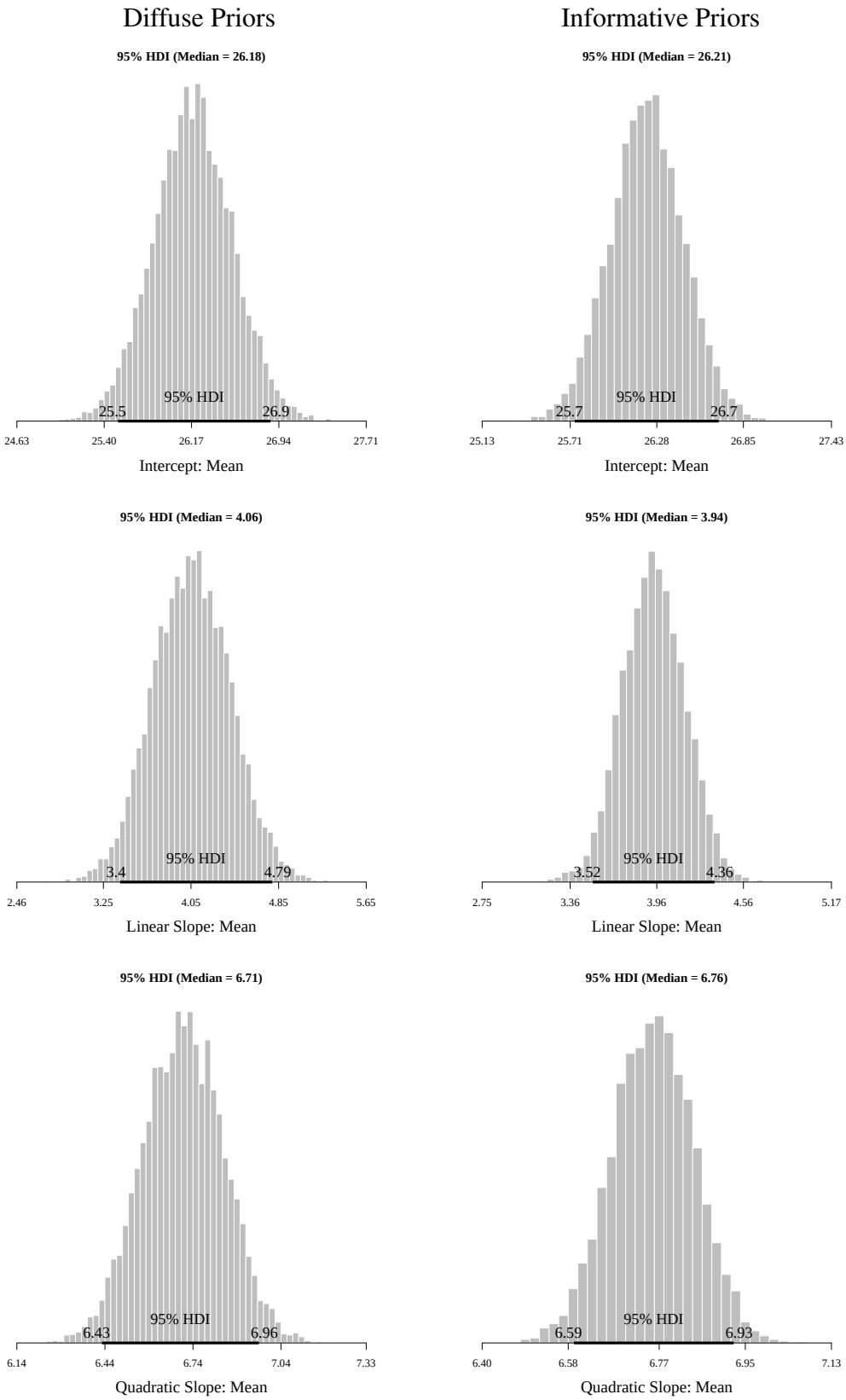


Figure 9: HDI histograms for Dataset 2. Results using diffuse priors are in the left-hand column, and results using informative priors (based on the data-splitting technique) are in the right-hand column.

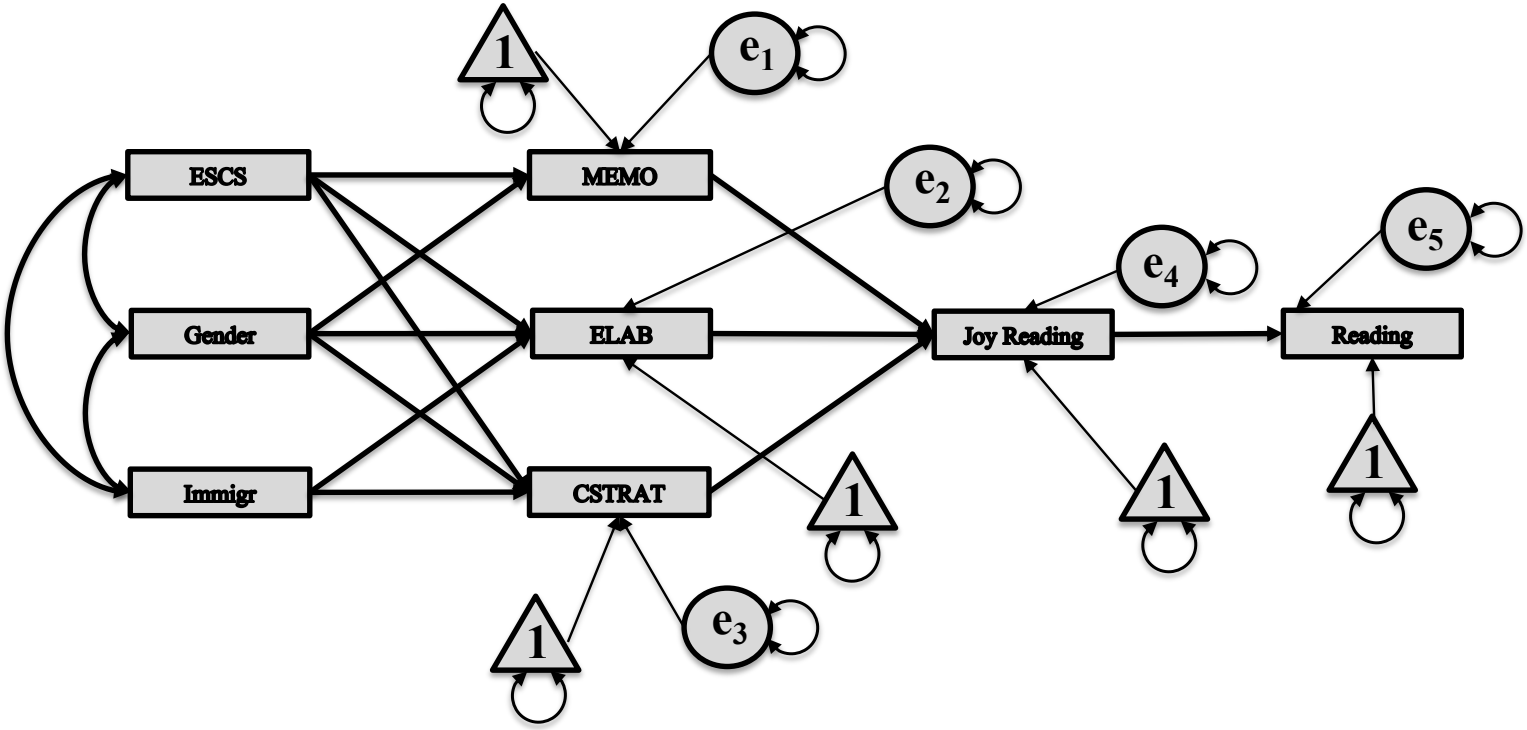


Figure 10: Prediction Model Estimated using Bayesian Model Averaging. ESCS = economic, social, and cultural status of the student; Gender: male = 0, female = 1; Immigr = immigrant status; MEMO = memorization strategies; ELAB = elaboration strategies; CSTRAT = control strategies; Joy Reading = joy in reading; Reading = reading assessment.

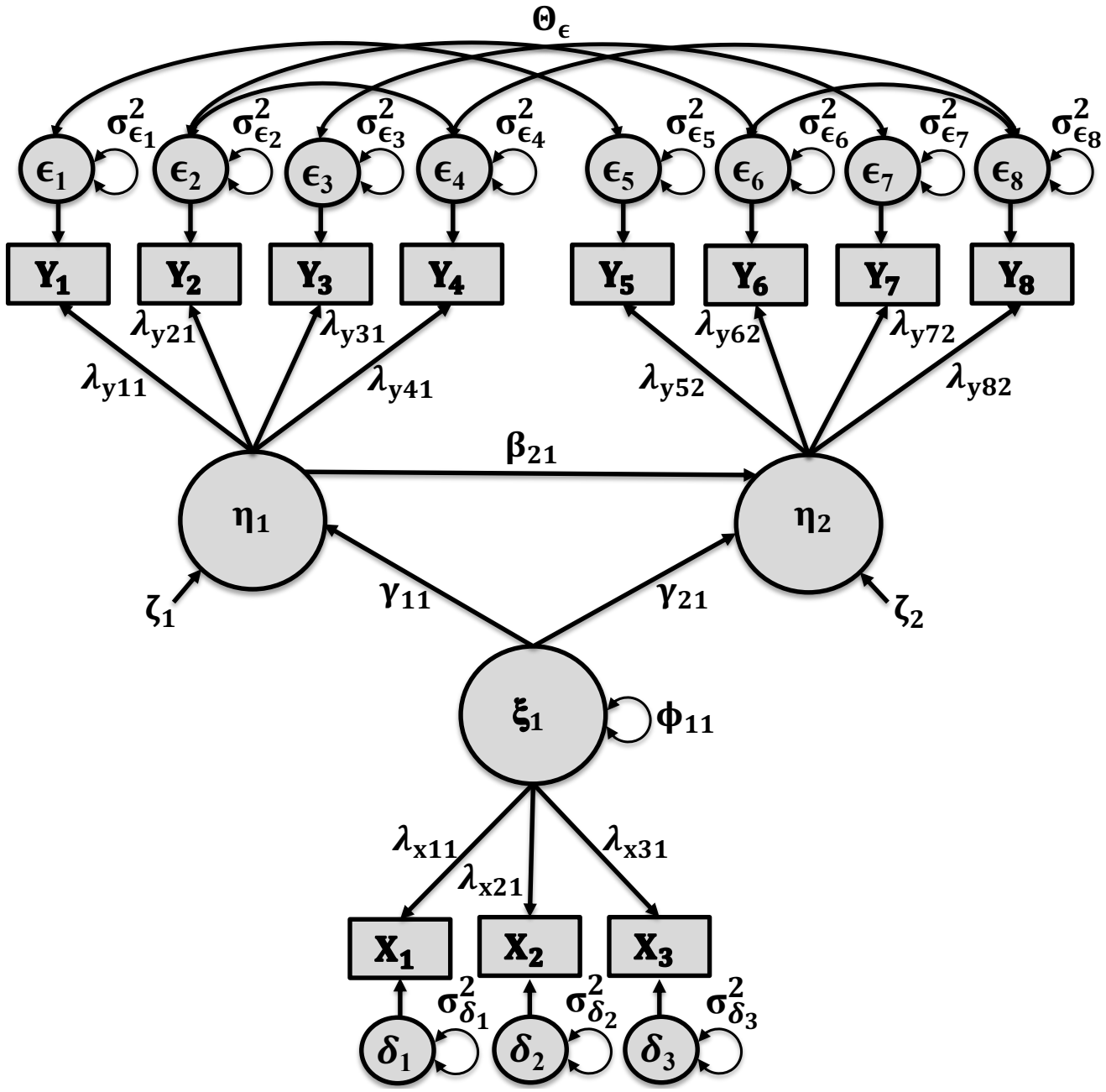


Figure 11: Structural Equation Modeling - Revisiting the Bollen (1989) Political Democracy Example. Notation follows basic LISREL notation as follows: ξ = exogenous latent factors, η = endogenous latent factors, Φ = latent factor covariance matrix, λ = factor loading in the Λ matrix, δ and ε = measurement errors, σ^2_δ and σ^2_ε = error variances, γ and β represent structural paths, θ = covariance matrix for ε elements, and ζ disturbances for η.

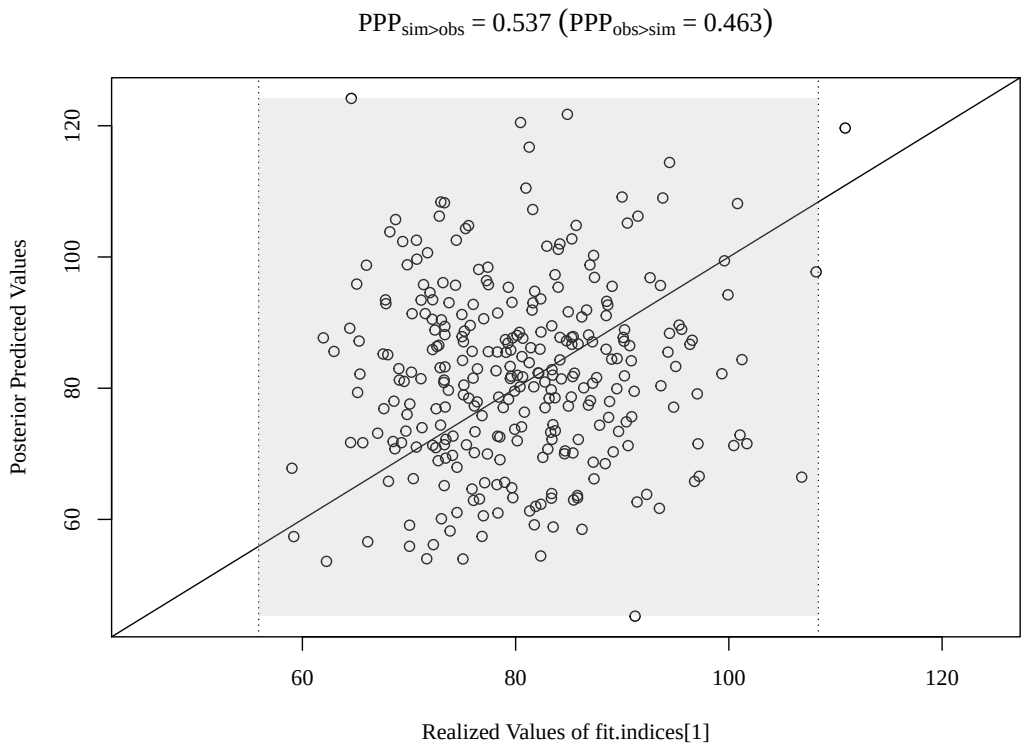


Figure 12: Posterior predictive *p*-value plots. These plots are for the SEM in Example 5.

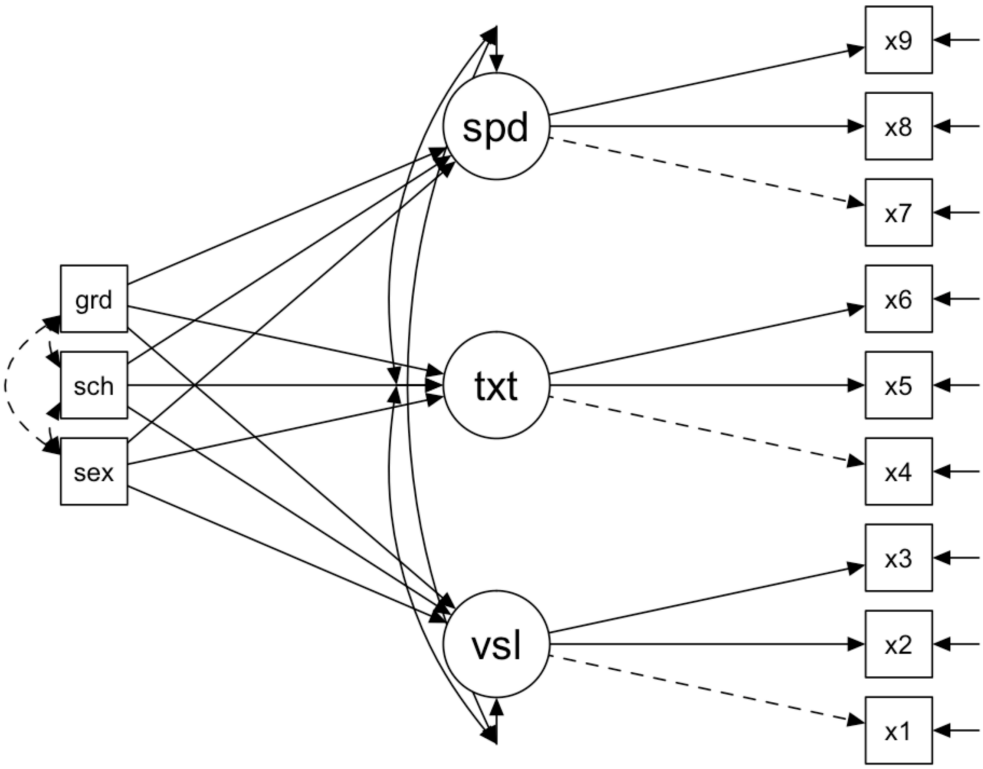


Figure 13: MIMIC model implemented in Example 6 with Holzinger and Swineford (1939) data.

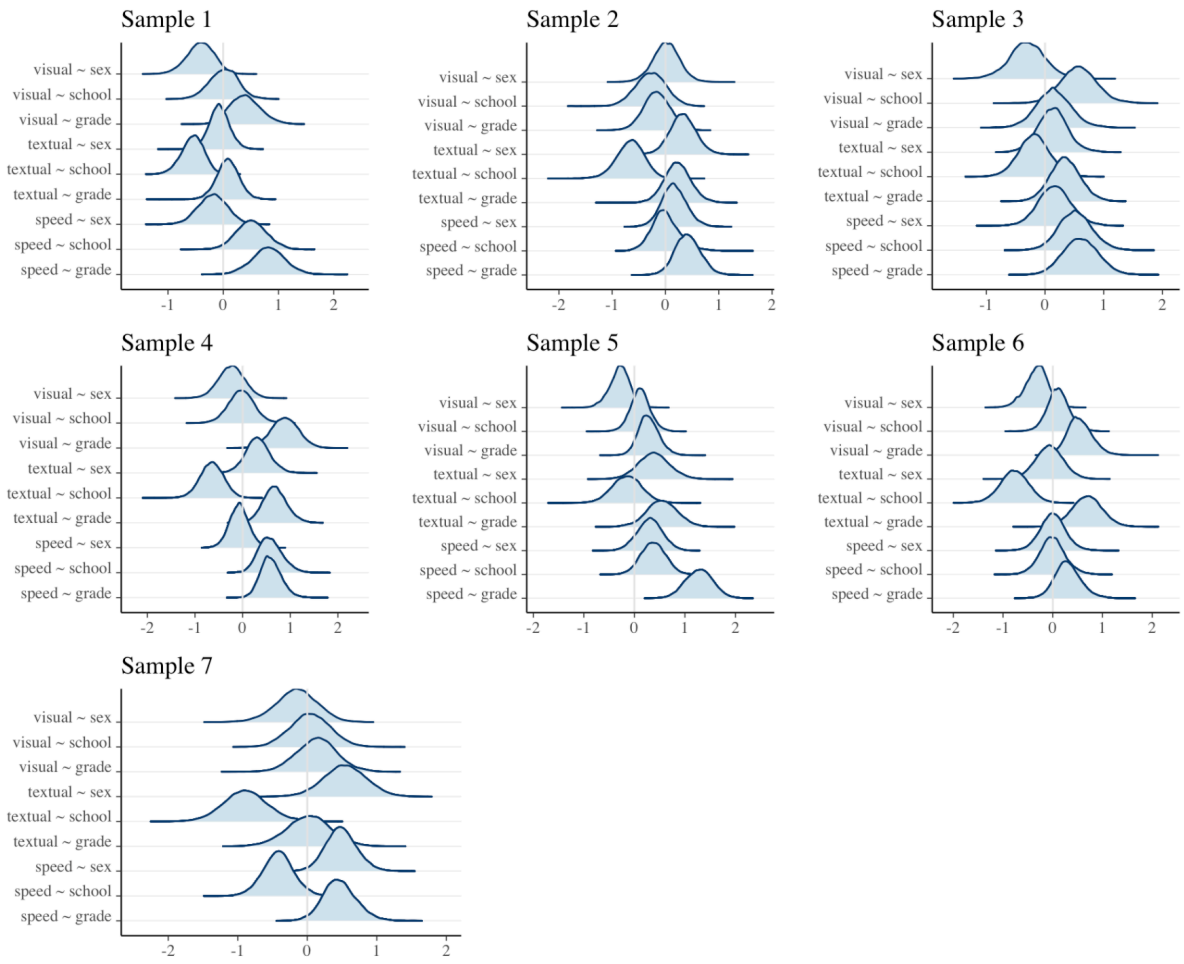


Figure 14: Posterior distributions for initial analysis. These plots are for the SEM in Example 6.

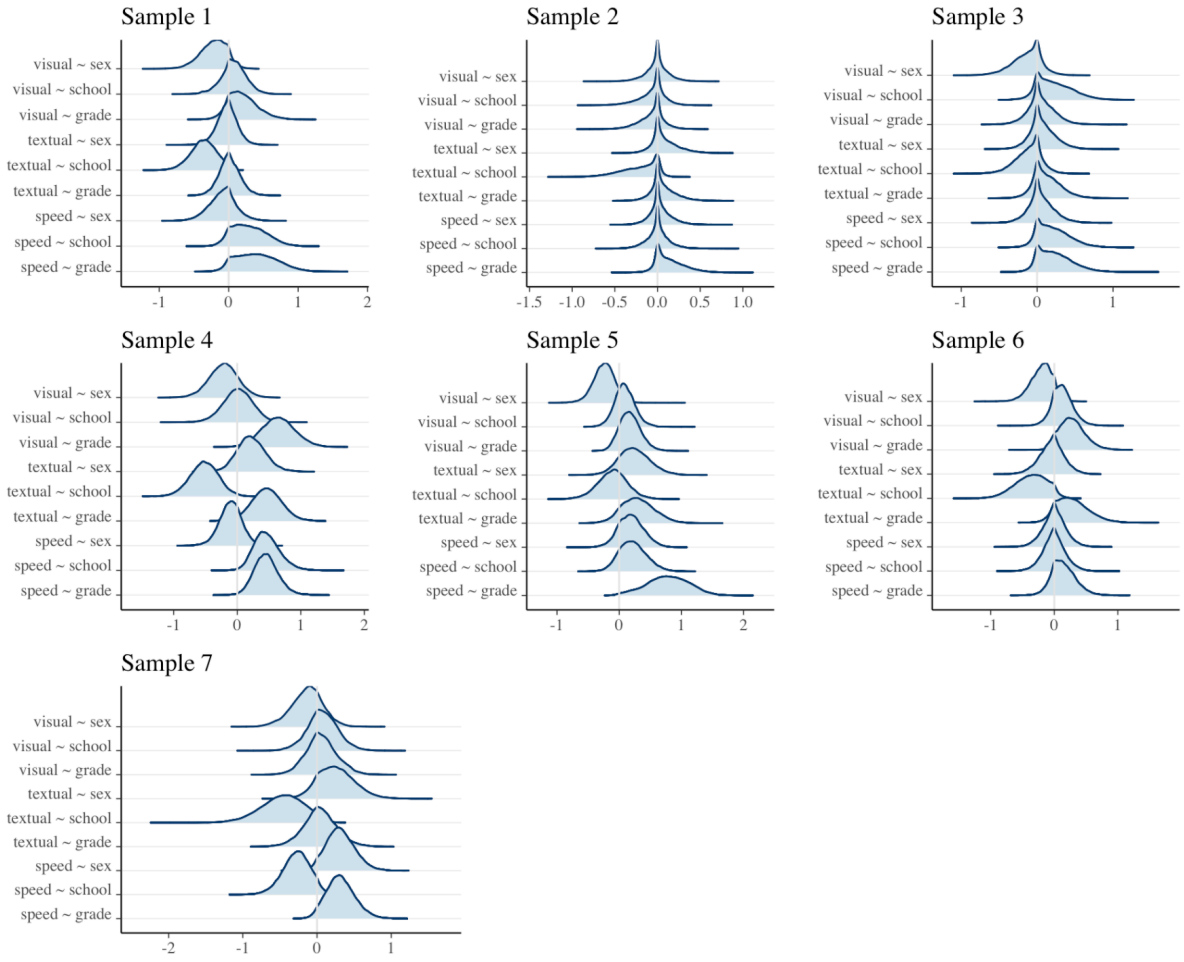


Figure 15: Posterior distributions for ridge analysis. These plots are for the SEM in Example 6.

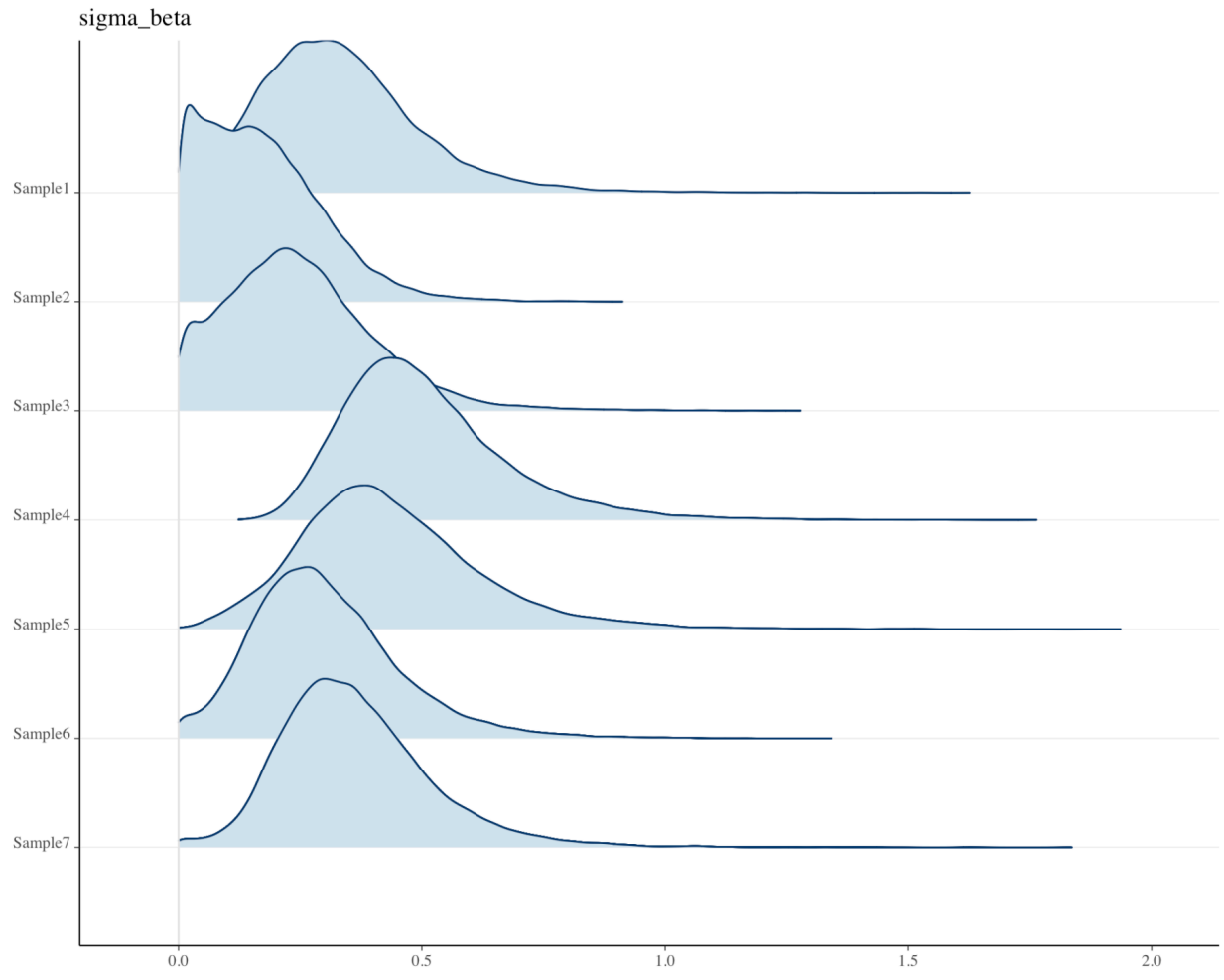


Figure 16: Penalty parameter differences across samples for ridge analysis. These plots are for the SEM in Example 6.

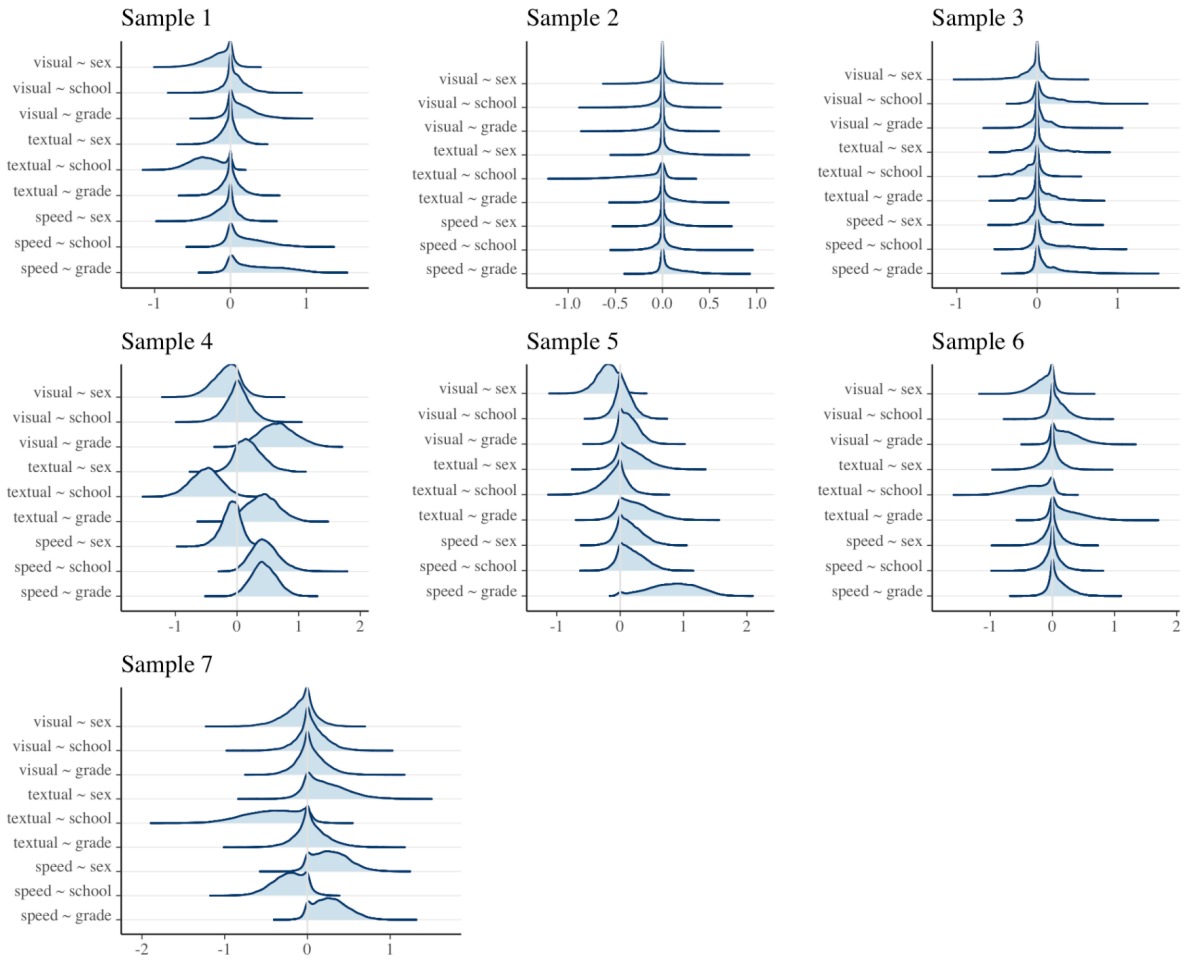


Figure 17: Posterior distributions for lasso analysis. These plots are for the SEM in Example 6.

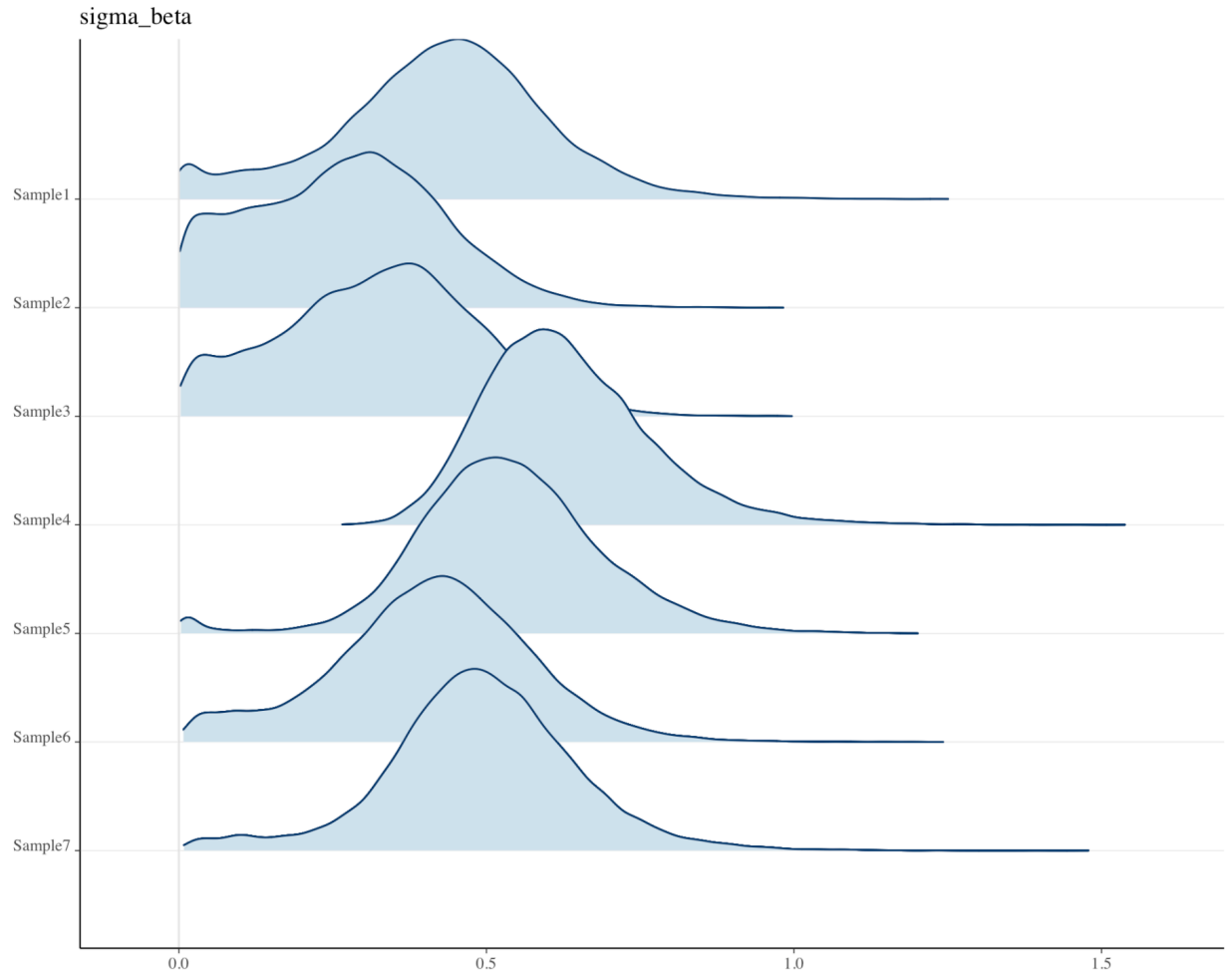


Figure 18: Penalty parameter differences across samples for lasso analysis. These plots are for the SEM in Example 6.

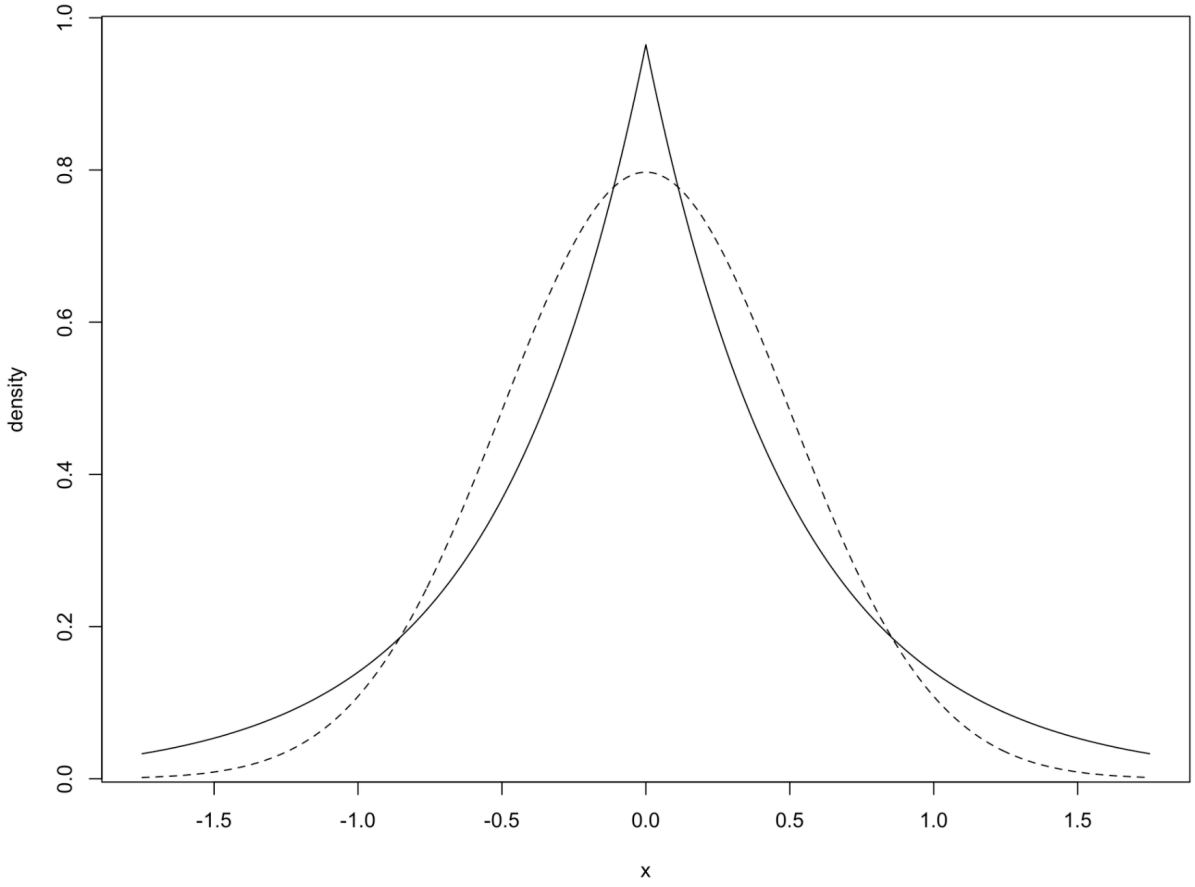


Figure 19: Comparing the priors across penalization methods. Ridge (normal prior)) is shown in the dashed line, and lasso (double exponential; Laplace) is shown in the solid line. These plots are for the SEM in Example 6.

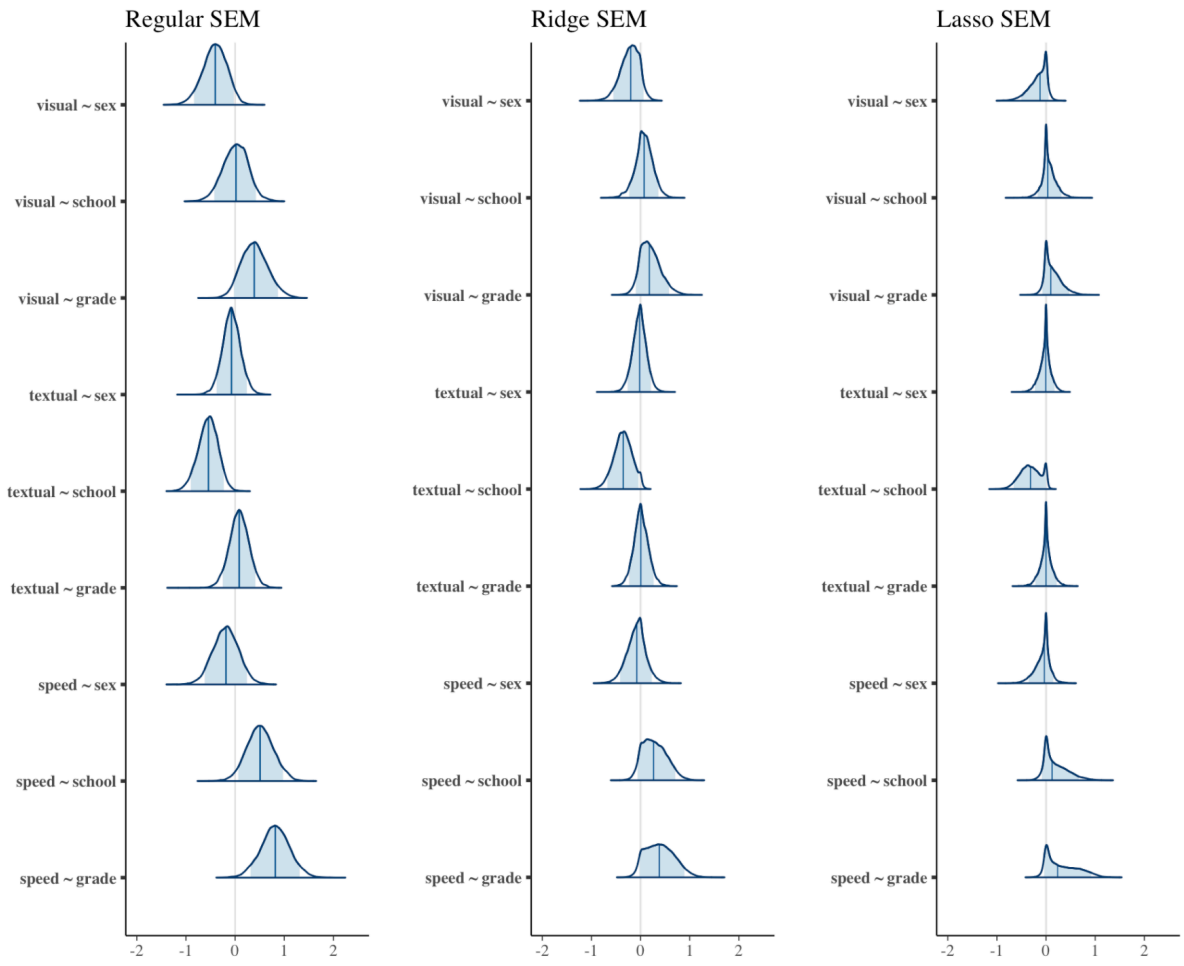


Figure 20: Posterior densities for all regression parameters for the initial analysis (left column), ridge analysis (middle column), and lasso analysis (right column). These plots are for the SEM in Example 6.

Footnotes

[1]Default priors with Stan and NUTS are placed on the standard deviations and correlations (as opposed to variances and covariances). Then, `blavaan` transforms the parameters back to the variance/covariance scale for the output.