Proceedings of the 89th Annual International Meeting of the Psychometric Society, Prague, Czech Republic (2024).

Incorporating Sparsity into Bayesian Stacking Procedures

Kjorte Harra* and David Kaplan

Department of Educational Psychology, University of Wisconsin-Madison, Madison, WI, United States
*Corresponding author. Email: harra@wisc.edu

Abstract

Bayesian stacking is a procedure adapted from machine learning that allows researchers to combine multiple unique models and optimize overall predictions, with the added benefit of not relying on strong assumptions necessary for Bayesian model averaging (BMA). For individual models, Bayesian regularization methods via sparsity-inducing priors elicit stronger predictive accuracy than unregularized modeling approaches. While model stacking is not intended to serve as a method for performing variable selection, we are unaware of any systematic investigation examining how sparsity-inducing priors applied to member models in a stack could conceivably lead to more accurate predictions. The present work investigates whether the addition of Bayesian regularization via sparsity-inducing priors of individual member models can be a worthwhile practice when using Bayesian stacking procedures. Against our expectations, we find that inducing sparsity in stacking member models does not improve predictive performance. Other results and limitations of this work are also discussed.

Keywords: Bayesian regularization, Bayesian stacking, predictive performance

To optimize predictive performance for a given outcome, there are many approaches researchers can take. Bayesian stacking, a model ensembling procedure adopted from machine learning, optimizes predictions by combining multiple unique models (Breiman, 1996; Clyde & Iversen, 2013; Wolpert, 1992; Yao et al., 2018). Bayesian stacking forms a weighted mixture of predictive distributions from an ensemble of individual models. This Bayesian model ensembling method is an improvement over the more classical approach of *Bayesian model averaging* (BMA) (Draper, 1995; Hoeting et al., 1999; Madigan & Raftery, 1994) in that Bayesian stacking does not assume that the true data generating model is in the space of models being averaged, and is theoretically expected to yield stronger predictive performance than that of any single model chosen for predictive purposes.

Another approach known to boost predictive performance is Bayesian regularization. Otherwise known as sparsity-inducing priors, these methods have demonstrated improved model accuracy and predictive performance under many modeling methods as compared to unregularized approaches, particularly with small samples (Harra & Kaplan, 2023; Jacobucci & Grimm, 2018; van Erp et al., 2019). Sparsity-inducing priors, or shrinkage priors, such as the lasso (Tibshirani, 1996) and horseshoe priors (Carvalho et al., 2009, 2010; Piironen & Vehtari, 2017) can perform variable selection and introduce model simplicity without sacrificing model performance. Although these methods have been well studied for individual model performance, it remains unclear whether these methods could also benefit modeling ensembling methods such as Bayesian stacking.

While incorporating sparsity through Bayesian regularization has been hypothesized to improve prediction accuracy (Breiman, 1996; Vehtari & Gabry, 2023; Yao et al., 2018), this remains an open question, particularly with the use of newer priors such as the regularized horseshoe prior (Piironen & Vehtari, 2017). Our present work seeks to investigate the potential benefits, if any, of incorporating

sparsity into member models within Bayesian stacking procedures for improving predictive accuracy. The following sections will provide the necessary context for this work, and then we will explore this via a full simulation study comprised of a stack of Bayesian linear regression models.

1. Bayesian Stacking

Model stacking is essentially a weighted combination of predictions from a set of specified K models ($k = 1, 2, ..., K$). Model predictions are combined (stacked) to yield a weighted combination of predictive distributions (Kaplan et al., 2025). This method of model ensembling was originally developed in the machine learning literature by (Wolpert, 1992) and (Breiman, 1996) and brought into the Bayesian framework by Clyde and Iversen (2013).

We can define a set of weights on a simplex as

$$\mathcal{W}_1^K = \left\{ w \in [0,1]^K : \sum_{k=1}^{K} w_k = 1 \right\}. \tag{1}$$

To approximate the full predictive distribution, $p(\tilde{y}_i|y_i, M_k)$, we use the leave-one-out (LOO) predictive distribution where (Yao et al., 2018)

$$\hat{p}_{k,-1}(y_i) = \int p(y_i|\theta_k, M_k) p(\theta_k|y_{-i}, M_k) d\theta_k. \tag{2}$$

The stacking weights using the log score are the solution to

$$\max_{w \in \mathcal{W}_1^K} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{k=1}^{K} w_k \hat{p}(y_i|y_{-i}, M_k). \tag{3}$$

Various weighting methods are available. ELPD_{loo} weighting is based on the ELPD (expected log point-wise predictive density) of a model, which is our primary focus for this paper. Other weighting methods include Pseudo-BMA (PBMA) and Pseudo-BMA+ (PBMA+) (Yao et al., 2018). However, preliminary analyses for this work demonstrated no noteworthy differences in performance between weighting strategies, so the remainder of this work will implement ELPD_{loo} weighting.

2. Overview of Bayesian Regularization

Bayesian regularization penalizes small regression coefficients by attaching a prior distribution to model parameters (Jacobucci & Grimm, 2018). Many regularization priors are available, beginning with the ridge prior (Hsiang, 1975) that seeks to shrink parameters close to zero and minimize collinearity. The Bayesian lasso (Park & Casella, 2008) improves upon the ridge prior as it enables shrinkage of coefficients to zero, allowing for variable selection.

The Bayesian ridge and lasso priors, described below, are extensions of frequentist methods to the Bayesian context. Strictly Bayesian approaches include the horseshoe prior (Carvalho et al., 2009, 2010), which allows for greater shrinkage than the ridge and the lasso while maintaining unregularized large coefficients. The regularized horseshoe (Piironen & Vehtari, 2017) prevents large coefficients from escaping shrinkage, allows further flexibility than the original horseshoe prior, and has been shown to further improve model predictive performance (Harra & Kaplan, 2023; Piironen & Vehtari, 2017).

Previous research has shown that Bayesian regularization can perform as well as, if not better than, classical methods of regularization in linear regression (van Erp et al., 2019). This finding has not been extended to ensemble modeling methods such as Bayesian stacking, particularly with a focus on optimizing out-of-sample predictive performance. Thus, this paper focuses on the performance of three Bayesian regularization priors, particularly the regularized horseshoe, in the context of Bayesian stacking procedures for linear regression. We investigate this via a simulation study comparing

several regularized model stacks to unregularized model stacks in terms of the amount of shrinkage induced and out-of-sample predictive performance.

2.1 Priors to be investigated

Figure 1 shows the density plots for the three regularization priors that we will be studying in this paper.

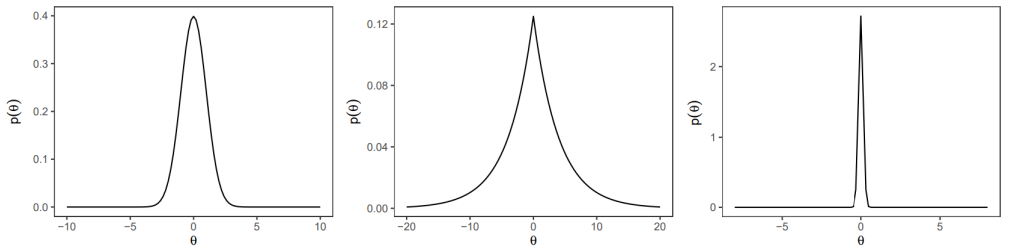

Figure 1. Regularization priors used in this paper. From left to right: Ridge normal prior $N(0,1)$, Lasso Laplace prior with location = 0, scale = 4, and the regularized horseshoe prior with $\beta_j|\lambda_j, \tau, c \sim \mathrm{N}(0, \tau^2\tilde{\lambda}_j^2)$, where $\tilde{\lambda}_j^2 = \frac{c^2\lambda_j^2}{c^2+\tau^2\lambda_j^2}$, and $\lambda_j \sim \mathcal{C}^+(0,1)$.

2.1.1 The Ridge Prior

Frequentist ridge regression (A. Hoerl & Kennard, 1970; R. Hoerl, 1985) aims to yield a parsimonious regularized regression model in the presence of highly correlated variables. The Bayesian specification of ridge regression was suggested by Hsiang (1975), who showed that if the ridge estimator, $\boldsymbol{\beta}$, has a mean of zero and covariance matrix $\boldsymbol{\Sigma} = (\sigma^2/\lambda)\mathbf{I}$, and if $\epsilon \sim N(0, \sigma_\epsilon^2\mathbf{I})$, then the posterior mean of $\boldsymbol{\beta}$ is $(\boldsymbol{x'x} + \lambda\mathbf{I})^{-1}\boldsymbol{x'y}$, which is an alternative specification of the ridge estimator. The penalty term (λ) is captured through normally distributed independent priors placed on the regression slope parameters. These normal priors have mean hyperparameter values fixed at zero in order to control shrinkage toward zero. The variance hyperparameter is typically rescaled to be in standard deviation form and is set to define the degree of spread that the distribution exhibits. Note that we specify a half-Cauchy prior distribution, denoted as $\mathcal{C}^+(0,1)$, for the residual standard deviation, but other conjugate priors could be specified as well. A representation of the ridge prior is given in the left of Figure 1.

2.1.2 The Lasso Prior

A drawback of ridge regression is that it does not improve parsimony in that all of the variables still remain in the model after penalization (Zou & Hastie, 2005). A method that appears similar to ridge regression but can yield a parsimonious model is the *least absolute shrinkage and selection operator* (Tibshirani, 1996).

The Bayesian lasso (Park & Casella, 2008) uses a double exponential or Laplace prior where

$$p(\beta_j) = \frac{1}{2\tau}\exp\left(-\frac{|\beta_j|}{\tau}\right), \tag{4}$$

where $\tau = 1/\lambda$.

The middle of Figure 1 shows the double exponential distribution. We see that this distribution is ideal because it peaks at zero, shrinking small coefficients toward zero. However, the double exponential can be set to have thick tails, allowing larger coefficients to remain large. Given that the distribution is centered at zero to control shrinkage toward zero, the mean hyperparameter setting is fixed to zero. The scale, or dispersion, of the double exponential distribution configurable

hyperparameter when implementing the lasso. This defines the amount of spread and the thickness of the tails, which controls the degree of shrinkage in coefficients. Again, a $\mathcal{C}^+(0,1)$ prior can be specified on the standard deviation of the residuals, if desired.

2.1.3 The Regularized Horseshoe Prior

The regularized horseshoe is a variant of the original horseshoe prior (Carvalho et al., 2009, 2010). The original horseshoe prior can be characterized as a scale mixture of normals with half-Cauchy tails offering unique features in enacting shrinkage that distinguish it other regularization priors. More specifically, the tails of its $\mathcal{C}^+$ distribution permit large parameters to remain unregularized, while the global shrinkage parameter τ severely shrinks parameters that are small.

A limitation of the original horseshoe prior relates to cases where large coefficients can transcend the global scale set by τ_0 with the impact being that the posteriors of these large coefficients can become quite diffused, particularly in the case of weakly-identified coefficients (Betancourt, 2018; Kaplan, 2023; Piironen & Vehtari, 2017). To remedy this issue, Piironen and Vehtari (2017) proposed a *regularized* version of the horseshoe prior. Following the notation used in Betancourt (2018) the regularized horseshoe prior takes the form of the following:

For $j = 1, ..., p$, where p are the number of predictors,

$$\beta_j \sim \mathcal{N}(0, \tau^2\tilde{\lambda}_j^2), \tag{5a}$$

$$\tilde{\lambda}_j = \frac{c\lambda_j}{\sqrt{c^2 + \tau^2\lambda_j^2}}, \tag{5b}$$

$$\lambda_j \sim \mathcal{C}^+(0, 1), \tag{5c}$$

$$c^2 \sim \mathcal{IG}\left(\frac{\nu}{2}, \frac{\nu}{2}s^2\right), \tag{5d}$$

$$\tau \sim \mathcal{C}^+(0, \tau_0), \tag{5e}$$

where $c > 0$ and s^2 is the variance for each of the p predictor variables. Those variables that have large variances would be considered more relevant a priori, and while it is possible to provide predictor-specific values for s^2, generally we scale the variables ahead of time so that $s^2 = 1$. Finally, c^2 is the slab width, which controls the size of the large regression coefficients (Piironen & Vehtari, 2017). The density plot for the regularized horseshoe is given on the right of Figure 1.

3. Present Study

A Monte Carlo simulation was conducted to evaluate shrinkage and out-of-sample predictive performance across 6 prior distributions and 4 sample size conditions ($n = 50, 100, 500, 1000$). For each iteration, a population of 10,000 observations was generated with 40 normal predictors grouped into 5 Bayesian linear regression models and an intercept-only model. Each model had half the coefficients set as small (ranging from 0 to 1), and half large (ranging from 10 to 20) with coefficient values varying across models. The outcome variable, y, was generated using these coefficients and the full population data, from which a standardized random sample of size n was drawn for model fitting and analyses. The same sample data were used for all prior conditions.

Hyperparameters for the regularized prior conditions were selected based on previous literature recommendations to control shrinkage toward zero, as detailed in previous sections (Hsiang, 1975; Park & Casella, 2008; Piironen & Vehtari, 2017). Each of the 24 study conditions was run for 500 iterations.

Prior Condition	Specification for β_j
Non-Informative	$\mathcal{N}(0, 100)$
rstanarm *Default*	$\mathcal{N}(0, 2.5)$
Informative	$\mathcal{N}(\bar{x_j}, 1)$
Ridge	$\mathcal{N}(0, 1)$
Lasso	$\frac{1}{2\tau}\exp\left(-\frac{\lvert\beta_j\rvert}{\tau}\right)$
Reg. Horseshoe	$\mathcal{N}(0, \tau^2\tilde{\lambda}_j^2)$

3.1 Evaluating Predictive Performance

For this paper, we use Bayesian leave-one-out cross-validation (LOO-CV) to evaluate model out-of-sample predictive performance (Vehtari et al., 2017). Bayesian LOO-CV is a special case of *k-fold* cross-validation, in which the data set is divided into k folds. The model of interest is fit with the training set and then compared to the i^{th} observation in the test set to measure predictive performance. LOO-CV is a k-fold cross-validation procedure where $k = n$.

The LOO-CV is uniquely suited to the question of out-of-sample predictive performance (Allen, 1974; Stone, 1974). The LOO-CV is quite similar to the *widely applicable information criterion* (WAIC) as a fully Bayesian counterpart to the AIC (Watanabe, 2010).

The *expected log point-wise predictive density* (ELPD) for LOO-CV, the ELPD_{loo}, is defined as:

$$\text{ELPD}_{loo} = \sum_{i=1}^{n} \log p(y_i \mid y_{-i}), \tag{6}$$

where

$$p(y_i \mid y_{-i}) = \int p(y_i \mid \theta) p(\theta \mid y_{-i}) d\theta \tag{7}$$

is the LOO predictive density given the data with the i^{th} data point left out (Vehtari et al., 2017). The log sum of these predictive densities in Equation (6) is the LOO-CV estimate of the ELPD (Gelman et al., 2014; Gronau & Wagenmakers, 2019; Vehtari et al., 2017).

An information criterion based on LOO, referred to as the *LOO-IC*, can be derived as

$$\text{LOO-IC} = -2\,\widehat{\text{ELPD}}_{loo} \tag{8}$$

which places the LOO-IC on the deviance scale. Among a set of competing models, the one with the smallest LOO-IC is considered the best from an out-of-sample point-wise predictive point of view. We use the LOO-IC for the comparison of our regularization priors in our simulation study.

4. Results

For this study, we aimed to examine differences in member model-induced shrinkage and model stack predictive performance across simulation conditions.

Induced shrinkage for the linear member models were compared by prior and sample size conditions, seen in Figure 2, which depicts the sum of coefficient estimates for each linear model across conditions. We observe that for small samples in particular, the lasso and regularized horseshoe induced the greatest amount of shrinkage for the linear member models compared to the other prior conditions. This expected finding demonstrates that prior distribution selection is influential in the amount of sparsity introduced into the stack member models, particularly for small samples where priors are more influential.

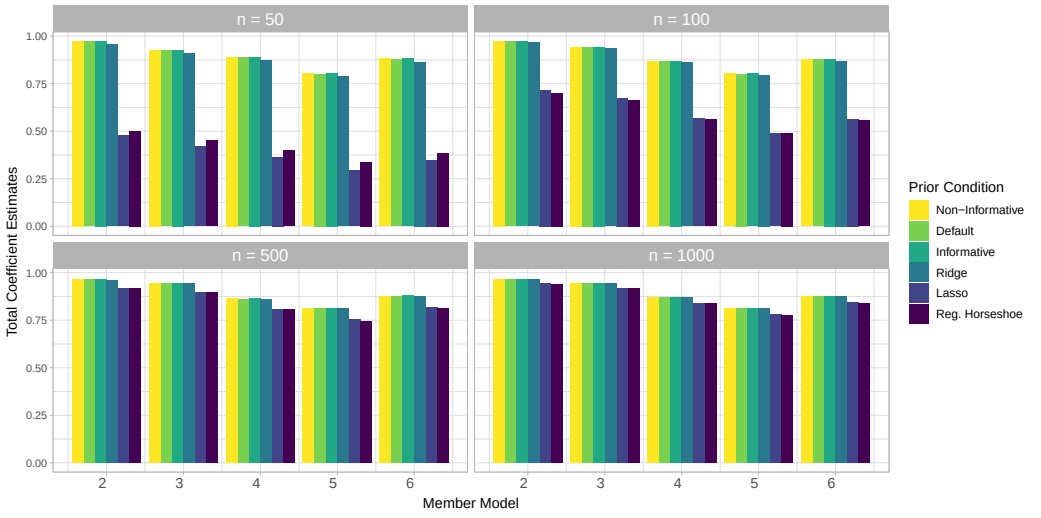

Figure 2. Mean total coefficient estimates for each linear member model across prior and sample size conditions, demonstrating induced shrinkage via regularized priors. Note: Model 1 is omitted as it is an intercept model with no regularized coefficients or variation across conditions.

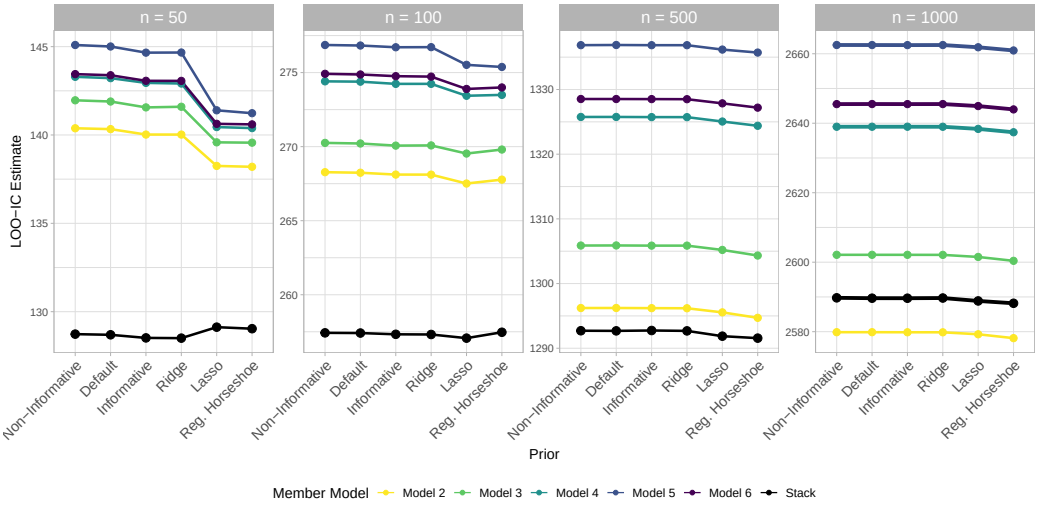

Figure 3. Comparison of mean LOO-IC estimates for individual models and model stacks across prior distribution and sample size conditions. The black line represents the stack's mean LOO-IC. Note: Model 1 is omitted as it is an intercept model with no regularized coefficients or variation across conditions.

Lastly, we compared the out-of-sample predictive performance of each linear member model to the stacked prediction across conditions, visualized in Figure 3. We find that for the linear member models, the lasso and regularized horseshoe demonstrated a boost in predictive performance in the form of the LOO-IC, particularly when samples are small. We also observed with small samples especially that the stacked predictions outperform all the member models. However, we saw no improvement in predictive performance from regularization via the lasso and regularized horseshoe prior for the model stack. Prior selection did not impact LOO-IC estimates for the model stack despite benefiting the linear member models.

5. Discussion

In this work, we found that inducing sparsity in Bayesian stacking member models boosts individual model out-of-sample predictive performance, especially when n is small, as expected (Harra & Kaplan, 2023). However, there appears to be no meaningful boost in predictive performance for the stacked models. In line with previous work on this topic, we observed that stacked predictions have stronger predictive accuracy than any individual member model (Kaplan et al., 2025; Yao et al., 2018). We also found that introducing regularization priors to the linear member models introduced sparsity and improved the predictive performance of the individual member models.

Our work here aligns with previous research demonstrating that stacked models dominate in predictive performance over any individual model (Breiman, 1996; Kaplan et al., 2025; Yao et al., 2018). As we expected, particularly with small samples, the model stack demonstrated improved out-of-sample predictive performance over the member models. This finding, and similar previous findings, demonstrate the effectiveness of Bayesian stacking.

While introducing sparsity can help with variable selection for individual models, there is insufficient evidence that sparsity can also help improve stacked predictions. It is possible that the stacking procedures negate gains in predictive performance that regularization introduces. Or, that the stacked models outperform any individual model to the extent that regularization via priors like the regularized horseshoe do not further that improvement in performance.

Our findings are limited to this particular simulation study. It's possible that other model ensemble scenarios, such as those with highly correlated variables, variables with vastly different effect sizes, or others, may find benefits in incorporating sparsity-inducing priors into Bayesian stacking. Alternatively, situations where the number of predictors p outnumber observations n may be worth investigating, as cases where $p > n$ has been shown to be when regularization is particularly useful (van Erp et al., 2019). Future research may aim to focus on what scenarios, if any, introducing Bayesian regularization into Bayesian stacking may prove useful. However, given the findings of this paper, we still recommend that researchers explore a variety of priors and weighting methods to optimize prediction for their models.

Funding Statement The research reported in this paper was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305D220012 to the University of Wisconsin-Madison. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Competing Interests The authors of this work claim no competing interests.

References

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, *16*, 125–127.

Betancourt, M. (2018). Bayes sparse regression. https://betanalpha.github.io/assets/case%5C_studies/bayes%5C_sparse%5C_regression.html

Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*, 49–64.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2009). Handling sparsity via the horseshoe. *Artificial Intelligence and Statistics*, 73–80.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*, 465–480.

Clyde, M., & Iversen, E. S. (2013). Bayesian model averaging in the M-open framework. In *Bayesian theory and applications* (pp. 483–498). Oxford University Press.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society (Series B)*, *57*, 55–98.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*, 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Gronau, Q. F., & Wagenmakers, E. J. (2019). Limitations of Bayesian leave-one-out cross-validation for model selection. *Computational Brain and Behavior*, *2*. https://doi.org/10.1007/s42113-018-0011-7

Harra, K., & Kaplan, D. (2023). On the performance of horseshoe priors for inducing sparsity in structural equation models. *Structural Equation Modeling*. https://doi.org/10.1080/10705511.2023.2280895
Hoerl, A., & Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.
Hoerl, R. (1985). Ridge analysis 25 years later. *The American Statistician*, *39*(3), 186–192.
Hoeting, J. A., Madigan, D., Raftery, A., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*, 382–417.
Hsiang, T. C. (1975). A Bayesian view on ridge regression. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *24*(4), 267–268.
Jacobucci, R., & Grimm, K. J. (2018). Comparison of frequentist and Bayesian regularization in structural equation modeling. *Structural Equation Modeling*, *25*, 639–649. https://doi.org/10.1080/10705511.2017.1410822
Kaplan, D. (2023). *Bayesian statistics for the social sciences* (2nd). Guilford Press.
Kaplan, D., Harra, K., Stampka, J., & Jude, N. (2025). Stacking models of growth: Implications for predicting the pace of progress to the education sustainable development targets using longitudinal large-scale assessment. *Psychometrika*, 1–29. https://doi.org/10.1017/psy.2025.2
Madigan, D., & Raftery, A. (1994). Model selection and accounting for model uncertainly in graphical models using Occam's window. *Journal of the American Statistical Association*, *89*, 1535–1546.
Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*, 681–686.
Piironen, J., & Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*, 5018–5051. https://doi.org/10.1214/17-EJS1337SI
Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*, 111–147.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267–288.
van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50.
Vehtari, A., & Gabry, J. (2023). Bayesian stacking and pseudo-BMA weights using the loo package [Accessed: 2024-06-04]. https://mc-stan.org/loo/articles/loo2-weights.html
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. https://doi.org/10.1007/s11222-016-9696-4
Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.
Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*, 241–259.
Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*, 917–1007. https://doi.org/10.1214/17-BA1091
Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x