Honey, I Shrunk the Irrelevant Effects! Simple and Fast Approximate Bayesian Regularization

Diana Karimova[a], Sara van Erp[d],
Roger Th.A.J. Leenders[b,c], & Joris Mulder[a,b]

[a] Department of Methodology and Statistics, Tilburg University
[b] Jheronimus Academy of Data Science
[c] Department of Organization Studies, Tilburg University
[d] Department Methodology and Statistics, Utrecht University

Abstract

Statistical models are becoming increasingly complex with more parameters to explain complex dependency structures among larger sets of variables. Regularization techniques (such as penalized regression) are ideal to identify the most important parameters by shrinking negligible effects to zero. The resulting regularized solutions are parsimonious and often show good predictive performance. Currently however regularization techniques have mainly been developed for standard modeling designs even though regularization techniques are also very useful for more complex modeling designs. Moreover, even though Bayesian regularization algorithms are competitive (and sometimes superior) to their classical counterpart, classical regularization techniques (such as the lasso) are still most common in applied research. To address these shortcomings, the current paper presents a fast and flexible approximate Bayesian regularization procedure. A Gaussian approximation is used for the integrated likelihood of the (large) set of parameters which is then combined with a Bayesian shrinkage prior to obtain a parsimonious solution with many (approximately) zero estimates. The method is implemented in the R package 'shrinkem'. The general applicability of the methodology is illustrated in various applications including linear regression models, relational event models, mediation models, factor analytic models, and Gaussian graphical models.

1 Introduction

In the current digital era, it has become relatively easy to acquire large data with many variables. Consequently, statistical models are becoming increasingly complex with larger numbers of parameters to explain possible complex relations between the variables (Azmak et al., 2015; Gomez-Cravioto et al., 2022). For example, to study temporal social interaction behavior in social networks, Perry & Wolfe (2013) considered a relational event model to analyze 21,635 email messages among 156 employees using 230 predictor variables. van Kesteren & Oberski (2019a) employed a high-dimensional mediation model for a genetic study of relationships between childhood trauma and cortisol stress activity (Houtepen, Vinkers, Carrillo-Roa, Hiemstra, Van Lier, et al., 2016) with 1,000 potential mediators (based on DNA methylation in the blood). Van Erp et al. (2019) presented a model for explaining crime rates using 125 potentially important predictor variables based on the community and law enforcement properties, such as the median family income, the percentage of housing that is occupied, the number of police officers, the police operating budget, in 319 communities. Furthermore, to study psychological networks (Armour et al., 2017; Williams & Mulder, 2020), graphical models are used to identify the most important conditional dependencies in a network of PTSD symptoms. In such applications, the challenge is to identify the most important dependency relations out of the many parameters. Because many of the parameters may be zero due to possible spurious relations, there is a high risk of inflated type I errors where possible spurious effects are incorrectly deemed to be important. Regularization algorithms are a class of statistical methods which are suitable for these problems as they result in (i) parsimonious solutions where negligible effects are shrunk to zero, and (ii) good predictive performance.

In a classical framework, these approaches aim to minimize the sum of squared residuals together with a penalization on the magnitude of the free parameters. Thereby, most important (large) parameters are freely estimated while shrinking negligible parameters to zero. Many different types of penalty functions have been proposed resulting in different parsimonious solutions such as the L_2 norm (resulting in so-called ridge regression, Marquardt & Snee, 1975), which constraints the sum of the squared parameters, the L_1 norm (known as the 'least absolute shrinkage and selection operator' or lasso; Tibshirani, 1996), which constraints the sum of the absolute values of the parameters, or a linear combination of L_1 norm and L_2 norm (also referred to as the elastic net model Zou & Hastie, 2005), to name only a few. Uncertainly quantifications of the regularized estimates are often obtained using bootstrapping.

Alternatively, in a Bayesian framework, the prior on the key parameters has a similar role as the penalty function in penalized regression (Van Erp et al., 2019; Korobilis, 2013). Typically, these priors are symmetrical around zero (so that negative values are shrunk in the same way as positive values), peaked around zero (causing small effects to be shrunk towards zero), and have thick tails (causing little to no shrinkage for large effects). There is a vast literature on possible shrinkage priors for Bayesian regularization purposes such as Gaussian prior distributions, resulting in a Bayesian alternative to ridge regression (Hsiang, 1975), Laplace priors, resulting in the Bayesian lasso (Park & Casella, 2008a; Tibshirani, 1996), or nonconcave priors, such as the horseshoe prior Carvalho et al. (2010). For an overview of possible priors for Bayesian regularization, see Van Erp et al. (2019),

for example.

Despite the tremendous potential of statistical regularization methods to provide applied researchers with interpretable (parsimonious) explanations and yielding good predictions when fitting complex models to data with many variables, the literature on statistical regularization has mainly focused on rather standard modeling designs. Statistical papers (and software) generally consider regularization algorithms which are tailored to very specific modeling designs (e.g., simple regression designs, standard graphical models, or specific measurement levels of the dependent variables (often Gaussian)). Another limitation is that most applications make use of standard regularization methods (such as the classical lasso) even though Bayesian regularization methods can be viewed to be superior as (i) Bayesian MCMC algorithms can easily handle (superior) nonconcave priors/penalties (**?**Park & Casella, 2008b) and (ii) Bayesian regularization results in the full posterior of the key parameters generally resulting in better quantifications of statistical uncertainty. For example, the classical lasso may result in standard errors of zero, yielding problematic overestimation of our certainty (Park & Casella, 2008b). Bayesian regularization methods however tend to be slower and fewer software packages have implemented these Bayesian algorithms.

The goal of the current paper is to address these shortcomings by presenting a generally applicable approximate Bayesian regularization (ABR) technique. As input, only a vector of the (unregularized) estimates (e.g., MLEs) is required together with its error (or posterior) covariance matrix. Thereby, the method can be used for any type of model and any type of parameter as long as a vector of the estimates and the corresponding errors are available, e.g., from existing software or published literature. To regularize the estimates, the errors are assumed to follow a multivariate Gaussian distribution which are combined with a specific shrinkage prior via Bayes' theorem. Statistical inference is based on the (approximated) full posterior. The methodology is readily available in the new R package '`shrinkem`'.

The paper is organized as follows. Section 2 introduces existing estimation methods as a stepping stone for the ABR method that is presented in Section 3. Section 3 also presents a simple illustration of the induced shrinkage behavior of the ABR method when using different priors and it introduces the R package `shrinkem`. Section 4 presents several applications of the methodology for various modeling designs where the results are compared with existing (tailored) regularization algorithms. The paper ends with a discussion in Section 5.

2 Statistical methods for model fitting

2.1 Maximum likelihood estimation

Let us consider a statistical model for a given data set where the parameters of interest are denoted by the vector $\boldsymbol{\beta}$ of length K and the nuisance parameters are denoted by the vector $\boldsymbol{\phi}$ of length L. The combined vector will be denoted by $\boldsymbol{\theta}' = (\boldsymbol{\beta}', \boldsymbol{\phi}')$. Maximum likelihood estimation belongs to the commonly used methods in statistical practice for fitting statistical models. The likelihood function, which will be denoted by $p(D|X, \boldsymbol{\theta})$,

quantifies the probability of the observed data of the dependent variables, denoted by D, given the unknown model parameters $\boldsymbol{\theta}$ and possible covariates X. The parameter values that maximize the (log) likelihood are called the maximum likelihood estimates (MLEs), i.e.,

$$\hat{\boldsymbol{\theta}}_{MLE} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\{\log p(D|X, \boldsymbol{\theta})\}. \tag{1}$$

The MLEs are the parameter values for which the observed data are most likely under the model at hand.

Under fairly general conditions, it can be shown that the shape of the likelihood function can be well approximated using a multivariate normal distribution (e.g., Gelman et al., 2014, Ch. 4) where the mean is centered around the MLE and the covariance matrix is equal to the inverse of the observed Fisher information matrix, denoted by $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}$ (also known as the "error covariance matrix"), i.e.,

$$\hat{p}(D|X, \boldsymbol{\theta}) \approx N(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}). \tag{2}$$

The Gaussian approximation of the likelihood is a commonly used technique in statistical inference, such as for constructing (approximate) confidence intervals, for Wald type significance testing, or in the derivation of the Bayesian information criterion (BIC), to name a few. The motivation of this approximation is that the sampling distribution of the MLE becomes concentrated around the true parameter value under mild conditions as the sample size grows (Wald, 1949; Wolfowitz, 1949), i.e.,

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}}^{-1/2}(\hat{\boldsymbol{\theta}}_{MLE} - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \boldsymbol{I}_K), \tag{3}$$

as $n \to +\infty$, where $\boldsymbol{\theta}_0$ is the true parameter value in the population.

Despite the ubiquity of MLE in statistical practice, MLEs (similarly as OLS estimates) may not be preferred in statistical problems where the model at hand contains many parameters out of which many may be potentially zero and when the sample sizes are relatively small. In this case, MLEs may result in overfitting as all parameters are freely estimated, including possible spurious effects, as well as poor predictions (refs?). Moreover nonsparse solutions using MLEs complicate the interpretation of the results as it becomes difficult which nonzero estimated effects (out of many) truly matter to predict one or more outcome variables of interest. Penalized regression methods have been developed to resolve these limitations.

2.2 Penalized regression

In the case of a large number of parameters in the model, penalization methods provide a way to obtain a more parsimonious model with many (approximate) zero estimates for negligible effects. Penalized regression constrains the magnitude of all estimates such that small, unimportant effects become (approximately) zero, while leaving large, important effects largely unaltered. This allows to eliminate negligible "noisy" effects from the analysis and produce parsimonious solutions which are easier to interpret in the case of complex models with many potentially important effects in the case of relatively small samples. A

penalized estimate can be obtained by adding a penalty term (typically a norm) on the (many) key parameters to the optimization function (Hastie et al., 2015):

$$\hat{\boldsymbol{\theta}}_{penalty} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\{-\frac{1}{n}\log p(D|X,\boldsymbol{\theta}) + \lambda||\boldsymbol{\beta}||_q\}, \tag{4}$$

where $||\boldsymbol{\beta}||_q$ denotes the L_q norm of the coefficients (excluding the intercept) and λ is a penalty parameter. As can be seen, when setting the penalty parameter to zero, i.e., $\lambda = 0$, the optimization problem becomes equivalent to (1) yielding the standard MLE solution.

The most well-known choices for the penalty term are the L_1 norm, which yields the so-called lasso (Tibshirani, 1996) solution, and the L_2 norm, which is known as the ridge solution Hoerl & Kennard (1970). A striking feature of the lasso is that it can yield exact zero's for the penalized estimates (unlike the ridge solution for instance), thus yielding a sparse solution with potentially few non-zero parameters. This property has made the lasso a popular choice for penalized regression. For norms with $q < 1$ the solutions are also sparse but the optimization problem is not convex, which makes the computation challenging (Hastie et al., 2015, Ch.3).

Despite its popularity, the lasso also suffers from important drawbacks, such as underestimation of the standard errors of the penalized estimates (Casella et al., 2010), and not abiding the oracle property, which comes down to inconsistent estimation behavior in certain scenario's (Zou, 2006a). Furthermore, to determine the penalty parameter λ which is crucial as it specifies the size of the penalty, computationally intensive methods are typically used such as cross-validation or graphical elbow methods. Due to these potential drawbacks, Bayesian approaches for statistical regularization are becoming increasingly popular.

2.3 Bayesian regularization

In a Bayesian framework, regularization occurs naturally via the prior distribution which is a standard element of a Bayesian model. The prior distribution quantifies which values of the parameters are likely or unlikely before observing the data. Thus, if a sparse solution is expected with many zero effects, priors can be specified which contain most probability mass around zero. Such priors result in sparse solutions where most posterior probability of negligible effects is is concentrated around zero. Thereby, the role of the prior in Bayesian regularization is comparable to the role of the penalty function in penalized regression. This can also be shown mathematically as the posterior is given by

$$p_{Bayes}(\boldsymbol{\theta}|D,X,\lambda) = \frac{p(D|X,\boldsymbol{\theta})p_{Bayes}(\boldsymbol{\theta}|\lambda)}{p(D|X)} \propto p(D|X,\boldsymbol{\theta})\ p_{Bayes}(\boldsymbol{\theta}|\lambda) \tag{5}$$

Consequently, the logarithm of the posterior can be written as:

$$\log p_{Bayes}(\boldsymbol{\theta}|D,X) = \log p(D|X,\boldsymbol{\theta}) + \log p_{Bayes}(\boldsymbol{\theta}|\lambda) + constant,$$

where the constant does not depend on $\boldsymbol{\theta}$. Hence, the posterior mode (a Bayesian point estimate) can be expressed as:

$$\hat{\boldsymbol{\theta}}_{Bayes} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\{\log p(D|\boldsymbol{\theta}) + \log p_{shrinkage}(\boldsymbol{\theta}|\lambda)\}. \tag{6}$$

Equation (6) illustrates the similarity of the Bayesian posterior mode with the penalized estimate in Equation (4)[1]. Furthermore, it can be shown that specific prior choices result in Bayesian counterparts of common penalty functions, such as the Bayesian lasso, which uses a Laplace prior on the coefficients (Park & Casella, 2008b).

The full posterior can be estimated using MCMC algorithms which result in accurate quantifications of the statistical uncertainty of the model parameters, thereby avoiding certain issues of the classical lasso for instance. Moreover, the penalty parameter can be jointly estimated with the model parameters using a noninformative prior so that computational intensive resampling methods such as bootstrapping can be avoided:

$$p_{Bayes}(\boldsymbol{\theta}, \lambda | D, X) \propto p(D|X, \boldsymbol{\theta})\ p_{Bayes}(\boldsymbol{\theta}|\lambda)\ p(\lambda).$$

From the full posterior, the posterior of the key parameters $\boldsymbol{\beta}$ can be extracted. Furthermore, using flexible MCMC algorithm, many different types of priors can be considered including nonconcave horseshoe priors (Carvalho et al., 2010). Despite these useful properties, Bayesian MCMC algorithms can be slow as the exploration of the high-dimensional posterior space may consist of many complex dependencies between the (possibly nuisance) parameters. This may be an important reason for its limited use in statistical practice.

3 A fast and flexible two-step procedure for approximate Bayesian regularization

3.1 Methodology

The current section presents a generally applicable two-step procedure for approximate Bayesian regularization (ABR). The methodology can be applied to virtually any statistical model and any set of parameters by first replacing the likelihood function with a Gaussian approximation, which is then combined with a prior distribution on the key parameters using a MCMC algorithm to obtain a sparse regularized solution. The procedure is fast due to its reliance on Gaussian approximations of the likelihood, thereby simplifying possible complex dependency structures between the parameters (including nuisance parameters). Furthermore, the procedure is flexible as the Bayesian MCMC algorithm allows virtually any type of prior (penalization) on the key parameters. Thereby the procedure avoids important limitations of classical penalization (underestimated uncertainty quantification, inability to use nonconcave penalty functions) and Bayesian regularization (computationally slow).

ABR consists of the following two steps:

Step 1. Extract the unregularized estimates of the key parameters and the corresponding error covariance matrix. Due to large sample theory (Equation (2)), and by

[1]Note that the posterior median or the posterior mean are more common to use as point estimates than the posterior mode. Here we simply write the posterior mode to illustrate the similarity with the penalized estimate $\hat{\boldsymbol{\theta}}_{penalty}$.

marginalizing over the key parameters, the integrated likelihood of the key parameters can be approximated by $N(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}})$.

Step 2. Fit an approximate regularized Bayesian model by combining the approximate Gaussian likelihood with a shrinkage prior:

$$\hat{p}_{Bayes}(\boldsymbol{\beta}, \lambda|D, X) \propto N(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}_{MLE}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}})\ p_{Bayes}(\boldsymbol{\beta}|\lambda)\ p(\lambda).$$

Note that in Step 1, one could for instance use the output from a classical analysis (e.g., MLEs and error covariance matrix) or from a Bayesian analysis (e.g., posterior estimates and covariance matrix using noninformative priors) for the estimate and covariance matrix of the key parameters. Fitting this approximate Bayesian model is computationally cheap because many shrinkage priors can be written as scaled mixtures of normals (Van Erp et al., 2019), which are conditionally conjugate with the Gaussian approximation of the likelihood.

Moreover, as prior for the squared penalty parameter, λ^2, a F prior is specified having two degrees of freedom parameters and a scale parameter. Because a F distribution can be written as a gamma scale mixture of inverse gamma distributions, the F prior is conditionally conjugate and therefore it can be implemented in a MCMC algorithm relatively straightforwardly (e.g., Mulder & Pericchi, 2018). Furthermore, when setting the first degrees of freedom parameter to 1, the implied prior for the penalty parameter λ follows a half-Student t distribution, which is becoming an increasingly common prior for scale parameters (Gelman, 2006; Polson & Scott, 2012). By setting the second degrees of freedom to 1 and together with a very large scale, a virtually flat prior can be constructed for λ.

3.2 Illustration of the regularization behavior of ABR

The goal of this section is to illustrate the induced shrinkage behavior of the ABR method for a given estimate and given error variance. This is done for a Gaussian prior (corresponding to ridge regression), when using a Laplace prior (corresponding to the Bayesian lasso), and when using the nonconcave horseshoe prior.A plot of the prior distributions is given in Figure 1, which shows that the Laplace prior is more peaked at zero and has thicker tails which induces heavier shrinkage near zero and less shrinkage away from zero in comparison to the Gaussian prior. The horseshoe prior has a pole at zero and thicker than the Laplace prior (it even has thicker tails than a Cauchy distribution; Mulder & Pericchi, 2018) inducing heavier shrinkage near zero and less shrinkage away from zero in comparison to the Laplace prior. These three priors were chosen as they are well-known in the literature and because they show clear differences regarding their shrinkage behavior.

The shrinkage behavior of these priors using ABR is analyzed by varying an unregularized estimate, $\hat{\beta}_{MLE}$, on a grid from 0 to 10 using a fixed error variance of 1. Moreover the penalty parameter λ is fixed at 1 to clearly see the induced shrinkage behavior of the different priors. Figure 2 shows the estimated posterior medians (upper left panel) and estimated posterior mode (upper right panel) with 95% credibility range as a function of the unregularized estimate as well as the difference between these estimates and the unregularized estimates (right panels). The figures show that the approximated regularization

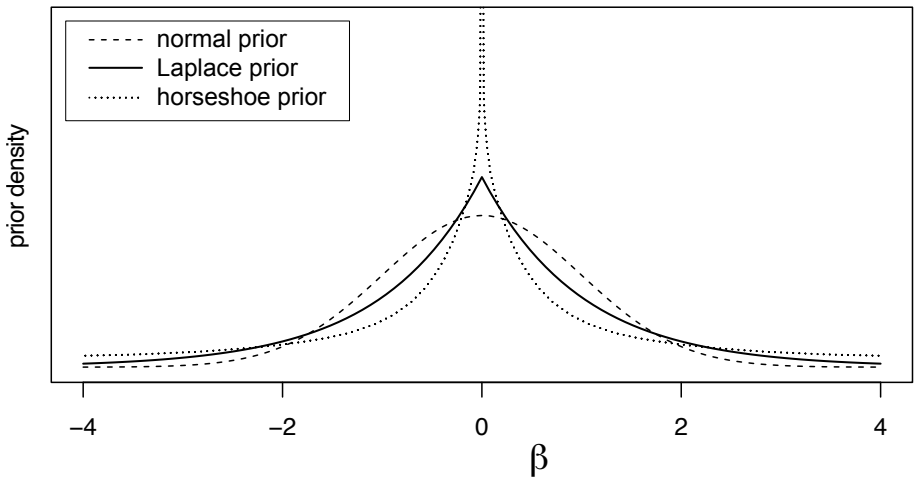

Figure 1: Plot of a normal prior (corresponding to ridge regression; dashed line), when using a Laplace prior (corresponding to the Bayesian lasso; solid line), and a nonconcave horseshoe prior (dotted line).

methods result in comparable shrinkage behavior as full Bayesian regularization methods: The Gaussian ridge prior results in a constant level of shrinkage, the Laplace prior first results in considerable shrinkage near zero and then moves along the estimate with an equal distance, and the horseshoe results in the heaviest shrinkage near zero which diminishes as the estimate further moves away from 0. Moreover, we see that the posterior median results in smoother shrinkage behavior in comparison to the posterior mode. We refer the interested reader to Tibshirani (1996) and Carvalho et al. (2009) to see that the shrinkage behavior is comparable with the original ridge, the lasso, and the horseshoe. In the applications in the following section the approximate regularized solutions are compared with the original regularized results in empirical data.

3.3 The R software package `shrinkem`

The ABR method is implemented in the R package `shrinkem` to enable fast and flexible shrinkage in the case unregularized estimates and their (dependent) errors are known. The main function `shrinkem` requires a numerical vector of the unregularized estimates, `x`, its error covariance matrix, `Sigma`, and a specific `type` of prior which can (currently) be `"ridge"` (Gaussian), `"lasso"` (Laplace), or `"horseshoe"`, e.g.,

```
shrinkem(x = estimates, Sigma = error.covariance.matrix, type = "horseshoe")
```

where `estimates` and `error.covariance.matrix` denote the objects in R containing the unregularized estimates and its error covariance matrix.

To optimize the penalty parameter, computationally intensive methods, such as cross-validation, are avoided as the penalty parameter is jointly estimated with the other param-

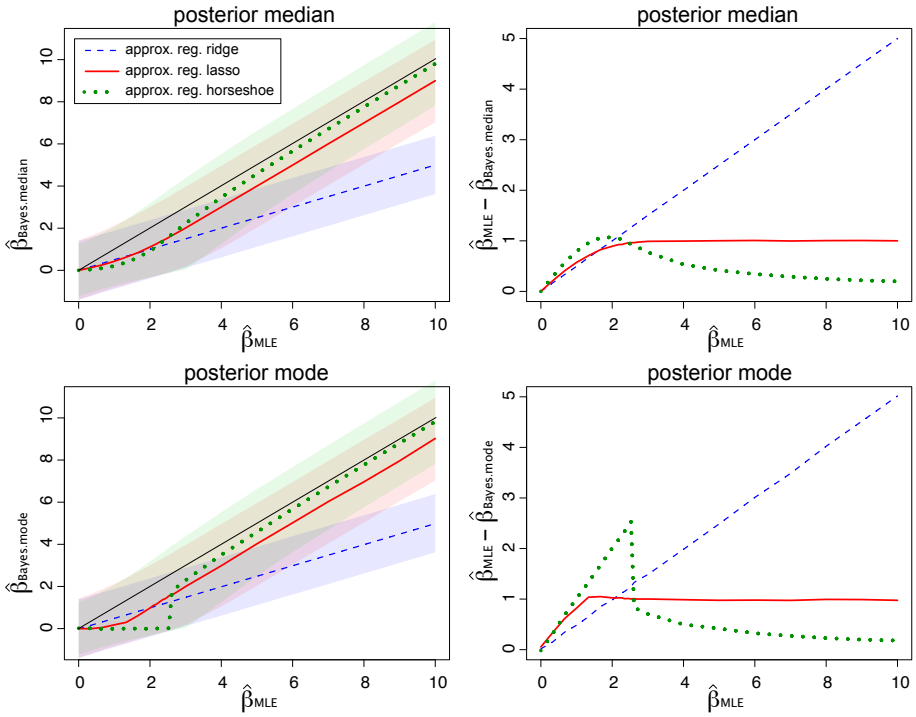

Figure 2: The Bayesian posterior median (upper panels) and the Bayesian posterior modes (lower panels) as a function of an unregularized estimate $\hat{\beta}_{MLE}$ with an error variance of 1 when using a Gaussian prior (dashed blue), a Laplace prior (solid red), and a horseshoe prior (dotted green). The right panels shows the difference of the Bayesian estimates with $\hat{\beta}_{MLE}$.

eters. For the squared penalty parameter λ^2, a F prior is chosen with degrees of freedom parameters `df1` and `df2`, which have default values of 1, and the squared scale `s2`, which has default value 1e3. This prior is equivalent to a half-Cauchy prior with scale $\sqrt{\texttt{s2}}$, which is virtually flat if `s2` is set large enough. The advantage of the F parameterization on λ^2 is that the prior is conditionally conjugate (e.g. Mulder & Pericchi, 2018), resulting in easy and fast posterior sampling.

It is also possible to fix the penalty parameters (via the arguments `lambda2.fixed` and `lambda2`. This may be useful when a user want to apply cross-validation techniques or for Bayesian regularization using an empirical Bayes estimate for λ^2. For the latter case, one could first fit the ABR model by freely estimating the penalty parameter λ using a (approximately) flat prior. The resulting posterior mode of λ then maximizes the marginal likelihood (Van Erp et al., 2019). Its value can then be used as fixed penalty parameter in `shrinkem`. Other optional arguments include `group`, which can be used to apply grouped regularization where the penalty parameter is separately optimized for the

different subsets of parameters (as in the group lasso, Yuan & Lin, 2006), and the number of posterior draws via `iterations`. Throughout this paper, the default choices will be used for these optional arguments to keep the focus on the general ABR method.

4 Empirical applications

The current section explores the shrinkage behavior and the predictive performance of the two-step approximation method using various empirical data sets and different types of models from the literature. The results of the ABR method are compared with their full, non-approximated counterparts. The goal is to illustrate that ABR often result in very similar results as existing regularization algorithms which are tailored to a specific modeling framework.

4.1 Small linear regression model - diabetes data

Park & Casella (2008a) considered data of the diabetes data of (see Efron et al., 2004) consisting of 442 diabetes patients on ten medical baseline variables. A standard linear model By fitting the Bayesian lasso on a grid of penalty parameters, the behavior of their Bayesian lasso was compared with the orginal lasso (Tibshirani, 1996) and with the ordinary ridge. Here we compare the shrinkage behavior of the ABR methodology using the Laplace (lasso) prior and the Gaussian (ridge) prior with the results from Park & Casella (2008a).

The shrinkage behavior was explored by fixing the penalty parameter λ on a grid of values to induce extreme shrinkage with only zero estimates to practically no shrinkage. Figure 3 shows the full Bayesian lasso (upper left panel) and the ordinary ridge (lower left panel) as well as the results of the ABR method using the Laplace (lasso) prior (upper right) panel and the Gaussian (ridge) prior (lower right panel). The left plots were taken from (Park & Casella, 2008a). The figure shows that the ABR lasso and the ABR ridge result in virtually identical shrinkage behavior as the full Bayesian lasso and the original ridge.

4.2 Large linear regression model - Communities and crime

We illustrate the ABR method for a linear regression model on a data set containing 125 predictors of the number of violent crimes per 100,000 residents in communities in the United States[2] (Redmond, 2011). This data set has previously been analyzed using shrinkage priors in van Erp et al. (2019). We only consider the scaled continuous predictors in this example ($p = 121$) and we split the data set in a 90% training ($n = 287$) and test ($n = 32$) set. We compare the ABR method using the ridge, lasso, and horseshoe priors to an exact implementation using the Rstan (Stan Development Team, 2024) interface to

[2]We used the unnormalized data, available at `https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized`

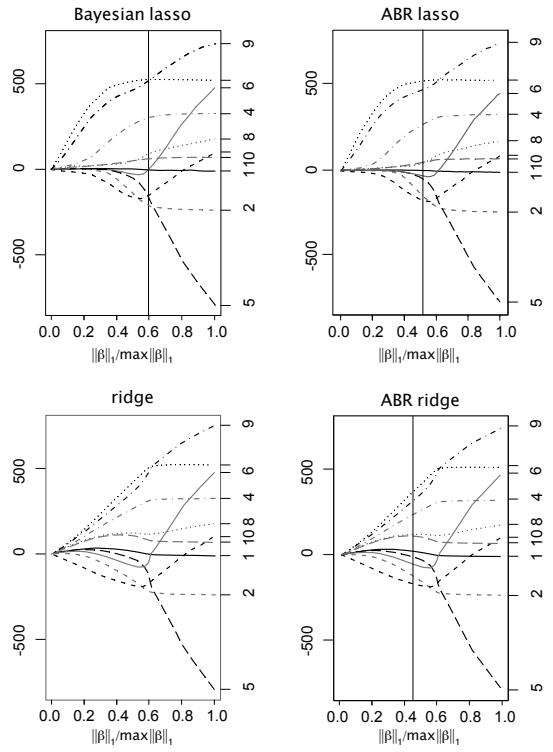

Figure 3: The regularization estimates when using the full Bayesian lasso (upper left), the ridge (lower left), the ABR lasso (upper right), and the ABR ridge (lower right) on the diabetes data (see Efron et al., 2004) while varying the penalty parameter λ on a grid of values. The panels on the right were derived from Park & Casella (2008a). Posterior medians were used as Bayesian point estimates.

Stan[3] (Carpenter et al., 2017). Note that for the lasso prior it is generally recommended to scale the prior to the error variance to avoid multimodal posteriors (Park & Casella, 2008b). However, in the approximate implementation, the error variance is not available so we only compare non-scaled lasso priors.

Figure 4 compares the posterior mean (circles) and mode (triangles) estimates and 95% credible intervals across priors and algorithms for the ten largest and smallest estimated regression coefficients. For the coefficients close to zero, the results were very comparable across priors and algorithms with only the horseshoe prior leading to slightly smaller

[3]Note that it is also possible to implement the ABR method itself in Stan. Stan model files to do so are available at `https://github.com/sara-vanerp/ApproxBR/tree/main`. These models can be adapted to other prior specifications.

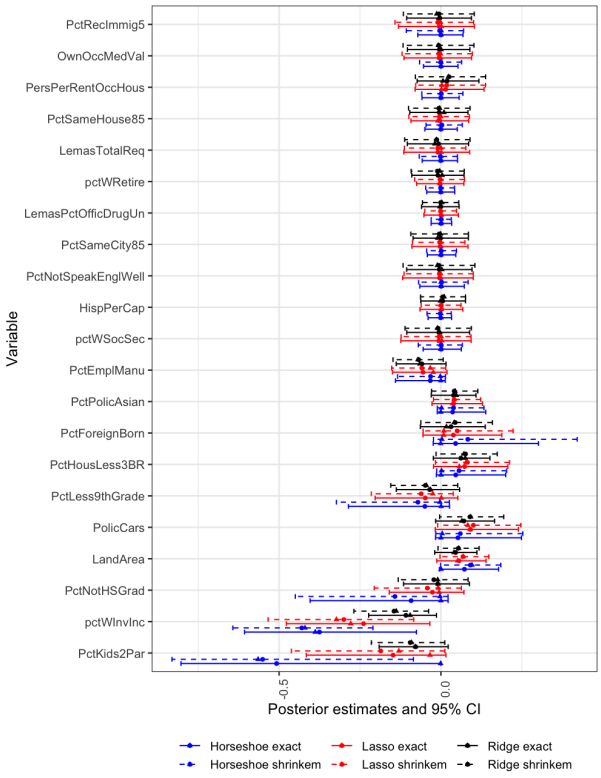

Figure 4: Posterior mean (circles) and mode (triangles) estimates, and 95% credible intervals for the ten smallest (top) and largest (bottom) regression coefficients in the crime application using either the approximate or the exact Bayesian regularization algorithm with different priors.

credible intervals overall. For the ten largest coefficients, it can be seen that the horseshoe prior leads to more shrinkage when the coefficient is relatively close to zero, especially when using the posterior mode (e.g., for *PctEmplManu*). However, as the regression coefficient becomes large enough, it escapes this shrinkage and remains larger in value for the horseshoe compared to the ridge and lasso (e.g., for *pctWInvInc* and *PctKids2Par*). Note that in the case of *PctKids2Par* the posterior mode for the exact algorithm is practically zero, while the posterior mean and the point estimates for the approximate algorithm are around -0.5. This is due to bimodality in the posterior distribution for this coefficient, which might be the result of divergent transitions arising in the exact horseshoe implementation, which might indicate non-convergence.

Finally, the prediction mean squared error (PMSE) did not differ substantially between methods, as can be seen in Table 1, but the error was considerably higher compared to the non-regularized model.

Table 1: Prediction mean squared error for the different priors and algorithms in the crime application

	unregularized	ridge	lasso	horseshoe
Exact	0.80	0.27	0.26	0.28
Approximate	-	0.26	0.25	0.27

4.3 Relational event model - Apollo 13 Mission's voice data

Relational event models are useful to study the drivers of social interaction behavior among actors in temporal social networks. Due to the complexity of social interaction behavior, relational event models can easily consist of a very large number of possible predictor variables as possible drivers of the social interaction in the network. Unregularized estimates are generally not parsimonious complicating the interpretation of the results and limiting our understanding of social interaction behavior in a network.

Here we consider a relational event sequence of voice loops taken from NASA's famous but disastrous Apollo 13 mission. The data consist of 5402 voice messages among 19 actors (three astronauts and sixteen members at mission control). Karimova et al. (2023) considered full Bayesian regularized relational event models for these data consisting of 103 possible predictor variables of drivers of the communication behavior. This relational event model can be seen as a special type of multinomial regression model. Gaussian (ridge) priors, Laplace (lasso) priors, and horseshoe priors were considered to obtain parsimonious solutions. Here we compare their results with the results when using the ABR method on the unregularized estimates using the same priors. The last 500 relational events were left out when fitting (training) the models and to evaluate the out-of-sample predictive performance of the models.

First we compare the estimates based on full Bayesian regularization of the relational event model using the three different shrinkage priors with the respective ABR counterparts. Figure 5 concisely shows the 95% credibility intervals of the 103 parameters based on the ridge (left panel), lasso (middel panel), and horseshoe prior (right panel) when using the full Bayesian analysis on the x-axis versus the approximate Bayesian analysis on the y-axis. The grey intervals depict intervals that contain the value 0. The plots show that the differences between the lasso and the horseshoe prior are extremely small. Here we also see that the horseshoe analyses result in less shrinkage of the larger estimates than the lasso analyses. Furthermore, we can see that the ridge analyses show some considerable differences between the full and the approximate analyses. For example, we see that the estimated effects of incoming two paths (itp) and outgoing two paths (otp) are quite different. Table 2 presents the number of 95% credibility intervals that do not contain zero. We again see that the ABR lasso and ABR horseshoe result in practically the same numbers of significant effects, which are considerably less than the unregularized analyses (which were based on a Bayesian analysis using flat priors). Furthermore, the ABR ridge results in considerably more significant effects than its full ridge counterpart.

Finally we consider the within-sample and the out-of-sample predictive performance

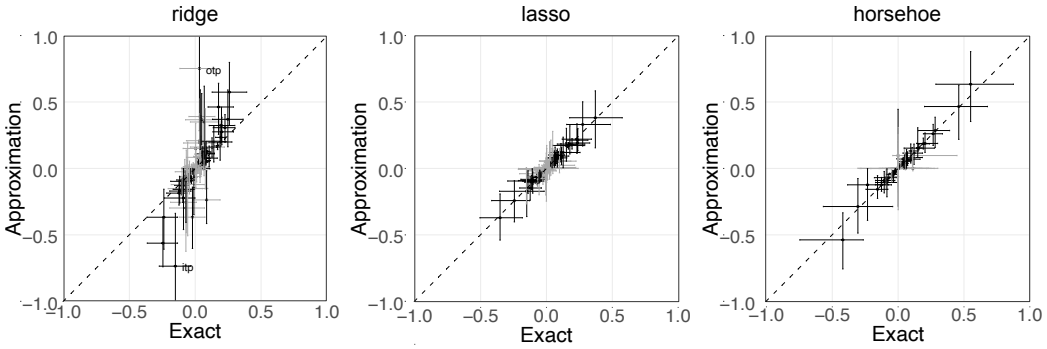

Figure 5: 95% credibility intervals of the 103 parameters for the REM of the Apollo 13 data based on the ridge (left panel), lasso (middel panel), and horseshoe prior (right panel) when using the full Bayesian analysis on the x-axis versus the approximate Bayesian analysis on the y-axis. Intervals that contain 0 are printed in grey.

	unregularized	Ridge	Lasso	Horseshoe
Exact	62	54	53	45
Approximation	–	62	54	45

Table 2: Number of 'significant' effects based on a 5% credible interval for the exact and the approximate regularized relational event model.

of the regularized solutions in this application. The predictive performance is assessed by checking the percentage of observed dyads that fall in the top 5%, top 10%, or top 20% most probable dyads according to the fitted model. Note that the network consists in total of $19 \times 18 = 342$ directed dyads. The results can be found in Table 3. Overall we can see that the predictive performance results of all models are quite good and also very similar. Only for certain out-of-sample scenarios, we can see that the percentages of certain approximated solutions are slightly lower.

4.4 Gaussian graphical models - PTSD symptoms data

Graphical models are used to study conditional dependency structures among the dependent variables in highly dimensional problems. These dependency structures are often depicted as networks where an edge implies a nonzero conditional dependence between two variables given all other variables and no edge implies conditional independence. Due to the large number of possible conditional dependencies, namely $P(P-1)/2$ in the case of P dependent variables, regularization methods are becoming increasingly popular to obtain parsimonious explanations of the dependency structure in a given data set. The graphical lasso (Friedman et al., 2008), abbreviated as 'glasso', belongs to the most commonly used techniques for this purpose where the elements of the precision matrix are

within-sample	5%				10%				20%			
	unreg.	ridge	lasso	HS	unreg.	ridge	lasso	HS	unreg.	ridge	lasso	HS
Exact	88.8	88.8	88.9	88.9	96.9	96.9	96.9	96.9	99.4	99.3	99.3	99.3
Approximation	–	88.8	88.8	88.8	–	96.9	96.9	96.8	–	99.4	99.3	99.3
out-of-sample	5%				10%				20%			
	unreg.	ridge	lasso	HS	unreg.	ridge	lasso	HS	unreg.	ridge	lasso	HS
Exact	92.7	92.7	92.8	92.9	97.7	97.7	97.8	97.8	99.5	99.3	99.4	99.4
Approximation	–	90.4	91.4	88.2	–	96.6	97.8	95.2	–	99.6	99.4	99.6

Table 3: Predictive performance of the full Bayesian regularized relational event models and their ABR simplifications. The results reflect the percentages of observed events that belong to the top 5%, the top 10%, and the top 20% of most likely event according to the fitted models using the full posterior for making predictions. 'HS' denotes the results of the horseshoe prior.

penalized using a L_1 norm. Bayesian alternatives of the graphical lasso are also available for continuous (Gaussian) dependent variables (Wang, 2012).

In this section we consider data of 221 people with a subthreshold posttraumatic stress disorder (PTSD) diagnosis. The network features 20 PTSD symptoms implying 190 possible conditional dependencies. A detailed description of the dataset can be found in Armour et al. (2017). These data were fitted using an unregularized Bayesian model using a noninformative Jeffreys prior on the precision matrix, the Bayesian glasso using the `BayesianGLasso` package in R (Trainor & Wang, 2022), and the ABR methods using the lasso and the horseshoe prior. For the ABR method, the MLEs of the off-diagonal elements of the precision matrix were taken (which quantify the conditional dependence among the 20 variables) and the corresponding error covariance matrix of these estimates were obtained using bootstrapping.

To assess the shrinkage behavior, it was checked how many 95% credibility intervals of the 190 off-diagonal elements of the precision matrix contained 0. To assess the predictive performance, leave-one-out cross-validation was performed where the data was split in a training set consisting of 220 observations and a test set consisting of 1 observation. For each trained model, 5,000 posterior draws were obtained for the unknown parameters, which were used to predict each variable of the test set given all other variables of the test set from which the mean squared error was determined. This resulted in 221 mean squared errors for all 221 observations. The distribution of these mean squared errors can be found in Figure 6. Moreover, Table 4 summarizes the 2.5%, 50%, and 97.5% quantiles of the errors. The figure and table show that all three regularized solutions result in tremendous improvements regarding the leave-one-out prediction errors in comparison to the unregularized solution. Moreover, we can see that the range of the mean squared prediction errors across the three regularization methods is very similar with slightly lower errors for the full Bayesian glasso method. Interestingly we see that the ABR solutions results in only 11 and 9 significant conditional dependencies (according to the 95% credibility intervals) out of 190 parameters which is considerably less than the unregularized solution (28 significant partial correlations) and the full Bayesian glasso (33 significant partial correlations).

Table 4: Quantiles of 221 leave-one-out mean squared prediction errors based on 221 observations consisting of 190 conditional dependencies and the number of 'significant' conditional dependencies.

	2.5%-quantile	50%-quantile	95%-quantile	number of 95%-CIs not containing 0
unregularized	1.321	15.1796	148.85	28
full Bayesian glasso	1.033	1.649	3.233	33
ABR lasso	1.098	1.798	3.051	11
ABR horseshoe	1.097	1.772	3.167	9

These results indicate that ABR can result in comparable predictive behavior as the full tailored regularization algorithm but resulting in much more parsimonious solutions.

4.5 Restricted factor analysis - Detecting measurement bias

We now illustrate the usefulness of the ABR method for structural equation models (SEMs). Specifically, we replicate the analysis of Liang & Jacobucci (2019) who used a restricted factor analysis model to detect measurement bias in 19 psychological tests administered to 7th and 8th grade students in two schools (Holzinger & Swineford, 1939). The tests aim to measure four correlated aspects of mental ability: spatial, verbal, speed, and memory. Uniform measurement bias with respect to age was assessed by regressing age on the 19 observed indicators. By regularizing the path coefficients from age to the indicators, we aim to detect substantial effects indicating measurement bias. The data set was obtained from the `psychTools` package and all code to reproduce the analysis is available at `https://github.com/sara-vanerp/ApproxBR`.

We compare the ABR method with a ridge and horseshoe prior to a classical regularized algorithm with a ridge penalty available in `regsem` (Jacobucci, 2023). For the classical ridge implementation, we use similar default settings as in Liang & Jacobucci (2019).

Figure 7 compares the posterior mean (circles) and mode (triangles) estimates and 95% credible intervals across priors and algorithms. The results for the classical regularized ridge algorithm are not shown because the optimal penalty parameter as chosen via cross-validation based on the BIC criterion in `regsem` was zero, such that the results are the same as the unregularized solution. All methods indicate the presence of measurement bias for certain indicators. For small effects, results are very similar for both the regularized and unregularized solutions, although the confidence interval for the unregularized solution is slightly wider compared to the regularized credible intervals and the posterior mode for the horseshoe prior is virtually zero for small effects. For larger effects, the shrinkage priors result in more shrinkage. In sum, the horseshoe prior leads to the most parsimonious solution without sacrificing predictive power as can be seen from the PMSEs in Table 5.

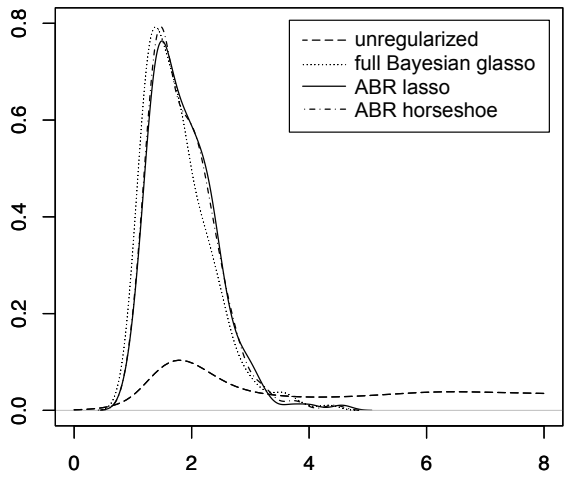

Figure 6: Distribution of 221 leave-one-out mean squared prediction errors using either the exact Bayesian glasso, the approximate Bayesian regularized (ABR) lasso, and ABR horseshoe, and the using unregularized solution.

Table 5: Prediction mean squared error for the different priors and algorithms in the measurement bias application

Unregularized	Classical ridge	Approximate ridge	Approximate horseshoe
0.95	0.95	0.95	0.95

4.6 Mediation analysis - Methylation data

We illustrate the usefulness of the algorithm in the context of exploring relevant mediators using data from (Houtepen, Vinkers, Carrillo-Roa, Hiemstra, van Lier, et al., 2016). The goal is to identify which locations in the genome mediate the relation between childhood trauma and stress reactivity at a later age. An advantage of the proposed Bayesian regularization method for a mediation analysis is its ability to specify a shrinkage prior on a function of parameters, such as the product of two effects to quantify the indirect effects while treating the direct effects as a nuisance which are integrated out and thus are not regularized.

The data set can be downloaded from the repository of the European Bioinformatics Institute[4] and consists of 85 healthy individuals. The independent variable childhood trauma exposure was measured using the short version of the Childhood Trauma Questionnaire (CTQ). The dependent variable stress reactivity was based on the increase of cortisol after administering the Trier Social Stress Test (TSST). A total of 385 882 DNA

[4] `https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-GEOD-77445`

Table 6: Prediction mean squared error for the different priors and algorithms in the mediation application

Unregularized	Approximate ridge	Approximate horseshoe
0.91	0.90	0.90

methylation loci were considered as possible mediators. We take the same preprocessing steps as in (van Kesteren & Oberski, 2019b) who select the top 1000 potential mediators based on their absolute product of correlations with the independent and dependent variable. However, since the ABR algorithm requires fitting the unregularized mediation model as a first step, we cannot consider more potential mediators than observations. We therefore select the top 45 potential mediators following the same approach as (van Kesteren & Oberski, 2019b) and we additionally include the five mediators that they selected based on the Coordinate-wise Mediaton Filter (CMF). In addition to centering the independent variable and potential mediators, we scale them to ensure a similar influence of the shrinkage priors on all indirect effects. To avoid extreme differences in the scales of the dependent and independent variables, we also scale the dependent variable by a factor 100. We split the data in a 90% training and 10% test set and compare the ABR algorithm with the unregularized solution using uninformative priors in `blavaan` (Merkle et al., 2021). To obtain the estimates and error covariance matrix for the indirect effects, we first run the mediation model in `blavaan` using uninformative priors and subsequently multiply the relevant posterior draws of the two relevant effects to obtain posterior draws for the indirect effects. This way, the ABR method can regularize the indirect effects directly instead of regularizing the direct effects separately, as is the case in `blavaan`. All code to reproduce the analysis is available at `https://github.com/sara-vanerp/ApproxBR`.

Figure 8 compares the posterior mean (circles) and mode (triangles) estimates and 95% credible intervals across priors and algorithms. It can be seen that the results are very similar across the horseshoe and ridge priors. Note how the credible intervals for the unregularized solution overlap with zero for every indirect effect and are much wider compared to the regularized solutions which pull all effects to zero with more certainty. As can be seen in Table 6, the average prediction error did not differ substantially between algorithms.

5 Discussion

Statistical regularization is a leading technique when working with statistical models with a large number of parameters out of which many may be zero. These regularization techniques aim to obtain parsimonious solutions by shrinking small, negligible effects to zero while leaving large, important parameters largely unaltered. Currently, regularization algorithms are mainly available for specific and rather standard modeling designs. Moreover, most of these regularization algorithms rely on classical penalized techniques even though Bayesian alternatives are known to avoid certain limitations of their classical counterparts (such as the overestimation of statistical certainty). To address these shortcomings, the

current paper presented a generally applicable method for approximate Bayesian regularization (ABR). As input, only a vector of unregularized (standard) estimates is required together with its error covariance matrix. Subsequently, the estimates are shrunk according to a specific prior. Various numerical illustrations showed that the method often behaves comparable as their true counterparts although sometimes there are very slight differences. These true counterparts however are only applicable for specific designs and parameters while the ABR method is generally applicable for any model and any set of parameters with known errors.

The ABR method is readily available using the R package `shrinkem`. Currently, a Gaussian (ridge), a Laplace (lasso), and a horseshoe prior are implemented. The choice of the specific prior for a given data set will depend on the specific application. By applying different priors, users can choose which solution is preferred depending on the interpretability (parsimony) of the solution and/or depending on the predictive performance of the solution for the data at hand.

A limitation of the method is that it may not be usable when the sample size is smaller than the number of parameters, i.e., $p > n$ (a scenario where regularization are commonly used), as unregularized estimated with a positive definite error covariance matrices would not be available. The illustrations in this paper showed however that also in the case $n > p$, (a scenario that is most common in social science research), regularization solution generally result in better predications than unregularized estimates while providing a more interpretable solution by shrinking unimportant effects to zero.

Moreover, the difference between the shrinkage behavior of ABR and its exact counterpart will depend on the accuracy of the Gaussian approximation. For example, in the Gaussian graphical model the integrated likelihood of the off-diagonal elements of the precision matrix will be skewed, which explains possible differences in shrinkage behavior between the approximation and the exact method. Interestingly, ABR resulted in considerably more parsimonious solutions (based on the 95%-CIs) while the predictive performance was only lower with a very slight degree in this application. Thus even if the induced shrinkage behavior is different, ABR can still result in useful results. Also note that other well-known methods which also rely on Gaussian approximations, e.g., the Wald test or the Bayesian information criterion (BIC), also give useful results in the case of deviations from normality. Still, it will be useful to study (theoretical) properties of the accuracy of ABR. Moreover, other (computationally efficient) multivariate distributions could also be considered for approximating the marginalized likelihood instead of the multivariate Gaussian distribution.

Finally, the literature on priors for Bayesian regularization goes well beyond the priors that were considered in this paper, such as the Bayesian elastic net (Li & Lin, 2010; Bornn et al., 2010), the spike-and-slab prior (George & McCulloch, 1993; Ishwaran & Rao, 2005), the adaptive lasso (Zou, 2006b), and spike-and-slab lasso (Ročková & George, 2018). It is relatively straightforward to apply ABR with these more advanced priors resulting in possible better solutions depending on the application. We leave these topics for future research.

Acknowledgements

This research was supported by an ERC Starting Grant 'TIMEISNOW' (758791) to DK, RL, and JM and by a NWO Veni Grant (Vl.Veni.221G.005) to SvE.

References

Armour, C., Fried, E. I., Deserno, M. K., Tsai, J., & Pietrzak, R. H. (2017). A network analysis of dsm-5 posttraumatic stress disorder symptoms and correlates in us military veterans. *Journal of anxiety disorders*, *45*, 49–59.

Azmak, O., Bayer, H., Caplin, A., Chun, M., Glimcher, P., Koonin, S., & Patrinos, A. (2015). Using big data to understand the human condition: the kavli human project. *Big data*, *3*(3), 173–188.

Bornn, L., Gottardo, R., & Doucet, A. (2010). Grouping priors and the Bayesian elastic net. *arXiv preprint arXiv:1001.4083*. Retrieved from `https://arxiv.org/abs/1001.4083`

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). doi: 10.18637/jss.v076.i01

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

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480.

Casella, G., Ghosh, M., Gill, J., & Kyung, M. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian analysis*, *5*(2), 369–411.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, *32*(2), 407–499.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, *1*(3), 515–534.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis (vol. 2).* Taylor & Francis Boca Raton.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881. doi: 10.2307/2290777

Gomez-Cravioto, D. A., Diaz-Ramos, R. E., Hernandez-Gress, N., Preciado, J. L., & Ceballos, H. G. (2022). Supervised machine learning predictive analytics for alumni income. *Journal of Big Data*, *9*(1), 1–31.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). Statistical learning with sparsity. *Monographs on statistics and applied probability*, *143*, 143.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67. Retrieved from `https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634` doi: 10.1080/00401706.1970.10488634

Holzinger, K. J., & Swineford, F. (1939). A study in factor analysis: The stability of a bi-factor solution. *Supplementary educational monographs*.

Houtepen, L. C., Vinkers, C. H., Carrillo-Roa, T., Hiemstra, M., Van Lier, P. A., Meeus, W., ... others (2016). Genome-wide dna methylation levels and altered cortisol stress reactivity following childhood trauma in humans. *Nature communications*, *7*(1), 1–10.

Houtepen, L. C., Vinkers, C. H., Carrillo-Roa, T., Hiemstra, M., van Lier, P. A., Meeus, W., ... Boks, M. P. M. (2016, mar). Genome-wide DNA methylation levels and altered cortisol stress reactivity following childhood trauma in humans. *Nature Communications*, *7*(1). doi: 10.1038/ncomms10967

Hsiang, T. C. (1975). A bayesian view on ridge regression. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *24*(4), 267–268. Retrieved from `http://www.jstor.org/stable/2987923`

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773. doi: 10.1214/009053604000001147

Jacobucci, R. (2023). regsem: Regularized structural equation modeling [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=regsem (R package version 1.9.5)

Karimova, D., Leenders, R. T. A., Meijerink-Bosman, M., & Mulder, J. (2023). Separating the wheat from the chaff: Bayesian regularization in dynamic social networks. *Social Networks*, *74*, 139–155.

Korobilis, D. (2013). Hierarchical shrinkage priors for dynamic regressions with many predictors. *International Journal of Forecasting*, *29*(1), 43–59.

Li, Q., & Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis*, *5*(1), 151–170. doi: 10.1214/10-ba506

Liang, X., & Jacobucci, R. (2019, December). Regularized structural equation modeling to detect measurement bias: Evaluation of lasso, adaptive lasso, and elastic net. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(5), 722–734. Retrieved from http://dx.doi.org/10.1080/10705511.2019.1693273 doi: 10.1080/10705511.2019.1693273

Marquardt, D. W., & Snee, R. D. (1975). Ridge regression in practice. *The American Statistician*, *29*(1), 3–20.

Merkle, E. C., Fitzsimmons, E., Uanhoro, J., & Goodrich, B. (2021). Efficient bayesian structural equation modeling in stan. *Journal of Statistical Software*, *100*(6). Retrieved from http://dx.doi.org/10.18637/jss.v100.i06 doi: 10.18637/jss.v100.i06

Mulder, J., & Pericchi, L. R. (2018). The matrix-F prior for estimating and testing covariance matrices. *Bayesian Analysis*, *13*(4), 1193–1214.

Park, T., & Casella, G. (2008a). The Bayesian Lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

Park, T., & Casella, G. (2008b). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. doi: 10.1198/016214508000000337

Perry, P. O., & Wolfe, P. J. (2013). Point process modelling for directed interaction networks. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(5), 821–849.

Polson, N. G., & Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*(4), 887–902.

Redmond, M. (2011). *Communities and Crime Unnormalized.* UCI Machine Learning Repository. (DOI: https://doi.org/10.24432/C5PC8X)

Ročková, V., & George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, *113*(521), 431–444.

Stan Development Team. (2024). *RStan: the R interface to Stan.* Retrieved from `https://mc-stan.org/` (R package version 2.32.6)

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

Trainor, P., & Wang, H. (2022). *Bayesianglasso: Bayesian graphical lasso.* Retrieved from `https://cran.r-project.org/web/packages/BayesianGLasso/index.html` (R package version 0.2.0)

Van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50.

van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50. doi: 10.1016/j.jmp.2018.12.004

van Kesteren, E.-J., & Oberski, D. L. (2019a). Exploratory mediation analysis with many potential mediators. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 710–723.

van Kesteren, E.-J., & Oberski, D. L. (2019b, apr). Exploratory mediation analysis with many potential mediators. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 710–723. doi: 10.1080/10705511.2019.1588124

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, *20*(4), 595–601. Retrieved from `http://www.jstor.org/stable/2236315`

Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, *7*(4), 867–886.

Williams, D. R., & Mulder, J. (2020). Bayesian hypothesis testing for gaussian graphical models: Conditional independence and order constraints. *Journal of Mathematical Psychology*, *99*, 102441.

Wolfowitz, J. (1949). On wald's proof of the consistency of the maximum likelihood estimate. *The Annals of Mathematical Statistics*, *20*(4), 601–602. Retrieved from `http://www.jstor.org/stable/2236316`

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(1), 49–67.

Zou, H. (2006a). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418–1429.

Zou, H. (2006b). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. doi: 10.1198/016214506000000735

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, *67*(2), 301–320.

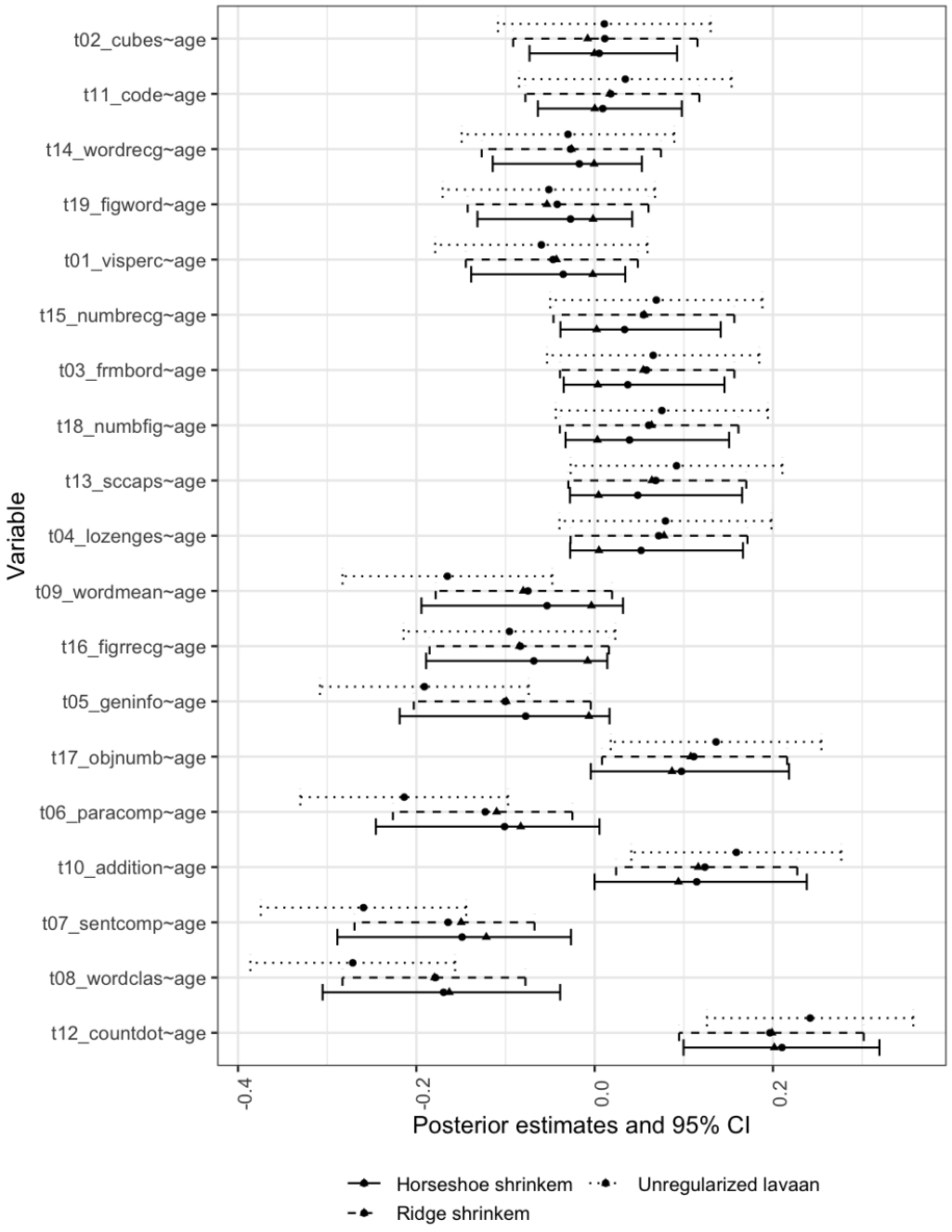

Figure 7: Posterior mean (circles) and mode (triangles) estimates and 95% credible intervals for the paths from age to the observed indicators in the measurement bias application using either the classical frequentist, approximate Bayesian, or exact Bayesian regularization algorithm with different priors and penalty functions.

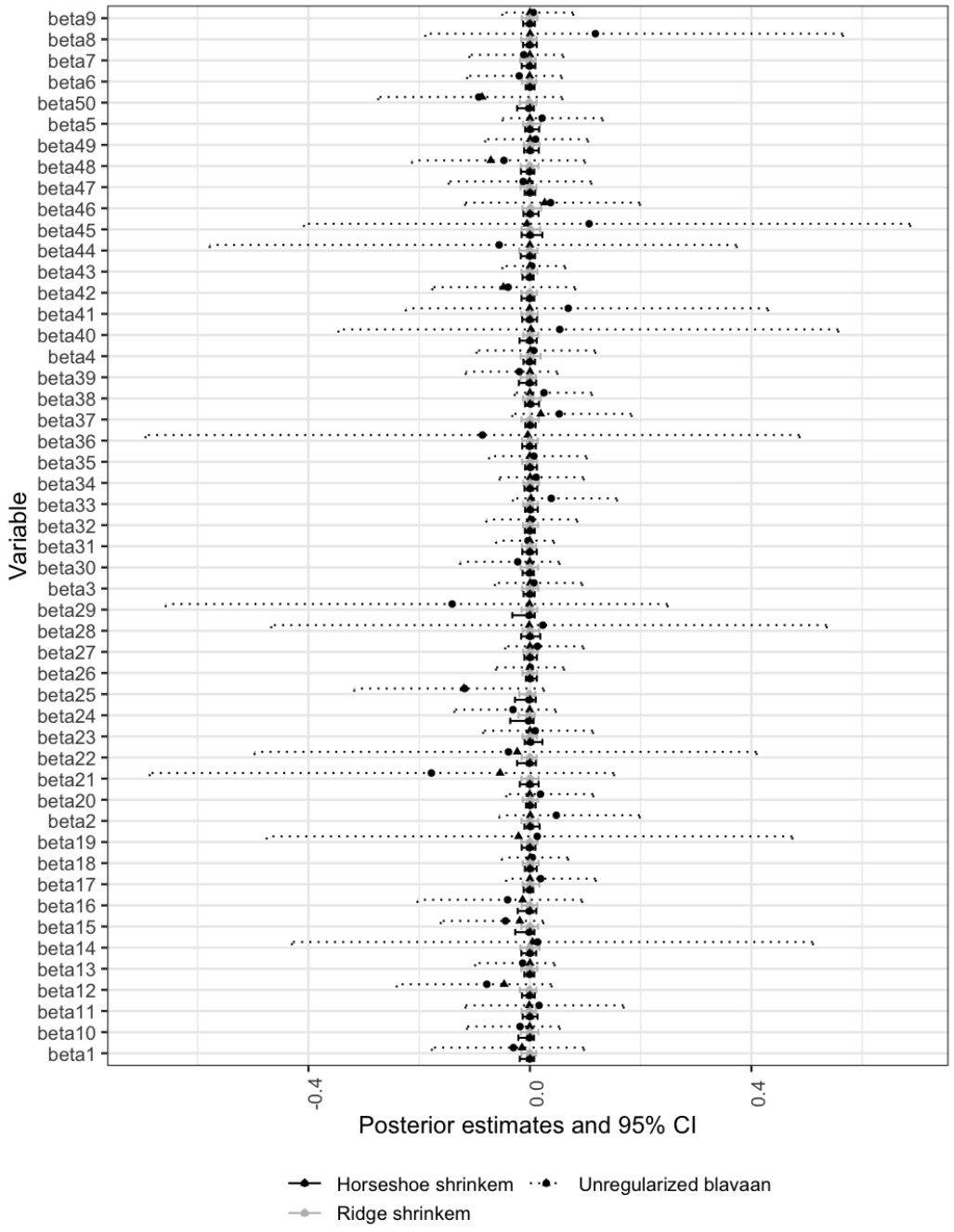

Figure 8: Posterior mean (circles) and mode (triangles) estimates and 95% credible intervals for the indirect effects in the mediation model using either the approximate Bayesian or exact Bayesian algorithm with different priors.