Bayesian global-local shrinkage methods for regularisation in the high dimension linear model

Jim E. Griffin
University College London

Philip J Brown*
University of Kent

20th January 2021

Abstract

This paper reviews global-local prior distributions for Bayesian inference in high-dimensional regression problems including important properties of priors and efficient Markov chain Monte Carlo methods for inference. A chemometric example in drug discovery is used to compare the predictive performance of these methods with popular methods such as Ridge and LASSO regression.

1 Introduction

This paper builds on a body of Bayesian approaches to variable selection and regularisation in regression from the Statistics literature and compares their predictive performance using some chemometric data used in drug discovery. Chemometrics and variable selection brought one of us (PJB) into close collaboration with Cliff Spiegelman in the late 1980's with exchange visits to Texas A&M, USA and Liverpool University, UK. Two joint papers Brown and Spiegelman (1991) and Brown et al. (1991) resulted, and periodic visits to Texas, following Marina Vannucci's appointment at Texas A&M. For an overview of this regression variable selection work see chapter 7 of Brown (1993). A paper in an issue in honour of Professor Luc Massart (Vannucci et al., 2005) was commissioned by Cliff.

*P. J Brown was supported in this work by a Leverhulme Emeritus Fellowship EM-2018-059/9.

By way of setting the scene, we will describe a few recent developments in the Bayesian Statistics literature and apply one of these to some chemometric drug discovery data arising from collaborations with computational chemists at GlaxoSmithKline during the last decade and also involving Professor I. Poli's group at the European Centre for Living Technology, Universita Ca'Foscari, Venice.

In chemometrics some Bayesian applications have emerged, for example in calibration, see Fearn et al. (2010) and in classification Fearn et al. (2019), but we will focus more on the area of variable selection in regression, especially when the number of explanatory variables is large, of the order of several thousand and when the number of observations is quite modest, numbering in tens or hundreds. This area has mushroomed in the last decade and a half. We will consider some examples of prior distributions which are sufficiently flexible to be useful for a variety of examples. For the Bayesian, the prior distribution assumed is fundamental and inference amounts to combining it with the data likelihood to form an 'after the data' or posterior probability distribution.

One of the earliest ways to cope with ill-conditioned and highly correlated regression data was developed in the early 1970's and assumes a spherical independent normal prior distribution on the regression coefficients, the ridge prior, see for example Marquardt (1970). The aim here was not so much variable selection as regularisation for achieving good predictions.

Vannucci et al. (2005) considered chemometric NIR and mid IR data and used parametrized curves made up of wavelets in several component multi-compositional setting. This lead to a regression problem where the wavelet bases act as regressors. The key idea for variable selection is that each regression coefficient can be present, with some regularisation through a ridge or other continuous prior, or is set to zero. But because a priori one doesn't know which coefficients should be set to zero there is a mixture prior with a spike of probability at zero leading to the so-called 'slab and spike' prior. Implementation is fairly straightforward with iterative Markov chain Monte Carlo (MCMC) techniques. At first sight, it is perhaps surprising that MCMC can effectively search such a vast space of possibilities (2^p, with a large number, p of regression coefficients) but Yang et al. (2016) show why pessimism is not warranted.

However, there are drawbacks to the 'slab and spike' approach, both in convergence issues with MCMC samplers, a task that becomes more difficult for very high dimensional problems, and in *fidelity of inference*, that is the ability to identify the right active coefficients and assign valid confidence sets, as developed in Van der Pas et al. (2016). The computational issues in awkward high dimensional problems have been mitigated by Griffin et al. (2020), in effect learning about promising di-

rections of search. Fidelity issues may be improved by getting away from the 'slab and spike' mixture prior to a 'one group' prior which is flexible enough to form a spike or leave the coefficient regularised depending on the information in the data. A computational environment STAN, see Gelman et al. (2014) has been developed and is continually evolving and based on Hamiltonian MCMC which enriches the MCMC approach by allowing the inclusion of directions of search and a type of conjugate gradients exploration with similarities to PLS iterative methods popular in chemometrics.

An alternative approach to Bayesian variable selection is to use a utility function and decision theory formulation. Two-stage approach to this uses a baseline weak regularising prior with the posterior then projected on to submodels defined by omitting covariates, so that predictions change as little as possible, see Lindley (1968), Goutis and Robert (1998), Brown et al. (2002), and Piironen et al. (2020).

When we talk of high dimensions and relatively little data, obviously many inferential issues will remain vague and uncertain, the challenge is to find likely submodels for which there is some hope of reflecting what might happen in future. Aside from such inferential issues there are technical aspects for getting fast algorithms that can reveal such fidelities as described in section 4.

2 The regression model and prior distributions

The linear regression model is often written as

$$Y = X\beta + \epsilon \tag{1}$$

where Y is a $(n \times 1)$-dimensional vector of responses, X is a $(n \times p)$-dimensional matrix of explanatory variables, β is a $(p \times 1)$-dimensional vector of regression coefficients and $\epsilon \sim \mathrm{N}(0, \sigma^2 I_n)$. Broadly we assume a scale mixture of normals formulation for the prior distibution of the regression parameters so that $\beta \sim \mathrm{N}(0, \Psi)$ with idiosyncratic hyperparameters $\Psi = \mathrm{diag}(\psi_1, \ldots, \psi_p)$ and then $\psi_i \overset{i.i.d.}{\sim} G$ for some distribution G. The properties of the posterior distribution are determined by the choice of hyperprior G. Typically G will involve both local and global hyperparameters and their distributions will characterise the behaviour of the resultant posterior distribution. One further dichotomy is whether the prior for the regression parameters is structured to be *conjugate* or not. In a conjugate formulation $\beta \sim \mathrm{N}(0, \sigma^2\Psi)$ so that when the likelihood from the regression model (1) is multiplied by the prior distribution the two blend together and the posterior distribution is of the same form as the prior. It is as if the prior can be thought of as pseudo data from the model. The

appeal is one of automatic scaling, if you change the scale of the response Y then inference will remain the same. This will not be the case for a non-conjugate formulation but in the conjugate formulation with number of explanatory variables p very large, there can be overfitting and poor estimates of σ^2, see Moran et al. (2019).

With the specification of normal prior and hyper priors, all that needs now to be done in the Bayesian paradigm is to multiply likelihood by prior distributions and organise into an appropriate form for MCMC sampling.

3 Global-Local Priors

In Bayesian analysis of the linear regression model the prior G is often explicitly structured into a set of local idiosyncratic parameters and a global parameter. There has been increasing interest in the use of global-local priors for regression coefficients, (see Bhadra et al., 2019, for a comprehensive review).

Each regression parameter has its own ideosyncratic scale, ψ_j for the jth regression coefficient, and there is also a global (across all regression coefficients) scale parameter which we will designate as τ. In notation with usual conventions, the conjugate global-local model for the regression coefficients, $\beta_j, j = 1 \ldots, p$ is

$$\beta_j \sim \mathrm{N}(0, \sigma^2\psi_j^2\tau^2), \qquad \psi_j \sim f, \qquad \tau \sim g, \tag{2}$$

and the non-conjugate global-local model is

$$\beta_j \sim \mathrm{N}(0, \psi_j^2\tau^2), \qquad \psi_j \sim f, \qquad \tau \sim g, \tag{3}$$

where f and g are general probability densities over the positive real line. It is specification of f and g which allows the rich class of different prior distributions. We will write $\Psi = (\psi_1, \ldots, \psi_p)$ and consider a few promising representatives in our chemometric example. The prior aims to shrink out small less important regression coefficients whilst leaving largely untouched important large coefficients, a hard act for all occasions! Some reassurance in this direction is provided by Van der Pas et al. (2016) who provide a general treatment of posterior consistency for these priors when the number of variables diverges to infinity and Ghosh et al. (2016) who consider asymptotic properties of Bayes risk.

3.1 Horseshoe prior and variants

The horseshoe prior, Carvalho et al. (2010) uses the choice

$$\psi_j \sim \mathcal{C}^+(0, 1), \tag{4}$$

where $\mathcal{C}^+(0, a^2)$ denotes a Cauchy distribution with location zero and scale a truncated to the positive real line, which is called the half-Cauchy distribution (Gelman, 2006). This popular prior aims to achieve minimal shrinkage of important large coefficients whilst shrinking out smaller regression coefficients. This is achieved by assuming a half-Cauchy prior on the local scales used to construct the prior distribution and as such it has had good successes but can overshrink large coefficients when the data likelihood is weak. We will use it to investigate the chemometric problem in section 5. The fidelity aspects have been investigated by Van der Pas et al. (2014). Their paper assumes τ is known. One variant puts a hyperprior on τ, and Van der Pas et al. (2017) investigate whether this affects the good minimax shrinkage properties of the horseshoe. Another variant due to Piironen and Vehtari (2017) seeks to regularise this horseshoe, which is called the regularized horseshoe or the 'Finnish horseshoe', so as to make it less severe when information is weaker as for example in logistic regression for binary data. This introduces an extra hyperparameter c and uses the hierarchy

$$\psi_j^2 = \frac{c^2\,\tilde{\psi}_j^2\tau^2}{c^2 + \tau^2\tilde{\psi}_j^2}, \qquad \tilde{\psi}_j \sim \mathcal{C}^+(0,1),. \tag{5}$$

This choice implies that the prior variance of β_j is in $(0, c^2)$. The prior variance controls the amount of shrinkage and the regularized horseshoe puts a lower limit on the amount of shrinkage (unlike the horseshoe where the prior variance is in $(0, \infty)$). The regularized horseshoe converges to the horseshoe as $c/\tau \to \infty$.

3.2 Normal-gamma prior

Griffin and Brown (2010) propose a prior where the distribution G has gamma density function $\text{Ga}(x|\lambda, b) \propto x^{\lambda-1}\exp\{-bx\}$ with expectation λ and λ is the shape. Thus we have

$$\psi_j^2 \sim \text{Ga}(\lambda, 1).$$

It forms a natural extension of the lasso which implicitly applies an exponential prior and is popular in finance for long tailed distributions. It has exponential tails rather than the polynomial tails of the horseshoe. The error variance σ^2 can also be incorporated to form a conjugate prior. This prior is investigated for 'fidelity' of inference in Van der Pas et al. (2016) and is shown to be able to achieve the optimal minimax rate of convergence to truth. By 'fidelity' of inference the authors are aiming for two things, *recovery* of the true underlying regression vector and secondly *uncertainty quantification* .

3.3 Normal-gamma-gamma distribution

A further elaboration of the normal-gamma provides a distribution with much fatter tails

$$\psi_j^2 \sim \text{Ga}(\lambda, \gamma_j), \qquad \gamma_j \sim \text{Ga}(c, 1). \tag{6}$$

Thus $\psi_j^2 \sim \text{GG}(\lambda, c, \tau)$ a gamma-gamma distribution. The shape parameter λ controls the behaviour close to zero, whereas shape c controls the behaviour in the tails, with $\lambda = c = 1/2$ being the horseshoe. It thus offers more flexibility than the horseshoe. It was introduced in Griffin and Brown (2017) when wishing to shrink interaction parameters differently from main effects. It considered levels of coefficients, corresponding to main effects, first order interactions and higher order interactions which are usually more numerous and need more shrinkage. Some order on the levels of shrinkage can be imposed by the notion of *heredity*, whereby in *strong* heredity both main effects should be important for the interaction to be important. *Weak* heredity on the other hand requires just one of the two main effects to be important for the interaction to be important. See Cadonna et al. (2020) for a comprehensive review of this prior and other uses of this structure in the literature.

3.4 Dirichlet-Laplace distribution

Like the normal-gamma this prior avoids the very fat-tailed distributions of the normal-gamma-gamma and the horseshoe. It modifies the exponential lasso prior differently from the normal-gamma, chopping up the regression parameter into random chunks as determined by a Dirichlet distribution on the p-dimensional simplex.

$$\beta_j|\phi, \tau \sim \text{DE}(\phi_j, \tau), \phi \sim \text{Dir}(a, \ldots, a), \tau \sim \text{gamma}(pa, 1/2) \tag{7}$$

where $\text{DE}(\phi_j, \tau)$ is a double exponential (i.e. Laplace) distribution with pdf $f(\beta_j) = (2\tau)^{-1} \exp -|\beta_j - \phi_j|/\tau$, and Dir(a,. . . , a) is a Dirichlet distribution over the simplex with equal parameter values. Typically the parameter a is prespecified or given a prior distribution. This Dirichlet-Laplace distribution is introduced and motivated in Bhattacharya et al. (2016). See Zhang et al. (2020) for an informative prior distribution on model fit with a similar structure.

4 Fast algorithm and vague priors

Computational speed becomes paramount in high dimensional Bayesian analysis. A number of tricks have been introduced and we consider ones described by Bhattacharya et al. (2016) and Makalic and Schmidt (2016), the latter for the horseshoe

prior. We will employ a clever insight by Makalic and Schmidt (2016) who re-parameterized the prior in (4) so that the MCMC updating scheme involves sampling ψ_j from inverse gamma distributions rather than the more problematic generalised inverse Gaussian which would be necessary under the original parameterisation. Naive sampling from $\beta|\Psi, \sigma^2$ can be computationally prohibitive if p is large since the standard form of the posterior distribution involves inversion of the matrix $X^T X + \Psi^{-1}$ which will be challenging unless $X^T X$ has special structure. If n is much smaller than p, Bhattacharya et al. (2016) develop an algorithm for sampling $\beta|\Psi, \sigma^2$ which avoids inverting the $(p \times p)$-dimensional matrix $X^T X + \Psi^{-1}$. Their algorithm incorporates a data augmentation step

1. Sample $u \sim \mathrm{N}(0, \Psi)$ and $\delta \sim \mathrm{N}(0, I_n)$ independently.
2. Set $v = Xu + \delta$.
3. Solve $(X\Psi X^T + I_n)w = Y - v$.
4. Set $\beta = u + \Psi X^T w$.

This involves solving an n-dimensional set of linear equations in step 3, a far quicker task than matrix inversion. We refer to this as the "Fast algorithm".

In some practical problems, we may not want to use the same prior for all variables but allow different priors on exclusive subsets of the variables. If all priors are proper, the fast algorithm can be used but there may be cases where we wish to use a non-informative prior for regression coefficients associated with some "key" explanatory variables. There is a long tradition of using unnormalisable 'vague' priors, see for example the Jeffreys prior using invariance arguments, Jeffreys (1961), especially for parameters that are important and we don't wish to pre-judge in any probabilistic way. In many contexts there may be a few such important parameters. We could set variance to a large value in the algorithm but we have found in sensitive problems this leads to further instability. If an improper prior is used then u will have an improper distribution in Step 1 of the fast algorithm and the algorithm cannot be implemented. This is easily bypassed by removing the effect of the key variables from both X and Y. Let Z denote the $n \times q$ matrix of carriers for the explanatory variables with vague priors, assumed few in number or singularites will appear. Then $G_Z = I_n - Z(Z^T Z)^{-1} Z^T$ is the orthogonal projection that removes the effect of these covariates and is idempotent. We just set $\tilde{X} = G_Z X, \tilde{Y} = G_Z Y$ and and apply the algorithm to $\tilde{X}$ and $\tilde{Y}$ (Chipman et al., 2001, section 3, last paragraph before section 3.1).

The posterior can be easily sampled using a Gibbs sampler which updates sets or blocks of parameters successively conditional on the other sets of parameters $\beta|\psi, \sigma^2$,

$\psi|\beta, \sigma^2$ and $\sigma^2|\beta, \psi$. Here there are three blocks and this could be called a three block sampler. In the case of a conjugate formulation however Pal et al. (2017) show that jointly updating β and σ^2 in the Gibbs sampler can lead to better theoretical properties for the resulting Markov chain. The joint sampling is achieved by sampling σ^2 from $\sigma^2|\psi$ and β from $\beta|\sigma^2, \psi$. This idea is extended to the more general class of global-local priors by Rajaratnam et al. (2019). This scheme is termed the 'two-block' sampler and we use this sampler in our examples for conjugate models.

In both the two-block and three-block sampler, the computational complexity of the fast algorithm will scale linearly in p for large enough p whereas the naive algorithm will scale polynomially in time (with the exact type of scaling determined by the method used to solve the system of linear equations in Step 3). This point and the potential for substantially faster algorithms in small n, large p settings are comprehensively illustrated in Bhattacharya et al. (2016).

5 Drug Discovery Data

The original experimental data generated was analysed by Pickett et al. (2011). The data were intended to be typical of data arising in the process of *lead optimisation* in drug development, where a promising compound (the *lead*) is improved by chemical modifications. In this application the compound can be modified at two sites, A and B, and 50 possible modifications (chemical reagents) were considered at each site. Thus A and B are the factors and the different reagents are the levels of the factors. The basic compound was an inhibitor and the aim was to synthesize all $50 \times 50 = 2500$ possible modifications and measure their inhibitory strength (pIC_{50}) by means of an assay. However, the 'complete' data matrix has 796 missing values (32%), mostly because the modified compound could not be synthesized (23%) or because it was found to be inactive (7%), but also because occasionally the assay failed (1%) or was not undertaken (1%). The same data was analysed by Borrotti et al. (2014). Optimal designs have been investigated and illustrated in Brown and Ridout (2016).

We will analyse data using a different molecular representation as constructed by Glaxo SmithKline by in house algorithms, see Hussain and Rea (2010) giving rise to different 'fragments'. In our case the full set of 1704 compounds could be represented by 3149 'attributes' (an attribute giving the presence or absence of a fragment). We have the full set of 1704 synthesised molecules each with an associated activity level ranging from 3.7 to 8, where high values are desirable. The model matrix X comprises a 1704 by 3149 matrix of zeros and ones identifying whether a

fragment is absent (0) or present (1). Whilst we are in the fortunate position of having the full set of 1704 activities, in practice assays are expensive and slow and GSK wanted to investigate 140 assayed from the 1704 available. Thus we can measure the effectiveness of different methods of prediction by comparing predictions and actual on the (1704 - 140)=1564 other compounds. To avoid peculiar or poor choice several random choices were made.

We compare the predictive performance of the different prior distributions described in Section 3 (in both conjugate and non-conjugate form) and the classical ridge and the lasso. The regularisation parameters in the latter two methods are selected using 10-fold cross-validation. In the Bayesian methods, we consider two priors for the observation variance σ^2: the vague prior $\pi(\sigma^2) \propto \sigma^{-2}$ and the half-Cauchy prior (Gelman, 2006). The hyperparameters are either fixed at pre-determined values or inferred from the data by giving hyperpriors to the hyperparameters. In order to compare the performance of the methods we generate 20 random cross-validation sets with 140 training samples and 1564 testing samples. We calculate the mean absolute error for all methods which is

$$\text{MAE} = \frac{1}{20}\frac{1}{N}\sum_{j=1}^{20}\sum_{i=1}^{N}\left|y_i^{(j)} - \hat{y}_i^{(j)}\right|$$

where $y_i^{(j)}$ is the i-th observation in the j-th testing set and $\hat{y}_i^{(j)}$ is the prediction of $y_i^{(j)}$ calculated using the j-th training set. This measure is less sensitive to the occasional outlier than mean square error. In the Bayesian methods, $\hat{y}_i^{(j)}$ is the posterior predictive median and, in the classical methods, $\hat{y}_i^{(j)} = x_i^{(j)}\hat{\beta}^{(j)}$ where $x_i^{(j)}$ are the observed variables for the i-th observation in the j-th testing set and $\hat{\beta}^{(j)}$ are the estimated regression coefficients calculated using the j-th training sample. This measures the accuracy of these point estimates. For the Bayesian methods, we also calculate the log predictive score, which measures the accuracy of the posterior predictive distribution as a density estimate. The log predictive score is

$$\text{LPS} = -\frac{1}{20}\frac{1}{N}\sum_{j=1}^{20}\sum_{i=1}^{N}\log p\left(y_i^{(j)}\middle|\, x_i^{(j)}, X_j^{(train)}, y_j^{(train)}\right)$$

where $X_j^{(train)}$ and $y_j^{(train)}$ are the design matrix and responses for observations in the j-th training sample. We exclude the lasso and ridge estimators from this measure since these estimators are only designed to provide point estimates.

The methods compared and the hyperprior choices are given below

- Horseshoe (HS) (fixed): $\tau = p_0/(p - p_0)$ where p_0 is a prior guess at the number of important variable and we set $p_0 = 3$ (Piironen and Vehtari, 2017).

- Horseshoe (HS) (hyperprior): $\tau(p-p_0)/p_0 \sim \mathcal{C}^+(0,1)$ and we set $p_0 = 3$ Piironen and Vehtari (2017).
- Regularized Horseshoe (RHS): $\tau(p-p_0)/p_0 \sim \mathcal{C}^+(0,1)$, $c^{-2} \sim \text{Ga}(2,8)$ and we set $p_0 = 3$ (Piironen and Vehtari, 2017).
- Normal-Gamma (NG): $\lambda \sim \text{Ga}(1, p/5)$ and $\tau \sim \mathcal{C}^+(0,1)$.
- Normal-Gamma-Gamma (NGG): $\lambda \sim \text{Ga}(1, p/5)$, $c \sim \mathcal{C}^+(0,1)$ and $\tau \sim \mathcal{C}^+(0,1)$.
- Dirichlet-Laplace: $a = 1/2$ (Bhattacharya et al., 2016).

	Conjugate				Non-Conjugate			
Method	Vague		Half-Cauchy		Vague		Half-Cauchy	
	MAE	LPS	MAE	LPS	MAE	LPS	MAE	LPS
DL	**0.34**	0.83	**0.34**	0.83	0.36	5.07	0.37	1.88
NG	0.36	0.82	0.36	0.82	0.37	>1000	0.36	0.82
HS	0.39	0.87	0.39	0.87	0.37	0.85	0.38	0.85
RHS	0.38	0.85	0.39	0.85	**0.35**	0.98	**0.35**	0.81
NGG	0.41	0.91	0.41	0.92	0.40	0.90	0.40	0.89
Ridge	0.46							
Lasso	0.49							

Table 1: MAE and LPS results for the regression with fragments only (bold best on MAE).

We consider two possible choices of variables. Firstly, a regression which only uses the 3149 fragments and the second regression which includes the effect of molecular weight (as a "fixed" variable), the effects of the 3149 fragments and the interaction between molecular weight and the fragments (which are given global-local shrinkage priors).

The prediction results for the regression with only the fragments are shown in Table 1. The Bayesian methods perform well. The conjugate Dirichlet-Laplace prior provides the smallest MAE. The MAE is about 31% lower for the Dirichlet-Laplace compared to the Lasso with a slightly smaller improvement compared to the ridge estimator. It is useful to divide the methods into two groups: priors with polynomial tails (HS, RHS and NGG) and priors with exponential tails (DL and NG). The conjugate setting outperforms the non-conjugate for the priors with exponential tails whereas the ordering is reversed for the priors with polynomial tails. The exponential tailed priors provide poor density forecasts (high LPS) if the non-conjugate prior

is used with a vague prior for σ^2. This is improved by using a half-Cauchy prior for σ^2 but the performance with the DL is still poor and so the use of a non-conjugate DL prior is not recommended (in contrast, the NG provides excellent performance). The poor density forecasting performance is due to underestimation of σ^2.

	Conjugate				Non-Conjugate			
Method	Vague		Half-Cauchy		Vague		Half-Cauchy	
	MAE	LPS	MAE	LPS	MAE	LPS	MAE	LPS
DL	**0.36**	0.90	**0.36**	0.90	0.43	7.19	0.43	2.56
NG	0.36	0.95	0.37	0.87	**0.36**	1.15	**0.37**	0.87
HS	0.45	1.00	0.45	0.99	0.42	0.94	0.43	0.95
RHS	0.44	0.96	0.44	0.97	0.41	0.92	0.41	0.91
NGG	0.43	0.97	0.44	0.98	0.42	0.95	0.43	0.96
Ridge	0.58							
Lasso	0.63							

Table 2: MAE and LPS results for the regression with molecular weight, fragments and the interactions of fragments and molecular weight (bold best on MAE).

The prediction result for the regression with molecular weight, fragments and the interactions of molecular weight and fragments are shown in Table 2. Here there are many explanatory variable, $p = 3149+3150 = 6299$. Again the Bayesian methods perform better than the classical methods. The Bayesian methods generally perform slightly worse with this data than with the regression on the fragments only (with the exception of the normal-gamma methods). Both the performance of the classical methods deteriorate but a much greater degree than the Bayesian methods. The improvement of the Dirichlet-Laplace prior over the Lasso is now 43% and slightly less than the ridge estimator. This suggests that the Bayesian methods are more robust to the inclusion of "noise" variables with little information value compared to the classical estimator (even though, the lasso is designed to eliminate variables with little predictive power). The relative performance of different Bayesian set-ups is similar to fragments only case.

We now consider inference on the regression coefficients using these priors on the full data set of 1704 observations. For each prior, the combination conjugate/non-conjugate set-up and prior for σ^2 which gives the smallest LPS is used. Figure 1 shows the posterior distribution of the regression coefficients summarized by their posterior median and 95% credible interval. All methods are able to identify a very

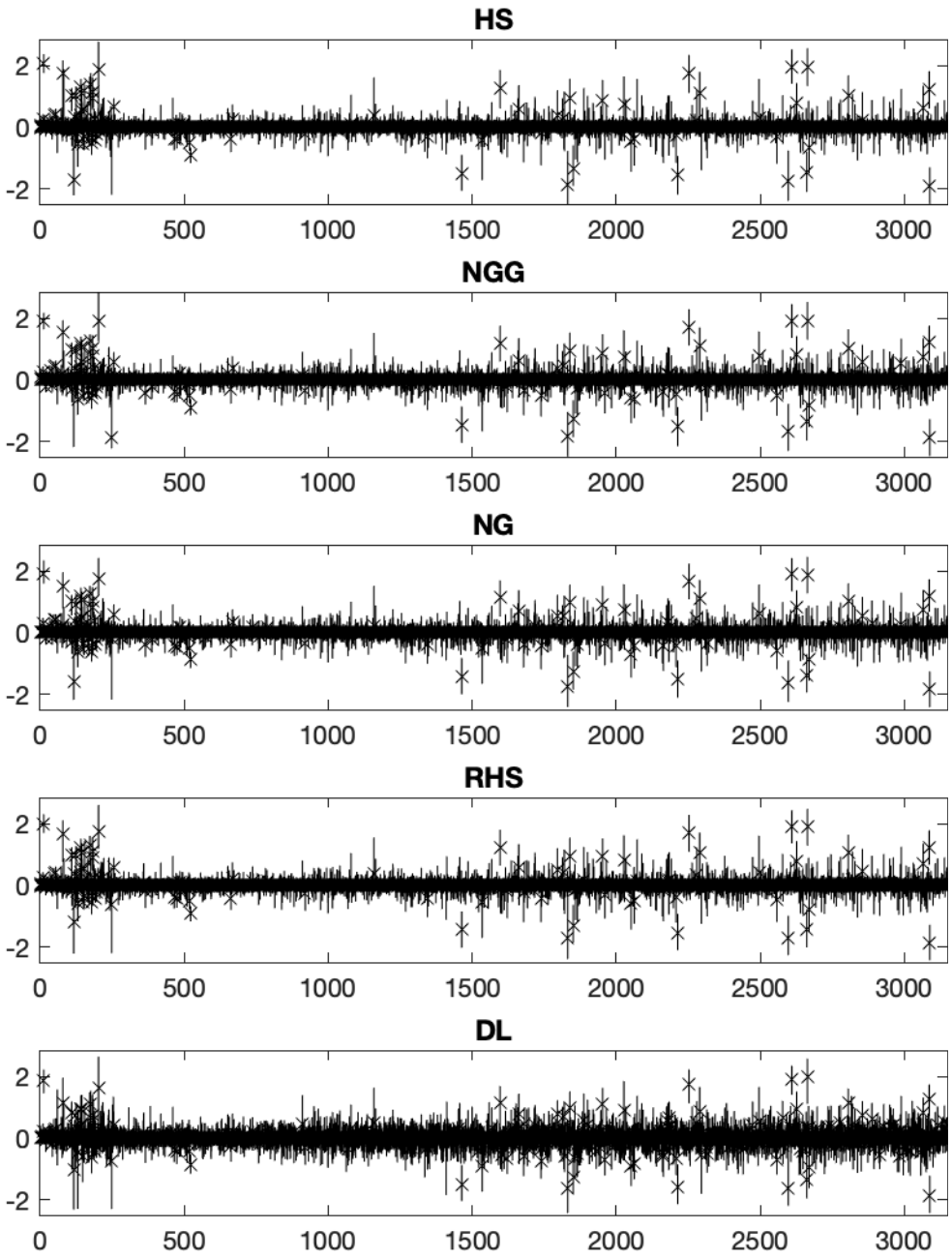

Figure 1: Fragments only, 1704 observations, the posterior distributions of the regression coefficients shown using the posterior median (cross) and 95% credible interval (line)

similar set of regression coefficients whose 95% credible interval do not cross zero which suggests the corresponding fragments have strong effects on activity. The

exponential tailed priors (DL and NG) tend to lead to slighter wider credible intervals for the other regression coefficients than the polynomial tailed priors (HS, RHS and NGG). The difference in predictive performance leads to the conclusion that the polynomial tailed priors are over-selecting regression coefficients (and so setting too many too close to zero). The DL and NG priors do a better job in this data of balancing selection and shrinkage. This illustrates the importance of the tail properties of the prior.

If we consider the regression model with molecular weight, fragments and interactions of molecular weight and interactions, the results for the main effects of fragments are shown in Figure 2 and results of the interactions between molecular weight and fragments are shown in Figure 3. The inference about the main effects is similar to the inference from the model without interactions but with larger credible intervals for a lot of regression coefficients. Interestingly, the NG and RHS priors provide inference which is closest to the inference in the model without interactions. The inference about the interaction shows that there is little evidence of substantial effects. The inference with the HS and DL priors shows that there is some evidence of sizeable effects for a few interactions. However, the corresponding posterior medians are very close to zero. This combined with the cross-validation results suggest that these interactions only worsen predictive performance.

6 Discussion

Global-local prior distributions offer effective alternatives to the more classical options of ridge and lasso regression. They give accurate predictions in the drug discovery example and offer estimates which can be examined by the usual array of inferential and display diagnostics. They can show the important variables in the linear model and are able to be scaled up to large problems of several thousand parameters.

In comparing different prior distributions the characteristic of importance is the tail behaviour and the spike at zero. Tails should be at least as fat as an exponential but not too fat. The normal distribution itself, leading to ridge shrinkage, has tails that are too thin and inference will be pushed towards vagaries in the data. In this chemometric study the two prior distributions that come out best are those with exponential tails (NG and DL), and not those with polynomial tails that tend to shrink out regression variables too readily in the case of HS and NGG as anticipated in Griffin and Brown (2011). The classical lasso and ridge do not fare well. Whilst the lasso has the fat exponential tail its classical formulation requires maximising

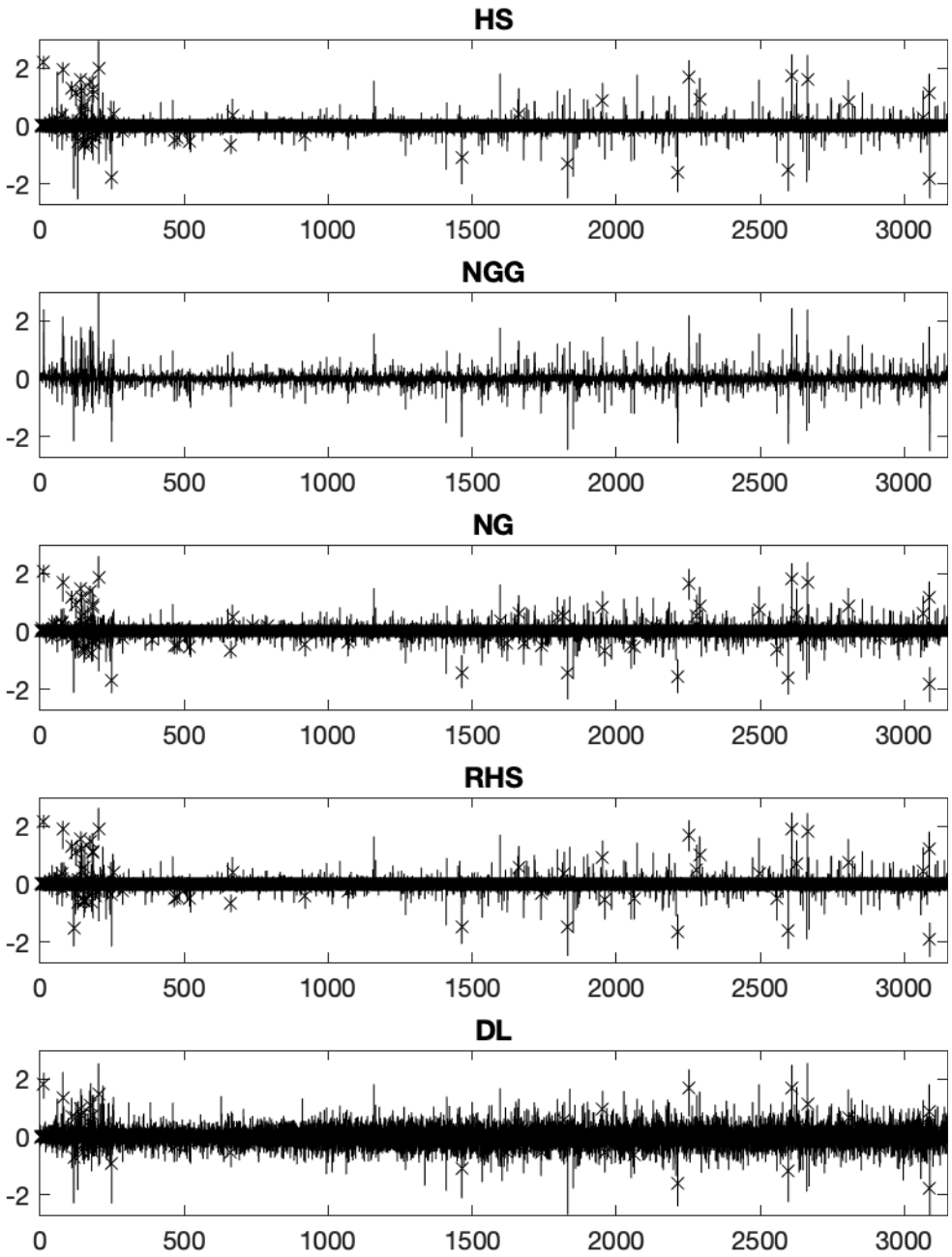

Figure 2: Full model (fragments plus MW interactions), 1704 observations, the posterior distributions of the fragments main effects using the posterior median (cross) and 95% credible interval (line)

rather than integration and such averaging is beneficial for inference.

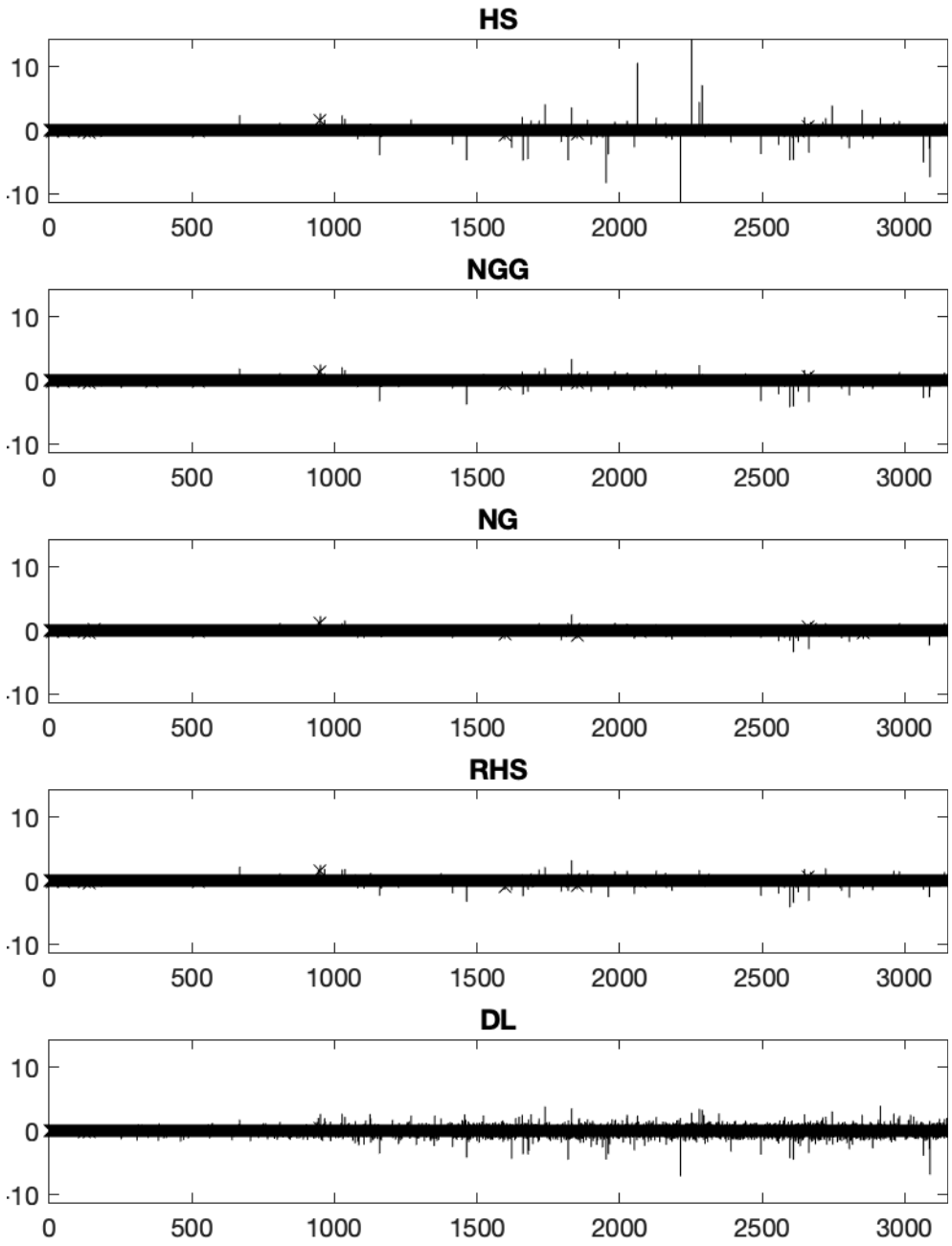

Figure 3: Full model (fragments plus MW interactions), 1704 observations, the posterior distributions of the interaction with MW coefficients using the posterior median (cross) and 95% credible interval (line)

Acknowledgement

We are grateful to Dr Darren Green of GSK for introducing us to this drug discovery data and for numerous interactions and advice from him and his group. We would

also like to thank Professor Irene Poli of ECLT, Venice for providing a welcoming forum for discussion of these problems and a rich research environment and numerous incisive discussions.

References

Bhadra, A., J. Datta, N. G. Polson, and B. T. Willard (2019). Lasso Meets Horseshoe : A Survey. *Statistical Science 34*, 405–427.

Bhattacharya, A., A. Chakraborty, and B. K. Mallick (2016). Fast sampling with Gaussian scale-mixture priors in high-dimensional regression. *Biometrika 103*, 985–991.

Bhattacharya, A., D. Pati, N. S. Pillai, and D. B. Dunson (2016). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association 111*, 1479–1490.

Borrotti, M., D. de March, D. Slanzi, and I. Poli (2014). Designing lead optimisation of MMP-12 inhibitors. *Computational and Mathematical methods in Medicine* (Article ID 258627).

Brown, P. J. (1993). *Measurement, Regression,and Calibration*. Oxford: Clarendon Press.

Brown, P. J. and M. S. Ridout (2016). Level-screening designs for factors with many levels. *Annals of Applied Statistics 10*, 864–883.

Brown, P. J. and C. H. Spiegelman (1991). Mean squared error and selection in multivariate calibration. *Statistics and Probability Letters 12*, 157–159.

Brown, P. J., C. H. Spiegelman, and M. C. Denham (1991). Chemometrics and spectral frequency selection. *Philosophical Transactions of the Royal Society, A 337*, 311–322.

Brown, P. J., M. Vannucci, and T. Fearn (2002). Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, B 64*, 519–536.

Cadonna, A., S. Frühwirth-Schnatter, and P. Knaus (2020). Triple the gamma – a unifying shrinkage prior for variance and variable selection in sparse state space and TVP models. *Econometrics 8*, 20.

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

Chipman, H., E. I. George, and R. E. McCulloch (2001). In T. W. Anderson, K. T. Fang, and I. Olkin (Eds.), *The Practical Implementation of Bayesian Model Selection*, Volume 38. IMS Lecture-Notes Monograph Series, Hayward, California.

Fearn, T., D. Perez-Marin, A. Garrido-Varo, and J. E. Guerrero-Ginel (2010). Inverse, classical, empirical and non-parametric calibrations in a Bayesian framework. *Journal of Near Infrared Spectroscopy 18*, 27–38.

Fearn, T., D. Perez-Marin, A. Garrido-Varo, and J. E. Guerrero-Ginel (2019). Classifying with confidence using Bayes rule and kernel density estimation. *Chemometrics and Intelligent Laboratory Systems 189*, 81–87.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis 1*, 515–533.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin (2014). *Bayesian Data Analysis* (3rd Edition ed.). CRC Press.

Ghosh, P., X. Tang, M. Ghosh, and A. Chakrabarti (2016). Asymptotic properties of Bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. *Bayesian Analysis 11*, 753–796.

Goutis, C. and C. P. Robert (1998). Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections. *Biometrika 85*, 29–37.

Griffin, J. and P. J. Brown (2017). Hierarchical shrinkage priors for regression models. *Bayesian Analysis 12*, 135–159.

Griffin, J. E. and P. J. Brown (2010). Inference with Normal-Gamma prior distributions in regression problems. *Bayesian Analysis 5*, 171–188.

Griffin, J. E. and P. J. Brown (2011). Discussion of 'Shrink globally, act locally: sparse Bayesian regularization and prediction' by Polson and Scott". In M. J. Bernardo J. M., Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics 9*, pp. 539–540. Oxford: Clarendon Press.

Griffin, J. E., K. Latuszynski, and M. F. J. Steel (2020). In search of lost (mixing) time: Adaptive markov chain schemes for Bayesian variable selection with very large p. *Biometrika 107*, ???–???

Hussain, J. and C. Rea (2010). Computationally efficient algorithm to identify matched molecular pairs(MMPs) in large data sets. *J. Them. Inf. Model. 50*, 339–348.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Clarendon Press, Oxford.

Lindley, D. V. (1968). The choice of variables in multiple regression (with discussion). *Journal of the Royal Statistical Society, B 30*, 31–66.

Makalic, E. and D. F. Schmidt (2016). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters 23*, 179–182.

Marquardt, D. W. (1970). Generalised inverses, ridge regression, biased linear estimation and nonlinear estimation. *Technometrics 12*, 591–612.

Moran, G. E., V. Ročkov*á*, and E. I. George (2019). Variance prior forms for high dimensional variable selection. *Bayesian Analysis 14*, 1091–1119.

Pal, S., K. Khare, and J. P. Hobert (2017). Trace class Markov chains for Bayesian inference with generalized double Pareto shrinkage priors. *Scandinavian Journal of Statistics 44*, 307–323.

Pickett, S. D., D. V. S. Green, D. L. Hunt, D. A. Pardoe, and I. Hughes (2011). Automated lead optimisation of MMP-12 inhibitors using a genetic algorithm. *ACS Medicinal Chemistry Letters 2*, 28–33.

Piironen, J., M. Paasiniemi, and A. Vehtari ((2020)). Projective inference in high - dimensional problems: Prediction and feature selection. *Electonic Journal of Statistics 14*, 2155–2197.

Piironen, J. and A. Vehtari (2017). On the hyperprior choice for global shrinkage parameter in the horseshoe prior. In *20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Volume 54, pp. 905–913. Fort Lauderdale, Florida, USA. JMLR.

Rajaratnam, B., D. Sparks, K. Khare, and L. Zhang (2019). Uncertainty quantification for modern high-dimensional regression via scalable Bayesian methods. *Journal of Computational and Graphical Statistics 28*, 174–184.

Van der Pas, S. L., B. J. K. Kleijn, and A. W. van der Vaart (2014). The horseshoe estimator: posterior concentration around nearly black vectors. *Electronic Journal of Statistics 8*, 2585–2618.

Van der Pas, S. L., J. B. Salomond, and J. Schmidt-Hieber (2016). Conditions for posterior contraction in sparse normal means problems. *Electronic Journal of Statistics 8*, 976–1000.

Van der Pas, S. L., B. Szabo, and A. W. van der Vaart (2017). Adaptive posterior contraction rates for the horseshoe. *Electronic Journal of Statistics 11*, 3196–3225.

Vannucci, M., N. Sha, and P. J. Brown (2005). NIR and mass spectra classification: Bayesian methods for wavelet-based feature selection. *Chemometrics and Intelligent Laboratory Sysyems 77*, 139–148. Festschrift in honour of Professor L. M. Massart.

Yang, Y., M. J. Wainwright, and M. I. Jordan (2016). On the computational complexity of high -dimensional Bayesian variable selection. *Annals of Statistics 44*, 2497–2532.

Zhang, Y. D., B. P. Naughton, H. D. Bondell, and B. J. Reich (2020). Bayesian regression using a prior on the model fit: The R2-D2 shrinkage prior. *Journal of the American Statistical Association*, to appear.