mathematics

MDPI

Article

Multivariate Bayesian Global–Local Shrinkage Methods for Regularisation in the High-Dimensional Linear Model

Valentina Mameli [1], Debora Slanzi [2,3], Jim E. Griffin [4] and Philip J. Brown [5,*]

[1] Department of Economics and Statistics, University of Udine, 33100 Udine, Italy; valentina.mameli@uniud.it
[2] Venice School of Management, Ca' Foscari University of Venice, 30121 Venice, Italy; debora.slanzi@unive.it
[3] European Centre for Living Technology, Ca' Foscari University of Venice, 30123 Venice, Italy
[4] Department of Statistical Science, University College London, London WC1E 6BT, UK; j.griffin@ucl.ac.uk
[5] Department of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury CT2 7NZ, UK
* Correspondence: philip.j.brown@kent.ac.uk

Abstract: This paper considers Bayesian regularisation using global–local shrinkage priors in the multivariate general linear model when there are many more explanatory variables than observations. We adopt priors' structures used extensively in univariate problems (conjugate and non-conjugate with tail behaviour ranging from polynomial to exponential) and consider how the addition of error correlation in the multivariate set-up affects the performance of these priors. Two different datasets (from drug discovery and chemometrics) with many covariates are used for comparison, and these are supplemented by a small simulation study to corroborate the role of error correlation. We find that structural assumptions of the prior distribution on regression coefficients can be more significant than the tail behaviour. In particular, if the structural assumption of conjugacy is used, the performance of the posterior predictive distribution deteriorates relative to non-conjugate choices as the error correlation becomes stronger.

Keywords: seemingly unrelated multivariate normal regression; structured priors; global–local priors; exponential-tailed and polynomial-tailed priors; drug discovery; chemometrics

MSC: 62F15; 62H12; 62J07

check for updates

Academic Editors: Víctor Amor-Esteban and David Almorza-Gomar

Received: 21 February 2025
Revised: 16 May 2025
Accepted: 22 May 2025
Published: 29 May 2025

Citation: Mameli, V.; Slanzi, D.; Griffin J.E.; Brown, P.J. Multivariate Bayesian Global–Local Shrinkage Methods for Regularisation in the High-Dimensional Linear Model. *Mathematics* **2025**, *13*, 1812. https://doi.org/10.3390/math13111812

1. Introduction

Multivariate regression modelling is an important means for understanding the effect of predictor variables on a set of responses, which can improve prediction by exploiting the relationship between responses. Breiman and Friedman [1] emphasises the gains that can be made in prediction when embracing the multiresponse nature of problems, or 'marrying' responses as described by M. Stone in the Discussion.

Bayesian methods achieve shrinkage through the choice of priors embodied both in the structure of the model, typically the conjugacy of the prior on the regression coefficients and the extent of 'borrowing strength' through shared hyperparameters, and the distributional assumptions on the priors for the regression coefficients. We concentrate on multivariate Bayesian regression in the large-p, small-n setting, where there are many explanatory variables common to explaining the variation and covariation of the responses. For illustrative purposes, we focus on just a few responses and, indeed, our examples involve just two responses, although our models and theory are quite general. We concentrate on popular global–local or 'one group' priors where the regression coefficients are given a continuous prior [2,3] as opposed to the 'two group' priors constituting 'slab and spike' priors. A comprehensive review of Bayesian variable selection and regularisation is provided by Tadesse and Vannucci [4]. Much of our work extends models used in univariate multiple regression but, at the same time, encounters difficulties and insights not present in the univariate formulations. There is a parallel body of literature in econometrics in the form of (time-varying) vector autoregressive models, where conjugacy is often not feasible; see, for example, Hauzenberger et al. [5].

Global–local priors for regression coefficients are mixtures of normal distributions that have heavier tails than a normal distribution and, typically, depend on the error variance (the conjugate choice). Popular univariate choices range from the horseshoe [6] with its polynomial tails [7] to priors for the regression coefficients with the exponential tails, as in the Normal–Gamma [8,9] and the Dirichlet–Laplace [10]. Van der Pas et al. [11] developed general conditions on the global–local prior for the optimal minimax rate of contraction in large-p, small-n problems and show that priors spanning polynomial to exponential tails can meet these conditions. Generally, these optimality results assume a known error variance and this is also the case with multivariate responses. If the error variance is unknown, Moran et al. [12] showed that conjugate priors for the regression coefficients can lead to an underestimation of the error variance in both 'one group' and 'slab and spike' settings.

An extension of conjugate 'one group' priors to multivariate modelling was discussed by Bai and Ghosh [13] using a class of prior distributions with polynomial tails. Kundu et al. [14] considered further extensions, including non-conjugate modelling, which is closest to our approach. This paper goes further by comparing the full range of polynomial to exponential tails in both conjugate and non-conjugate settings, and judges performance both in estimation and in both point and density prediction. The full predictive distribution requires a multivariate formulation, and we show that the popular conjugate Bayesian model can perform poorly if there are strong correlations between the errors.

Other authors have developed Bayesian approaches to the multivariate regression model, mainly sticking with 'slab and spike' priors but employing various covariance structures on the regression coefficients. Brown et al. [15] and Brown et al. [16] demonstrated the feasibility of Bayesian variable selection in multivariate regression and the benefits compared to separate univariate regressions. Various simplification have been adopted to exploit conjugacy with respect to regression coefficients and residual covariance. Petretto et al. [17] assumed the same set of predictors for each response in the context of an expression Quantitative Trait Locus (eQTL) analysis. Bhadra and Mallick [18] used the same simplification on predictors and also included sparse covariance selection using graphical models on decomposable graphs. In both papers, the regression coefficients and residual covariances are integrated out, thus reducing the search for Markov chain Monte Carlo (MCMC) compared to that of variable selection indicators as in Brown et al. [19]. Bottolo et al. [20] relaxed the assumption of a common set of predictors for each response, which reduces the computational burden with a large number of responses. They exploited a Bartlett factorisation of the covariance matrix and partial conjugacy in a decomposable graph. See also Brown et al. [21] and Le et al. [22], who used the same factorisation to generalise the Inverted Wishart distribution to allow varying precision within the matrix rather than being stuck with a single degree of freedom parameter. Our data typically do not involve such a large number of responses as that in Bottolo et al. [20], but do potentially involve many more covariates.

Another scale mixture of normal priors for the multivariate case—specifically, a Bayesian Lasso or double-exponential prior with a group structure on regression coefficients—is implemented by Liquest et al. [23]. They also added a 'slab and spike' indicator selection to make sure that small coefficients are driven to zero.

There are two issues in exploring priors for Bayesian multivariate regression. Firstly we explore the relative merits of different tail behaviours in the regression coefficients' prior distribution. Secondly, the role of conjugate priors that incorporate the error covariance matrix, as opposed to non-conjugate models, in their effect on the predictive distribution is assessed. We investigate these issues in the context of two examples: from chemometrics, a regression of the weight of a mixture of two sugars (glucose and sucrose) in aqueous solution on a near-infrared spectrum discretised at 700 wavelengths (giving rise to 700 highly correlated explanatory variables); and a drug discovery example with 140 observations and 3149 functional attributes. All theoretical results that we know of, at least in the conjugate case, restrict attention to known error variance or covariance in the case of multivariate regression, so that any unintended consequences of putting the error covariance into the prior on the regression coefficients have not been fully explored.

Van der Pas et al. [11] gave conditions for the posterior to contract at the optimal minimax rate in single-response regression. The general characteristics are that the tails of the prior distribution on the regression coefficients should be at least as heavy as the Laplace distribution, but not too heavy, and they should have a large amount of mass close to zero. We have chosen to extend two widely used single-response, representative members of the class of global–local prior distributions to the multivariate model: the horseshoe [6] with its polynomial tails, and the Normal–Gamma [8] with exponential tails, together with the polynomial-tailed Normal–Gamma–Gamma [24], not previously employed in multivariate modelling. The horseshoe prior has been thoroughly investigated in 'nearly black' settings and is a popular reference point, as it only has a single scale hyperparameter to specify, but questions arise concerning its ability to estimate weaker signals, when it can over-shrink wiping out the signal completely. A review of global–local priors in univariate regression was given in [3], and included the Dirichlet–Laplace [10], which, with its exponential tails, performed particularly well. This is not a true global–local model and so is harder to incorporate into our structural assumptions. Kundu et al. [14] found that this leads to slow mixing for large numbers of covariates as a result of an extra Metropolis–Hastings step and, consequently, they abandoned it for larger problems.

The Supplementary Materials provide details of extensions to the multivariate regression of the efficient algorithms for the large-p, small-n univariate regression of Bhattacharya et al. [25]. The full conditionals of hyperparameters are also detailed there, as are the trace class calculations for the conjugate case.

The aim of this paper is two-fold: firstly, to explore the extent that structural assumptions and, in particular, the conjugacy of the prior intrude on predictions, and, secondly, to discover whether the fatness of the tails of the prior (polynomial versus exponential, via examples of each) is critical. There is one example of exponential tails and two of polynomial tails, a factor thus involving three levels. The structural assumptions we examine are also three-fold: conjugate, non-conjugate with separate parameters for each response, and non-conjugate with parameters shared across responses. Sharing parameters allows the 'borrowing of strength' and is a natural part of the conjugate model; we investigate how these choices affect inference.

This paper is organised as follows. The Introduction presents the multivariate models discussed and the existing literature. The Materials and Methods details the specific multivariate regression models and Bayesian models employed with both conjugate and non-conjugate formulations, as well as hyperparameters that encourage communality (sharing) and those that emphasise rich separate parameterisation. The distributional assumptions and hyperparameter distributions are presented. Then, we describe the MCMC algorithm used for inference, and the consequences of conjugacy for error covariance estimation and prediction are explored. Two real datasets are introduced, one from chemometrics, and the other from drug discovery, both with large numbers of explanatory variables. The Results describe the application of the method to these two applications and a small simulation study (emerging from these examples), which further investigates the role of error correlation. Lastly, Conclusions are drawn and suggestions for further work are made.

2. Materials and Methods

2.1. Standard Multivariate Regression

In the standard multivariate regression, we have a response matrix Y, $(n \times q)$, and a matrix of explanatory variables common to all responses X, $(n \times p)$, modelled by

$$Y - M - XB \ \sim \mathcal{N}_{n,q}(I_n, \Sigma), \tag{1}$$

where B is a $(p \times q)$ matrix of regression coefficients, $M(n \times q) = 1[\alpha_1, \dots, \alpha_q]$ are the q intercepts, 1 is a $n \times 1$ unit vector, and Σ is a $(q \times q)$ error covariance matrix. We use $\mathcal{N}_{p,q}(A, B)$ to denote an $(p \times q)$ matrix variate normal with mean zero, the jth row covariances $a_{jj}B$, and kth column covariance matrix $b_{kk}A$ for $j = 1, \dots, p$ and $k = 1, \dots, q$. The notation used here is often more convenient and illuminating, originating in Dawid [26], and further developed in Brown [27], providing similar forms of Bayesian updates of priors with changing dimensions. The equivalent vectorised form for the multivariate normal linear regression model can be represented as follows: for $Y^{*T} = (Y_1^T, \dots, Y_q^T)$, a $(1 \times nq)$ vector of q responses strung out. The composite $(nq \times pq)$-dimensional design matrix $X^* = \mathrm{DIAG}[X, \dots, X]$ is a block diagonal matrix with the q response design matrices on the diagonal, zero elsewhere. Now, with intercepts $\alpha_k(q \times 1)$ strung out conformally as an $(nq \times 1)$ vector μ and with a $(pq \times 1)$ vector of coefficients β, we have the regression model

$$Y^* = \mathrm{vec}(1\alpha^T) + X^*\beta + e, \tag{2}$$

where the $(nq \times 1)$ vector e of residual errors have a multivariate normal distribution

$$N_{nq}(0, \Sigma \otimes I_n) \tag{3}$$

$\mathrm{vec}(\cdot)$ is an $nq \times 1$ vectorised form of the $(n \times q)$ matrix. and $\otimes$ is the Kronecker product of two matrices, so that observations are correlated within q responses but independent across n sets of explanatory variables.

We always assume that we have common explanatory variables available, even if some can be effectively removed by shrinkage. Also, we standardise all predictors, that is, the p explanatory variables are all zero-centred and scaled to have variance 1, such that $\sum_{i=1}^n x_{ij} = 0$ and $\sum_{i=1}^n x_{ij}^2 = n$ for $j = 1, \dots q$, as in Brown et al. [15] and Brown et al. [19]. The centring equation naturally accompanies a vague unnormalised prior for the intercept, which will then be decoupled from the other covariates, and its distribution can be separately specified given Σ as a multivariate normal with mean $\bar{Y}$ and covariance matrix Σ/n. The variance standardisation is natural to scaling: it ensures that parameter and hyperparameter values should not necessarily change with sample size. We cover both a conjugate prior and non-conjugate priors for matrix B, or equivalently, a $(pq \times 1)$ vector for β.

2.2. Scale Mixture of Normals' Priors for Regression Coefficients in the Multivariate Model

The natural scale mixture model for the non-conjugate case begins with

$$\beta_{jk} \sim N(0, \omega_{jk}) \tag{4}$$

with local variances ω_{jk} for $j = 1, \ldots, p$ and $k = 1, \ldots, q$.

We broadly consider three different structural assumptions on the scale hyperparameters. The most general global–local model has

$$\omega_{jk} = \psi_{jk}^2 \tau_k^2 \tag{5}$$

with q global scales $(\tau_1, \tau_2, \ldots, \tau_q)$, and pq local idiosyncratic variances ψ_{jk}^2.

This is the separated non-conjugate form. In terms of structural form, this is akin to the saturated model discussed by D. V. Lindley in Breiman and Friedman [1].This also corresponds to the 'naive' model of Kundu et al. [14], but provides shrinkage not tied to correlation across responses and can be more flexible. A very small local idiosyncratic variance drives the corresponding regression coefficient to zero and, in effect, removes the corresponding explanatory variable for one response [28]. Small, but not very small, idiosyncratic variances lead to a more nuanced effect with the regression coefficient shrunk towards (rather than close to) zero.

The shared non-conjugate form borrows strength in communality across responses by assuming a single local parameter for each predictor across responses, giving the multi-outcome form of Kundu et al. [14], and would correspond to a main-effects model for responses and explanatory variables:

$$\omega_{jk} = \psi_j^2 \tau_k^2. \tag{6}$$

There are $p + q$ primary hyperparameters (rather than $(p + 1)q$ in the separate non-conjugate form), such that the idiosyncratic variances are common across the q responses but otherwise vary and are local to each explanatory variable.

The shared conjugate model aims for a similar level of elaboration to the shared non-conjugate, although, with the natural scaling properties of conjugacy, it is less compelling to vary the global scales by response. Therefore, we use p idiosyncratic variances $\psi_1^2, \ldots, \psi_p^2$ and a single global scale τ. The shared conjugate model for $p \times q$ matrix B is

$$B \sim \mathcal{N}_{p,q}(\Omega^*, \Sigma) \tag{7}$$

with $p \times p$ matrix $\Omega^* = \tau^2 \Psi$,

$$\Psi = \mathrm{DIAG}[\psi_1^2, \ldots, \psi_p^2] \tag{8}$$

and Σ is the $q \times q$ error covariance matrix. If we write $B^T(q \times p) = (\beta_1, \ldots, \beta_p)$, so that β_j is the vector of regression coefficients for the jth explanatory variable across the q responses, then

$$\beta_j \sim N_q(0, \psi_j^2 \tau^2 \Sigma), \qquad j = 1, \ldots, p, \tag{9}$$

and the covariance of β_j will reflect the covariance matrix of errors Σ.

In both the shared non-conjugate and conjugate forms, the shared idiosyncratic scale ψ_j for the jth explanatory variable plays the same role in driving all q regression coefficients to zero, as in Equation (6) if all q coefficients are unimportant and small (irrespective of correlation). More important and sizeable coefficients can be accommodated by a larger value of ψ_j. This leads to a less severe shrinkage towards zero of all regression coefficients for the jth explanatory variable, including those which have no effect. In the conjugate form, the prior covariance of β_j will reflect the correlation structure in Σ. This may differ strongly from a correlation structure, which favours regression coefficient values reflecting the main drivers of each response. Whilst the idiosyncratic variances are focused on regularisation through shrinkage on a response by response basis, the injection of Σ into the prior can have significant unintended spoiling features, as we demonstrate.

More elaborate structures are natural in different hierarchical settings, as in Griffin and Brown [24]. The intercepts of the q responses are assumed to have vague priors. A wide range of priors borrowed from univariate regression are suitable candidates as global–local priors; aside from the three representatives we have worked with here, there are many others, for example, the Dirichlet–Laplace [10], the horseshoe + [29], the graphical horseshoe [30], the R2D2 [31], and the factor shrinkage models [32]. We now list the specific priors that we consider for global and local hyperparameters for the different Bayesian priors that we compare. These are priors for non-negative hyperparameters that involve either the gamma distribution or a half-Cauchy prior. Notationally, we denote $Ga(x|c,d) \propto x^{c-1}\exp\{-dx\}$ with expectation c/d and shape c so that d provides a form of 'scale'. Also, let $\mathrm{C}^+(0,w)$ be a half-Cauchy with scale w that has pdf $g(x) = \frac{2w}{\pi(w^2+x^2)}$ for $x > 0$. To avoid notational overload, we have omitted conditioning arguments in what follows where it is obvious.

2.2.1. Normal–Gamma Prior

Griffin and Brown [8] proposed a gamma hyperprior on the variance of the normal distribution of regression coefficients where the mixing distribution has the gamma density function. It includes the Bayesian Lasso [33] with shape $c = 1$ but, by having more flexibility in shape, is able to achieve optimum minimax rates of convergence to the 'truth' [11] when this is not possible for the Bayesian Lasso itself.

We consider the following models, for $j = 1, \ldots, p$ and $k = 1, \ldots, q$.

We start examining the separate non-conjugate model, whose hierarchy is given here:

- Separated Non-Conjugate Model
  $\beta_{jk} \sim N(0, \psi_{jk}^2 \tau_k^2), \quad \psi_{jk}^2 \sim Ga(\lambda_k, 1), \quad \tau_k \sim \mathrm{C}^+(0, \tau_0), \quad \lambda_k \sim Ga(1, p/5).$
  We then focus on the shared non-conjugate model, which is based on the following hierarchy:
- Shared Non-Conjugate Model
  $\beta_{jk} \sim N(0, \psi_j^2 \tau_k^2), \quad \psi_j^2 \sim Ga(\lambda, 1), \quad \tau_k \sim \mathrm{C}^+(0, \tau_0), \quad \lambda \sim Ga(1, p/5).$
  Finally,
- Shared Conjugate Model
  $\beta_j \sim N_q(0, \psi_j^2 \tau^2 \Sigma), \quad \psi_j^2 \sim Ga(\lambda, 1), \quad \tau \sim \mathrm{C}^+(0, \tau_0), \quad \lambda \sim Ga(1, p/5).$

The conjugate model is similar to the shared model in its p idiosyncratic variance parameters, ψ_j^2, for $j = 1, \ldots, p$. The Normal–Gamma has exponential tails rather than the polynomial tails of the horseshoe, which we examine next.

2.2.2. Horseshoe Prior

This has become a popular prior in single-response regression [6], see also [34]. It has polynomial tails and a sharp spike at zero, so that small regression coefficients are strongly shrunk to zero, but larger coefficients are not shrunk. Our elaborations of the horseshoe prior for multivariate regression are, for $j = 1, \ldots, p$ and $k = 1, \ldots, q$, as follows:

- Separated Non-Conjugate Model
  $\beta_{jk} \sim N(0, \psi_{jk}^2 \tau_k^2), \quad \psi_{jk} \sim \mathrm{C}^+(0, 1), \quad \tau_k \sim \mathrm{C}^+(0, \tau_0);$
- Shared Non-Conjugate Model
  $\beta_{jk} \sim N(0, \psi_j^2 \tau_k^2), \quad \psi_j \sim \mathrm{C}^+(0, 1), \quad \tau_k \sim \mathrm{C}^+(0, \tau_0);$
- Shared Conjugate Model
  $\beta_j \sim N_q(0, \psi_j^2 \tau^2 \Sigma), \quad \psi_j \sim \mathrm{C}^+(0, 1), \quad \tau \sim \mathrm{C}^+(0, \tau_0).$

Typically, we are guided by Piironen and Vehtari [35] for the default choice of scale $\tau_0 = p_0/(p - p_0)$, with p_0 a prior guess for the number of important variables; typically, we set $p_0 = 3$. The most extensively researched estimate of τ is provided by Van der Pas et al. [36], following on from Van der Pas et al. [37], who suggested that empirical Bayes with marginal maximum likelihood or hierarchical Bayes with a truncated Cauchy give good results based on theoretical considerations and simulations.

2.2.3. Normal–Gamma–Gamma Prior

This offers some extra flexibility compared with the horseshoe. A mixture of the Normal–Gamma provides a distribution with much fatter tails [24]. We consider the following priors, for $j = 1, \dots, p$ and $k = 1, \dots, q$:

- Separated Non-Conjugate Model
  $\beta_{jk} \sim N(0, \psi_{jk}^2 \tau_k^2), \quad \psi_{jk}^2 \sim Ga(\lambda_k, \gamma_{jk}), \quad \gamma_{jk} \sim Ga(c_k, 1), \quad \tau_k \sim \mathrm{C}^+(0, \tau_0),$
  $\lambda_k \sim Ga(1, p/5);$
- Shared Non-Conjugate Model
  $\beta_{jk} \sim N(0, \psi_j^2 \tau_k^2), \quad \psi_j^2 \sim Ga(\lambda, \gamma_j), \quad \gamma_j \sim Ga(c, 1), \quad \tau_k \sim \mathrm{C}^+(0, \tau_0),$
  $\lambda \sim Ga(1, p/5);$
- Shared Conjugate Model
  $\beta_j \sim N_q(0, \psi_j^2 \tau^2 \Sigma); \quad \psi_j^2 \sim Ga(\lambda, \gamma_j); \quad \gamma_j \sim Ga(c, 1); \quad \tau \sim \mathrm{C}^+(0, \tau_0);$
  $\lambda \sim Ga(1, p/5).$

There is probably not a lot of information to distinguish tail behaviour between responses, so that the elaboration from c to c_k may not be warranted. We leave it in its most elaborate form to derive the formulae necessary for MCMC. 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. Further elaborations of the Normal–Gamma are given in Cadonna et al. [38], which provides a comprehensive review of this prior and other uses of this structure in the literature.

2.3. *Error Covariance Prior*

The covariance matrix Σ is given a suitably weak but proper prior for example within the Inverse Wishart family

$$\Sigma \sim \mathcal{IW}(\delta, Q).$$

We differ from standard notation by defining the shape parameter to be $\delta = \nu - q + 1$, where ν is the degrees of freedom parameter. Typically, a vague specification of this shape parameter is $Q \propto I_q$ and $\delta = 3$, the lowest integer value for which the prior expectation of the Inverse Wishart exists [15]. This shift from degrees of freedom to shape means that, under marginalisation, the distributional form does not change.

2.4. *MCMC Algorithm*

A Gibbs sampler can be used to make inference in the multivariate regression model with the priors described in this paper. We update the blocks of parameters, β, $(pq \times 1)$ (or, equivalently, B), Σ, $(q \times q)$, ψ_{jk}, τ_k (or, equivalently, τ), for $j = 1, \dots, p$ and $k = 1, \dots, q$, and the scale hyperparameters from their full conditional distributions. We have assumed that the explanatory variables are centred and scaled so that they have a mean of zero and a variance of one. We use a vague invariant prior for the q intercepts, α. For the non-conjugate model in (4), the posterior for the blocks of parameters is structured as

$$f_1(Y \mid B, \Sigma) f_2(B \mid \Omega) f_3(\Sigma) f_4(\Psi, \tau), \tag{10}$$

where Ω indexes parameters Ψ and τ. For the conjugate model in (9), the second term above is augmented with Σ to $f_2(B \mid \Omega, \Sigma)$. The MCMC algorithm then consists of a stepwise algorithm constructed from the posterior obtained from (10):

1. $[B \mid Y, \Sigma, \Omega]$
2. $[\Sigma \mid \Omega, Y, B]$
3. $[\Omega \mid Y, B, \Sigma]$

We detail the first two below, as they are substantively important for the development. The conditional distribution of B is matrix–variate Normal and the conditional distribution of Σ is Inverse Wishart. For the full conditional for hyperparameters Ω, see the Supplementary Materials.

2.4.1. Full Conditional for the Regression Coefficients

The full conditional distribution of the regression coefficients is available analytically. However, as in the univariate case, this is expensive to sample since it involves the inversion of a $(pq \times pq)$-dimensional matrix in the multivariate case, a sizeable increase from the univariate p-variable regression. The inversion scales cubically with dimension. Since we are interested in the $n < p$ case, the data augmentation device in Bhattacharya et al. [25] is used to speed up sampling from the full conditional by solving a set of equations with linear complexity. We note below that, at least as far as the conjugate model is concerned, the inversion dimension is p rather than pq. The Gibbs full conditionals for conjugate and non-conjugate forms are derived below.

A convenient form of the likelihood for β is obtained by reducing the covariance matrix of the q responses of the original nq errors to the identity by premultiplying (2) by a square root matrix of the inverse. Let

$$H = \Sigma^{-1/2} \otimes I_n$$

with I_n as the $(n \times n)$ identity matrix. The log of the full conditional for β is

$$-\frac{1}{2}(z - \Phi\beta)^T(z - \Phi\beta) - \frac{1}{2}\beta^T D^{-1}\beta + C$$

where $\Phi = HX^*$, $z = HY^\star$; the form of D depends on the prior form and C is the log of the normalising constant.

In the non-conjugate assignment of a prior for B as in Equation (4),

$$D = \mathrm{DIAG}[\omega_{11}, \dots, \omega_{p1}, \omega_{12} \dots \omega_{p2} \dots \omega_{1q}, \dots \omega_{pq}]$$

is the $pq \times pq$ diagonal matrix of global–local hyperparameters, ordered response by response. The possible forms of the entries ω_{jk} are given by Equations (5) and (6), where an alternative form for the latter is $D = T \otimes \Psi$, where $T = \mathrm{DIAG}(\tau_1^2, \dots, \tau_q^2)$ and Ψ is given by Equation (8).

With the conjugate form,

$$D = \tau^2 \Sigma \otimes \Psi,$$

with again Ψ as the $(p \times p)$ diagonal matrix of shared local parameters. This matrix is readily invertible and, hence, still conforms to the desirable characteristics of the fast algorithm in [25].

In both the conjugate and the non-conjugate cases, after 'completing the square', the full conditional is proportional to

$$\exp\left\{-\frac{1}{2}(\beta - \Sigma^* a)^T \Sigma^{*-T}(\beta - \Sigma^* a)\right\}$$

with

$$\Sigma^* = (\Phi^T\Phi + D^{-1})^{-1},$$
$$a = \Phi^T z.$$

This is the same form as the proposed algorithm for speeding up MCMC in Bhattacharya et al. [25], and we are able to use the same data augmentation algorithm linear in dimension pq, avoiding inverting a $(pq \times pq)$ matrix with its cubic complexity. The steps in implementing this sped-up algorithm to update the block of regression parameters are given in the Supplementary Materials.

We note that the critical $pq \times pq$ matrix Σ^* can be formed from a smaller $p \times p$ matrix in the conjugate case with

$$\Sigma^* = [\Sigma^{-1} \otimes \{X^T X + (\tau^2\Psi)^{-1}\}]^{-1}.$$

The required algorithm is best seen from the matrix variate formulation, when, for the conjugate model,

$$B - \{X^T X + (\tau^2\Psi)^{-1}\}^{-1} X^T Y \sim \mathcal{N}_{p,q}[\{X^T X + (\tau^2\Psi)^{-1}\}^{-1}, \Sigma], \tag{11}$$

as in Brown [27], so that the column means of B, referring to one response, just require cycling through the q columns of Y and the inverse of a $p \times p$ matrix. Calculations of regression coefficients are as if there is just one response, as the matrix notation brings out.

The intercepts α are decoupled from the regression parameters β under the posterior distribution by virtue of centring the explanatory variables and having a simple normal-form conditional on Σ, as in [39] see the Supplementary Materials.

2.4.2. Full Conditional for the Error Covariance Matrix

The full conditional for Σ is best described by moving to the matrix variate normal formulation of the model in Equation (1). With the non-conjugate forms of a prior, the full conditional for Σ is constructed by combining the probability density function of the matrix normal and the form of the Inverse Wishart prior density.

This gives

$$\Sigma \sim \mathcal{IW}(\delta^*, Q^*), \tag{12}$$

where

$$Q^* = Q + (\tilde{Y} - XB)^T(\tilde{Y} - XB), \tag{13}$$

$\tilde{Y} = Y - \bar{Y}$, each response is corrected by its mean, and $\delta^* = \delta + n - 1$. The expectation of Σ under the full conditional is $Q^*/(\delta^* - 2)$. This is a natural estimate, closely related to the 'oracle' estimate if $B = B^*$ were known,

$$\hat{\Sigma} = (Y - XB^*)^T(Y - XB^*)/n.$$

On the other hand, the dependence of Σ on B in the conjugate prior leads to a full conditional distribution $\Sigma \mid B, \Omega$, proportional to

$$f_1(Y \mid B, \Sigma)\, f_2(B \mid \Omega, \Sigma)\, f_3(\Sigma),$$

from (10).

The last two terms can be seen as an induced prior on Σ conditional on B, which has the form $\Sigma \sim \mathcal{IW}(\delta^{**}, Q^{**})$ with $\delta^{**} = \delta + p$ and

$$Q^{**} = Q + \tau^{-2} B^T \Psi^{-1} B. \tag{14}$$

The conjugate prior effectively adds p extra elements of data given B. When p is large, such as several hundred or several thousand, as in our examples, this can lead to unintended consequences. The prior 'data' $\tau^{-2} B^T \Psi^{-1} B$, which arise from the conjugate prior chosen to achieve a reduced parameter model, eschewing cost or utility considerations, may distort inference for Σ and even affect the estimation of the regression coefficients.

With the conjugate prior on the regression coefficients, a further term enters the full conditional relative to the non-conjugate case. From the pdf of Equation (7), we have the additional terms

$$|\Psi|^{-q/2} |\Sigma|^{-p/2} \exp\Big\{-\tfrac{1}{2}\operatorname{trace}(\Sigma^{-1}[\tau^{-2} B^T \Psi^{-1} B])\Big\}.$$

Thus, for the conjugate case, we have the full conditional with the exponential term $-\frac{1}{2}\operatorname{trace}(\Sigma^{-1} Q^*)$, where

$$Q^* = Q + (\tilde{Y} - XB)^T (\tilde{Y} - XB) + \tau^{-2} B^T \Psi^{-1} B, \tag{15}$$

and now the shape parameter of the Inverse Wishart in Equation (12) is $\delta + p + n - 1 = \delta^* + p$.

The fact that the divisor is $\delta^* + p$ means that the conditional expectation of Σ given B, with many parameters, can significantly underestimate Σ (relative to the non-conjugate form in (13)) unless the additional term in (15) is roughly $p\Sigma$. The extra term in (15) can be written as the $q \times q$ matrix with (ij)-th element

$$\tau^{-2} \beta_i^T \Psi^{-1} \beta_j, i, j = 1, 2, \ldots, q, \tag{16}$$

where β_i is vector of p regression coefficients for the ith response. Now, with sparsity, all but a few, say, r, of these coefficients will be zero or near zero. What is more, the sharing assumption of the conjugate prior distribution encourages similarity between the regression coefficients across responses, so that effectively many of the explanatory variables will have little explanatory value, with only $r \ll p$ being sizeable. Thus, heuristically, the entries of the $q \times q$ in (16) are $O(r/p)$, and so are much smaller than $p\Sigma$. In the univariate case, Proposition 3 of Moran et al. [12] argues that large regression coefficients are unlikely to be blown up sufficiently to compensate for the increased divisor since the divisors $\tau^2 \psi_j^2$ are naturally bounded away from zero in the global–local setting, and the same could be argued in the multivariate case for all the contributions to (16). In our multivariate case, we investigate this also in a simulation study based on the style of data for the 'Sugars' application. Since there is the potential to underestimate Σ in the conjugate case, as argued above, in any simulation, we judge the 'decreased size' of the error covariance matrix through a likelihood ratio deviance measure.

2.4.3. Deviance for Covariance Matrix

As we have argued, there is the potential for an underestimation of Σ in the conjugate case, so it is natural to seek to measure the difference in size of the estimated covariance matrix from its true value. We achieve this through a deviance or goodness of fit measure.

Imagine that we have a sample of n observations from a multivariate normal distribution with mean zero, $Y \sim \mathrm{N}_q(0, \Sigma)$; then, -2loglikelihood is

$$n \log\det(\Sigma) + n\,\mathrm{trace}[\hat{\Sigma}\Sigma^{-1}], \tag{17}$$

where, strictly, $\hat{\Sigma}$ is the cross-products matrix divided by the number of observations, n. If, in fact, $\hat{\Sigma}$ is any estimate of Σ, the value of Σ that minimises (17) is $\hat{\Sigma}$ ([40], Theorem 4.2.1). Thus, the notional log likelihood ratio test for the comparison of a true covariance Σ with one estimated with $\hat{\Sigma}$ is

$$DEV = n\,\mathrm{trace}[\hat{\Sigma}\Sigma^{-1}] - n \log\det\,[\hat{\Sigma}\Sigma^{-1}] - nq. \tag{18}$$

Nominally, asymptotically, this will have a Chi Squared distribution on $q(q+1)/2$ degrees of freedom (which equals 3 in our examples where $q = 2$).

2.4.4. Performance Measures

For each dataset, we define a training dataset $(X^{(train)}, y^{(train)})$ and a validation dataset $(X^{(valid)}, y^{(valid)})$ with N observations. Performance is measured for the jth response using the prediction mean square error (PMSE):

$$\mathrm{PMSE}_j = \frac{1}{N}\sum_{i=1}^{N}\left(y_{i,j}^{(valid)} - \hat{y}_{i,j}\right)^2$$

and the mean absolute error (MAE):

$$\mathrm{MAE}_j = \frac{1}{N}\sum_{i=1}^{N}\left|y_{i,j}^{(valid)} - \hat{y}_{i,j}\right|$$

where the point prediction $\hat{y}_{i,j}$ is calculated using regression coefficients fitted with the training data. Predictions for the Bayesian methods are calculated using the posterior median of the regression coefficients. For the Bayesian methods, we also calculate the Log Predictive Score, which measures the accuracy of the posterior predictive distribution and, in the bivariate response case, uses the bivariate normal formula for plugging in median predictions. The Log Predictive Score is

$$\mathrm{LPS} = -\frac{1}{N}\sum_{i=1}^{N}\log p\left(Y_i^{(valid)}\middle| X_i^{(valid)}, X^{(train)}, y^{(train)}\right),$$

with $p(\cdot)$ being the bivariate normal density. We also provide the likelihood ratio discrepancy measure in Equation (18), as described above.

2.5. Two Datasets

2.5.1. Sugars Data

The original dataset involved mixtures of three sugars (glucose, fructose, and sucrose) in aqueous solution in designed concentrations, together with their near-infrared spectra (2nd-difference absorbance spectra at 700 contiguous wavelengths). It was originally described in Brown [27] and subsequently analysed in various papers—see, for example, [15]. The data are challenging in that they involve a large number of explanatory variables (700) and these are highly correlated, forming spectral curves. Water has a strong spectrum, which adds to difficulties in the prediction of mixture concentrations. Here, we address these difficulties by only considering two of the three sugars, glucose and sucrose, the first and third variables in the original dataset. We have not attempted any transformation of the response since the Beer–Lambert law (Beer's law) asserts that absorbance is linearly related to concentration, a fundamental principle of spectroscopy and a consequence of molecular vibration modes.

The original data are designed as 125 observations in a 5^3 experiment, with the 3 factors at 5 levels of composition by weight. A separate challenging validation sample is available, comprising an incomplete 3^3 design with 21 observations largely outside the range of concentrations in the training data. We measure the effectiveness of different methods of prediction from the 125 observational training data by comparing predictions and actual values for these 21 units.

2.5.2. Drug Discovery Data

Drug discovery, a primary objective of the pharmaceutical industry, involves the design of molecules that must achieve the highest levels of bio-activity to potentially become successful drugs. Aside from activity, the molecule must reach its target in the body, solubility being important, and small molecules are often the most marketable drugs with the fewest side effects. The original data analysed by Pickett et al. [41] contained 1704 synthesizable compounds that were exhaustively tested for both activity and solubility. They analysed all synthesisable compounds within a drug family involved in inflammatory diseases such as asthma. Aspects of design were discussed in Brown and Ridout [42]. The data have subsequently been re-expressed in terms of the presence or absence of fragments or functional complexes of atoms— see [3,28,43]—with $p = 3149$ attributes or fragments.

In practice, one would want to analyse relatively few compounds to predict the characteristics of all synthesisable compounds. Previous analyses used 'activity' as the response and the fragments as regressors in a univariate regression model. We also consider a calculated measure of the solubility of the molecule as an additional response. Our goal is a prediction of the bivariate responses (activity and solubility) using only a subset of 140 compounds as a training sample, which was envisaged as economically feasible by collaborators at GSK plc, using a multivariate regression. This leaves 1704 − 140 = 1564 compounds to be used as a validation set to examine the performance of our methods.

2.5.3. Design of Simulation Study

We wanted to further investigate the differences between the structural assumptions of the Bayesian methods and error covariance inflation for the conjugate models. For this, we used the output from the Normal–Gamma shared non-conjugate fit as the representative truth for the regression coefficients. We then superimposed bivariate errors on the multivariate model with these 'true' regression coefficients and errors with the same variances but three different correlations: $\rho = 0, 0.5$, and 0.99. We created 30 datasets, and we executed the MCMC chain for 100,000 iterations with a burn-in of 50,000 iterations for each dataset and model choice. As in the sugar data example, there are 125 observations in the training set and 21 observations in the validation set.

2.6. MCMC Convergence

Pilot runs were used to establish a suitable burn-in period and number of samples for each prior and dataset (or data-generating process in the simulation study). We considered trace plots of both the regression coefficients and the hyperparameters, finding that the chains had converged in the burn-in period and that the number of samples were sufficiently large.

2.7. Statistical Learning Methods

We also include versions of Lasso and Ridge to analyse the data. For this, we used the statistical software by Friedman et al. [44] in the Matlab version R2013, which includes a multi-response version of both Lasso and Ridge [45]. This is very much a pared-down implementation relative to versions of Ridge as described in [1,46]: it involves just a single shrinkage parameter, λ, but does give a non-Bayesian benchmark and, in the case of the Lasso, implicitly gives an alternative exponential-tailed prior distribution to contrast with the Bayesian exponential-tailed prior (Normal–Gamma) and the fatter-tailed priors of the horseshoe and Normal–Gamma–Gamma, in contrast to the thinner-tailed prior of Ridge regression. The *glmnet* methodology, although pseudo-Bayes, does not offer conjugate scaling. The single scaling parameter was chosen to minimise the 10-fold cross-validatory fit within the training data.

3. Results

We provide the results of two examples and a simulation study. We fit multivariate regression models using our priors as described (in both conjugate and non-conjugate form) to data from chemometrics and drug discovery. The predictive performance is compared to the statistical learning multivariate Ridge and Lasso.

3.1. Sugars Data for Two Responses

We consider eleven models: the separated non-conjugate versions, the shared non-conjugate versions, and the shared conjugate versions for the Horseshoe (HS), the Normal–Gamma–Gamma (NGG), and the Normal–Gamma (NG), as well as the the classical Lasso and Ridge. The results are given in Table 1.

These findings were obtained by retaining 100,000 samples after a burn-in period of 900,000 steps for all the methods (we do not thin the chain). The exponential-tailed Normal–Gamma is best in LPS, MAE, and PMSE within each structural assumption. The 21 validation sample has values of 0, 12, and 25 (seven of each) on both responses so that the overall prediction variance is more than 100 and, hence, the predictions are excellent, reducing this to less than 0.5. In other words, the models explain more than 99% of the validation sample variation.

The most striking feature is the difference between conjugate and non-conjugate Bayesian models on deviance and correlation and, to a lesser extent, the performance measures. The deviance measure is consistently large for all three priors in the conjugate models. Notable is the nature of the underestimation of Σ for the conjugate methods relative to the separated and shared non-conjugate methods; consistently, *RHO* is 0.05–0.07 for the conjugate methods but 0.37–0.40 for the non-conjugate methods. It would seem that, although, in the mixture experiment, the two sugars are correlated when it comes to the conjugate model, the cross-products term in Equation (14) is small and, together with the variance terms, is not able to compensate for the huge increase in degrees of freedom divisor, which is incremented by $p = 700$. In fact, the correlations extracted from Equation (14) are, respectively, -0.017, -0.00001, and -0.138 for horseshoe, Normal–Gamma, and Normal–Gamma–Gamma. Although the estimation of correlation is badly affected and the Log Predictive Score rose modestly, the estimates of regression coefficients are still good. We show later in the simulation study that it is only when correlation is very high that regression estimates deteriorate badly.

Table 1. Sugar data: Prediction Mean Square Error, Mean Absolute Error, and Log Predictive Score. $PMSE_1$, MAE_1, and LPS_1 stand for Prediction Mean Square Error, Mean Absolute Error, and Log Predictive Score for the first response. $PMSE_2$, MAE_2, and LPS_2 stand for Prediction Mean Square Error, Mean Absolute Error, and Log Predictive Score for the second response. MAE is the average of the MAE of the two responses and $PMSE$ is the average of the $PMSE$ of the two responses. LPS stands for Log Predictive Score. DEV stands for deviance. Based on Equation (18), as baseline, we consider the horseshoe shared non-conjugate. RHO is the median MCMC estimated correlation for Σ. The best performing model(s) for each structural group is (are) shown in bold. Validation set standard deviations in brackets.

	Separated			**Shared NC**			**Shared C**				
	HS	**NGG**	**NG**	**HS**	**NGG**	**NG**	**HS**	**NGG**	**NG**	**Lasso**	**Ridge**
$PMSE_1$	0.25	0.24	0.28	0.78	0.47	0.30	0.65	0.55	0.24	**0.16**	34.69
	(0.34)	(0.40)	(0.46)	(1.78)	(0.91)	(0.52)	(1.35)	(1.09)	(0.30)	(0.25)	(34.15)
MAE_1	0.41	0.37	0.43	0.61	0.51	0.43	0.59	0.56	0.41	**0.30**	4.99
	(0.30)	(0.32)	(0.32)	(0.66)	(0.47)	(0.34)	(0.56)	(0.50)	(0.29)	(0.26)	(3.2)1
LPS_1	0.65	0.55	0.67	2.13	1.21	0.70	1.67	1.25	0.76		
	(0.49)	(0.85)	(0.74)	(2.98)	(1.54)	(0.83)	(2.04)	(1.49)	(0.42)		
$PMSE_2$	0.50	0.44	0.19	0.33	0.27	0.18	0.36	0.26	**0.10**	0.47	20.88
	(0.58)	(0.54)	(0.23)	(0.33)	(0.33)	(0.21)	(0.40)	(0.31)	(0.14)	(0.59)	(22.50)
MAE_2	0.58	0.54	0.35	0.49	0.42	0.35	0.50	0.41	**0.26**	0.56	3.91
	(0.42)	(0.40)	(0.27)	(0.31)	(0.31)	(0.25)	(0.34)	(0.31)	(0.19)	(0.41)	(3.21)
LPS_2	1.37	1.22	0.65	1.45	0.92	0.57	1.02	0.73	**0.71**		
	(1.38)	(1.34)	(0.50)	(1.73)	(1.26)	(0.56)	(1.10)	(0.71)	(0.26)		
$PMSE$	0.37	0.34	0.24	0.56	0.37	0.24	0.50	0.41	**0.17**	0.32	27.79
	(0.33)	(0.31)	(0.24)	(0.88)	(0.46)	(0.27)	(0.70)	(0.55)	(0.16)	(0.30)	(23.43)
MAE	0.49	0.45	0.39	0.55	0.47	0.39	0.54	0.49	**0.33**	0.43	4.45
	(0.25)	(0.23)	(0.19)	(0.36)	(0.26)	(0.20)	(0.34)	(0.28)	(0.17)	(0.22)	(2.31)
LPS	2.04	1.76	1.42	2.57	1.89	**1.36**	2.22	1.89	1.55		
	(1.38)	(1.31)	(0.87)	(2.44)	(1.51)	(0.95)	(1.77)	(1.46)	(0.54)		
DEV	0.91	0.53	1.11	0	0.28	0.99	15.40	13.25	11.91		
RHO	0.40	0.40	0.40	0.38	0.37	0.40	0.05	0.06	0.07		

The Normal–Gamma, with its exponential tails, is consistently better on MAE than the Lasso, even sharing its exponential tails, and is better than all the other Bayesian methods, HS and NGG, with their polynomial tails. In fact, the Lasso, also with exponential tails, improves on both HS and NGG with their polynomial tails. Validation sample boxplots of the prediction performance for the Bayes methods are shown in Figure 1.

Of the statistical learning methods for the sugars data, whilst Lasso performs well, Ridge behaves poorly, leaving around one quarter of the variation unexplained. From minimising the 10-fold cross-validatory fit, the ridge shrinkage parameter is 39.803 (log value, 3.6839) and cannot avoid confusion in areas of the spectrum where water has a strong signal. The Lasso, on the other hand, with a 10-fold cross-validatory choice of shrinkage parameter of 0.0398 (log value −3.2239), concentrates on just a few regions of the spectrum—see Figure 2.

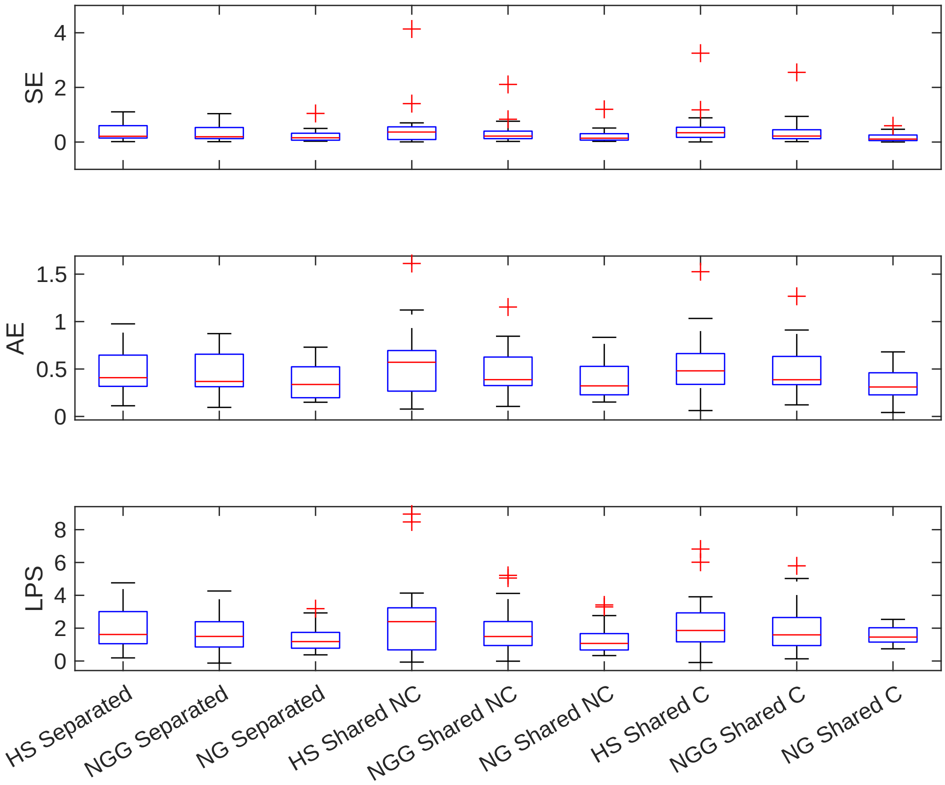

Figure 1. Sugar data. Top panel: Predicted squared error between observed and predicted values; Middle panel: Absolute error between observed and predicted values; Bottom panel: Log Predictive Scores of the observed values under the predicted distributions. All computations are performed across the 21 observations for the various methods displayed on the bottom panel.

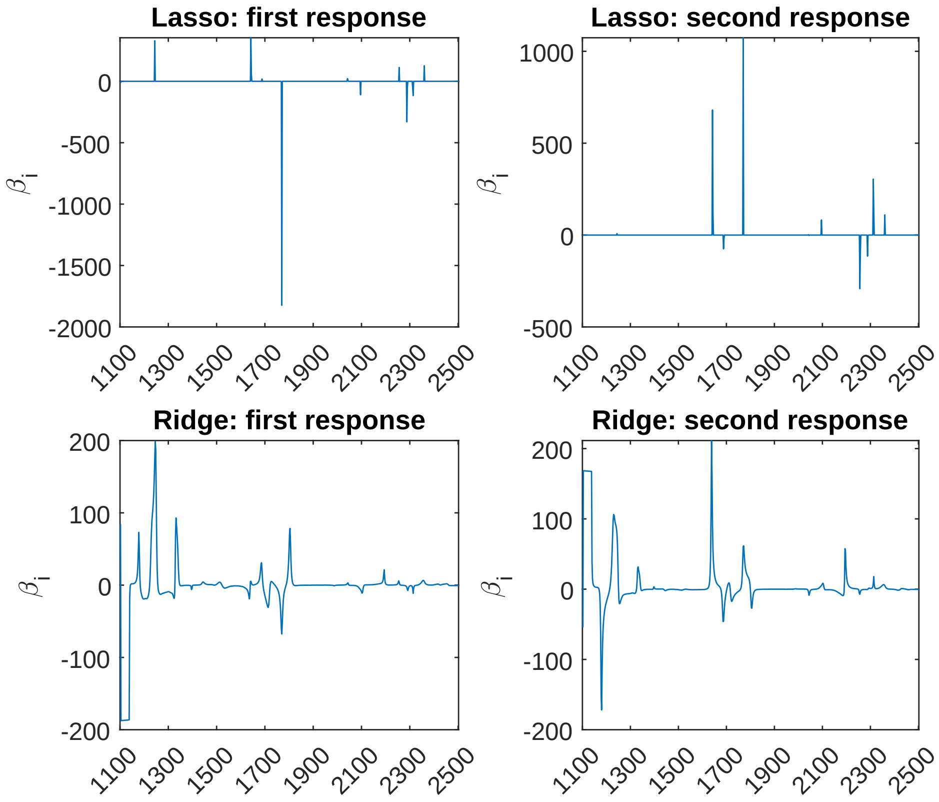

Figure 2. Sugar data. Top panel: Lasso-fitted spectrum from 1100 to 2500 nm for the two responses; Bottom panel: Ridge-fitted spectrum from 1100 to 2500 nm for the two responses.

3.2. Drug Discovery

We consider eleven models: the separated non-conjugate versions, the shared non-conjugate versions, and the shared conjugate versions for the horseshoe (HS), the Normal–Gamma–Gamma (NGG), and the Normal–Gamma (NG), along with the classical Lasso and Ridge. The results are given in Table 2. These findings were obtained by retaining 40,000 samples after a burn-in period of 100,000 steps for the NGG shared non-conjugate and conjugate, for the three HS, and for the NG separated. The findings for the other remaining approaches are produced by retaining 30,000 samples after a burn-in period of 70,000 steps (we thin the chain by removing all except the 10th sampled value).

The two responses each have a variance of around 1, so that an overall prediction Mean Square Error of about 0.3 for all nine Bayesian methods have reasonable predictive accuracy, with around 70% validation variance explained. But this leaves 30% variance unexplained, far worse than with the sugars data.

In fact, Table 2 shows remarkable consistency of predictions for all three Bayesian priors across the different structural assumptions. The Lasso, with its implicit exponential tails, also performs well, as well as the Bayesian methods. But the other classical method, Ridge regression, implicitly with its thin tails, again behaves much worse than all the methods. The two responses have relatively small correlation, but still the Bayesian conjugate model has a tendency to underestimate this, with small RHO for each of the three shared conjugate priors. Although there is evidence of reduced correlation between the responses, the Log Predictive Score does not highlight any elevated scores. The deviance measure of the covariance matrix arbitrarily compared with the horseshoe does not show excessive values, although conjugate values are somewhat elevated, but so are the separated values.

Table 2. Drug discovery example: Prediction Mean Square Error, Mean Absolute Error, and Log Predictive Score. $PMSE_1$, MAE_1, and LPS_1 stand for Prediction Mean Square Error, Mean Absolute Error, and Log Predictive Score for Activity. $PMSE_2$, MAE_2, and LPS_2 stand for Prediction Mean Square Error, Mean Absolute Error, and Log Predictive Score for Solubility. MAE is the average of the MAE of the two responses, and $PMSE$ is the average of the $PMSE$ of the two responses. LPS stands for Log Predictive Score. DEV stands for deviance. Using Equation (18) as baseline, we consider the horseshoe shared non-conjugate. RHO is the median MCMC estimated correlation for Σ. Validation data standard deviations are in brackets.

	Separated			**Shared NC**			**Shared C**				
	HS	**NGG**	**NG**	**HS**	**NGG**	**NG**	**HS**	**NGG**	**NG**	**Lasso**	**Ridge**
$PMSE_1$	0.24	0.25	0.24	0.27	0.29	0.25	0.27	0.26	0.27	0.27	0.46
	(0.43)	(0.43)	(0.42)	(0.55)	(0.59)	(0.48)	(0.54)	(0.51)	(0.54)	(0.49)	(0.51)
MAE_1	0.38	0.39	0.38	0.39	0.40	0.39	0.38	0.38	0.39	0.40	0.58
	(0.31)	(0.31)	(0.31)	(0.34)	(0.35)	(0.33)	(0.34)	(0.33)	(0.34)	(0.33)	(0.37)
LPS_1	0.75	0.77	0.72	0.73	0.76	0.71	0.72	0.70	0.71		
	(0.92)	(0.96)	(0.84)	(0.94)	(1.01)	(0.91)	(0.90)	(0.88)	(0.89)		
$PMSE_2$	0.34	0.34	0.35	0.33	0.34	0.33	0.34	0.33	0.33	0.32	0.73
	(0.83)	(0.83)	(0.85)	(0.75)	(0.79)	(0.72)	(0.75)	(0.76)	(0.76)	(0.77)	(1.10)
MAE_2	0.39	0.39	0.40	0.39	0.40	0.40	0.40	0.40	0.39	0.39	0.67
	(0.44)	(0.44)	(0.44)	(0.42)	(0.43)	(0.41)	(0.42)	(0.42)	(0.42)	(0.42)	(0.53)
LPS_2	0.72	0.68	0.72	0.72	0.71	0.72	0.75	0.70	0.71		
	(1.01)	(1.03)	(1.07)	(1.03)	(1.08)	(0.98)	(1.02)	(1.00)	(1.03)		
$PMSE$	0.29	0.30	0.30	0.30	0.31	0.29	0.30	0.30	0.30	0.29	0.60
	(0.48)	(0.49)	(0.49)	(0.49)	(0.52)	(0.45)	(0.49)	(0.48)	(0.49)	(0.46)	(0.58)
MAE	0.39	0.39	0.39	0.39	0.40	0.39	0.39	0.39	0.39	0.39	0.63
	(0.29)	(0.29)	(0.28)	(0.30)	(0.30)	(0.29)	(0.29)	(0.29)	(0.30)	(0.27)	(0.30)
LPS	1.48	1.44	1.45	1.45	1.45	1.44	1.49	1.44	1.44		
	(1.16)	(1.17)	(1.16)	(1.28)	(1.31)	(1.19)	(1.29)	(1.26)	(1.27)		
DEV	4.37	8.34	5.32	0	1.34	2.85	7.22	6.41	6.23		
RHO	0.20	0.22	0.20	0.11	0.15	0.17	0.002	0.002	0.004		

Drug discovery data have two responses that are very different, activity and solubility, and show little correlation in any of the methods. As we further demonstrate for the conjugate model in the simulation, it is when correlation is large that the estimation of Σ really deteriorates and predictions are affected.

To further enhance prediction using the fragments constructed by GSK from refining functional molecular groups, perhaps one should experiment with priors and methods that can take advantage of such hierarchies of fragments by methods akin to generalised wavelets or even factor models [32].

3.3. Simulation Using Sugars Data as Baseline

The design of the small simulation study is as described. Fitting the nine models delineated by two factors at three levels: Method [HS, NGG, NG] and Structure [Separated (S), Shared Non-Conjugate (Shared NC), Shared Conjugate (Shared C)]. The simulation provides ground truth for the error covariance matrices and the deviance measure.

The results are given in Table 3.

Table 3. Simulation results: Prediction Mean Square Error (PMSE); Mean Absolute Error (MAE), Log Predictive Score (LPS), Deviance (DEV), and *RHO* for the estimated correlation (standard error in brackets).

$\rho = 0$									
	Separated			**Shared NC**			**Shared C**		
	HS	**NGG**	**NG**	**HS**	**NGG**	**NG**	**HS**	**NGG**	**NG**
$PMSE$	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.003)	(0.002)	(0.002)
MAE	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21	0.21
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.006)	(0.004)	(0.004)
LPS	0.15	0.14	0.14	0.14	0.14	0.14	0.13	0.14	0.14
	(0.027)	(0.027)	(0.027)	(0.030)	(0.027)	(0.027)	(0.029)	(0.027)	(0.027)
DEV	0.03	0.02	0.02	0.02	0.02	0.02	0.01	0.02	0.02
	(0.00018)	(0.00013)	(0.00017)	(0.00009)	(0.00016)	(0.00018)	(0.00017)	(0.00031)	(0.00027)
RHO	−0.02	−0.02	−0.02	−0.02	−0.02	−0.02	−0.00	−0.00	−0.01
$\rho = 0.5$									
	Separated			**Shared NC**			**Shared C**		
	HS	**NGG**	**NG**	**HS**	**NGG**	**NG**	**HS**	**NGG**	**NG**
$PMSE$	0.08	0.08	0.08	0.08	0.11	0.08	0.09	0.10	0.15
	(0.003)	(0.003)	(0.005)	(0.003)	(0.003)	(0.003)	(0.003)	(0.005)	(0.006)
MAE	0.23	0.23	0.23	0.23	0.25	0.23	0.23	0.26	0.31
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)	(0.006)
LPS	0.21	0.20	0.20	0.20	0.20	0.20	0.38	0.73	0.88
	(0.028)	(0.028)	(0.028)	(0.029)	(0.028)	(0.027)	(0.022)	(0.021)	(0.020)
DEV	0.48	0.48	0.48	0.48	0.48	0.48	7.89	6.84	5.48
	(0.00092)	(0.00080)	(0.00094)	(0.00041)	(0.00098)	(0.00075)	(0.0056)	(0.00390)	(0.0045)
RHO	0.57	0.57	0.57	0.56	0.57	0.57	0.09	0.10	0.14
$\rho = 0.99$									
	Separated			**Shared NC**			**Shared C**		
	HS	**NGG**	**NG**	**HS**	**NGG**	**NG**	**HS**	**NGG**	**NG**
$PMSE$	0.07	0.07	0.07	0.07	0.07	0.07	0.12	0.17	0.18
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.005)	(0.006)	(0.007)
MAE	0.21	0.21	0.21	0.21	0.21	0.21	0.28	0.34	0.34
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.007)	(0.007)	(0.007)
LPS	−0.93	−0.92	−0.92	−1.00	−0.92	−0.92	0.63	1.20	1.35
	(0.025)	(0.0245)	(0.024)	(0.025)	(0.024)	(0.024)	(0.036)	(0.019)	(0.017)
DEV [1]	0.13	0.13	0.13	0.13	0.13	0.13	1.04	0.98	0.87
RHO	0.89	0.89	0.89	0.89	0.89	0.89	0.14	0.16	0.22

[1] refers to (1.0×10^4) of the value.

3.4. Commentary on the Simulation

Let us examine the three cases, $\rho = 0, 0.5$ and 0.99, which show steadily increasing deviance as correlation increases.

When $\rho = 0$, the conjugate model just has error variances as part of the generation of regression coefficients, so there is no pressure to correlate the responses. The model is then somewhat like the univariate model, although the dormant correlation could intrude, i.e., ρ is still in the model even though it is zero in the simulation. Unlike the separate non-conjugate model, the local parameters ψ_j are shared across the responses, and whether there is deviance inflation comes down to the arguments in Moran et al. [12] as to whether there are sufficient larger coefficients in one or both responses as in the second term in Equation (15) to compensate for the increased denominator for the estimation of Σ. From the simulation, there is no degradation of the error covariance matrix.

However, when $\rho = 0.5$, the story is different. The deviance measure is consistently large for all the conjugate methods. The *RHO* is also consistently much smaller for the shared conjugate model and much smaller than the simulated true value of 0.5, evidently a result of the prior data having a large influence both on variance and correlation. The Log Predictive Score also confirms the poorer fit of the conjugate models. The correlation is reduced in the pseudo-data given in Equation (14). The difference for conjugate models versus non-conjugate models for Σ is dramatic.

When $\rho = 0.99$, the estimation of both Σ and the regression coefficients is significantly affected and all the conjugate PMSE are larger by a factor of more than two compared to the non-conjugate results. The non-conjugate models provide a reasonably faithful estimation of *RHO*, whereas the conjugate models provide more deflated estimates, as anticipated by the prior data correlations. The cross-product correlation term in prior data, Equation (14) is very small, no more than 0.002 across the three conjugate models. The deviance measure, although large for all methods, is between six and eight times larger for the conjugate models.

There is also conflicting messaging: the pull of correlation versus the need to discriminate and shrink coefficients to zero. The drawing together of regression effects for the two responses is now stronger and affects the stability of estimation. Within any one of the structural assumptions—Separated, Shared NC, or Shared C—the differences between the priors are relatively small, whereas the structural assumptions have a much more marked effect.

4. Discussion

We have resisted commenting in detail on the computational speed of the different methods, as these depend so much on how the programming was implemented and size of data. Looking at the 30 replicate simulation timings, for the Normal–Gamma and the horseshoe, they were very similar, whereas, for the Normal–Gamma–Gamma, they were about 30% more. The biggest difference was for conjugate versus non-conjugate: the conjugate calculations were about 30% longer, but this was because, for the ease of programming, we chose to use the same method of fast updating involving pq regression coefficients without implementing the p parameter update, as described following Equation (11). The correlation, ρ, did not seem to influence the speed of computation.

We have reviewed the multivariate linear model with many explanatory variables common to the q responses. Using global–local priors' shrinkage on the regression coefficients allows us to find variables that are influential and tailored specifically and differently for the prediction of each response. The models can have *shared* hyperparameters or *separated* hyperparameters and be conjugate or not, constituting *structural* prior aspects. We find that such structural assumptions can be more influential than the detailed tail behaviour of the prior, be it with polynomial or exponential tails. In our examples in drug discovery and chemometrics, with data that are far from 'nearly black', although more marginal, exponential tails, such as those provided by the Normal–Gamma, are particularly good.

The assumption of conjugacy in particular, though useful for automatic scaling and sped-up 'trace-class' algorithms, can have severe implications for the estimation of the error covariance matrix and its correlation structure and co-variability. What is more, it can impinge on the estimation of the regression coefficients, strongly when correlation is strong. We therefore recommend that conjugate forms of the model be avoided in these multivariate contexts, at least when the estimation of Σ is needed and correlation is high. It would be strange indeed if such conjugate priors, where the regression coefficient distribution depends directly on the error characteristics, which is like pseudo-data with a similar generating mechanism as the data, but directed rather to variable selection and shrinkage, were without consequences. Here, the consequences can be quite extreme and belie the need for a prior to be weak or non-informative.

We have demonstrated these aspects in two examples, one in chemometrics with a significant covariance structure and 700 contiguous spectral measurements and another in drug discovery with more than three thousand covariates. These aspects have been corroborated in a small simulation study.

5. Conclusions

We aimed to look at two different aspects of the prior distribution: tail influence and structural assumptions. In the chemometric sugars data, across most measures, we found that the prior with exponential tails, in the form of the Normal–Gamma, outperformed priors with polynomial tails and both statistical learning methods, Lasso and Ridge regression. Regarding structural assumptions, the assumption of conjugacy strongly affected the estimation of the covariance matrix of errors. However, its effect on actual predictions was more nuanced, with large effects on the predictive distributions only seen when error covariance was sizeable. Sharing hyperparameters across the two responses was beneficial in the case of the sugars data but less so for the drug discovery data due to their small error correlation and more complex structure of explanatory variables. The drug discovery data generally showed far fewer differences between priors and the statistical learning Lasso procedure, although the conjugate structural assumption still showed a reduced estimation of correlation. Ridge performed poorly all around. The present study is limited to just two datasets, but we feel confident that other datasets will corroborate the findings, given the underpinning theory. There is scope for comparison to other prior distributions used in the univariate literature that have been extended to multivariate regression. Also, the methodology developed is not constrained to two responses and could be applied to examples with many responses and may indeed show richer instances of the problems with the conjugate Bayesian formulation.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/math13111812/s1, File S1: Supplement to 'Multivariate Bayesian global-local shrinkage methods for regularisation in the high dimensional linear model'.

Author Contributions: Conceptualisation, P.J.B.; methodology, P.J.B. and V.M.; formal analysis, V.M., D.S. and J.E.G.; data curation, V.M. and D.S.; writing—original draft preparation, P.J.B. and V.M.; writing—review and editing, P.J.B., V.M., D.S. and J.E.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The Sugars data may be obtained from the authors. Restrictions apply to the availability Drug Discovery data. These data were obtained from GlaxoSmithKline and are available from the authors with the permission of GlaxoSmithKline.

Acknowledgments: We thank Irene Poli and the European Centre for Living Technology for providing opportunities for interaction and discussions about this research. We especially would like to thank Darren Green and his team at GlaxoSmithKline Medicines Research Centre (UK) for their support and discussions over a number of years. Philip Brown was supported in this work by a Leverhulme Emeritus Fellowship EM-2018-059/9. V. Mameli was partially supported by the Departmental Strategic Plan (PSD) of the University of Udine, Dept. of Economics and Statistics (2022–2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Breiman, L.; Friedman, J.H. Predicting multivariate responses in multiple linear regression (with Discussion). *J. R. Stat. Soc. B* **1997**, *59*, 3–54. [CrossRef]
2. Bhadra, A.; Datta, J.; Polson, N.G.; Willard, B.T. Lasso meets horseshoe: A survey. *Stat. Sci.* **2019**, *34*, 405–427. [CrossRef]
3. Griffin, J.; Brown, P.J. Bayesian global-local shrinkage methods for regularisation in the high dimensional linear model. *Chemom. Intell. Lab. Syst.* **2021**, *210*, 104255. [CrossRef]
4. Tadesse, M.G.; Vannucci, M. *Handbook of Bayesian Variable Selection*; Chapman and Hall/CRC: New York, NY, USA, 2022.
5. Hauzenberger, N.; Huber, F.; Koop, G. Macroeconomic Forecasting Using BVARs. Book Chapter. Available online: https://www.researchgate.net/publication/376688025_Macroeconomic_Forecasting_Using_BVARs (accessed on 1 March 2025).
6. Carvalho, C.M.; Polson, N.G.; Scott, J.G. The horseshoe estimator for sparse signals. *Biometrika* **2010**, *97*, 465–480. [CrossRef]
7. Polson, N.G.; Scott, J.G. Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics 9*; Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., West, M., Eds.; Clarendon Press: Oxford, UK, 2011; pp. 501–538.
8. Griffin, J.E.; Brown, P.J. Inference with Normal-Gamma prior distributions in regression problems. *Bayesian Anal.* **2010**, *5*, 171–188.
9. Zhang, L.; Khare, K.; Xing, Z. Trace class Markov chains for Normal-Gamma Bayesian shrinkage model. *Electron. J. Stat.***2019**, *13*, 167–207. [CrossRef]
10. Bhattacharya, A.; Pati, D.; Pillai, N.S.; Dunson, D.B. Dirichlet-Laplace priors for optimal shrinkage. *J. Am. Stat. Assoc.* **2016**, *111*, 1479–1490. [CrossRef] [PubMed]
11. Van der Pas, S.L.; Salomond, J.B.; Schmidt-Hieber, J. Conditions for posterior contraction in sparse normal means problems. *Electron. J. Stat.* **2016**, *8*, 976–1000. [CrossRef]
12. Moran, G.E.; Ročková, V.; George, E.I. Variance prior forms for high dimensional variable selection. *Bayesian Anal.* **2019**, *14*, 1091–1119. [CrossRef]
13. Bai, R.; Ghosh, M.L. High-Dimensional Posterior Consistency under Global-Local Shrinkage Priors. *J. Multivar. Anal.* **2018**, *167*, 157–170. [CrossRef]
14. Kundu, D.; Mitra, R.; Gaskins, J.T. Bayesian variable selection for multi-outcome models through shared shrinkage. *Scand. J. Stat.* **2021**, *48*, 295–320. [CrossRef]
15. Brown, P.J.; Vannucci, M.; Fearn, T. Multivariate Bayesian Variable Selection and Prediction. *J. R. Stat. Soc. Ser. B* **1998**, *60*, 627–641. [CrossRef]
16. Brown, P.J.; Fearn, T.; Vannucci, M. The Choice of Variables in Multivariate Regression: A Non-conjugate Bayesian Decision Theory Approach. *Biometrika* **1999**, *86*, 635–648. [CrossRef]
17. Petretto, E.; Bottolo, L.; Langley, S.R.; Heinig, M.; McDermott-Roe, C.; Sarwar, R.; Pravenec, M.; Hübner, N.; Aitman, T.; Cook, S.; et al. New Insights into Genetic Control of Gene Expression using a Bayesian Multi-tissue Approach. *PLoS Comput. Biol.* **2010**, *6*, e1000737. [CrossRef]
18. Bhadra, A.; Mallick, B.K. Joint High-Dimensional Bayesian Variable and Covariance Selection with an application to eQTL analysis. *Biometrics* **2013**, *69*, 447–457. [CrossRef]
19. Brown, P.J.; Vannucci, M.; Fearn, T. Bayes model averaging with selection of regressors. *J. R. Stat. Soc. B* **2002**, *64*, 519–536. [CrossRef]
20. Bottolo, L.; Banterle, M.; Richardson, S.; Ala-Korpela, M.; Järvelin, M.; Lewin, A. A computationally efficient Bayesian seemingly unrelated regressions model for high-dimensional quantitative trait loci discovery. *J. R. Stat. Soc. C* **2021**, *70*, 886–908. [CrossRef]
21. Brown, P.J.; Le, N.D.; Zidek, J.V. Inference for a covariance matrix. In *Aspects of Uncertainty: A Tribute to D.V. Lindley*; Freeman, P.R., Smith, A.F.M., Eds.; Wiley: Chichester, UK, 1994; pp. 77–92.

22. Le, N.D.; Sun, L.; Zidek, J.V. Spatial interpolation and temporal backcasting for environmental fields having monotone data patterns. *Can. J Stat.* **2001**, *29*, 529–554. [CrossRef]
23. Liquest, B.; Mengerson, K.; Pettitt, A.N.; Sutton, M. Bayesian variable selection regression of Multivariate responses for group data. *Bayesian Anal.* **2017**, *12*, 1039–1067.
24. Griffin, J.; Brown, P.J. Hierarchical shrinkage priors for regression models. *Bayesian Anal.* **2017**, *12*, 135–159. [CrossRef]
25. Bhattacharya, A.; Chakraborty, A.; Mallick, B.K. Fast sampling with Gaussian scale-mixture priors in high-dimensional regression. *Biometrika* **2016**, *103*, 985–991. [CrossRef] [PubMed]
26. Dawid, A.P. Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika* **1981**, *68*, 265–274. [CrossRef]
27. Brown, P.J. *Measurement, Regression, and Calibration*; Clarendon Press: Oxford, UK, 1993.
28. Slanzi, D.; Mameli, V.; Brown, P.J. A comparative study on high-dimensional Bayesian regression with binary predictors. *Commun. Stat. Simul. Comput.* **2023**, *52*, 1979–1999. [CrossRef]
29. Bhadra, A.; Datta, J.; Poison, N.G.; Scott, J.G. The horseshoe+ estimator of ultra-sparse signals. *Bayesian Anal.* **2017**, *12*, 1105–1131. [CrossRef]
30. Li, Y.; Datta, J.; Craig, B.A.; Bhadra, A. Joint mean-covariance estimation via the horseshoe. *J. Multivar. Anal.* **2021**, *183*, 104716. [CrossRef]
31. Zhang, Y.D.; Naughton, B.P.; Bondell, H.D.; Reich, B.J. Bayesian Regression using a prior on the model fit: The R2D2 Shrinkage prior. *J. Am. Stat. Assoc.* **2022**, *117*, 862–874. [CrossRef]
32. Frühwirth-Schnatter, S.; Hosszejni, D.; Lopes, H.F. Sparse Bayesian Factor Analysis when the number of factors is unknown (with Discussion). *Bayesian Anal.* **2025**, *20*, 213–344. [CrossRef]
33. Park, T.; Casella, G. The Bayesian Lasso. *J. Am. Stat. Assoc.* **2008**, *103*, 681–686. [CrossRef]
34. Makalic, E.; Schmidt, D.F. A simple sampler for the horseshoe estimator. *IEEE. Signal Process. Lett.* **2010**, *23*, 179–182. [CrossRef]
35. Piironen, J.; Vehtari, A. On the Hyperprior Choice for global shrinkage parameter in the horseshoe prior. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), Fort Lauderdale, FL, USA, 20–22 April 2017; Volume 54, pp. 905–913.
36. Van der Pas, S.L.; Szabo, B.; van der Vaart, A.W. Adaptive posterior contraction rates for the horseshoe. *Electron. J. Stat.* **2017**, *11*, 3196–3225. [CrossRef]
37. Van der Pas, S.L.; Kleijn, B.J.K.; van der Vaart, A.W. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electron. J. Stat.* **2014**, *8*, 2585–2618. [CrossRef]
38. Cadonna, A.; Frühwirth-Schnatter, S.; Knaus, P. Triple the gamma—A unifying shrinkage prior for variance and variable selection in sparse state space and TVP models. *Econometrics* **2020**, *8*, 20. [CrossRef]
39. Chipman, H.; George, E.I.; McCulloch, R.E. *The Practical Implementation of Bayesian Model Selection*; Anderson, T.W., Fang, K.T., Olkin, I., Eds.; IMS Lecture-Notes Monograph Series; Institute of Mathematical Statistics: Hayward, CA, USA, 2001; Volume 38.
40. Mardia, K.V.; Kent, J.T.; Bibby, J.M. *Multivariate Analysis*; Academic Press: London, UK, 1979.
41. Pickett, S.D.; Green, D.V.S.; Hunt, D.L.; Pardoe, D.A.; Hughes, I. Automated lead optimisation of MMP-12 inhibitors using a genetic algorithm. *ACS Med. Chem. Lett.* **2011**, *2*, 28–33. [CrossRef]
42. Brown, P.J.; Ridout, M.S. Level-screening designs for factors with many levels. *Ann. Appl. Stat.* **2016**, *10*, 864–883. [CrossRef]
43. Mameli, V.; Slanzi, D.; Poli, I.; Green, D.V.S. Search for relevant subsets of binary predictors in high dimensional regression for discovering the lead molecule. *Pharm. Stat.* **2021**, *20*, 898–915. [CrossRef]
44. Friedman, J.; Hastie, T.; Tibshirani, R. Regularisation paths for Generalised Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]
45. Qian, J.; Hastie, T.; Friedman, J.; Tibshirani, R.; Simon, N. Glmnet for Matlab. 2013. Available online: http://hastie.su.domains/glmnet_matlab/ (accessed on 1 March 2025).
46. Brown, P.J.; Zidek, J.V. Adaptive multivariate ridge regression. *Ann. Stat.* **1980**, *8*, 64–74. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.