Bayesian Analysis (2017) **12**, Number 1, pp. 135–159

Hierarchical Shrinkage Priors for Regression Models

Jim Griffin* and Phil Brown†

Abstract. In some linear models, such as those with interactions, it is natural to include the relationship between the regression coefficients in the analysis. In this paper, we consider how robust hierarchical continuous prior distributions can be used to express dependence between the size but not the sign of the regression coefficients. For example, to include ideas of heredity in the analysis of linear models with interactions. We develop a simple method for controlling the shrinkage of regression effects to zero at different levels of the hierarchy by considering the behaviour of the continuous prior at zero. Applications to linear models with interactions and generalized additive models are used as illustrations.

Keywords: Bayesian regularization, interactions, structured priors, strong and weak heredity, generalized additive models, normal-gamma prior, normal-gamma-gamma prior, generalized beta mixture prior.

1 Introduction

Regression modelling is an important method of understanding the effect of predictor variables on a response. These effects can be hard to estimate and interpret if the predictor variables are highly correlated (the problem of collinearity) or there are many predictor variables. These problems are often addressed by variable selection or regularization which can lead to more interpretable models and better out-of-sample prediction. If the regression effects are considered related, it is natural to include this information in the variable selection or regularization to improve inference.

In a Bayesian framework, regression effects can be regularized using zero-mean scale mixtures of normals to give a wide class of priors for regression coefficients (see *e.g.* Polson and Scott, 2011) in which the prior density can be expressed as

$$\pi(\beta_i) = \int \mathrm{N}(0, \Psi_i)\, dG(\Psi_i)$$

where G is a distribution function with density g (if it exists). Many priors fit into this class. "Two group" priors, as classified by Polson and Scott (2011), assume that G is a discrete mixing distribution with two possible values. This class includes the spike-and-slab prior (Mitchell and Beauchamp, 1988) where G has an atom at zero and the stochastic search variable selection prior (George and McCulloch, 1993) where G has atoms at two non-zero values. Alternatively, the class of "one group" priors assume

*School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7NF, J.E.Griffin-28@kent.ac.uk

†School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7NF

 DOI: 10.1214/15-BA990

that G is absolutely continuous and encourage shrinkage of regression coefficients close to zero. Examples include the double exponential (Park and Casella, 2008; Hans, 2009) (leading to the Bayesian Lasso), the normal-gamma (Caron and Doucet, 2008; Griffin and Brown, 2010) the Bayesian elastic net (Hans, 2011), the horseshoe prior (Carvalho et al., 2010), the normal-exponential-gamma (NEG) (Griffin and Brown, 2011), the generalized Beta mixtures (Armagan et al., 2011), the generalized t (Lee et al., 2012) or double Pareto prior (Armagan et al., 2013) and the exponential power prior (Polson et al., 2013).

Priors for regression models often assume independence between the regression coefficients. This assumption is questionable if there are known or suspected relationships between the predictor variables. In linear models with interactions, one common classical heuristic (*strong heredity*) for variable selection is that a two-way interaction term can only be included if both main effects terms are included. Chipman (1996) and Chipman et al. (1997) use a spike-and-slab prior with strong heredity interpreted as a belief that the prior probability of inclusion of a two-way interaction coefficient is related to inclusion of the two associated main effects. Of course, other assumptions could be made but it is clear that it is often natural to assume a relationship between the usefulness of the interaction term and the usefulness of the main effects. More generally, a Bayesian version of the group Lasso (Yuan and Lin, 2006) was developed by Kyung et al. (2010) and Raman et al. (2009). A different approach is taken by Griffin and Brown (2012) who defined priors which allow correlation between the effects rather than dependence through the absolute effect sizes (as implied by the group Lasso). This idea has also been applied to unifying and robustifying ridge and g-priors for regression in Griffin and Brown (2013). Structured priors have also been proposed in biological application, *e.g.* Yi et al. (2007), Stingo et al. (2011), Li and Zhang (2010) and Rockova and Lesaffre (2014). Hierarchical shrinkage priors have been also used in other areas such as factor models (Bhattacharya and Dunson, 2011).

In this paper, we concentrate on hierarchical priors for regression problems where relationships between the predictor variables can be assumed and regression coefficients can be arranged in levels. Regression coefficients in higher levels will usually add additional complexity to the model and so need to be, both, more aggressively shrunk to zero to avoid over-fitting and dependent on the importance of regression coefficients at lower levels. Specifically, we consider priors in which regression coefficients at one level depend on a subset of the effect sizes at lower levels. This is a fairly general structure which can include different grouping structures (see *e.g.* Yuan and Lin, 2006; Jacob et al., 2009) in a simple way, whilst also expressing much more complicated structures. The methodology gives a general and relatively simple way of controlling the shrinkage at different levels of the hierarchy.

Our methods are distinguished from earlier approaches to hierarchical structure through:

(i) use of a continuous prior distribution (the one group problem) that does not attach an extra premium on regression coefficients being exactly zero as would arise from some implementations of spike-and-slab priors as in (Chipman, 1996) and some citations above;

(ii) use of Bayesian tools unlike Bien et al. (2013) or Yuan et al. (2009);

(iii) use of flexible and possibly heavy tailed priors, unlike Chipman et al. (1997) or Yi et al. (2007);

(iv) providing a simple method to control shrinkage at various levels of the hierarchy.

The paper is organized as follows. Section 2 explains the use of normal-gamma and normal gamma-gamma (or generalized beta mixture) priors. Section 3 considers hierarchical priors for regression models and develops a simple way to understand the shrinkage at different levels. Section 4 briefly describes computational strategies for inference in models using these priors. Section 5 includes applications of hierarchical shrinkage priors to linear models with interactions, general additive models and general additive models with interactions. A discussion follows in Section 6. The Supplementary Material (Griffin and Brown, 2016) contains, (A) a proposition which aids graphing through standardisation and (B) proofs of the theorems.

2 Continuous priors for sparse regression

The normal linear regression model for an $(n \times 1)$-dimensional vector of responses y and an $(n \times p)$-dimensional design matrix $\mathbf{X}$ is

$$y = \alpha 1 + \mathbf{X}\beta + \epsilon \tag{1}$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2 \mathbf{I}_n)$, 1 is a $(n \times 1)$-dimensional vector of 1's, α is an intercept and β is a $(p \times 1)$-dimensional vector of regression coefficients. We assume that the variables have been measured on comparable scales (or transformed to comparable scales). The prior is assumed to have the form $p(\alpha, \sigma^2, \beta) \propto \sigma^{-2} p(\beta)$, which is scale-invariant for α and σ^2, and we will concentrate on the choice of $p(\beta)$.

A common prior for β assumes independence conditional on $\Psi = (\Psi_1, \ldots, \Psi_p)$ and

$$\beta_j \sim \mathrm{N}(0, \Psi_j), \qquad j = 1, \ldots, p. \tag{2}$$

The parameter Ψ_j is the conditional variance of β_j and smaller values of Ψ_j imply that the prior favours smaller values of $|\beta_j|$. The value of Ψ_j can be seen as measuring the importance of the j-th variable with larger values of Ψ_j representing more importance.

In this paper, we will consider two specific priors. Firstly, the normal-gamma prior (Caron and Doucet, 2008; Griffin and Brown, 2010) has the form

$$\beta_j \sim \mathrm{N}(0, \Psi_j), \qquad \Psi_j \sim \mathrm{Ga}(\lambda, \gamma).$$

Here λ is a shape parameter and γ the rate parameter and the prior variance is $\mathrm{V}[\beta_j] = \mathrm{E}(\Psi_j) = \frac{\lambda}{\gamma}$ with exponential tails. Secondly, for a heavier tailed alternative, the generalized beta mixture prior distribution (Armagan et al., 2011) can be expressed as a hierarchical extension of the normal-gamma prior

$$\beta_j \sim \mathrm{N}(0, \Psi_j), \qquad \Psi_j \sim \mathrm{Ga}(\lambda, \gamma_j), \qquad \gamma_j \sim \mathrm{Ga}(c, d). \tag{3}$$

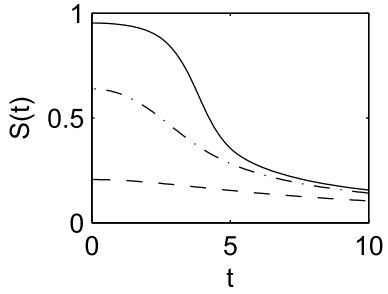

Figure 1: Shrinkage profiles for an NG prior with $\lambda = 0.1$, (solid line), $\lambda = 1$ (dot-dashed line) and $\lambda = 5$ (dashed line) with $\gamma = 1/\text{SE}^2$.

and the prior variance is $\text{V}[\beta_j] = \frac{\lambda d}{c-1}$ if $c > 1$. We will refer to this distribution as the normal-gamma-gamma (λ, c, d) prior distribution to emphasize the link to the normal-gamma distribution. The hyperparameters have simple interpretations: d is a scale parameter, λ controls the behaviour of the distribution close to zero and c controls the tail behaviour of the distribution. The marginal density of β_j is not available in closed form but the marginal distribution of Ψ_j is a gamma-gamma distribution which has the density

$$g(\Psi_j) = \left(\frac{1}{d}\right)^{\lambda} \frac{\Gamma(\lambda + c)}{\Gamma(\lambda)\Gamma(c)} \Psi_j^{\lambda - 1} \left(1 + \frac{\Psi_j}{d}\right)^{-(\lambda + c)}.$$

This prior will be written $\Psi_j \sim \text{GG}(\lambda, c, d)$; and corresponds to the inverted-beta-2 distribution of Raiffa and Schlaifer (1961, Section 7.4.2). The monotone transformation $\frac{\Psi_j}{\Psi_j + d}$ has a beta distribution with parameters λ and c implying that the median of Ψ_j is d if $\lambda = c$. This is a useful characterisation if $c \leq 1$ and the mean does not exist. In particular, this is true for the horseshoe prior (Carvalho et al., 2010) which occurs if $\lambda = c = 1/2$. Several of the absolutely continuous priors for regression coefficients described in Section 1 can be written as special cases of the normal-gamma-gamma distribution including the NEG distribution which arises when $\lambda = 1$ and the normal-gamma distribution which arises if $c/d = \mu$ as $c \to \infty$.

Results for linear regression models which express the posterior expectation and variance in terms of the least squares estimate of β and the variance of its sampling distribution (for $n > p$) have been derived by several authors including Griffin and Brown (2010) and Polson and Scott (2012). We consider a linear regression model with one regressor and write $\text{E}[\beta|\hat{\beta}] = (1 - S(t))\hat{\beta}$ where $\hat{\beta}$ is the least squares estimate of β with standard error SE and $S(t)$ is a function of the t-statistic, $t = \hat{\beta}/\text{SE}$. The function $0 \leq S(t) \leq 1$ is referred to as the shrinkage profile since it measures the amount that the least squares estimate is shrunk to zero. We say that a regression coefficient is more aggressively shrunk to zero if $S(t)$ is closer to one for small t.

Figure 1 show shrinkage profiles for a normal-gamma prior with different values of λ. Smaller values of λ increasingly favour more aggressive shrinkage of small least squares

estimates. This is intuitively reasonable since this parameter controls the shape of the distribution of Ψ_i at small values. A normal-gamma-gamma prior will also give similar results. Consequently, we define an adaptive shrinkage parameter for a prior distribution in terms of the prior density of Ψ_i as $\sup\{z|p(\Psi_i) = O(\Psi_i^{z-1}) \text{ as } \Psi_i \to 0\}$ where $p(\Psi_i)$ is the prior density of Ψ_i which characterises the range of $S(t)$. This will be simply λ in the case of both the normal-gamma and normal-gamma-gamma prior distributions.

3 Hierarchical shrinkage priors

3.1 Motivating examples

Before looking at shrinkage within a general hierarchical structure, it is useful to set the context by considering two statistical models: the linear models with interactions and the generalized additive model. These illustrate the need for priors which can express relationships between regression coefficients with different levels of adaptive shrinkage for some regression coefficients.

Linear models with interaction terms

Variable selection and regularization methods for linear models with interactions have received attention in the literature (Chipman, 1996; Chipman et al., 1997; Yuan et al., 2007). The model assumes that response y_i which is observed with covariates $X_{i1}, \ldots, X_{ip}$ can be expressed as

$$y_i = \alpha + \sum_{j=1}^{p} X_{ij}\beta_j + \sum_{j=1}^{p}\sum_{k=1}^{j-1} X_{ij}X_{ik}\delta_{jk} + \epsilon_i, \qquad \text{for } i = 1, \ldots, n \tag{4}$$

where $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. It is often considered natural to make the inclusion of an interaction contingent on the inclusion of main effects. Chipman et al. (1997) formalize this idea using two forms of the heredity principle. *Strong heredity* states that an interaction can only be included if both main effects are included. *Weak heredity* states that an interaction can be included if at least one main effects is included. The use of strong or weak heredity suggests beliefs which are inconsistent with an assumption of prior independence between the regression coefficients. It is also natural to assume that, *a priori*, the signs of the interactions are not related to the signs of the main effects with the coefficients of the interactions being shrunk more aggressively to zero than the coefficients of the main effects.

Generalized additive models

The generalized additive model (GAM) (Hastie and Tibshirani, 1993) is a non-linear regression model which represents the mean of the response as a linear combination of potentially non-linear functions of each variable so that

$$y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i$$

where $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ and f_j are function to be estimated from the data. Reviews of Bayesian analysis of these models are given by Kohn et al. (2001) and Denison et al. (2002). A common approach assumes that each non-linear function can be represented as a linear combination of basis functions so that, *e.g.*,

$$f_j(X_{ij}) = \theta_j X_{ij} + \sum_{k=1}^{K} \gamma_{jk}\, g(X_{ij}, \tau_{jk})$$

where $g(x, \tau_{j1}), \ldots, g(x, \tau_{jK})$ are a set of basis functions with knot points $\tau_{j1}, \ldots, \tau_{jK}$. This leads to a linear model for the responses

$$y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i = \sum_{j=1}^{p} \theta_j X_{ij} + \sum_{j=1}^{p} \sum_{k=1}^{K} \gamma_{jk}\, g(X_{ij}, \tau_{jk}) + \epsilon_i. \tag{5}$$

The set of knot points is often chosen to be relatively large and many γ_{jk}'s are set to zero to avoid over-fitting. In a Bayesian framework, this is usually approached as a variable selection problem and so we effectively have p different variable selection problems (one for each variable). We will refer to this as selection at the *basis level*. There is also the more standard variable selection problem of choosing a subset of the variables which are useful for predicting the response. The effect of the j-th variable is removed from the model if θ_j and $\gamma_{j1}, \ldots, \gamma_{jK}$ are all set to zero. We refer to this as selection at the *variable level*. In this model, prior independence between the coefficients for the j-th variable (θ_j) and $(\gamma_{j1}, \ldots, \gamma_{jK})$ seems unreasonable and dependence in size (rather than the sign) of these coefficients will be reasonable in many problems. Typically, we would like different types of adaptive shrinkage at the basis level and the variable level which suggests a prior with at least two adaptive shrinkage parameters.

3.2 General construction

The examples in Section 3.1 illustrate the need for priors which allow dependence in the size of regression coefficients but not their sign with hyperparameters that control the level of adaptive shrinkage implied by the prior for different regression coefficients. Hierarchical priors are a simple and useful way to build such a prior distribution. We assume that the regression coefficients can be arranged in L levels and that $\beta^{(l)}$ is the $(p_l \times 1)$-dimensional vector of regression coefficients in the l-th level. Intuitively, higher levels add additional flexibility (and complexity) to the model and so the inclusion of these regression coefficients would depend on regression coefficients at earlier levels. For example, the first level could refer to a linear regression with main effects only and higher levels add interactions of increasing order. The prior for the regression coefficients at a particular level are assumed to have the same adaptive shrinkage parameter *a priori* and typically become more aggressively shrunk to zero at higher levels. Our general prior, building on (2), assumes that

$$\beta_j^{(l)} \overset{ind.}{\sim} \mathrm{N}\left(0, \Psi_j^{(l)}\right), \qquad j = 1, \ldots, p_l, \quad l = 1, \ldots, L,$$

and

$$\Psi_j^{(l)} = a_l \frac{f_l\left(\Psi^{(1)}, \ldots, \Psi^{(l-1)}\right)}{\mathrm{E}[f_l\left(\Psi^{(1)}, \ldots, \Psi^{(l-1)}\right)]} \eta_j^{(l)}, \qquad j = 1, \ldots, p_l, \quad l = 1, \ldots, L \tag{6}$$

where f_l is a function only taking non-negative values, $\eta_1^{(l)}, \ldots, \eta_{p_l}^{(l)}$ are independent of $\Psi^{(1)}, \ldots, \Psi^{(l-1)}$, with $\eta_1^{(l)}, \ldots, \eta_{p_l}^{(l)} \overset{i.i.d.}{\sim} G_l$ where G_l is a distribution specific to the l-th level with $\mathrm{E}[\eta_j^{(l)}] = 1$ and $a_l = \mathrm{E}[\Psi_j^{(l)}] = \mathrm{V}[\beta_j^{(l)}]$ is a level-specific scale parameter. The prior allows correlation between $\Psi_i^{(l)}$ and $\Psi_j^{(m)}$ but $\beta_i^{(l)}$ and $\beta_j^{(m)}$ are uncorrelated (although, they can be dependent). The function f_l controls the effects of $\Psi^{(1)}, \ldots, \Psi^{(l-1)}$ on $\Psi^{(l)}$ and will usually be a simple function whose expectation can be easily calculated, *e.g.* a combination of additions and multiplications. Products have the useful property of being small if one element in the product is small and sum have the useful property of being small if all elements in the sum are small. Other choices of f_l, such as minimum or maximum are possible, but calculation of the expectation could be difficult. The structure is quite general. For example, a Bayesian group lasso (Kyung et al., 2010; Raman et al., 2009) arises from taking a single level, setting $\Psi_i^{(1)} = \Psi_j^{(1)}$ if i and j are in the same group and choosing $\eta_j^{(l)}$ to have a gamma distribution. The construction could be extended to a prior where the regression coefficients are correlated by assuming that β are dependent conditional on Ψ but this is not considered in this paper.

Example: Linear model with interaction terms

In our framework, we interpret *strong heredity* as a prior belief that δ_{jk} in (4) will be strongly shrunk to zero if either β_j or β_k are strongly shrunk to zero. We interpret *weak heredity* as a prior belief that δ_{jk} will be strongly shrunk to zero if both β_j and β_k are strongly shrunk to zero. These prior beliefs can be represented using the prior in (6) with $L = 2$. The first level contains the main effects and has $p_1 = p$ terms listed as $\beta_1, \ldots, \beta_p$. The second level contains the interactions and has $p_2 = p(p-1)/2$ terms listed as δ_{jk} for $k = 1, \ldots, j-1$, $j = 1, \ldots, p$.

In the case of strong heredity, we use the prior

$$\beta_j \sim \mathrm{N}\left(0, a_1 \eta_j^{(1)}\right) \text{ and } \delta_{jk} \sim \mathrm{N}\left(0, a_2 \eta_{jk}^{(2)} \eta_j^{(1)} \eta_k^{(1)}\right).$$

The prior variance of δ_{jk} is small if at least one of $\eta_j^{(1)}$, $\eta_k^{(1)}$ or $\eta_{jk}^{(2)}$ is small. Therefore, an interaction term δ_{jk} will tend to be small (since its variance is small) if either $\eta_{jk}^{(2)}$ is small or if at least one of β_j or β_k are small (which suggests that at least one of $\eta_j^{(1)}$ or $\eta_k^{(1)}$ is small). In the case of weak heredity, we use the prior

$$\beta_j \sim \mathrm{N}\left(0, a_1 \eta_j^{(1)}\right) \text{ and } \delta_{jk} \sim \mathrm{N}\left(0, a_2 \eta_{jk}^{(2)} \frac{1}{2}\left(\eta_j^{(1)} + \eta_k^{(1)}\right)\right).$$

The prior variance of δ_{jk} is small if $\eta_{jk}^{(2)}$ is small or *both* $\eta_j^{(1)}$ and $\eta_k^{(1)}$ are small. Therefore, the interaction terms will tend to be small if $\eta_{jk}^{(2)}$ is small or if both β_j and β_k are small (using similar reasoning to the strong heredity case).

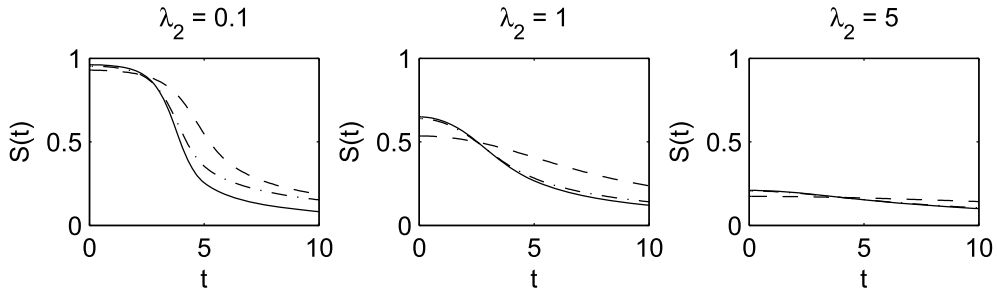

Figure 2: Shrinkage profiles for the model with the ShIS and ScIS priors with $\lambda_1 = 10\lambda_2$. The shrinkage profile are for: $\beta^{(2)}$ with the ShIS prior (solid line), $\beta^{(2)}$ with the ScIS prior (dashed line), and $\beta^{(2)}$ with the one-stage prior (dot-dashed line) with $d = 1/\mathrm{SE}^2$.

3.3 Comparative shrinkage propagation

The hierarchical prior induces a marginal posterior distribution of the regression coefficients at each level. Clearly, smaller values of scale a_l will tend to lead to more shrinkage of the marginal posterior mean of $\beta_j^{(l)}$. However, the influence of G_l is less clear. We will concentrate on the shape of the shrinkage profiles for the marginal posterior mean of $\beta_j^{(l)}$ and investigate its dependence on $G_1, \ldots, G_l$. This will provide a rationale for choosing a_l and G_l to give particular shrinkage profiles.

As an illustration, we consider a two level prior with one regression coefficient at each level where f_l is a product. It is assumed that $\mathrm{V}[\beta^{(1)}] = \lambda_1\, d$ and $\mathrm{V}[\beta^{(2)}] = \lambda_2\, d$ (where $\lambda_2 \leq \lambda_1$ to induce the same or greater shrinkage at the second level), G_1 is a $\mathrm{Ga}(\lambda_1, \lambda_1)$ distribution. To illustrate the importance of the choice of G_2, we consider two choices: G_2 is the same distribution as G_1, which we refer to as a *scale-induced shrinkage* (ScIS) prior, or, G_2 is a $\mathrm{Ga}(\lambda_2, \lambda_2)$ distribution, which we refer to as a *shape-induced shrinkage* (ShIS) prior. The shrinkage profile for $\beta^{(1)}$ depends only on G_1 and will be the same for both prior. However, the shrinkage profile of $\beta^{(2)}$ will differ. The marginal priors for $\beta^{(2)}$ have the forms

$$\beta_j^{(2)} \sim \mathrm{N}\left(0, \lambda_2\, d\, \Psi\right), \qquad \Psi \sim \mathrm{Ga}(\lambda_1, \lambda_1/\eta_j^{(1)}), \qquad \eta_j^{(1)} \sim \mathrm{Ga}(\lambda_1, \lambda_1)$$

with the ScIS prior and

$$\beta_j^{(2)} \sim \mathrm{N}\left(0, \lambda_2\, d\, \Psi\right), \qquad \Psi \sim \mathrm{Ga}(\lambda_2, \lambda_2/\eta_j^{(1)}), \qquad \eta_j^{(1)} \sim \mathrm{Ga}(\lambda_1, \lambda_1)$$

with the ShIS prior. For comparison, we consider the one-level prior

$$\beta_j^{(2)} \sim \mathrm{N}\left(0, \lambda_2\, d\, \Psi\right), \qquad \Psi \sim \mathrm{Ga}(\lambda_2, \lambda_2)$$

which has the same prior variance for $\beta^{(2)}$ as both priors and the same adaptive shrinkage parameter as the ScIS prior.

Figure 2 shows the shrinkage profile for both the ShIS and ScIS priors. The ShIS prior, leads to more adaptive shrinkage than the ScIS prior, that is more shrinkage

for small coefficients and less shrinkage for larger coefficients. This effect is more pronounced when λ_2 is small (which indicates greater adaptive shrinkage). The shape of the shrinkage profiles for $\beta^{(2)}$ with the ScIS prior more closely resembles the shrinkage profiles for $\beta^{(2)}$ for the one-level prior. This suggests the form of $G_1, \ldots, G_L$ as well as the scale parameters $a_1, \ldots, a_L$ play an important role in determining the shrinkage profile for the marginal posterior mean of the regression coefficients. In particular, the adaptive shrinkage parameter for the conditional distribution of $\beta^{(2)}$ gives a good guide to the type of adaptive shrinkage for the marginal posterior mean of $\beta^{(2)}$. We will only consider ScIS-type prior in the rest of this paper.

The following results express the adaptive shrinkage parameters for the marginal posterior mean in hierarchical priors where f_l is a product or sum in terms of the adaptive shrinkage parameter for the conditional distributions. We also consider the usefulness of this concept for characterising the shrinkage profile. The scale parameter a_l is assumed to be $a_l = s_l\, d$ where s_l is the adaptive shrinkage parameter of G_l and d is a global scale parameter. It follows that $\mathrm{E}[\Psi_j^{(l)}] = s_l\, d$ which mimics the normal-gamma prior distribution where the adaptive shrinkage parameter and d are the shape and scale parameters of the gamma distributions respectively.

We first consider the case where the underlying random variables are gamma distributed, in which case the adaptive shrinkage parameter is given by the shape parameter of the gamma distribution.

Theorem 1 (Gamma case)**.** *Suppose that* $\eta_i \sim Ga(\lambda_i, 1)$ *for* $i = 1, 2, \ldots, K$ *then*

1. *the adaptive shrinkage parameter of* Ψ *is* $\min\{\lambda_i\}$ *if* $\Psi = \prod_{i=1}^K \eta_i$.
2. *the adaptive shrinkage parameter of* Ψ *is* $\sum_{i=1}^K \lambda_i$ *if* $\Psi = \sum_{i=1}^K \eta_i$.

An interesting special case occur when there are two levels and $f_2(\Psi_j^{(1)}) = \Psi_j^{(1)}$ so that $\Psi_j^{(2)} = \eta_j^{(1)}\eta_j^{(2)}$ for which the density has the analytic expression,

$$g(\Psi) = \frac{2}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\Psi^{(\lambda_1+\lambda_2)/2-1}K_{|\lambda_1-\lambda_2|}(2\sqrt{\Psi})$$

where $K_\nu(\cdot)$ is the modified Bessel function of the third kind (Abramowitz and Stegun, 1964, p. 374). The distribution is referred to as the K-distribution (Jakeman and Pusey, 1978) in several areas of physics. Using a small value approximation (Abramowitz and Stegun, 1964, Eq. 9.6.9), this density at a value of Ψ near zero is approximately proportional to

$$\frac{\Gamma(|\lambda_1 - \lambda_2|)}{\Gamma(\lambda_1)\Gamma(\lambda_2)}\Psi^{\min\{\lambda_1,\lambda_2\}-1},$$

and the adaptive shrinkage parameter is $\min\{\lambda_1, \lambda_2\}$ which agrees with Theorem 1.

Theorem 1 can be extended to the case where the underlying random variables are gamma-gamma distributed as defined in (3):

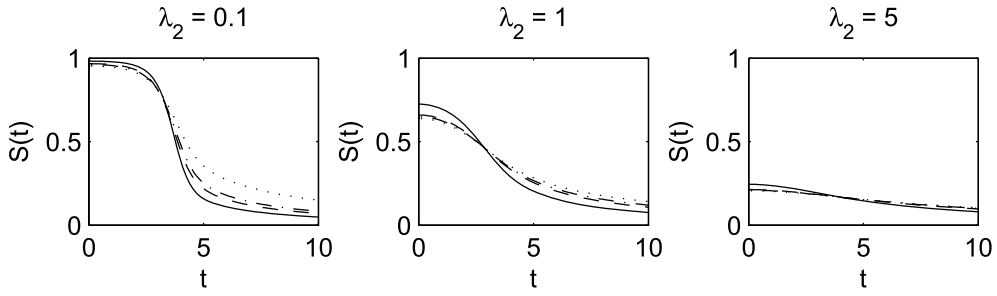

Figure 3: Shrinkage profiles for various choices of products of two normal-gamma prior distributions with: $\lambda_1 = \lambda_2$ (solid line), $\lambda_1 = 5\lambda_2$ (dashed line), $\lambda_1 = 10\lambda_2$ (dot-dashed line) with $d = 1/\mathrm{SE}^2$ compared to a normal-gamma(λ_1, d) with shape λ_1 (dotted line) with $d = 1/\mathrm{SE}^2$.

Theorem 2 (Gamma-gamma case)**.** *Suppose that* $\eta_i \sim GG(\lambda_i, c_i, 1)$ *for* $i = 1, 2, \ldots, K$ *then*

1. *the adaptive shrinkage parameter of* Ψ *is* $\min\{\lambda_i\}$ *if* $\Psi = \prod_{i=1}^{K} \eta_i$.
2. *the adaptive shrinkage parameter of* Ψ *is* $\sum_{i=1}^{K} \lambda_i$ *if* $\Psi = \sum_{i=1}^{K} \eta_i$.

Therefore, the shape close to zero of the products of either a normal-gamma or normal-gamma-gamma distribution is controlled by the shape parameters $\lambda_1, \ldots, \lambda_K$ rather than the other characteristics of the priors. Proofs of these theorems are in supplementary Appendix B.

Theorems 1 and 2 give expressions for adaptive shrinkage parameter, which is the shape close to zero of the density, of a product or sum of gamma or gamma-gamma distributed random variables. We now assess the ability of the adaptive shrinkage parameter to characterise the shrinkage profile for a hierarchical prior. Figure 3 shows the shrinkage profiles of $\beta_j^{(2)}$ when $L = 2$ and G_1 and G_2 are both gamma distributions. The marginal prior for $\beta_j^{(2)}$ is

$$\beta_j^{(2)} \sim \mathrm{N}\left(0, \lambda_2\, d\, \Psi\right), \qquad \Psi \sim \mathrm{Ga}(\lambda_2, \lambda_2/\eta_j^{(1)}), \qquad \eta_j^{(1)} \sim \mathrm{Ga}(\lambda_1, \lambda_1)$$

where $d = 1/\mathrm{SE}^2$ and SE is standard error of the least squares estimate of $\beta_j^{(2)}$. For comparison, we consider the one-level prior

$$\beta_j^{(2)} \sim \mathrm{N}\left(0, \lambda_2\, d\, \Psi\right), \qquad \Psi \sim \mathrm{Ga}(\lambda_2, \lambda_2).$$

In both cases, the adaptive shrinkage parameter is λ_2. Typically we want high shrinkage for small coefficients (t small) and little shrinkage of large coefficients (t large). The shape of the shrinkage curves are very similar for different choices of λ_1 with shrinkage decreasing slightly as λ_2 becomes larger. The effect is more pronounced if λ_2 is smaller.

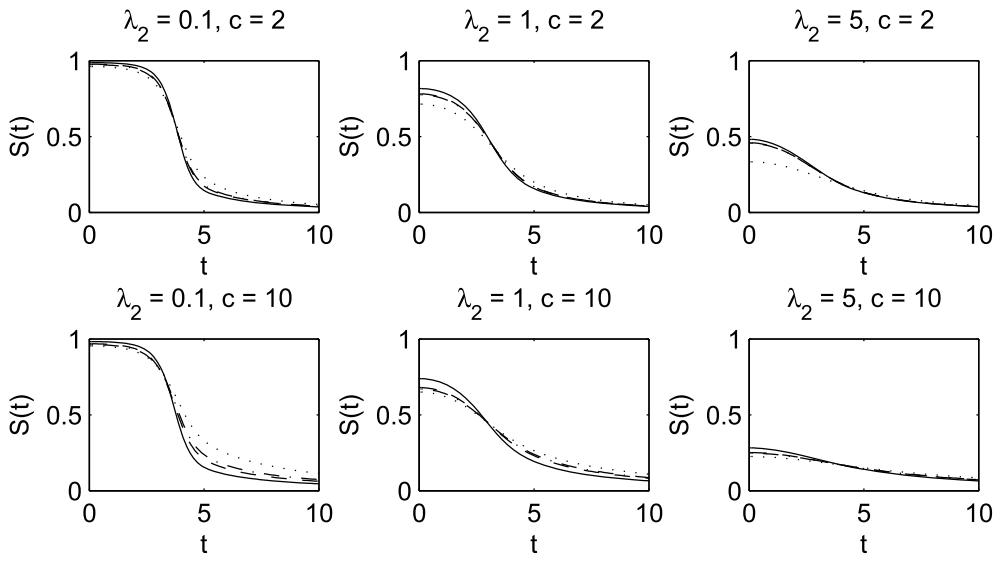

Figure 4: Shrinkage profiles for various choices of products of two normal-gamma-gamma prior distributions with: $\lambda_1 = \lambda_2$ (solid line), $\lambda_1 = 5\lambda_2$ (dashed line), $\lambda_1 = 10\lambda_2$ (dot-dashed line) with $d = 1/\mathrm{SE}^2$ compared to a normal-gamma-gamma with shape λ_1 (dotted line) and scale $d = 1/\mathrm{SE}^2$.

This suggests that the adaptive shrinkage parameter (although fairly crude) does give comparable forms of shrinkage for different values of t. In the NGG case, we consider marginal prior for $\beta_j^{(2)}$ with the form

$$\beta_j^{(2)} \sim \mathrm{N}\left(0, \lambda_2\, d\, \Psi\right), \qquad \Psi \sim \mathrm{GG}(\lambda_2, c, (c-1)\eta_j^{(1)}/\lambda_2), \qquad \eta_j^{(1)} \sim \mathrm{GG}(\lambda_1, c, (c-1)/\lambda_1)$$

and the one-level prior

$$\beta_j^{(2)} \sim \mathrm{N}\left(0, \lambda_2\, d\, \Psi\right), \qquad \Psi \sim \mathrm{GG}(\lambda_2, c, (c-1)/\lambda_2).$$

Figure 4 shows the shrinkage profiles with different values of c show results that are very similar to the normal-gamma case.

Example: Linear model with interaction terms

Returning to the linear model with interaction terms example in Section 3.1, we use the flexible gamma-gamma mixing distribution for the hierarchical prior. In the case of strong heredity, we assume that $a_1 = \lambda_1\, d$, $a_2 = \lambda_2\, d$, $\eta_j^{(1)} \sim \mathrm{GG}(\lambda_1, c, \frac{c-1}{\lambda_1})$ and $\eta_{jk}^{(2)} \sim \frac{c}{\lambda_2}\,\mathrm{GG}(\lambda_2, c, \frac{c-1}{\lambda_2})$ for $c > 1$. The adaptive shrinkage parameters are λ_1 for the main effects and $\min\{\lambda_1, \lambda_2\}$ for the interactions. More aggressive shrinkage of the interactions than the main effects corresponds to $\lambda_2 < \lambda_1$. In the case of weak heredity,

we assume that $a_1 = \lambda_1\, d$, $a_2 = \lambda_2\, d$, $\eta_j^{(1)} \sim \mathrm{GG}(\lambda_1, c, \frac{c-1}{\lambda_1})$ and $\eta_{jk}^{(2)} \sim \frac{c}{\lambda_2}\,\mathrm{GG}(\lambda_2, c, \frac{c-1}{\lambda_2})$ for $c > 1$. In this case, the adaptive shrinkage parameters are λ_1 for the main effects and $2\min\{\lambda_1, \lambda_2\}$ for the interactions. More aggressive shrinkage of the interactions than the main effects corresponds to $\lambda_2 < \lambda_1/2$.

4 Computational strategy

Posterior inference with these priors can be made using Markov chain Monte Carlo methods. In this section, we will describe the general strategy for inference rather than describe algorithms for specific models. We will assume the general model

$$y_i = \alpha + \sum_{l=1}^{L} X_i^{(l)} \beta^{(l)} + \epsilon_i, \qquad i = 1, \ldots, n$$

where $X_i^{(l)}$ is a $(n \times p_l)$-dimensional matrix whose columns are given by the variables in the l-th level, $\epsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$,

$$\beta_j^{(l)} \overset{i.i.d.}{\sim} \mathrm{N}\left(0, \Psi_j^{(l)}\right), \qquad j = 1, \ldots, p_l, \quad l = 1, \ldots, L$$

and

$$\Psi_j^{(l)} = s_j^{(l)}\, d\, \frac{f_{jl}\left(\Psi^{(1)}, \ldots, \Psi^{(l-1)}\right)}{\mathrm{E}[f_{jl}\left(\Psi^{(1)}, \ldots, \Psi^{(l-1)}\right)]}\, \eta_j^{(l)}, \qquad j = 1, \ldots, p_l, \quad l = 1, \ldots, L. \tag{7}$$

Typically, the distribution of $\eta_j^{(l)}$ has parameters which are denoted $\phi^{(l)}$. The Gibbs sampler will be used to sample from the posterior distribution of the parameters $(\alpha, \beta, \sigma, \Psi, d, \phi)$ where $\beta = \{\beta^{(l)} | l = 1, \ldots, L\}$, $\Psi = \{\Psi^{(l)} | l = 1, \ldots, L\}$ and $\phi = \{\phi^{(l)} | l = 1, \ldots, L\}$. The full conditional distributions of (α, β) and σ^2 follow from standard results for Bayesian linear regression models. The parameters Ψ, d and ϕ are updated one-element-at-a-time by adaptive Metropolis-Hastings random walk steps using a variation on the algorithm proposed by Atchadé and Rosenthal (2005). The output of adaptive Metropolis-Hastings algorithms are not Markovian (since the proposal distribution is allowed to depend on the previous values of the Markov chain) and so standard Markov chain theory cannot be used to show that the resulting chain is ergodic. Relatively simple conditions are given for the ergodicity of adaptive Metropolis-Hastings algorithms by Roberts and Rosenthal (2007). Our algorithms meet these conditions with the additional restriction that Ψ, d and ϕ are bounded above (at a very large value). Suppose that we wish to update $\phi^{(l)}$ at iteration i (the same idea will also be used to update the elements of Ψ and d). A new value $\phi^{(l)\,\prime}$ is proposed according to

$$\log \phi^{(l)\,\prime} = \log \phi^{(l)} + \epsilon^{(l)}$$

where $\epsilon^{(l)} \sim \mathrm{N}(0, \sigma^{2\,(i)}_{\phi^{(l)}})$. The notation $\sigma^{2\,(i)}_{\phi^{(l)}}$ makes the dependence on the previous values of the chain explicit and the induced transition density of the proposal is denoted

$q_{\sigma^{2\,(i)}_{\phi^{(l)}}}(\phi^{(l)}, \phi^{(l)\,\prime})$. The value $\phi^{(l)\,\prime}$ is accepted or rejected using the standard Metropolis-Hastings acceptance probability

$$\alpha\left(\phi^{(l)}, \phi^{(l)\,\prime}\right) = \frac{\prod_{j=1}^{p_l} p\left(\Psi_j^{(l)} \middle| \phi^{(l)\,\prime}\right) p\left(\phi^{(l)\,\prime}\right) q_{\sigma^{2\,(i)}_{\phi^{(l)}}}\left(\phi^{(l)\,\prime}, \phi^{(l)}\right)}{\prod_{j=1}^{p_l} p\left(\Psi_j^{(l)} \middle| \phi^{(l)}\right) p\left(\phi^{(l)}\right) q_{\sigma^{2\,(i)}_{\phi^{(l)}}}\left(\phi^{(l)}, \phi^{(l)\,\prime}\right)}.$$

The variance of the increment is updated by

$$\log \sigma^{2\,(i+1)}_{\phi^{(l)}} = \log \sigma^{2\,(i)}_{\phi^{(l)}} + i^{-a}\left(\alpha\left(\phi^{(l)}, \phi^{(l)\,\prime}\right) - \tau\right)$$

where $1/2 < a \leq 1$. This algorithm leads to an average acceptance rate which converges to τ. We choose $a = 0.55$ and $\tau = 0.3$ (following the suggestion of Roberts and Rosenthal (2009)) in our examples.

The posterior distribution can be highly multi-modal and so it is necessary to use parallel tempering to improve the mixing. An effective, adaptive implementation is described by Miasojedow et al. (2013).

5 Examples

5.1 Example 1: Blood glucose data

A blood glucose data set has been studied by Hamada and Wu (1992) amongst others. Yuan et al. (2007) analysed these data using their extension of the LARS algorithm which includes both strong and weak heredity. The data has one two-level factor and seven three-level factors. The experimental design and data are given in Yuan et al. (2007). We followed their analysis by fitting a linear model with interactions and by including the three levels as linear and quadratic effects using orthogonal polynomials. The model in Section 3.1 was extended to allow for quadratic effects and assumed that

$$y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \sum_{j=1}^{p} X_{ij}^2\gamma_j + \sum_{j=1}^{p}\sum_{k=1}^{j-1} X_{ij}X_{ik}\delta_{jk} + \epsilon_i, \qquad i = 1, \ldots, n.$$

The prior proposed in Section 3.2 was extended with $\gamma_j \overset{ind.}{\sim} \mathrm{N}(0, \lambda_1\, d\, \eta_j^{(1)})$. The parameter c was chosen to be 2 giving a heavy tail to the NGG distributions (but also a finite variance). The priors for the hyperparameters of the model were as follows. The adaptive shrinkage parameter for the main effects was given the prior $\lambda_1 \sim \mathrm{Ex}(1)$ which centred the prior over a heavy-tailed version of the Bayesian lasso. We defined $\lambda_2 = r\lambda_1$ where $0 < r < 1$ which implies that the interactions will be shrunk more aggressively than the main effects. We assumed that $r \sim \mathrm{Be}(2, 6)$ which implied that $\mathrm{E}[r] = 1/3$ suggesting that the interactions will be substantially more aggressively than the main effects. The scale parameter, d, was given the prior $p(d) \propto (1+d)^{-2}$ which implied that $\mathrm{E}[d] = 1$ with a heavy tail.

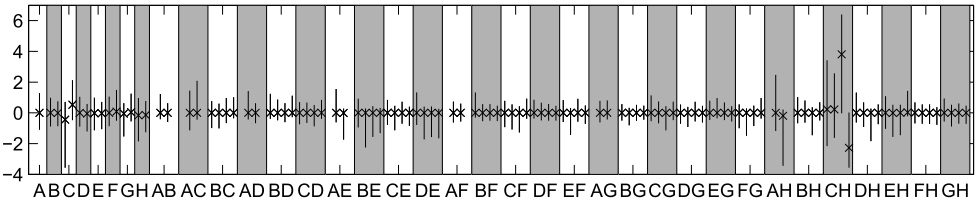

Figure 5: Blood glucose data – the posterior distribution of the regression coefficients with strong heredity shown as the posterior median (cross) and 95% credible interval (solid line). The main effects are ordered linear and quadratic, the interactions are ordered linear effect of first factor and linear effect of second factor, quadratic effect of first factor and linear effect of second factor, linear effect of first factor and quadratic effect of second factor, quadratic effect of first factor and quadratic effect of second factor.

The marginal posterior distributions of the regression coefficients using the strong heredity prior are presented in Figure 5. The most important terms were the interaction between C and H which had posterior medians which are furthest from zero and some 95% credibility intervals which did not include zero. In particular, the interaction between the linear and quadratic effects of C with the quadratic effect of H were the most important terms. The interactions of AH also showed some signs of being important since, although the posterior median was zero for both regression coefficients, the 95% posterior credibility intervals placed substantial mass on positive and negative values for the linear and quadratic effects respectively. The linear and quadratic effects of C also seemed important with posterior medians away from zero and support for a wide-range of values. All other effects had posterior medians which were very close to zero with a 95% credibility interval concentrated around 0.

The marginal posterior distributions of the Ψ's are shown in Figure 6. The variable C had the largest posterior median main effect followed by A and H. In terms of the interactions, it was clear that AH and CH had the largest upper point of the 95% credible interval which illustrated the importance of these interactions in the model. All these results were consistent with inference about the regression coefficients but gave a clearer picture of the importance of different variables.

The prior with weak heredity was also fitted and the results showed a very similar picture to those using the strong heredity prior. The Ψ's for the main effect of C and H were estimated to be slightly smaller than under strong heredity and the other main effects were estimated to be slightly larger. This reflected the importance of the interaction of CH in the model. Under strong heredity, there was stronger evidence of the importance of the main effects of C and H. The Ψ's for the importance of the interactions between AH and CH were estimated to be slightly smaller.

The inference about the adaptive shrinkage parameters λ_1 and λ_2 and the scale parameter d are shown in Table 1. The parameter λ_1 had a posterior median 0.48 which indicated that some effects were close to zero. The parameter λ_2 had a much smaller

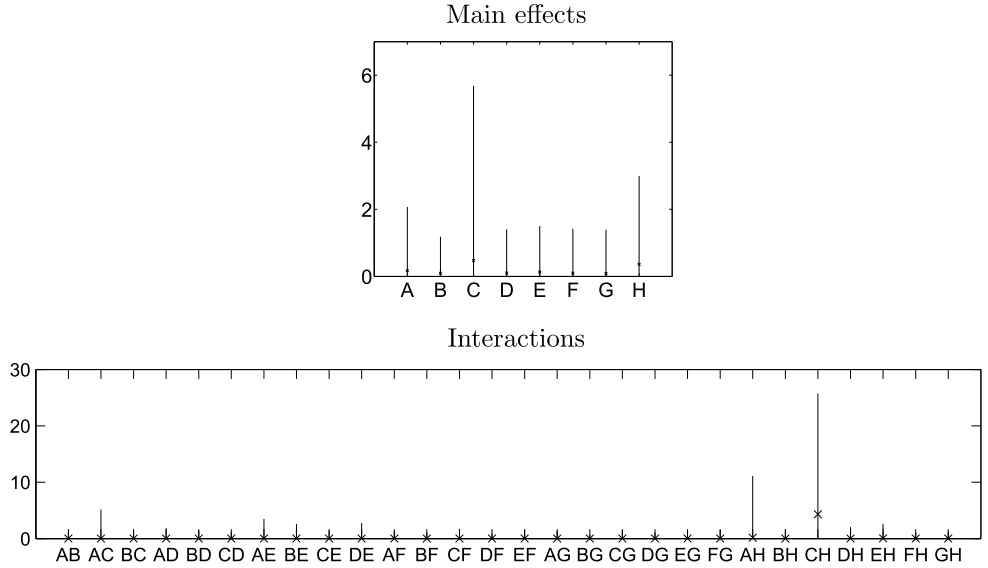

Figure 6: Blood glucose data – the posterior distribution of Ψ for each main effect and each interaction shown as the posterior median (cross) and 95% credible interval (solid line).

λ_1	0.48 *(0.15, 2.71)*
λ_2	0.054 *(0.018, 0.89)*
d	2.69 *(0.26, 27.89)*

Table 1: Blood glucose data – the posterior distribution of the hyperparameters summarised as posterior median and 95% credible interval.

posterior median which indicated that the data supported more aggressive shrinking of the interactions. These results were consistent with the inferences about the regression coefficients.

5.2 Out-of-sample predictive performance

The performance of the hierarchical prior introduced in this paper was compared to the hierarchical lasso (Bien et al., 2013) and three spike-and-slab priors for variable selection for interactions (Chipman, 1996) using five-fold cross-validation. Three data sets were used: ozone (Breiman and Friedman, 1985), Boston housing and blood glucose (Hamada and Wu, 1992). The results are summarized by the root mean squared error (RMSE) where the posterior predictive mean was used as the estimated prediction. The results suggest that the predictive performance of the hierarchical prior is competitive to both the hierarchical lasso and the spike-and-slab methods. The hierarchical prior gives the smallest RMSE for two data sets and is a close second on the ozone data.

	Blood glucose	Ozone	Boston housing
Bayesian strong heredity	13.0	4.0390	3.83
Bayesian weak heredity	11.4	4.0469	3.75
Bayesian relaxed heredity	12.2	4.0482	3.70
Hierarchical shrinkage prior	**10.5**	4.0272	**3.40**
Hierarchical lasso	14.4	**4.0181**	3.70

Table 2: Out-of-sample root mean squared errors for hierarchical shrinkage prior, hierarchical lasso and spike-and-slab priors for three data sets.

The effective sample sizes (ESS) were estimated for all methods using the initial monotone sequence estimator defined by Geyer (1992). With the hierarchical prior introduced in this paper, the estimates were 3317 for blood glucose, 249 for ozone and 57.2 for Boston housing. In contrast, the mixing was much worse for the Bayesian relaxed heredity method with estimates of 2.5 for blood glucose, 185 for ozone and 3.8 for Boston housing. The results were similar for the other two spike-and-slab priors. In all data sets, the computational times of the hierarchical prior is roughly ten times the computational time for the spike-and-slab prior. Therefore, the ESS per unit time is higher with the hierarchical prior than the spike-and-slab for the blood glucose and Boston housing data.

5.3 Example 2: Prostate cancer data

Data from a prostate cancer trial (Stamey et al., 1989) have become a standard example in the regularization literature. The response is the logarithm of prostate-specific antigen (*lpsa*). There are eight predictors: log(cancer volume) (*lv*), log(prostate weight) (*lw*), age (in years), the logarithm of the amount of benign prostatic hyperplasia (*lbph*, log(capsular penetration) (*lcp*), Gleason score (*gl*), percentage Gleason score 4 or 5 (*pg*)), and seminal vesicle invasion (*svi*). We considered all variables to be continuous apart from *svi* which is binary (it should be noted that Gleason score is ordinal and has 4 observed levels (scores of 6, 7, 8 and 9) in the data). Following Lai et al. (2012) the continuous effects are flexibly modelled using the GAM model in Section 3.1 with a piecewise linear spline basis function,

$$f_j(X_{ij}) = \theta_j X_{ij} + \sum_{k=1}^{K} [\gamma_{jk}(X_{ij} - \tau_k)_+]$$

where $(x)_+ = \max\{0, x\}$ and $\tau_k = \frac{k-1}{K-1}$ for $k = 1, \ldots, 60$. All continuous variables were normalized to have a minimum of 0 and a maximum of 1.

In Section 3.1, we discussed how inference in the GAM model could be seen as a two-level variable selection problem (at the basis level and at the variable level). We define a hierarchical shrinkage prior with two levels using a flexible gamma-gamma prior for the variance of the normal prior,

$$\theta_j \sim \mathrm{N}\left(0, \lambda_1 \, d \, \eta_j^{(1)}\right), \qquad \eta_j^{(1)} \sim \mathrm{GG}\left(\lambda_1, c, \frac{c-1}{\lambda_1}\right),$$

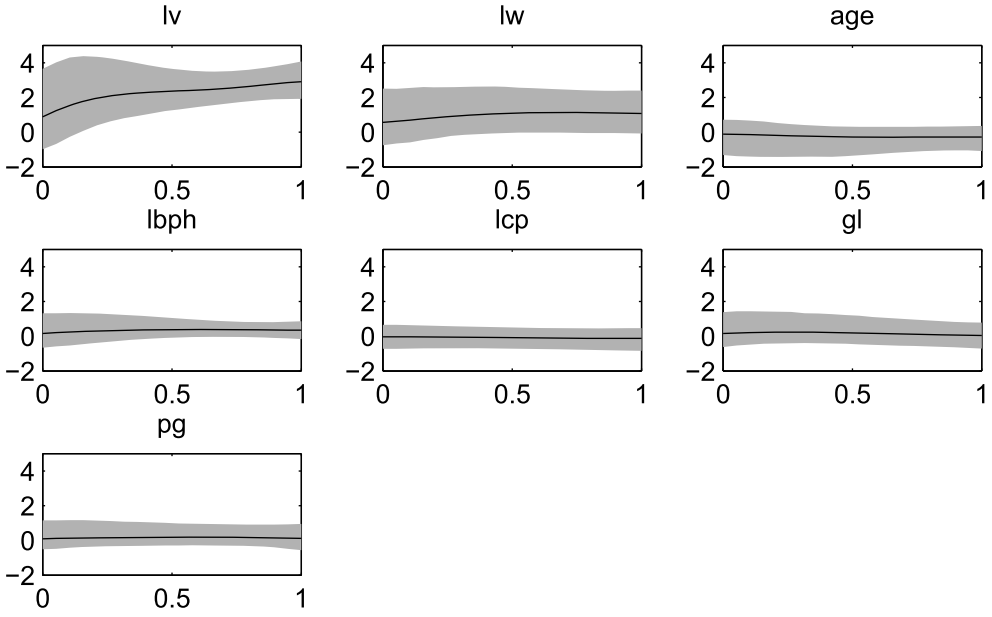

Figure 7: Prostate cancer data – the posterior distribution of the linear effects $\beta_j(x)$ for each variable summarized as the posterior median (solid line) and pointwise 95% credible interval (grey shading).

$$\gamma_{jk} \sim \mathrm{N}\left(0, \lambda_{2,j}\, d\, \eta_{jk}^{(2)}\, \eta_j^{(1)}\right), \text{ and } \eta_{jk}^{(2)} \sim \mathrm{GG}\left(\lambda_{2,j}, c, \frac{c-1}{\lambda_{2,j}}\right).$$

A small value of the parameter $\eta_j^{(1)}$ implies that the j-th variable is unimportant and will effect the shrinkage of both the linear effect θ_j and basis function coefficients $\gamma_{j1}, \ldots, \gamma_{jK}$ leading to shrinkage at the variable level. The variable selection problem at the basis level is achieved through the different values of $\eta_{jk}^{(2)}$ which allow some basis function coefficients to be set very close to zero. The prior allows different levels of adaptive shrinkage for the basis function coefficients for each variable (*i.e.* adaptive shrinkage parameter $\lambda_{2,j}$ for the j-th variable). The adaptive shrinkage parameters of the basis functions for the j-th variable are $\min\{\lambda_1, \lambda_{2,j}\}$. The adaptive shrinkage parameter of the variables is λ_1. The priors for the hyperparameters were: $\lambda_1 \sim \mathrm{Ga}(1,1)$, $\lambda_{2,j} \overset{i.i.d.}{\sim} \mathrm{Ga}(1,10)$, and common scale, d as heavy tailed with $p(d) \propto (1+d)^{-2}$. The parameter λ_1 controls adaptive shrinkage at the variable level and the choice centres the prior for the regression coefficients over the Bayesian lasso prior. The smaller prior mean for $\lambda_{2,j}$, $\mathrm{E}[\lambda_{2,j}] = 0.1$, implies greater levels of adaptive shrinkage at the basis level than the variable level and that only a few knots will be important for each variable.

The results of fitting the flexible regression model are shown in Figure 7. The inference about the regression effects are shown as $\beta_j(x) = \frac{f_j(x)}{x}$ and can be interpreted as the variable-dependent linear regression effect for the j-th variable. The effect of *lv* was

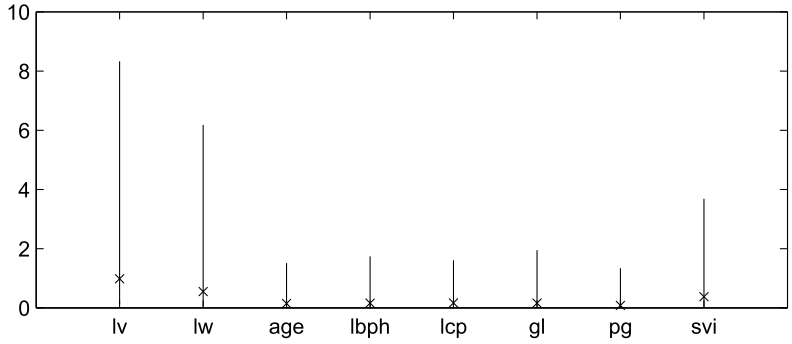

Figure 8: Prostate cancer data – the posterior distribution of Ψ for each variable summarized as the posterior median (cross) and 95% credible interval (solid line).

λ_1	0.96 *(0.31, 3.44)*	d	0.64 *(0.09, 6.10)*

Table 3: Prostate cancer data – the posterior distribution of the hyperparameters summarised as posterior median and 95% credible interval.

clearly important with an effect with the posterior median increasing from 0.88 to 2.91 over the range of the data. The effect of *lw* also seemed important and relatively constant over the range of the data. The other variables were clearly less important with a posterior median which is constant and close to zero and a narrower 95% credible intervals than the other variables. The effect of *svi* had a posterior median of 0.58 with a 95% credible interval of $(0.08, 1.06)$ which indicated the importance of this variable for the regression model.

The posterior distribution of the $\Psi_i^{(1)}$ is a measure of the overall strength of effect for the i-th variable. The distribution for each variable is shown in Figure 8. The results were consistent with the estimates of the regression effects. The *lv* variable gave the largest posterior median and had support at larger values of Ψ than other variables. The variables *lw* and *svi* also had important effects and had the next two largest values of the posterior median and were clearly useful as a scalar summary of the regression effects.

A summary of the posterior distribution of d and the adaptive shrinkage parameter for variables, λ_1, are shown in Table 3. The posterior median of λ_1 is close to 1 indicating that only some of the variables are important but that there is not a need for a lot of adaptive shrinkage.

5.4 Example 3: Computer data

Data on the characteristics and performance of 209 CPUs were considered by Ein-Dor and Feldmesser (1987) and subsequently analysed by Gustafson (2000) using Bayesian non-linear regression techniques. The response is performance of the CPU. In common

with Gustafson (2000), we consider 5 predictors: A, the machine cycle time (in nanoseconds); B, the average main memory size (in kilobytes); C, the cache memory size (in kilobytes); D, the minimum number of input channels; and E, the maximum number of input channels. In a similar spirit to Gustafson (2000), we modelled the data using a GAM with interactions which introduces bivariate functions, $f_{jl}(\cdot,\cdot)$, which allows modelling of non-linear interaction effects. In this case, the GAM model is extended to

$$\begin{aligned} y_i &= \sum_{j=1}^{p} f_j(X_{ij}) + \sum_{j=1}^{p}\sum_{k=1}^{j-1} f_{jk}(X_{ij}, X_{ik}) + \epsilon_i \\ &= \sum_{j=1}^{p} \theta_j^{(M)} X_{ij} + \sum_{j=1}^{p}\sum_{k=1}^{K} \gamma_{jk}^{(M)} g(X_{ij}, \tau_{jk}) + \sum_{j=1}^{p}\sum_{k=1}^{j-1} \theta_{jk}^{(I)} X_{ij} X_{ik} \\ &\quad + \sum_{j=1}^{p}\sum_{k=1}^{j-1}\sum_{l=1}^{K}\sum_{m=1}^{K} \gamma_{jklm}^{(I)} g(X_{ij}, \tau_{jl}) g(X_{ik}, \tau_{km}) + \epsilon_i \end{aligned} \tag{8}$$

where, again, $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$. The γ parameters for the nonlinear functions (splines) involve K knots. The bracketed superfixes (M) and (I) refer to main effects and interaction levels respectively. We used the model with $g_j(x, \tau) = (x - \tau)_+$ and $K = 10$ knots. The 5 main effects and 10 interactions lead to 1055 regression parameters in the model.

Gustafson (2000) used a square root transformation of the predictors since these data are highly skewed. In principle the distribution of variables shouldn't matter in non-linear regression modelling. However, knots are evenly spaced and so it would be useful to have data relatively evenly spread across the range of the knots. We found that a log transformation of the response lead to better behaved residuals than the untransformed response and also transformed the variables by $f(x) = \log(1 + x)$. All transformed variables were subsequently transformed to have a minimum of 0 and a maximum of 1.

A hierarchical shrinkage prior can be constructed for this problem by combining the prior for a GAM with only main effects and the prior for the linear model with interactions. The regression coefficients are organized into four levels: a main effects level, an interactions level, a basis level for main effects, and a basis level for interaction. The main effects level has $p_1 = p$ terms of the form $\theta_j^{(M)}$ for $j = 1, \ldots, p$. The interaction level has $p_2 = p(p-1)/2$ terms of the form $\theta_{jk}^{(I)}$ for $j = 1, \ldots, p$ and $k = 1, \ldots, j-1$. The basis level for main effects contains $\gamma_{jk}^{(M)}$ for $j = 1, \ldots, p$, $k = 1, \ldots, K$ and has $p_3 = pK$ terms. The basis level for interactions contains $\gamma_{jklm}^{(I)}$ for $j = 1, \ldots, p$, $k = 1, \ldots, j-1$, $l = 1, \ldots, K$, $m = 1, \ldots, K$ and contains $p_4 = (p-1)/2K^2$. The proposed prior, with strong heredity, is

$$\theta_j^{(M)} \sim \mathrm{N}\left(0, \lambda_1 \, d \, \eta_j^{(1)}\right), \quad \eta_j^{(1)} \sim \mathrm{GG}\left(\lambda_1, c, \frac{c-1}{\lambda_1}\right),$$

$$\theta_{jk}^{(I)} \sim \mathrm{N}\left(0, \lambda_2 \, d \, \eta_{jk}^{(2)} \, \eta_j^{(1)} \, \eta_k^{(1)}\right), \quad \eta_{jk}^{(2)} \sim \mathrm{GG}\left(\lambda_2, c, \frac{c-1}{\lambda_2}\right),$$

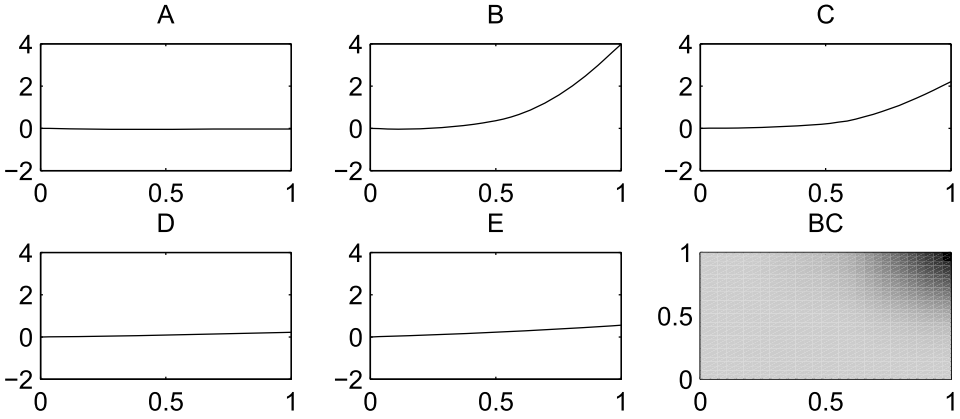

Figure 9: Computer data – the posterior mean of each main effect and each interaction. Darker colours represent lower values in the graphs for the interactions.

$$\gamma_{jk}^{(M)} \sim \mathrm{N}\left(0, \lambda_3\, d\, \eta_{jk}^{(3)}\, \eta_j^{(1)}\right), \quad \eta_{jk}^{(3)} \sim \mathrm{GG}\left(\lambda_{3,j}, c, \frac{c-1}{\lambda_3}\right),$$

$$\gamma_{jklm}^{(I)} \sim \mathrm{N}\left(0, \lambda_{4,j,k}\, d\, \eta_{jklm}^{(4)}\, \eta_{jk}^{(2)} \eta_j^{(1)} \eta_k^{(1)}\right), \qquad \eta_{jk}^{(4)} \sim \mathrm{GG}\left(\lambda_4, c, \frac{c-1}{\lambda_4}\right).$$

If $\eta_j^{(1)}$ is small then both the main effects $\theta_j^{(M)}$ and the basis function coefficients $\gamma_{jk}^{(M)}$ will tend to be small. Similarly, if $\eta_{jk}^{(2)}\eta_j^{(1)}\eta_k^{(1)}$ is small then both the interaction terms $\theta_{jk}^{(I)}$ and the basis function coefficients $\gamma_{jklm}^{(I)}$ will tend to be small. This allows variable selection at the main effect and interaction term levels. The prior also links the interaction and main effects terms (and, consequently, their associated basis function coefficients) since $\eta_{jk}^{(2)}\eta_j^{(1)}\eta_k^{(1)}$ is more likely to be small if both $\eta_j^{(1)}$ and $\eta_k^{(1)}$ are small. We assume that $\lambda_2 < \lambda_1$ and so the sparsities are λ_1 for the main effects level, λ_2 for the interactions level, $\min\{\lambda_1, \lambda_{3,j}\}$ for the basis level for the j-th main effects and $\min\{\lambda_2, \lambda_{4,j,k}\}$ for the basis level for interactions.

The priors for the hyperparameter of the model were as follows. The adaptive shrinkage parameters for the main effects and interaction terms were chosen as $\lambda_1 \sim \mathrm{Ex}(1)$ and $\lambda_2 = r\lambda_1$ where $r \sim \mathrm{Be}(2, 6)$ which implied that $\mathrm{E}[r] = 1/3$ suggesting that the interaction are *a priori* much sparser than the main effects. The conditional adaptive shrinkage parameters for the nonlinear terms were chosen to be $\lambda_{4,j,k} \overset{i.i.d.}{\sim} \mathrm{Ga}(1, 100)$ and $\lambda_{2,j} \overset{i.i.d.}{\sim} \mathrm{Ga}(1, 10)$ which implies that nonlinear terms were less likely to be included in the interaction function than the main effects function (which reflected the larger number of terms in the interaction function). The scale parameter, d, was given the prior $p(d) \propto (1+d)^{-2}$ which implied that $\mathrm{E}[d] = 1$ but with a heavy tail.

The estimated main effects and interactions are shown in Figure 9. The effect of A, D and E were small whereas B and C had an increasing, non-linear effect with a

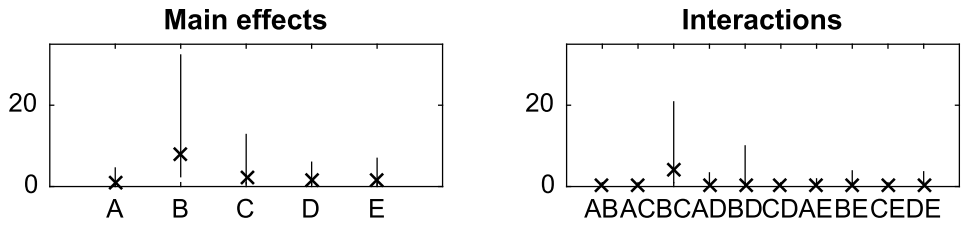

Figure 10: Computer data – the posterior distribution of Ψ for each main effect and each interaction summarized as the posterior median (cross) and 95% credible interval (solid line).

λ_1	1.96 *(0.41, 4.68)*	λ_2	0.40 *(0.13, 1.12)*	d	0.84 *(0.09, 10.06)*

Table 4: Computer data – the posterior distribution of the hyperparameters summarised as posterior median and 95% credible interval.

largest effect of roughly 4 for B and roughly 2 for C. The interaction effects mostly had a posterior median of zero. The main exception was the interaction between B and C which has a posterior median of -4 when both B and C are 1. This indicated that the effect of large values of B and C were over-estimated by the linear effects alone.

Figure 10 shows the posteriors for the Ψ's for the main effects and interactions. These results were consistent with the estimated effects. The variables B and C had the largest posterior medians and upper point of the 95% credible interval for the main effects. Similarly, the interaction between B and C had the largest posterior median and upper point of the 95% credible interval than the other interactions.

A summary of the posterior distribution of λ_1, λ_2 and d are shown in Table 4. The posterior median of λ_1 is close to 2 which indicates that most effects are relatively important (although this is estimated with a wide 95% credible interval due to the small number of regressors). The posterior median of λ_2 indicates that the interactions are much sparser than the main effects.

6 Discussion

This paper considers the specification of hierarchical priors in linear models where regression coefficients can be divided into levels and the relationship between the regression coefficients can be expressed hierarchically. We describe some methods for controlling the adaptive shrinkage of groups of regression coefficients at different levels of the prior. This is achieved through the shape rather than the scale of the gamma-gamma mixing density and an appropriate level of adaptive shrinkage can be chosen. These priors have applications in problems such as models with interactions and non-linear Bayesian regression models. Rather than impose blanket sharp heredity principles, our long-tailed normal-gamma-gamma priors are able to easily adapt if the data contradicts. We feel that these approaches will have the potential for many applications in future.

For example, Kalli and Griffin (2014) use a simple, two stage hierarchical prior in a regression model with time-varying regression coefficients. This allows the control of both adaptive shrinkage of the overall effect of a variable (where values of the regression coefficients at all times are shrunk to zero) and adaptive shrinkage of the effect of each regression coefficient over time.

Supplementary Material

Supplementary Material of "Hierarchical Shrinkage Priors for Regression Models" (DOI: 10.1214/15-BA990SUPP; .pdf).

References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions*. Dover. 143

Armagan, A., Dunson, D., and Clyde, M. (2011). "Generalized Beta Mixtures of Gaussians." In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems 24*, 523–531. 136, 137

Armagan, A., Dunson, D. B., and Lee, J. (2013). "Generalized double Pareto shrinkage." *Statistica Sinica*, 23: 119–143. MR3076161. 136

Atchadé, Y. F. and Rosenthal, J. S. (2005). "On adaptive Markov chain Monte Carlo algorithms." *Bernoulli*, 11: 815–828. MR2172842. doi: http://dx.doi.org/10.3150/bj/1130077595. 146

Bhattacharya, A. and Dunson, D. B. (2011). "Sparse Bayesian infinite factor models." *Biometrika*, 98: 291–306. MR2806429. doi: http://dx.doi.org/10.1093/biomet/asr013. 136

Bien, J., Taylor, J., and Tibshirani, R. (2013). "A lasso for hierarchical interactions." *Annals of Statistics*, 41: 1111–1141. MR3113805. doi: http://dx.doi.org/10.1214/13-AOS1096. 137, 149

Breiman, L. and Friedman, J. (1985). "Estimating optimal transformations for multiple regression and correlation." *Journal of the American Statistical Association*, 80: 580–598. MR0803258. 149

Caron, F. and Doucet, A. (2008). "Sparse Bayesian nonparametric regression." In: McCallum, A. and Roweis, S. (eds.), *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, 88–95. Omnipress. 136, 137

Carvalho, C., Polson, N., and Scott, J. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97: 465–480. MR2650751. doi: http://dx.doi.org/10.1093/biomet/asq017. 136, 138

Chipman, H. (1996). "Bayesian variable selection approach with related predictors." *Canadian Journal of Statistics*, 24: 17–36. MR1394738. doi: http://dx.doi.org/10.2307/3315687. 136, 139, 149

Chipman, H., Hamada, M., and Wu, C. F. J. (1997). "A Bayesian variable selection approach for analyzing designed experiments with complex aliasing." *Technometrics*, 39: 372–381. 136, 137, 139

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley. MR1962778. 140

Ein-Dor, P. and Feldmesser, J. (1987). "Attributes of the performance of Central Processing Units: A relative performance prediction model." *Communications of the Association for Computer Machinery*, 30: 308–317. 152

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88: 881–889. 135

Geyer, C. J. (1992). "Practical Markov chain Monte Carlo." *Statistical Science*, 7: 473–511. 150

Griffin, J. E. and Brown, P. J. (2010). "Inference with Normal-Gamma prior distributions in regression problems." *Bayesian Analysis*, 5: 171–188. MR2596440. doi: http://dx.doi.org/10.1214/10-BA507. 136, 137, 138

Griffin, J. E. and Brown, P. J. (2011). "Bayesian hyper-lassos with non-convex penalisation." *Australian and New Zealand Journal of Statistics*, 53: 423–442. MR2910027. doi: http://dx.doi.org/10.1111/j.1467-842X.2011.00641.x. 136

Griffin, J. E. and Brown, P. J. (2012). "Structuring shrinkage: Some correlated priors for regression." *Biometrika*, 99: 481–487. MR2931267. doi: http://dx.doi.org/10.1093/biomet/asr082. 136

Griffin, J. E. and Brown, P. J. (2013). "Some priors for sparse regression modelling." *Bayesian Analysis*, 8: 691–702. MR3102230. doi: http://dx.doi.org/10.1214/13-BA827. 136

Griffin, J. and Brown, P. (2016). "Supplementary Material of Hierarchical Shrinkage Priors for Regression Models" *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/15-BA990SUPP. 137

Gustafson, P. (2000). "Bayesian Regression Modeling with Interactions and Smooth Effects." *Journal of the American Statistical Association*, 95: 795–806. 152, 153

Hamada, M. and Wu, C. F. J. (1992). "Analysis of designed experiments with complex aliasing." *Journal of Quality Technology*, 24: 130–137. 147, 149

Hans, C. (2009). "Bayesian lasso regression." *Biometrika*, 96: 835–845. MR2564494. doi: http://dx.doi.org/10.1093/biomet/asp047. 136

Hans, C. (2011). "Elastic net regression modeling with the orthant normal prior." *Journal of the American Statistical Association*, 106: 1383–1393. MR2896843. doi: http://dx.doi.org/10.1198/jasa.2011.tm09241. 136

Hastie, T. J. and Tibshirani, R. J. (1993). *Generalized Additive Models*. Chapman and Hall. MR1082147. 139

Jacob, L., Obozinski, G., and Vert, J.-P. (2009). "Group Lasso with Overlaps and Graph Lasso." In: Bottou, L. and Littman, M. (eds.), *Proceedings of the 26th International Conference on Machine Learning*, 433–440. Montreal: Omnipress. 136

Jakeman, E. and Pusey, P. N. (1978). "Significance of K-distributions in scattering experiments." *Physical Review Letters*, 40: 546–550. 143

Kalli, M. and Griffin, J. E. (2014). "Time-varying sparsity in dynamic regression models." *Journal of Econometrics*, 178: 779–793. MR3144682. doi: http://dx.doi.org/10.1016/j.jeconom.2013.10.012. 156

Kohn, R., Smith, M., and Chan, D. (2001). "Nonparametric regression using linear combinations of basis functions." *Statistics and Computing*, 11: 313–322. MR1863502. doi: http://dx.doi.org/10.1023/A:1011916902934. 140

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). "Penalized Regression, Standard Errors, and Bayesian Lassos." *Bayesian Analysis*, 5: 369–412. MR2719657. doi: http://dx.doi.org/10.1214/10-BA607. 136, 141

Lai, R. C. S., Huang, H.-C., and Lee, T. C. M. (2012). "Fixed and random effects selection in nonparametric additive mixed models." *Electronic Journal of Statistics*, 6: 810–842. MR2988430. doi: http://dx.doi.org/10.1214/12-EJS695. 150

Lee, A., Caron, F., Doucet, A., and Holmes, C. (2012). "Bayesian sparsity-path-analysis of genetic association using generalised t priors." *Statistical Applications in Genetics and Molecular Biology*, 11 (2): Art 5. MR2935747. doi: http://dx.doi.org/10.2202/1544-6115.1712. 136

Li, F. and Zhang, N. R. (2010). "Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics." *Journal of the American Statistical Association*, 105: 1202–1214. MR2752615. doi: http://dx.doi.org/10.1198/jasa.2010.tm08177. 136

Miasojedow, B., Moulines, E., and Vihola, M. (2013). "An adaptive parallel tempering algorithm." *Journal of Computational and Graphical Statistics*, 22: 649–664. MR3173735. doi: http://dx.doi.org/10.1080/10618600.2013.778779. 147

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression (with discussion)." *Journal of the American Statistical Association*, 83: 1023–1036. MR0997578. 135

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103: 672–680. MR2524001. doi: http://dx.doi.org/10.1198/016214508000000337. 136

Polson, N. G. and Scott, J. G. (2011). "Shrink globally, act locally: Sparse Bayesian regularization and prediction." In: Bernardo J. M., M. J., Bayarri, Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics 9*, 501–538. Oxford: Clarendon Press. MR3204017. doi: http://dx.doi.org/10.1093/acprof:oso/9780199694587.003.0017. 135

Polson, N. G. and Scott, J. G. (2012). "Local shrinkage rules, Lévy processes and Regularized Regression." *Journal of the Royal Statistical Society, Series B*, 74: 287–311. MR2899864. doi: http://dx.doi.org/10.1111/j.1467-9868.2011.01015.x. 138

Polson, N. G., Scott, J. G., and Windle, J. (2013). "The Bayesian Bridge." *Journal of the Royal Statistical Society, Series B*, 76: 713–33. MR3248673. doi: http://dx.doi.org/10.1111/rssb.12042. 136

Raiffa, H. and Schlaifer, R. (1961). *Applied Statistical Decision Theory*. M.I.T. Press. MR0226757. 138

Raman, S., Fuchs, T., Wild, P., Dahl, E., and Roth, V. (2009). "The Bayesian Group-Lasso for analyzing contingency tables." In: Bottou, L. and Littman, M. (eds.), *Proceedings of the 26th International Conference on Machine Learning*, 881–888. Montreal: Omnipress. 136, 141

Roberts, G. O. and Rosenthal, J. S. (2007). "Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms." *Journal of Applied Probability*, 44: 458–475. MR2340211. doi: http://dx.doi.org/10.1239/jap/1183667414. 146

Roberts, G. O. and Rosenthal, J. S. (2009). "Examples of Adaptive MCMC." *Journal of Computational and Graphical Statistics*, 18: 349–367. MR2749836. doi: http://dx.doi.org/10.1198/jcgs.2009.06134. 147

Rockova, V. and Lesaffre, E. (2014). "Incorporating grouping information in Bayesian variable selection with applications in genomics." *Bayesian Analysis*, 9: 221–258. MR3188306. doi: http://dx.doi.org/10.1214/13-BA846. 136

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, N. (1989). "Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: Radical prostatectomy treated patients." *Journal of Urology*, 16: 1076–1083. 150

Stingo, F. C., Chen, Y. A., Tadesse, M. G., and Vannucci, M. (2011). "Incorporating biological information into linear models: A Bayesian approach to the selection of pathways and genes." *Annals of Applied Statistics*, 5: 1978–2002. MR2884929. doi: http://dx.doi.org/10.1214/11-AOAS463. 136

Yi, N., Shriner, D., Banerjee, S., Mehta, T., Pomp, D., and Yandell, B. S. (2007). "An efficient Bayesian model selection approach for interpreting quantitative trait loci models with many effects." *Genetics*, 176: 1865–1877. 136, 137

Yuan, M., Joseph, V. R., and Lin, Y. (2007). "An efficient variable selection approach for analyzing designed experiments." *Technometrics*, 49: 430–439. MR2414515. doi: http://dx.doi.org/10.1198/004017007000000173. 139, 147

Yuan, M., Joseph, V. R., and Zou, H. (2009). "Structured variable selection and estimation." *The Annals of Applied Statistics*, 3: 1738–1757. MR2752156. doi: http://dx.doi.org/10.1214/09-AOAS254. 137

Yuan, M. and Lin, Y. (2006). "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society B*, 68: 49–67. MR2212574. doi: http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x. 136