Hierarchical sparsity priors for regression models

J.E. Griffin and P.J. Brown
School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury CT2 7NF, UK

May 11, 2014

Abstract

Sparse regression problems, where it is usually assumed that there are many variables and that the effects of a large subset of variables are negligible, have become increasingly important. This paper describes the construction of hierarchical prior distributions when the effects are considered related. These priors allow dependence between the regression coefficients and the shrinkage to zero of different regression coefficients to be related. The properties of these priors are discussed and applications to linear models with interactions and generalized additive models are used as illustrations.

Keywords: Bayesian regularization; structured priors; generalized additive models; interactions; normal-gamma priors

1 Introduction

Regression modelling is an important means of understanding the effect of predictor variables on a response. These effects can be hard to estimate if the predictor variables are highly correlated (the problem of collinearity) or there are large numbers of predictor variables. Standard estimators such as least squares tend to have large standard errors in these cases which make interpretation difficult and can lead to over-fitting. These problems can be addressed
using traditional variable selection procedures, such as stepwise regression and subset selection in the classical paradigm (Hastie et al., 2001) or “spike-and-slab” priors in the Bayesian framework (Mitchell and Beauchamp, 1988). More recently, alternative regularization methods have been proposed which use penalized maximum likelihood methods in a classical framework or absolutely continuous priors in a Bayesian framework. These methods assume that the effects are sparse which means that only a subset (which is often considered small) of the predictor variables has an effect on the response. Many methods have been proposed which eliminate the subset of variables which have little or no effect and so allow better estimation of the important effects. These methods can, in turn, lead to more interpretable models and better out-of-sample prediction.

Most work in the area of regularization has not explicitly included any known relationships between the predictor variables in the analysis. Many penalty functions in penalized maximum likelihood methods are expressed as a sum of terms for each predictor variable whereas priors in Bayesian analyses are constructed as products of terms for each regression coefficient (implying a priori independence between the regression coefficients). However, in many data sets, there are known relationships between the predictor variables which we wish to include in the analysis. For example, suppose that we use a linear model with main effects and two-way interaction terms. One commonly used heuristic in variable selection is that a two-way interaction term can only be included if both main effects terms are included. This can be interpreted as a belief that the absolute size of the two-way interaction coefficient is related to the two associated main effect coefficients (if either main effect has a small absolute coefficient then the interaction term must also have a small absolute coefficient). 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. Several approaches have been developed in the literature to allow these relationships to be included in the analysis. Perhaps, the most popular is the group lasso (Yuan and Lin, 2006) where predictor variables are divided into groups and a penalty function is developed for which the penalized maximum likelihood estimates of the regression coefficients in a group are either all zero or all non-zero. Therefore, variable selection occurs at the level of the groups rather than the individual predictor variables. A similar approach for non-overlapping groups was developed by Jacob et al. (2009). The group lasso approach was extended to
linear models with two-way interactions by Yuan et al. (2007) and to more complicated problems by Yuan et al. (2009).

The Bayesian approach to regularization defines a prior distribution for the regression coefficients which encourages a proportion of those coefficients to be shrunk to a value close to zero (or exactly zero). Many priors have been proposed including: 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), 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). All these priors are scale mixtures of normal distributions and these papers have shown how the shape of the mixing distribution is critical for successful modelling of sparse effects. Some work has used maximum a posteriori (MAP) estimates whilst others have looked at full posterior inference. The former have usually been proposed for problems with large numbers of regressors and used the EM algorithm for inference. Full posterior inference is usually implemented using Markov chain Monte Carlo methods and this approach will be followed in this paper. Prior distributions which include known relationships between variables have also been considered. A Bayesian version of the group Lasso was developed by Kyung et al. (2010) and Raman et al. (2009). A rather different approach is taken by Griffin and Brown (2012) who defined priors which allow correlation between the effects rather than the absolute effect sizes (as implied by the group Lasso). It has also been applied to unifying and robustifying ridge and g-priors for regression in Griffin and Brown (2013). The variable selection problem in the linear model with interactions has been approached by Chipman et al. (1997) using “spike-and-slab” prior distributions. More recently, structured priors have been proposed in biological application, e.g. Stingo et al. (2011) and Li and Zhang (2010).

In this paper, a method for building prior distributions for structured regression problems (where relationships between the predictor variables can be assumed) is developed. The prior involves organising the regression coefficients in a hierarchical structure where the regression coefficients at one level depend on a subset of the effect sizes at a higher level. This is a fairly general structure which can include both overlapping and non-overlapping groups in a simple way, whilst also expressing much more complicated structures. The
paper is organized as follows. Section 2 explains the use of normal-gamma and normal gamma-gamma (or generalized beta mixture) priors for sparse regression problems. Section 3 develops hierarchical structured regression models using a hierarchical prior with motivating examples in section 3.1 of the linear model with interaction terms and the generalized additive model (GAM). The general construction is given in section 3.2 and its use in specific statistical models in section 3.3. Section 4 briefly discusses computational strategies for models using these priors. Section 5 includes applications of the models introduced in section 3.1 to data sets. Section 6 gives a brief discussion. Shrinkage characterisation and proofs of theorems are given in the Appendices A and B respectively.

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 \(X\) is

\[
y = \alpha 1 + X\beta + \epsilon
\]

where \(\epsilon \sim N(0, \sigma^2 I_n)\), 1 is a \((n \times 1)\)-dimensional vector of 1’s, \(\alpha\) is an intercept and \(\beta\) is a \((p \times 1)\)-dimensional vector of regression coefficients. The prior for \(\alpha\) and \(\sigma^{-2}\) is chosen to be the scale-invariant choice \(p(\alpha, \sigma^2) \propto \sigma^{-2}\) and we will concentrate on the choice of prior for the regression coefficients \(\beta\), which will be assumed independent of \(\alpha\) and \(\sigma^2\), in the rest of the paper. It is assumed that the variables have been measured on comparable scales (or scaled to have comparable scales).

A commonly chosen prior for regression coefficients is the normal distribution, which is conjugate for a linear regression model with normal errors. It is well-known that this prior is not suitable for problems where the regression coefficients are thought to be sparse (that is a proportion of the regression coefficients are zero or very close to zero). Scale mixtures of normals are typically used as priors for these problems (Polson and Scott, 2011) in which the prior density can be expressed as

\[
\pi(\beta_i) = \int N(0, \Psi_i) \, dG(\Psi_i)
\]

where \(G\) is a distribution function, whose density is \(g\) (if it exists). The “spike-and-slab” prior (Mitchell and Beauchamp, 1988) and the stochastic search variable selection prior (George and McCulloch, 1993) fit into this structure with
$G$ chosen to be a discrete mixing distribution with two possible values. These priors have been extensively applied in Bayesian analyses. Alternatively, $G$ can be chosen to be absolutely continuous. The conditional variance $\Psi_i$ has a simple interpretation as “importance” or relevance (Tipping, 2000; Bishop and Tipping, 2000) of a variable in the regression (larger scales are associated with larger absolute values of coefficients and so imply more importance). In this paper, we will look at generalizations of normal-gamma prior distribution (Caron and Doucet, 2008; Griffin and Brown, 2010) or a generalized beta mixture prior distribution (Armagan et al., 2011) for $\beta_i$.

The generalised beta mixture prior distribution, which we will refer to as the normal-gamma-gamma prior distribution, is expressed in hierarchical form:

$$
\beta_j \sim N(0, \Psi_j), \quad \Psi_j \sim \text{Ga}(\lambda, \gamma_j), \quad \gamma_j \sim \text{Ga}(c, d).
$$

The prior mean of $\beta_j$ is 0 and the prior variance is $V[\beta_j] = E(\Psi_j) = \frac{\lambda d}{c+1}$ if $c > 1$. The hyperparameters have simple interpretations: $d$ is a scale parameter, $\lambda$ controls the behaviour of the distribution close to zero and $c$ controls the tail behaviour of the distribution. The marginal density of $\beta_j$ is not available in closed form but the marginal distribution of $\Psi_j$ 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)^{-\left(\lambda+c\right)}.\]

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 authors showed that the monotone transformation $\Psi_j + d$ has a beta distribution with parameters $\lambda$ and $c$ implying that the median is $d$ if $\lambda = c$. This is a useful characterisation if the mean does not exist which happens if $c \leq 1$. In particular, this is true for the horseshoe prior 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 including the NEG prior which arises when $\lambda = 1$ and the normal-gamma prior distribution which arises if $c/d = \mu$ as $c \to \infty$.

Shrinkage results for regression models which express the posterior expectation and variance in terms of the least squares estimate of $\beta$ and the variance of its sampling distribution (for $n > p$) have been derived by several authors including Carvalho et al. (2010), Griffin and Brown (2010) and Polson and Scott (2012). We now show generally for scale mixtures of normals that the amount of shrinkage in a simple linear regression only depends on the $t$
statistic for the regression coefficients (rather than the least squares estimates and the standard error separately).

**Proposition 1** Suppose that we have the regression model in (1) with a single regressor which has been centred. The intercept is given the vague prior \( p(\alpha) \propto 1 \) and \( \beta \) has prior \( \pi_\beta(\beta) \). Let \( \tau = \beta/SE \) and \( t = \hat{\beta}/SE \) where \( \hat{\beta} \) is the least squares estimate of the regression coefficient and \( SE \) is its standard error then

\[
E[\beta|\hat{\beta}] = (1 - S(t))\hat{\beta}
\]

where \( S(t) = -\frac{1}{t} \left[ \frac{d}{ds} \log h(s) \right]_{s=t} \),

\[ h(s) = \int N(s|\tau,1)\pi_\tau(\tau)d\tau \] and \( \pi_\tau(\tau) = SE\pi_\beta(SE\tau) \).

The proof and some illustrative graphs of shrinkage are given in Appendix A.

In the case of the scale mixture of normals considered in this paper, we have \( \pi_\beta(\beta) = \int N(\beta|0,\Psi)g(\Psi)d\Psi \) and so \( \pi_\tau(\tau) = \int N(\tau|0,\Psi^*)g_{\Psi^*}(\Psi^*)d\Psi^* \) where \( g_{\Psi^*}(\Psi^*) = SE^2g(SE^2\Psi) \). Returning to our specific cases, the normal-gamma prior leads to \( \Psi^* \sim \text{Ga}(\lambda, \gamma_i/SE^2) \) and \( \Psi^* \sim \text{Ga}(\lambda, \gamma_i^*) \); with \( \gamma_i^* \sim \text{Ga}(c, dSE^2) \) for the normal-gamma-gamma prior. Therefore, the shrinkage induced by the posterior expectation (relative to the least squares estimate) can be expressed in terms of a scale defined relative to the standard error. This simplifies the presentation of the shrinkage function for different choices of prior as they can be presented relative to a standard scale. The effect of changing the standard error or the scale of the prior distribution is just to re-scale the \( x \)-axis of the graphs.

The sparsity of a set of regression coefficients can be considered to be the proportion which have values close to zero. The graphs in Figure 13 in Appendix A suggest that smaller values of \( \lambda \) will increasingly favour sparser sets of regression coefficients since small values of \( t \) are likely to be shrunk very close to zero. This is intuitively reasonable since this parameter controls the shape of the distribution of \( \Psi_i \) at small values for both priors, gamma and gamma-gamma. Consequently, we define the sparsity shape parameter for a prior distribution in terms of the prior density of \( \Psi_i \) as

\[
\sup \left\{ z \left| \frac{p(\Psi_i)}{\Psi_i^{-1}} \to \kappa \text{ as } \Psi_i \to 0 \text{ for finite } \kappa \right. \right\}
\]

where \( p(\Psi_i) \) is the prior density of \( \Psi_i \). This will be simply \( \lambda \) in the case of both the normal-gamma and normal-gamma-gamma prior distributions and
indicates the shape of the prior distribution of $\Psi_i$ close to zero. The use of the supremum or least upper bound leads to clearer results in some special cases discussed in section 3.2.

3 Hierarchical sparsity prior

3.1 Motivating Examples

Before developing our general hierarchical shrinkage results we first set the context in terms of linear models with interactions and GAMs. These illustrate the need for priors which can express relationships between regression coefficients with different levels of sparsity for some regression coefficients.

3.1.1 Linear models with interaction terms

Variable selection and regularization methods for linear models with interactions have received some attention in the literature. The model assumes that response $y_i$ observed with covariates $X_{i1}, \ldots, X_{ip}$ can be expressed as

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

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Many authors have argued for the principle of marginality or the idea of effect heredity (Chipman et al., 1997). These approaches assume that the inclusion of an interaction should be contingent on the inclusion of main effects. Chipman et al. (1997) introduced two forms of the heredity principle. The first is strong heredity which states that an interaction can only be included if both main effects are included. The second is weak heredity which states that an interaction can be included if at least one main effect is included. In models with higher order interactions, there is clearly potential for many different rules for including an interaction which would depend on the inclusion or exclusion of main effects and lower order interactions. Chipman et al. (1997) described a “spike-and-slab”-based approach to this problem which allows both strong and weak heredity. An extension of the LARS algorithm (Efron et al., 2004) to include these principles is described in Yuan et al. (2007), which is generalized to other regression problems in Yuan et al. (2009). 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 clear that, a priori, the scale of the interaction coefficient should depend
on the magnitude, but not the sign, of the main effect coefficients with the
coefficients of the interactions being sparser than the coefficients of the main
effects.

3.1.2 Generalized additive models

The 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 \( \epsilon_i \sim N(0, \sigma^2) \). 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

\[ f_j(X_{ij}) = \sum_{k=1}^{q} \theta_{jk} X_{ij}^{k} + \sum_{k=1}^{K} \gamma_{jk} g(X_{ij}, \tau_{jk}) \]

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

\[ y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \epsilon_i = \sum_{j=1}^{p} \sum_{k=1}^{q} \theta_{jk} X_{ij}^{k} + \sum_{j=1}^{p} \sum_{k=1}^{K} \gamma_{jk} g(X_{ij}, \tau_{jk}) + \epsilon_i. \]

The set of knot points is often chosen to be relatively large and many \( \gamma_{jk} \)'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 \( p \) dif-
f erent variable selection problems (one for each variable). We will refer to this
as selection at the basis level. There is also potentially the more standard vari-
able 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 \( \theta_{j1}, \ldots, \theta_{jq} \) 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 (\( \theta_{j1}, \ldots, \theta_{jq} \)) and (\( \gamma_{j1}, \ldots, \gamma_{jK} \)) seems unre-as
sional and dependence in size (rather than the sign) of these coefficients will
be reasonable in many problems. Typically, we would like different amounts
of sparsity at the basis level and the variable level which suggests a prior with
at least two sparsity 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 amount of sparsity implied by the prior for different regression coefficients. The Bayesian group lasso (Kyung et al., 2010; Raman et al., 2009) is one example of a prior which allows dependence between the size of regression coefficients but no correlation in the signs. It is assumed that the regression coefficients are divided into disjoint groups \( b_1, b_2, \ldots, b_G \) where \( b_i \) is the \((p_i \times 1)\)-dimensional vector of regression coefficients for the \( i \)-th group. The prior assumes that \( b_1, b_2, \ldots, b_G \) are independent and \( b_i \sim \mathcal{N}(0, \Psi_i D^{(i)}) \) where \( \Psi_i \) is given a gamma prior distribution and \( D^{(i)} \) is a \((p_i \times p_i)\)-dimensional matrix. This induces correlation in the conditional variances of the regression coefficients, \( \Psi_i D^{(i)}_{jj} \) for \( j = 1, \ldots, p_i \), but not necessarily in the regression coefficients (the correlation between \( b_{ij} \) and \( b_{ik} \) will be zero if \( D^{(i)}_{jk} = 0 \)).

The group lasso prior is a simple way of building dependence between regression coefficients if they can be divided into groups. We consider a more general structure for the prior of the regression coefficients, \( \beta = (\beta_1, \ldots, \beta_p) \), in (1). We assume that the elements of \( \beta \) are independent conditional on \( \Psi = (\Psi_1, \ldots, \Psi_p) \) and

\[
\beta_j \sim \mathcal{N}(0, \Psi_j), \quad j = 1, \ldots, p.
\]

The parameters \( \Psi_j \) is the conditional variance of \( \beta_j \) and smaller values of \( \Psi_j \) imply typically smaller values of \(|\beta_j|\). Building hierarchical priors for \( \Psi \) allows the construction of a prior with correlated \( \Psi \) but not \( \beta \). This form of dependence is important. The scale \( \Psi_j \) can be interpreted as the importance of the \( j \)-th variable in the regression and so correlation \( \Psi_j \) and \( \Psi_k \) implies a relationship between the importance of the \( j \)-th and \( k \)-th variables. Lack of correlation between the regression coefficient would imply, for example, no correlation in the sign of regression coefficients, which is a natural assumption in many regression problems. The construction could be extended to a model where the regression coefficients are correlated by assuming that \( \beta \) are dependent conditional on \( \Psi \).

We assume that the regression coefficients can be arranged in levels. Let there be \( L \) levels and \( \beta^{(l)} \) be the \((p_l \times 1)\)-dimensional vector of regression coefficients in the \( l \)-th level. The regression coefficients at a particular level will have the same sparsity \textit{a priori} and their scales will usually depend on scales of regression coefficients in lower levels. For example, in the linear model with
two-way interactions described in section 3.1.1, one level would include the
main effects and the other level would include the interactions. As we have
already discussed, it seems natural to allow the scale of a two-way interaction
to depend on the scale of the two associated main effects in the prior and and
that the interactions are a priori sparser than the main effect. Our general prior
assumes that
\[ \beta_j^{(l)} \overset{i.i.d.}{\sim} N \left( 0, \Psi_j^{(l)} \right), \quad j = 1, \ldots, p_l, \quad l = 1, \ldots, L, \]
and
\[ \Psi_j^{(l)} = s_j^{(l)} d f_{jl} \left( \Psi^{(1)}, \ldots, \Psi^{(l-1)} \right) \frac{\eta_j^{(l)}}{E[f_{jl} \left( \Psi^{(1)}, \ldots, \Psi^{(l-1)} \right)]}, \quad j = 1, \ldots, p_l, \quad l = 1, \ldots, L \quad (2) \]
where \( \eta_j^{(l)} \) are given independent prior distributions with mean 1 and \( s_j^{(l)} \) is the
sparsity shape parameter of \( \eta_j^{(l)} \). It follows that \( E[\Psi_j^{(l)}] = s_j^{(l)} d \) which mimics
the normal-gamma prior distribution where the sparsity shape parameter is
the shape parameter of the gamma distributions and \( d \) can be interpreted as
a scale parameter. The function \( f_{jl} \) will usually be a simple function using
combinations of additions and multiplications to allow easy calculation of its
expectation and clear understanding of the sparsity. 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. The
Bayesian group lasso 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.

A sensible choice of sparsity is important for good estimation of the re-
gression coefficient. Therefore, it is important to consider the sparsity shape
parameters of the distributions of the regression coefficients induced by this
prior. The sparsity within the \( l \)-th level is controlled by the sparsity shape parameter of the marginal distribution of \( \Psi_j^{(l)} \). It is also interesting to con-
sider the sparsity shape parameter of the distribution of \( \Psi_j^{(l)} \) conditional on
\( \Psi^{(1)}, \ldots, \Psi^{(l-1)} \). We refer to the sparsity shape parameter of the marginal
distribution of \( \Psi_j^{(l)} \) as the marginal sparsity shape parameter and the sparsity
shape parameter of the conditional distribution of \( \Psi_j^{(l)} \) given \( \Psi^{(1)}, \ldots, \Psi^{(l-1)} \)
as the conditional sparsity shape parameter. We similarly distinguish between the
shrinkage induced by the marginal and conditional distributions.

The conditional sparsity shape parameter and shrinkage are more easily
understood than the marginal sparsity shape parameter and shrinkage. The
conditional sparsity shape parameter is given by the sparsity shape parameter
of $\eta_j^{(l)}$ and the conditional shrinkage has scale of $d \frac{f_{jl}(\Psi^{(1)}, \ldots, \Psi^{(l-1)})}{E[f_{jl}(\Psi^{(1)}, \ldots, \Psi^{(l-1)})]}$. Therefore, smaller values of $f_{jl}(\Psi^{(1)}, \ldots, \Psi^{(l-1)})$ lead to larger amounts of shrinkage at all values of the $t$-statistic in the characterisation of Appendix A. To characterise the marginal sparsity shape parameter, we will consider functions, $f_{jl}(\Psi^{(1)}, \ldots, \Psi^{(l-1)})$, formed through products or sums.

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

1. the sparsity shape parameter of $\Psi$ is $\min\{\lambda_i\}$ if $\Psi = \prod_{i=1}^{K} \eta_i$.  
2. the sparsity shape parameter of $\Psi$ is $\sum_{i=1}^{K} \lambda_i$ if $\Psi = \sum_{i=1}^{K} \eta_i$.

An interesting special case is the product of two gamma random variables $\Psi = \eta_1\eta_2$ for which the density has an 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, pg. 374). The distribution is called the $K$-distribution (Jakeman and Pusey, 1978) in several areas of physics. Using a small value approximation (Abramowitz and Stegun, 1964, eqn 9.6.9), this density at a value of $\Psi$ 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 so the sparsity shape parameter is $\min\{\lambda_1, \lambda_2\}$ which is in agreement with Theorem 1. Theorem 1 can be extended to the gamma-gamma case giving:

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

1. the sparsity shape parameter of $\Psi$ is $\min\{\lambda_i\}$ if $\Psi = \prod_{i=1}^{K} \eta_i$.  
2. the sparsity shape parameter of $\Psi$ 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 parameters.

The previous results relate to the shape of the prior density for $\Psi_i$ close to zero when it is defined through products or sums. The appropriateness of the marginal sparsity shape parameter can be checked by comparing the shrinkage profiles for a product or a sum of normal-gamma (or normal-gamma-gamma) distributed random variables and for a single normal-gamma (or
Figure 1: Shrinkage profiles for various choices of products of two normal-gamma prior distribution with: $\lambda_2 = \lambda_1$ (solid line), $\lambda_2 = 5\lambda_1$ (dashed line), $\lambda_2 = 10\lambda_1$ (dot-dashed line) with $d = 1/\text{SE}^2$ compared to a normal-gamma with shape $\lambda_1$ (dotted line) with $d = 1/\text{SE}^2$.

Figure 2: Shrinkage profiles for various choices of product of three normal-gamma prior distributions where $\lambda_3 = \lambda_2$ with: $\lambda_2 = \lambda_1$ (solid line), $\lambda_2 = 5\lambda_1$ (dashed line), $\lambda_2 = 10\lambda_1$ (dot-dashed line) with $d = 1/\text{SE}^2$ compared with a normal-gamma with shape $\lambda_1$ (dotted line) with $d = 1/\text{SE}^2$.

A normal-gamma-gamma distributed random variable with the marginal sparsity shape parameter of the product or sum. If the concept of marginal sparsity is useful then we would expect the shrinkage profiles to be similar. Figure 1 and Figure 2 show the shrinkage curves for different choices of products of two and three normal-gamma distributions respectively with $d = 1/\text{SE}^2$. The
Figure 3: Shrinkage profiles for various choices of products of two normal-gamma-gamma prior distribution with: $\lambda_2 = \lambda_1$ (solid line), $\lambda_2 = 5\lambda_1$ (dashed line), $\lambda_2 = 10\lambda_1$ (dot-dashed line) with $d = 1/SE^2$ compared to a normal-gamma-gamma with shape $\lambda_1$ (dotted line) with $d = 1/SE^2$.

The marginal sparsity shape parameter is $\lambda_1$ and the shrinkage curve for a single normal-gamma prior with sparsity shape parameter of $\lambda_1$ is also shown. The shape of the shrinkage curves are very similar for different choices of $\lambda_2$ with shrinkage decreasing slightly as $\lambda_2$ becomes larger. The effect is more pronounced if $\lambda_1$ is smaller. The results with the product of three normal-gamma distributions are similar. This suggests that the sparsity shape parameter (although fairly crude) does give comparable forms of shrinkage for different values of $t$. Figures 3 and 4 show similar graphs for the NGG case with different values of $c$ which show results that are very similar to the normal-gamma case.
3.3 Specific examples

3.3.1 Linear model with interaction terms

In our framework, we interpret strong heredity as a prior belief that $\beta_{jk}$ will be strongly shrunk to zero if either $\beta_j$ or $\beta_k$ are strongly shrunk to zero. We interpret weak heredity as a prior belief that $\beta_{jk}$ will be strongly shrunk to zero if both $\beta_j$ and $\beta_k$ are strongly shrunk to zero. These prior beliefs can be represented using a hierarchical sparsity prior. First, we define two levels: the main effect level and the interaction level. The first level (the main effect level) has $p_1 = q$ terms listed as $\beta_1, \ldots, \beta_q$ and the second level (the interaction level) has...
\( p_2 = q(q - 1)/2 \) terms listed as \( \beta_{jk} \) for \( k = 1, \ldots, j - 1, j = 1, \ldots, p \). In the case of strong heredity, we use the prior

\[
\beta_j \sim N \left( 0, \lambda_1 d \eta_j^{(1)} \right), \quad \beta_{jk} \sim N \left( 0, \lambda_2 d \eta_j^{(1)} \eta_k^{(1)} \eta_{jk}^{(2)} \right)
\]

\[
\eta_j^{(1)} \sim GG \left( \lambda_1, c, \frac{c - 1}{\lambda_1} \right) \text{ and } \eta_j^{(2)} \sim \frac{c}{\lambda_2} GG \left( \lambda_2, c, \frac{c - 1}{\lambda_2} \right).
\]

The prior variance of \( \beta_{jk} \) is small if either \( \eta_j^{(1)} \) or \( \eta_k^{(1)} \) is small (and hence also the prior variances of \( \beta_j \) and \( \beta_k \)). Therefore, an interaction term \( \beta_{jk} \) will tend to be small (since its variance is small) if either \( \beta_j \) is small (which implies that its prior variance is small) or \( \beta_k \) is small (which implies that its prior variance is small). In the case of weak heredity, we use the prior

\[
\beta_j \sim N \left( 0, \lambda_1 d \eta_j^{(1)} \right), \quad \beta_{jk} \sim N \left( 0, \lambda_2 d \frac{1}{2} \left( \eta_j^{(1)} + \eta_k^{(1)} \right) \eta_{jk}^{(2)} \right)
\]

\[
\eta_j^{(1)} \sim GG \left( \lambda_1, c, \frac{c - 1}{\lambda_1} \right) \text{ and } \eta_j^{(2)} \sim \frac{c}{\lambda_2} GG \left( \lambda_2, c, \frac{c - 1}{\lambda_2} \right).
\]

The prior variance of \( \beta_{jk} \) is small if the prior variances of both \( \beta_j \) and \( \beta_k \) are small. Therefore, the interaction terms will tend to be small if and only if both \( \beta_j \) and \( \beta_k \) are small (using similar reasoning to the strong heredity case). In general we would assume that \( \lambda_2 < \lambda_1 \) since the interactions will tend to be sparser than the main effects. This implies that the marginal and conditional sparsity shape parameters of the main effects are \( \lambda_1 \) and the marginal and conditional sparsity shape parameters of the interactions are \( \lambda_2 \).

### 3.3.2 GAM models

In section 3.1.2, 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). This can be approached using a hierarchical sparsity prior by defining the first level (the variable level) by \( p_1 = pq \) terms \( \theta_{jk} \) for \( j = 1, \ldots, p, k = 1, \ldots, q \) and the second level (the basis level) by \( p_2 = pK \) terms \( \gamma_{jk} \) for \( j = 1, \ldots, p \) and \( k = 1, \ldots, K \). We propose the prior

\[
\theta_{jk} \sim N \left( 0, \lambda_1 d \eta_j^{(1)} \right), \quad \gamma_{jk} \sim N \left( 0, \lambda_2 d \eta_j^{(1)} \eta_{jk}^{(2)} \right),
\]

\[
\eta_j^{(1)} \sim GG \left( \lambda_2, c, \frac{c - 1}{\lambda_2} \right) \text{ and } \eta_{jk}^{(2)} \sim GG \left( \lambda_1, c, \frac{c - 1}{\lambda_2} \right).
\]

The prior implies that all polynomial term coefficients have the same conditional prior variance. 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 polynomial coefficients \( \theta_j, \ldots, \theta_j \) and basis function coefficients \( \gamma_j, \ldots, \gamma_j \) 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 results in section 3 suggest that the marginal sparsity shape parameters of the basis functions are \( \min\{\lambda_1, \lambda_2\} \), and the conditional sparsity shape parameters of the basis functions are \( \lambda_2 \). The marginal and conditional sparsity shape parameters of the variables are \( \lambda_1 \).

### 3.3.3 GAM models with interactions

A more elaborate form of GAM allows for functions, \( f_{ji}(\cdot, \cdot) \), modelling non-linear interaction effects. In this case, the GAM model is extended to

\[
y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \sum_{j=1}^{p-1} \sum_{k=1}^{q} f_{jk}(X_{ij}, X_{ik}) + \epsilon_i
\]

\[
= \sum_{j=1}^{p} \sum_{k=1}^{q} \theta_{jk}^{(M)} X_{ij}^k + \sum_{j=1}^{p} \sum_{k=1}^{q} \gamma_{jk}^{(M)} g(X_{ij}, \tau_{jk}) + \sum_{j=1}^{p-1} \sum_{k=1}^{q} \sum_{l=1}^{q} \sum_{m=1}^{q} \theta_{jklm}^{(I)} X_{ij}^l X_{ik}^m \\
+ \sum_{j=1}^{p-1} \sum_{k=1}^{q} \sum_{l=1}^{q} \sum_{m=1}^{q} \gamma_{jklm}^{(I)} g(X_{ij}, \tau_{jl}) g(X_{ik}, \tau_{km}) + \epsilon_i
\]

where, again, \( \epsilon_i \sim N(0, \sigma^2) \). A hierarchical sparsity 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. We define a first level (the variable level) with \( p_1 = pqM \) terms \( \theta_{jk}^{(M)} \) for \( j = 1, \ldots, p, k = 1, \ldots, qM \) and a second level (the interaction level) with \( p_2 = p(p-1)/2qI \) terms \( \theta_{jklm}^{(I)} \) for \( j = 1, \ldots, p, k = 1, \ldots, j-1, l = 1, \ldots, qI, m = 1, \ldots, qI \). The third and fourth levels contain the basis functions for the main effect and the interactions respectively. The third level has \( p_3 = pK \) terms \( \gamma_{jk}^{(M)} \) for \( j = 1, \ldots, p, k = 1, \ldots, K \) and the fourth level has \( p_4 = p(p-1)/2K^2 \) terms \( \gamma_{jklm}^{(I)} \) for \( j = 1, \ldots, p, k = 1, \ldots, j-1, l = 1, \ldots, K, m = 1, \ldots, K \). The proposed prior, with strong heredity, is

\[
\theta_{jk}^{(M)} \sim N\left(0, \lambda_1 d \eta_j^{(1)} \right), \quad \theta_{jklm}^{(I)} \sim N\left(0, \lambda_2 d \eta_j^{(1)} \eta_k^{(1)} \eta_j^{(2)} \right)
\]

\[
\gamma_{jk}^{(M)} \sim N\left(0, \lambda_3 d \eta_j^{(1)} \eta_j^{(3)} \right), \quad \gamma_{jklm}^{(I)} \sim N\left(0, \lambda_4 d \eta_j^{(1)} \eta_j^{(2)} \eta_j^{(4)} \right)
\]

\[
\eta_j^{(1)} \sim GG\left(\lambda_1, c, \frac{c-1}{\lambda_1} \right), \quad \eta_j^{(2)} \sim GG\left(\lambda_2, c, \frac{c-1}{\lambda_1} \right)
\]
\[ \eta_{jk}^{(3)} \sim \text{GG} \left( \lambda_3, c, \frac{c-1}{\lambda_3} \right) \text{ and } \eta_{jklm}^{(4)} \sim \text{GG} \left( \lambda_4, c, \frac{c-1}{\lambda_4} \right). \]

If \( \eta_j^{(1)} \) is small then both the main effects \( \theta_{jk}^{(M)} \) and the basis function coefficients \( \gamma_{jk}^{(M)} \) will be small. Similarly, if \( \eta_j^{(1)} \eta_k^{(1)} \eta_j^{(2)} \) is small then both the interaction terms \( \theta_{jklm}^{(I)} \) and the basis function coefficients \( \gamma_{jklm}^{(I)} \) will be small. This allows variable selection at the main effect and interaction term levels. The prior also links the priors for the interaction and main effects (and, consequently, their associated basis function coefficients) since \( \eta_j^{(1)} \eta_k^{(1)} \) will be small if both \( \eta_j^{(1)} \) and \( \eta_k^{(1)} \) are small. The marginal sparsities are \( \lambda_1 \) for level 1, \( \min\{\lambda_1, \lambda_2\} \) for level 2, \( \min\{\lambda_1, \lambda_3\} \) for level 3 and \( \min\{\lambda_1, \lambda_2, \lambda_3\} \) for level 4. The conditional sparsities are \( \lambda_2 \) for level 2, \( \lambda_3 \) for level 3 and \( \lambda_4 \) for level 4.

### 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

\[ y_i = \alpha + \sum_{l=1}^{L} X_{i}^{(l)} \beta^{(l)} + \epsilon_i, \quad i = 1, \ldots, n \]

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

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

and

\[ \Psi_j^{(l)} = s_j^{(l)} d \frac{f_{jj} (\Psi^{(1)}, \ldots, \Psi^{(l-1)})}{E[f_{jj} (\Psi^{(1)}, \ldots, \Psi^{(l-1)})]} \eta_j^{(l)}, \quad j = 1, \ldots, p_l, \quad l = 1, \ldots, L. \quad (4) \]

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 \((\alpha, \beta)\) and \( \sigma^2 \) follow from standard results for Bayesian linear regression models. The parameters \( \Psi, d \) and \( \phi \) 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 condition with the additional restriction that $\Psi$, $d$ and $\phi$ 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 $\Psi$ and $d$). A new value $\phi^{(l)}'$ is proposed according to

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

where $\epsilon^{(l)} \sim N\left(0, \sigma^{2(i)}_{\phi^{(l)}}\right)$. 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 $q^{2(i)}_{\phi^{(l)}}\left(\phi^{(l)}, \phi^{(l)}'\right)$. The value $\phi^{(l)}'$ is accepted or rejected using the standard Metropolis-Hastings acceptance probability

$$\alpha\left(\phi^{(l)}, \phi^{(l)}'\right) = \frac{\prod_{j=1}^{p} p\left(\Psi^{(l)}_j | \phi^{(l)}'\right) p\left(\phi^{(l)}', \phi^{(l)}\right) q^{2(i)}_{\phi^{(l)}}\left(\phi^{(l)}', \phi^{(l)}\right)}{\prod_{j=1}^{p} p\left(\Psi^{(l)}_j | \phi^{(l)}\right) p\left(\phi^{(l)}\right) q^{2(i)}_{\phi^{(l)}}\left(\phi^{(l)}, \phi^{(l)}'\right)}.$$

The variance of the increment is updated by

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

where $1/2 < a \leq 1$. This algorithm leads to an average acceptance rate which converges to $\tau$. 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-level factors as linear and quadratic effects using orthogonal polynomials. The model in section 3.1.1 was extended to allow for quadratic effects and assumed that

$$y_i = \sum_{j=1}^{p} X_{ij} \theta_j + \sum_{j=1}^{p} X_{ij}^2 \delta_j + \sum_{j=1}^{p} \sum_{k=1}^{j-1} X_{ij} X_{ik} \gamma_{jk} + \epsilon_i, \quad i = 1, \ldots, n.$$ 

The prior proposed in section 3.3.1 was extended with \( \delta_j \sim N\left(0, \lambda_1 d \eta_j^{(1)}\right) \). 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 sparsity parameter for the main effects was given the prior \( \lambda_1 \sim \text{Ex}(1) \) which centred the prior over a heavy-tailed version of the Bayesian lasso. We defined \( \lambda_2 = r \lambda_1 \) and assumed that the interactions were sparser than the main effects which implied that \( r < 1 \) and so we choose \( r \sim \text{Be}(2, 6) \) which implied that \( E[r] = 1/3 \) suggesting that the interactions will be much sparser than the main effects. The scale parameter, \( d \), was given the prior \( p(d) \propto (1 + d)^{-2} \) which implied that \( E[d] = 1 \) 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 well away 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 $\Psi$’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 $\Psi$’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.

|    |       |
|----|-------|
| λ₁ | 0.48 (0.15, 2.71) |
| λ₂ | 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

The inference about the sparsity shape parameters λ₁ and λ₂ and the scale parameter d are shown in Table 1. The parameter λ₁ had a posterior median 0.48 which indicated that some effects were close to zero. The parameter λ₂ had a much smaller posterior median which indicated that the interactions were much sparser than the main effects. These results were consistent with the inferences about the regression coefficients.

5.2 Example 2: Prostate cancer data

Data from a prostate cancer trial (Stamey et al., 1989) have become a standard example in the regularization literature (Tibshirani, 1996; Zou and Hastie, 2005; Kyung et al., 2010). 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). Previous modelling had often included the continuous variables as linear effects. An exception is Lai et al. (2012) who considered flexibly modelling their effects. We followed this approach using the GAM model in section 3.1.2 with the prior described in section 3.3.2. All continuous variables were normalized to have a minimum of 0 and a maximum of 1. A piecewise linear spline basis function was assumed for each
variable so that

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

where \((x)_{+} = \max\{0, x\}\) and \(\tau_k = \frac{k-1}{K-1}\) for \(k = 1, \ldots, K\). In this example, we use \(K = 60\). The priors for the hyperparameters were: \(\lambda_1 \sim \text{Ga}(1, 1)\), \(\lambda_2 \sim \text{Ga}(1, 10)\), and \(p(d) \propto (1 + d)^{-2}\). These priors were also used in Example 1 and the justification is the same. The prior for \(\lambda_2\) implied that \(\mathbb{E}[\lambda_2] = 0.1\) which suggested much greater sparsity in the basis function (since only a few are typically needed to model the functional effect).

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

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 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.

![Figure 8: Prostate cancer data – the posterior distribution of $\Psi$ for each variable shown as the posterior median (cross) and 95% credible interval (solid line)](image)

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 had the largest posterior median and had support at larger values of $\Psi$ 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.

| $\lambda_1$ | 0.96 (0.31, 3.44) |
|-------------|-------------------|
| $d$         | 0.64 (0.09, 6.10) |

Table 2: Prostate cancer data – the posterior distribution of the hyperparameters summarised as posterior median and 95% credible interval

A summary of the posterior distribution of $\lambda_1$ and $d$ are shown in Table 2 and a summary of the posterior distributions of variable-specific $\lambda_2$ are shown in Figure 9. The posterior median of $\lambda_1$ is close to 1 indicating that only some of the variables are important but that there is not a high degree of sparsity. The parameter $\lambda_2$ indicate the sparsity in the coefficients of the spline basis for each variable. A smaller value of $\lambda_2$ indicates that less splines are needed to model
Figure 9: Prostate cancer data – the posterior distribution of \( \lambda^2 \) for each variable shown as the posterior median (cross) and 95% credible interval (solid line)

the effect of the variable and, therefore, are a measure of the departure from linearity for each variable. The variable \( lv \) has the largest posterior median and so the largest departure from linearity whereas \( lcp \) has the smallest posterior median and so the smallest departure form linearity. This is consistent with the estimated effects shown in Figure 7.

5.3 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 as described in section 3.3.3 together with the prior structure. Gustafson (2000) used a square root transformation of the predictors since these a 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 unevenly 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. We assumed a
non-linear form for the effect of each variable and the interactions effects. We used the model in (3) with $q_M = 1$, $q_I = 0$ and $g_j(x, \tau) = (x - \tau)_+$. The number of knots was $K = 10$. There were 5 main effects and 10 interactions which leads to 1055 regression parameters in the model.

The priors for the hyperparameter of the model were as follows. The sparsity parameters for the main effects and interaction terms were chosen as in Example 1 with $\lambda_1 \sim \text{Ex}(1)$ and $\lambda_2 = r \lambda_1$ where $r \sim \text{Be}(2, 6)$ which implied that $E[r] = 1/3$ suggesting that the interaction were much sparser than the main effects. The conditional sparsity shape parameters for the nonlinear terms were chosen to be $\lambda_3 \sim \text{Ga}(1, 10)$ and $\lambda_4 \sim \text{Ga}(1, 100)$ which indicated that nonlinear terms were less likely to be included in the interaction function than the main effects function (which reflectd 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 $E[d] = 1$ but with a heavy tail.

The estimated main effects and interactions are shown in Figure 10. The effect of A, D and E were small whereas B and C had an increasing, non-linear effect with a 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 11 shows the posteriors for the $\Psi$’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.

|   |     |             |             |
|---|---|-------------|-------------|
| $\lambda_1$ | 1.96 | (0.41, 4.68) |             |
| $\lambda_2$ | 0.40 | (0.13, 1.12) |             |
| $d$          | 0.84 | (0.09, 10.06)|             |

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

A summary of the posterior distribution of $\lambda_1$, $\lambda_2$ and $d$ are shown in Table 3 and a summary of the posterior distributions of variable-specific $\lambda_3$ and $\lambda_4$ are shown in Figure 12. The posterior median of $\lambda_1$ is close to 2 and in-
Figure 10: Computer data – the posterior mean of each main effect and each interaction. Darker colours represent lower values for the interaction graphs.
Figure 11: Computer data – the posterior distribution of $\Psi$ for each main effect and each interaction shown as the posterior median (cross) and 95% credible interval (solid line).

Figure 12: Computer data – the posterior distribution of $\lambda_3$ and $\lambda_4$ for each main effect and each interaction respectively shown as the posterior median (cross) and 95% credible interval (solid line).

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 $\lambda_2$ indicates that the interactions are much sparser than the main effects. The sparsity parameters $\lambda_3$ and $\lambda_4$ indicate the amount that the effects deviate from linearity. The posterior medians of $\lambda_3$ for B and C
are much larger than for the other variables. This indicates a departure from linearity which is confirmed by the estimate regression effects in Figure 10. The posterior distributions of $\lambda_4$ are fairly similar indicating little difference in the level of departure from normality for the interactions (the relatively small effect of the interactions leads to a small amount of information about these parameters).

6 Discussion

This paper describes a hierarchical approach to prior construction in sparse regression problems. We assume that variables can be divided into levels and the relationship between the regression coefficients can be expressed hierarchically. The framework allows control of both the conditional sparsity and marginal sparsity of groups of regression coefficients at different levels of the prior. These priors have natural applications in problems with such structure such as a model with interactions and non-linear Bayesian regression models. These priors are able to find sparse estimates in situations where there are large numbers of parameters. We feel that these approaches will have many applications. For example, Kalli and Griffin (2012) use a simple, two stage prior hierarchical prior in a regression model with time-varying regression coefficients. This allows the control of both sparsity of the variables (where are values in the series are shrunk to zero) and sparsity within a series for a particular variable.

References

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

Armagan, A., D. Dunson, and M. Clyde (2011). Generalized beta mixtures of Gaussians. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (Eds.), Advances in Neural Information Processing Systems 24, pp. 523–531.

Armagan, A., D. B. Dunson, and J. Lee (2013). Generalized double Pareto shrinkage. Statistica Sinica 23, 119–143.
Atchadé, Y. F. and J. S. Rosenthal (2005). On adaptive Markov chain Monte Carlo algorithms. *Bernoulli* 11, 815–828.

Bishop, C. M. and M. E. Tipping (2000). Variational relevance vector machines. In C. Boutilier and M. Goldszmidt (Eds.), *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pp. 46–53. Morgan Kaufmann.

Caron, F. and A. Doucet (2008). Sparse bayesian nonparametric regression. In A. McCallum and S. Roweis (Eds.), *Proceedings of the 25th Annual International Conference on Machine Learning (ICML 2008)*, pp. 88–95. Omnipress.

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

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

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

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics* 32, 407–499.

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

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

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

Griffin, J. E. and P. J. Brown (2011). Bayesian hyper-lassos with non-convex penalisation. *Australian and New Zealand Journal of Statistics* 53, 423–442.

Griffin, J. E. and P. J. Brown (2012). Structuring shrinkage: some correlated priors for regression. *Biometrika* 99, 481–487.

Griffin, J. E. and P. J. Brown (2013). Some priors for sparse regression modelling. *Bayesian Analysis* 8, to appear.
Gustafson, P. (2000). Bayesian regression modeling with interactions and smooth effects. *Journal of the American Statistical Association* 95, 795–806.

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

Hans, C. (2009). Bayesian lasso regression. *Biometrika* 96, 835–845.

Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. *Journal of the American Statistical Association* 106, 1383–1393.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer.

Hastie, T. J. and R. J. Tibshirani (1993). *Generalized additive models*. Chapman and Hall.

Jacob, L., G. Obozinski, and J.-P. Vert (2009, June). Group lasso with overlaps and graph lasso. In L. Bottou and M. Littman (Eds.), *Proceedings of the 26th International Conference on Machine Learning*, Montreal, pp. 433–440. Omnipress.

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

Kalli, M. and J. E. Griffin (2012). Time-varying sparsity in dynamic regression models. Technical Report UKC/SMSAS/12/002, University of Kent.

Kohn, R., M. Smith, and D. Chan (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing* 11, 313–322.

Kyung, M., J. Gill, M. Ghosh, and G. Casella (2010). Penalized Regression, Standard Errors, and Bayesian Lassos. *Bayesian Analysis* 5, 369–412.

Lai, R. C. S., H.-C. Huang, and T. C. M. Lee (2012). Fixed and random effects selection in nonparametric additive mixed models. *Electronic Journal of Statistics* 6, 810–842.

Lee, A., F. Caron, A. Doucet, and C. Holmes (2012). Bayesian sparsity-path-analysis of genetic association using generalised t priors. *Statistical Applications in Genetics and Molecular Biology* 11: iss2, Art 5.
Li, F. and N. R. Zhang (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association* 105, 1202–1214.

Miasojedow, B., E. Moulines, and M. Vihola (2013). An adaptive parallel tempering algorithm. *Journal of Computational and Graphical Statistics*, forthcoming.

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

Park, T. and G. Casella (2008). The Bayesian Lasso. *Journal of the American Statistical Association* 103, 672–680.

Polson, N. G. and J. G. Scott (2011). Shrink globally, act locally: sparse Bayesian regularization and prediction. In M. J. Bernardo J. M., Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West (Eds.), *Bayesian Statistics* 9, pp. 501–538. Oxford: Clarendon Press.

Polson, N. G. and J. G. Scott (2012). Local shrinkage rules, Lévy processes and regularized regression. *Journal of the Royal Statistical Society, Series B* 74, 287–311.

Polson, N. G., J. G. Scott, and J. Windle (2013). The Bayesian bridge. *Journal of the Royal Statistical Society, Series B*, forthcoming.

Raiffa, H. and R. Schlaifer (1961). *Applied statistical decision theory*. M.I.T. Press.

Raman, S., T. Fuchs, P. Wild, E. Dahl, and V. Roth (2009, June). The Bayesian group-lasso for analyzing contingency tables. In L. Bottou and M. Littman (Eds.), *Proceedings of the 26th International Conference on Machine Learning*, Montreal, pp. 881–888. Omnipress.

Roberts, G. O. and J. S. Rosenthal (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *Journal of Applied Probability* 44, 458–475.

Roberts, G. O. and J. S. Rosenthal (2009). Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics* 18, 349–367.
A Shrinkage characterisation

Figure 13 shows the shrinkage curves, $1 - S(t)$, for the normal-gamma and normal-gamma-gamma prior distributions with different choices of $\lambda$ (for the normal-gamma prior) and $\lambda$ and $c$ (for the normal-gamma-gamma prior). Since the prior mean of $\Psi^*$ is fixed, the NGG curves will tend to the NG curve as $c \to \infty$. The graphs show that the shape of the shrinkage curve for small values of $t$ is largely controlled by the scale and the value of $\lambda$. Decreasing $\lambda$ leads to larger amounts of shrinkage for small $t$ and the change from high to low levels of shrinkage happens more abruptly. The scale changes the position of the shrinkage curve with a larger mean of $\Psi_i$ leading to more shrinkage.
Figure 13: Shrinkage profiles for various choices of normal-gamma (NG) and normal-gamma-gamma (NGG) prior distributions where $\frac{V[\beta]}{\lambda} = \frac{1}{SE^2}$ (solid line), $\frac{V[\beta]}{\lambda} = \frac{10}{SE^2}$ (dashed line) and $\frac{V[\beta]}{\lambda} = \frac{0.1}{SE^2}$ (dot-dashed line).

for all values of $t$. The parameter $c$ controls the tails of the prior distribution and effects the shrinkage at large values of $t$. Smaller values of $c$ lead to less shrinkage in the tails.
B Proofs

B.1 Proof of Proposition 1

Let $\tau = \beta / \text{SE}$ then the sampling distribution of $t$ is $N(\tau / \text{SE}, 1)$ and Proposition 1 of Griffin and Brown (2010) shows that

$$E[\tau|t] = (1 - S(t))t$$

where

$$S(t) = -\frac{1}{t} \left[ \frac{d}{ds} \log h(s) \right]_{s=t},$$

$h(s) = \int N(s|\tau, 1)\pi_\tau(\tau)\,d\tau$ and $\pi_\tau(\tau) = \text{SE} \pi(\text{SE}\tau)$.

It follows that

$$E[\beta|\hat{\beta}] = (1 - S(t))\hat{\beta}$$

The prior $\pi_\tau$ is just the prior for $\beta$ induced through rescaling by a factor of the standard error. $\square$

B.2 Proof of Theorem 1

Part (i)

Suppose that $\lambda_1 = \min\{\lambda_i\}$ then

$$p(\Psi) = \prod_{i=1}^{K} \frac{1}{\Gamma(\lambda_i)} \Psi^{\lambda_i-1} \int_{0}^{\infty} \cdots \int_{0}^{\infty} \exp \left\{ -\Psi \sum_{i=2}^{K} \eta_i \right\} \prod_{i=2}^{K} \eta_i^{\lambda_i-\lambda_1-1} \exp \left\{ -\sum_{i=2}^{K} \eta_i \right\} \,d\eta_2 \cdots d\eta_K$$

Thus

$$C(\Psi) = p(\Psi) / \Psi^{\lambda_1-1}$$

$$= \prod_{i=1}^{K} \frac{1}{\Gamma(\lambda_i)} \int_{0}^{\infty} \cdots \int_{0}^{\infty} \exp \left\{ -\Psi \sum_{i=2}^{K} \eta_i \right\} \prod_{i=2}^{K} \eta_i^{\lambda_i-\lambda_1-1} \exp \left\{ -\sum_{i=2}^{K} \eta_i \right\} \,d\eta_2 \cdots d\eta_K$$

By the dominated convergence theorem

$$\lim_{\Psi \to 0} C(\Psi) = C(0) = \prod_{i=1}^{K} \frac{1}{\Gamma(\lambda_i)} \prod_{i=2}^{K} \int_{0}^{\infty} \eta_i^{\lambda_i-\lambda_1-1} \exp \left\{ -\sum_{i=2}^{K} \eta_i \right\} \,d\eta_i = \frac{1}{\Gamma(\lambda_1)} \prod_{i=2}^{K} \frac{\Gamma(\lambda_i - \lambda_1)}{\Gamma(\lambda_i)}$$

since $\lambda_i \geq \lambda_1$. Therefore, the sparsity shape parameter is $\min\{\lambda_i\}$. $\square$
Part (ii)

In this case, \( \Psi \sim \text{Ga}(\sum_{i=1}^{K} \lambda_i, 1) \) and so the sparsity shape parameter is \( \sum_{i=1}^{K} \lambda_i \).

B.3 Proof of Theorem 2

Part (i)

Suppose that \( \lambda_1 = \min \{ \lambda_i \} \) then

\[
p(\Psi) \propto \Psi^{\lambda_1 - 1} \int_0^{\infty} \cdots \int_0^{\infty} \left\{ 1 + \Psi \prod_{i=2}^{K} \eta_i \right\}^{-(\lambda_1 + c_1)} \left\{ \prod_{i=2}^{K} \eta_i^{\lambda_i - 1} \left( 1 + \eta_i \right)^{-(\lambda_i + c_i)} \right\} d\eta_2 \cdots d\eta_K
\]

Thus

\[
C(\Psi) = p(\Psi) / \Psi^{\lambda_1 - 1}
\]

\[
\propto \int_0^{\infty} \cdots \int_0^{\infty} \left\{ 1 + \Psi \prod_{i=2}^{K} \eta_i \right\}^{-(\lambda_1 + c_1)} \left\{ \prod_{i=2}^{K} \eta_i^{\lambda_i - 1} \left( 1 + \eta_i \right)^{-(\lambda_i + c_i)} \right\} d\eta_2 \cdots d\eta_K
\]

By the dominated convergence theorem

\[
\lim_{\Psi \to 0} C(\Psi) = C(0) \propto \int_0^{\infty} \cdots \int_0^{\infty} \left\{ \prod_{i=2}^{K} \eta_i^{\lambda_i - 1} \left( 1 + \eta_i \right)^{-(\lambda_i + c_i)} \right\} d\eta_2 \cdots d\eta_K
\]

a constant, since we are integrating kernels of \( \text{GG}(\lambda_i - \lambda_1, \lambda_1 + c_i, 1) \) distribution and \( \lambda_i \geq \lambda_1 \). Therefore, the sparsity parameter of the marginal distribution of \( \Psi_i \) is given by the simple form of \( \min \{ \lambda_i \} \). \( \square \)

Part (ii)

Suppose \( \Psi_i \sim \text{GG}(\lambda_i, c, d), i = 1, 2 \) then \( Y = \Psi_1 + \Psi_2 \) has a density

\[
f_y(y) \propto \int_0^{y} (y - w)^{\lambda_1 - 1} \left[ 1 + \frac{(y - w)}{d} \right]^{-(\lambda_1 + c)} w^{\lambda_2 - 1} \left[ 1 + \frac{w}{d} \right]^{-(\lambda_2 + c)} dw
\]

\[
= y^{\lambda_1 + \lambda_2 - 1} \int_0^{1} (1 - z)^{\lambda_1 - 1} z^{\lambda_2 - 1} \left[ 1 + \frac{y(1 - z)}{d} \right]^{-(\lambda_1 + c)} \left[ 1 + \frac{yz}{d} \right]^{-(\lambda_2 + c)} dz
\]

\[
= y^{\lambda_1 + \lambda_2 - 1} C(y)
\]

and by dominated convergence theorem

\[
\lim_{y \to 0} C(y) = f_0^1 (1 - z)^{\lambda_1 - 1} z^{\lambda_2 - 1} dz
\]

so the sparsity of the convolution is \( \lambda_1 + \lambda_2 \). This result can be easily generalised to the sum of \( K \) independent \( \text{GG}(\lambda_i, c, d), i = 1, \ldots, K \) random variables.