Biometrika (2012), *pp.* 1–7
doi: 10.1093/biomet/asr082

Printed in Great Britain

Structuring shrinkage: some correlated priors for regression

By J. E. Griffin and P. J. Brown

School of Mathematics, Statistics and Actuarial Science, Cornwallis Building, University of Kent, Canterbury CT2 7NF, U.K.

J.E.Griffin-28@kent.ac.uk Philip.J.Brown@kent.ac.uk

Summary

This paper develops a rich class of sparsity priors for regression effects that encourage shrinkage of both regression effects and contrasts between effects to zero whilst leaving sizeable real effects largely unshrunk. The construction of these priors uses some properties of normal-gamma distributions to include design features in the prior specification, but has general relevance to any continuous sparsity prior. Specific prior distributions are developed for serial dependence between regression effects and correlation within groups of regression effects.

Some key words: Fused prior; Grouped prior; Lasso; Multiple regression; Normal-gamma prior; Sparsity.

1. Introduction

The construction of sparse estimators in regression models has increasingly been seen as an important problem in statistics, given data with large numbers of variables where the effects of only a few are assumed to be far from zero. By encouraging sparsity, these estimators can avoid overfitting and so lead to better out-of-sample predictions than traditional estimators. There are two popular Bayesian paradigms for sparsity in regression problems: the so-called slab and spike paradigm with priors that are not absolutely continuous, which was initially proposed by Mitchell & Beauchamp (1988); and an absolutely continuous prior that places a lot of mass close to zero. The idea is often motivated by a Bayesian interpretation of the lasso estimator (Tibshirani, 1996), popular in machine learning and classical statistics. We focus on this use of a single continuous prior where the posterior mean is sparse in the sense that many regressors are shrunk close to zero if the prior density has a spike at zero. This idea is made more precise by the idea of super-efficiency (Polson & Scott, 2011) or the results of Fan & Li (2001). Examples of such priors are the horseshoe prior of Carvalho et al. (2010) and the normal-gamma or variance-gamma prior of Griffin & Brown (2010). This latter normal-gamma prior, popular in finance for robust modelling, is the main focus of this paper. It has been demonstrated that it adaptively shrinks the least squares estimate to zero with shrinkage decreasing as the effect size increases and real sizeable effects left largely unshrunk.

Often we focus not only on whether an effect is close to zero but also on whether subsets of effects have similar values. If a variable in our model is categorical, then we might be interested in whether effects for the categories are close or clustered. Alternatively, if a variable is ordinal, then we might be interested in whether the effects are close for consecutive levels of the variable. In both cases, the design suggests that particular differences of effects have special interest and we should consider how our prior shrinks these differences, as well as shrinking the effects to zero. In the classical literature, the first problem has been approached using variations of the lasso by Yuan & Lin (2006), who suggest the popular group lasso. The Bayesian interpretation of the penalty function of Yuan & Lin (2006) has been discussed by Raman et al. (2009) and Kyung et al. (2010). The second problem has been approached using the fused lasso (Tibshirani et al., 2005; Kyung et al., 2010). In our Bayesian approach, the particular technical difficulty of incorporating design features with spikey continuous prior densities is that sparsity will typically also be required for linear combinations of the regression effects. We have found that densities derived from the normal-gamma model are particularly useful since the shape close to zero of the density of sums

of normal-gamma random variables can be easily understood. In fact, particular sums of such variables will also have normal-gamma distributions. Controlling the parameters of this distribution for the differences allows us to encourage shrinkage between the posterior means of prespecified pairs or groups of regression effects. It is not just important that certain contrasts be shrunk towards zero, but also that those real effects that remain are also left unshrunk in directions encouraged by the design.

The double exponential prior of the Bayesian lasso (Park & Casella, 2008; Hans, 2009) is a particular case of the normal-gamma prior for which the gamma-mixing distribution has shape parameter $\lambda = 1$, which can be restrictive. For example, this does not allow a density with an infinite spike at zero, which only occurs when $\lambda \leqslant 0{\cdot}5$. In general, we consider mixing densities that behave like $x^{\lambda-1}$ close to 0 and refer to λ as the sparsity shape parameter, since it is a key characteristic in determining the degree of shrinkage of small effects; smaller values of λ encourage more shrinkage to zero of small effects. Larger effects will still have relatively little attenuation.

2. A MULTIVARIATE NORMAL-GAMMA DISTRIBUTION

The normal-gamma distribution has proved a useful prior in sparse regression models. We write $\beta_i \sim \mathrm{NG}\{\lambda, 1/(2\gamma^2)\}$ to mean that $p(\beta_i) = \int N(\beta_i \mid 0, \psi_i)\mathrm{Ga}\{\psi_i \mid \lambda, 1/(2\gamma^2)\}\, d\psi_i$, where $N(x \mid \mu, \sigma^2)$ represents a normal density with mean μ and variance σ^2 and $\mathrm{Ga}(y \mid a, b)$ represents a gamma density with mean a/b and variance a/b^2. A multivariate version can be induced by taking linear combinations of independent normal-gamma distributions, enabling us to induce useful design features in our priors.

DEFINITION 1. *Suppose $\beta = C\phi$, where $C = (C_{ik})$ is a $(p \times q)$-dimensional matrix with real entries and ϕ is a q-dimensional vector of independent random variables for which $\phi_k \sim \mathrm{NG}(\lambda_k, \frac{1}{2})$. Then β is said to have a p-variate correlated normal-gamma distribution, written* $\mathrm{CNG}(\lambda, C)$, *where $\lambda = (\lambda_1, \ldots, \lambda_q)$, $\lambda_k \geqslant 0$ $(k = 1, \ldots, q)$; with the understanding that $\lambda_k = 0$ implies a point mass at zero. Here to avoid singular degeneracy we assume that the $p \times p$ covariance matrix of β, $C\mathrm{diag}(\lambda)C^{\mathrm{T}}$, is of full rank p.*

Let $S_i(C)$ be the subset of $\{1, \ldots, q\}$ for which C_{ik} are nonzero $(k = 1, \ldots, q)$. The marginal density of β_i can be expressed as a scale mixture of normal densities

$$p(\beta_i) = \int N(\beta_i \mid 0, \psi_i) g(\psi_i)\, d\psi_i,$$

where $\psi_i = \sum_{k \in S_i(C)} \zeta_{ik}$ and $\zeta_{ik} \sim \mathrm{Ga}\{\lambda_k, 1/(2C_{ik}^2)\}$, so ψ_i is the convolution of independent gamma random variables having different scales. The density of ψ_i is given by Moschopoulos (1985, Theorem 1) as an infinite sum

$$g(\psi_i) \propto \psi_i^{\eta_i - 1} \exp(-b_i^* \psi_i) \sum_{l=0}^{\infty} \delta_l \psi_i^l,$$

and so the shape of the density is gamma-like with the shape close to zero controlled by the aggregate shape $\eta_i = \sum_{k \in S_i(C)} \lambda_k$, with a spike for $\eta_i \leqslant 1$. Thus η_i qualifies as the sparsity shape of this density. The tails of the density are exponential with scale $b_i^* = \min_{k \in S_i(C)} C_{ik}^2$.

The choice $C = \gamma^2 B$, where γ is a scalar and B is a $(p \times q)$-dimensional matrix, where B_{ik} is either 0 or 1, introduces a scale parameter and gives the exact marginal distribution of β_i as $\mathrm{NG}\{\sum_{k=1}^q B_{ik}\lambda_k, 1/(2\gamma^2)\}$ with $\mathrm{var}(\beta_i) = 2\gamma^2 \sum_{k=1}^q B_{ik}\lambda_k$ $(i = 1, \ldots, p)$. If, in addition, β_i and β_j are identically distributed, then the correlation simplifies to

$$\mathrm{corr}(\beta_i, \beta_j) = \frac{\sum_{k=1}^q B_{ik}B_{jk}\lambda_k}{\sum_{k=1}^q B_{ik}\lambda_k} \quad (i, j = 1 \ldots, p).$$

To understand the properties of the correlated normal-gamma prior distribution, we look at the simplest possible case with two regressors, $p = 2$, and $q = 3$. A general model sets

$$B = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

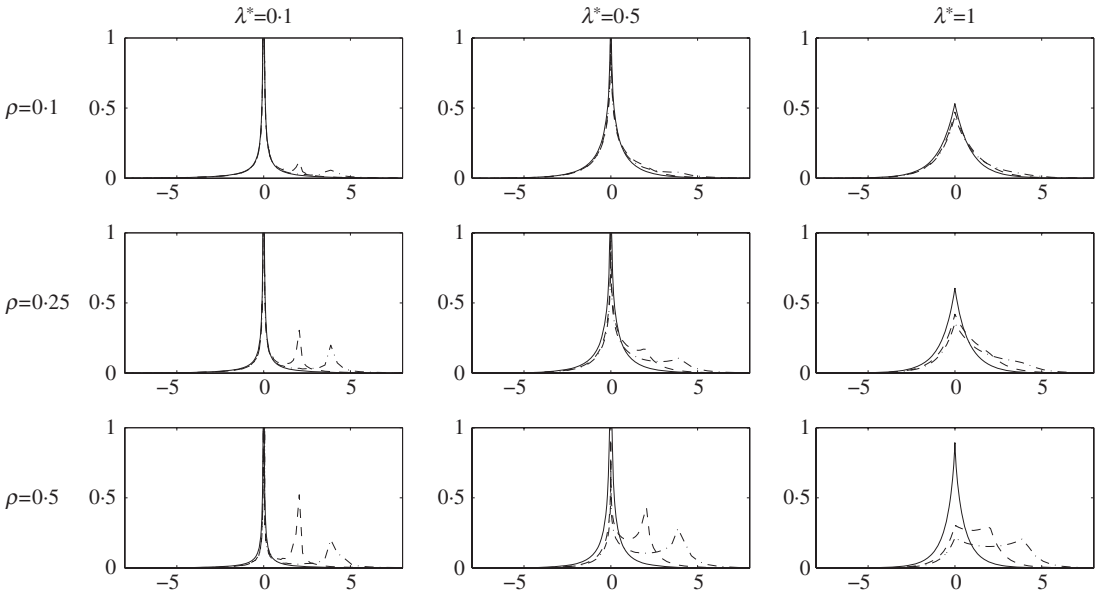

Fig. 1. The conditional density of β_1 given $\beta_2 = 0{\cdot}0001$ (solid), $\beta_2 = 2$ (dashed) and $\beta_2 = 4$ (dot-dashed) for a correlated normal-gamma prior with parameters $\lambda^\star$, ρ and $\gamma = 1/\lambda^{\star\,1/2}$.

and $\lambda = \{\rho\lambda^\star, (1-\rho)\lambda^\star, (1-\rho)\lambda^\star\}$. This parameterization implies that the marginal distributions of β_1 and β_2 are $\mathrm{NG}(\lambda^\star, \gamma)$ and the correlation between β_1 and β_2 is ρ. Figure 1 shows the conditional density of β_1 given β_2. In each case, the mass for β_1 is increased around the value of β_2. This effect becomes more pronounced as ρ increases for a fixed value of $\lambda^\star$ and as $\lambda^\star$ decreases for a fixed value of ρ. The effect is so strong that the density is bimodal when $\rho = 0{\cdot}5$ for all three chosen values of $\lambda^\star$ and when $\rho = 0{\cdot}25$ for $\lambda^\star = 0{\cdot}1$ and $\lambda^\star = 0{\cdot}5$. The implication for shrinkage of the posterior mean is computable from Griffin & Brown (2010, Lemma 1).

3. Correlated priors in regression problems

3·1. *Introduction*

The standard linear regression model assumes that an $n \times 1$ vector of observations y is related to a $n \times p$ matrix of explanatory variables X by

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

where 1 is an $(n \times 1)$-dimensional vector of 1s, β is an $(p \times 1)$-dimensional vector of regression effects and $\epsilon \sim N(0, \sigma^2 I)$. We assume that the intercept, α, and observational error variance, σ^2, have been given prior $p(\alpha, \sigma^2) \propto \sigma^{-2}$ and that these parameters are a priori independent of β. The properties of the normal-gamma priors in regression problems are discussed in Griffin & Brown (2010).

We are interested in exploring prior densities for β that have specified properties for the marginal prior of weighted sums of regression effects as well as the regression effects themselves. This will be achieved by assuming that $\beta \sim \mathrm{CNG}(\lambda, \gamma^2 B)$ and choosing appropriate values for B and λ. Let A be an $a \times p$ matrix for which $\theta = A\beta$ is an $a \times 1$ vector of weighted sums of interest. Defining $\beta \sim \mathrm{CNG}(\lambda, \gamma^2 B)$ implies that $\theta \sim \mathrm{CNG}(\lambda, \gamma^2 M)$, where $M = AB$ and the sparsity shape of the prior marginal density of θ_i is $\eta_i = \sum_{k \in S_i(M)} \lambda_k$. If we restrict attention to weighted sums that are simple differences, for which A_{ik} is either -1, 0 or 1, A_{ik} is nonzero for only two values of k for each i and $\sum_{k=1}^{q} A_{ik} = 0$, then the marginal distribution of θ_i is $\mathrm{NG}\{\eta_i, 1/(2\gamma^2)\}$.

Contrasts are weighted sums of regression effects whose coefficients sum to zero. They are important for structuring shrinkage, since shrinking a contrast to zero implies that the effects in the contrast are shrunk towards each other. For example, shrinking the contrast $\beta_1 - \beta_2$ towards zero implies that β_1 is shrunk to β_2. The shape of the prior density of the contrast controls the amount of shrinkage, with a small sparsity parameter encouraging aggressive shrinkage when the least squares estimate of the contrast is small, but leaving large estimated differences largely unaffected. We will define a non-contrast to be a weighted sum of regression effects whose coefficients sum to a nonzero value. The correlated normal-gamma prior leads to a different marginal distribution for contrasts and non-contrasts. The indices in the set $S_i(M)$ are those for which $\sum_j A_{ij} B_{jk} \neq 0$. If $B_{jk} = 1$ when $A_{ij} \neq 0$, then $\sum_j A_{ij} B_{jk} = \sum A_{ij}$, which is zero for a contrast but nonzero for a non-contrast, so the set $S_i(M)$ and hence the sparsity shape must be larger for a contrast than for a non-contrast. This implies that, for linear combinations resulting in small effects, correlated normal-gamma priors will tend to shrink contrasts more than non-contrasts.

3·2. *A general construction*

In the correlated normal-gamma prior, the forms of B and λ can be derived that lead to appropriate sparsity parameters for the elements of θ and β. Suppose that the sparsity shapes $\eta_1, \ldots, \eta_a$ are associated with the contrasts θ and that $\eta_{a+1}, \ldots, \eta_{a+p}$ are associated with the regression effects β. We allow some λ_i to be zero, as in the general definition of the correlated normal-gamma. We also define $B^{(p)}$ to be a $p \times (2^p - 1)$ matrix whose columns are the p-variate binary expansions of 2^p, omitting the zero vector. This matrix spans the full set of interactions in p dimensions. Defining $\beta \sim \mathrm{CNG}(\lambda, \gamma^2 B^{(p)})$ implies that $\theta \sim \mathrm{CNG}(\lambda, \gamma^2 M)$ where $M = AB^{(p)}$. It is useful to define a function $^+$ which indicates whether or not each entry of a matrix is 0. If Z is a $p \times q$ matrix, then Z^+ is also a $p \times q$ matrix with $Z^+_{ij} = 0$ if $Z_{ij} = 0$ and $Z^+_{ij} = 1$ otherwise. The conditions on β and θ can only be met if we can solve $R^+\lambda = \eta$ with $\lambda_k \geqslant 0$ for $k = 1, \ldots, 2^p - 1$, where

$$R^+ = \left[\begin{pmatrix} AB^{(p)} \\ B^{(p)} \end{pmatrix} \right]^+ .$$

There are potentially multiple solutions if $a + p < 2^p - 1$, but the general solution is $\lambda = (R^+)^- \eta + Ve$, where Z^- represents the Moore–Penrose inverse of Z, V are eigenvectors of $R^{+\mathrm{T}} R^+$ in the null space and e is a vector whose length is the number of such eigenvectors.

3·3. *Examples when* $p = 3$

The simplest interesting priors occur when $p = 3$. In this case,

$$B^{(3)} = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{pmatrix}.$$

We assume that $\beta_i \sim \mathrm{NG}\{\lambda^\star, 1/(2\gamma^2)\}$ for $i = 1, 2, 3$ and we consider the simple contrasts

$$A = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & -1 \end{pmatrix}.$$

If the regressors are observed over time, then we might assume that β_1 and β_2 and β_2 and β_3 are correlated but that β_1 and β_3 are uncorrelated, leading to a first-order dependence structure. Tibshirani et al. (2005) highlight several statistical problems where this type of structure would be appropriate. This would imply choosing $\eta_1 = \eta_2 = \lambda^\star$, which implies that the differences of consecutive regression effects have the same prior as the regression effects themselves, and $\eta_3 = 2\lambda^\star$, which leads to the distribution of θ_3 being $\mathrm{NG}\{2\lambda^\star, 1/(2\gamma^2)\}$, which is the distribution of the difference of two independent normal-gamma distributed random variables. The potential solutions are $\lambda = \lambda^\star[(0{\cdot}5, 0, 0{\cdot}5, 0{\cdot}5, 0, 0{\cdot}5, 0) + v(1, 1, -1, 1, -1 - 1, 1)]$ but the only solution for which all $\lambda_i \geqslant 0$ occurs

when $v = 0$. This solution for λ contains zeros and so we can write the prior in a more compressed form as $\beta \sim \mathrm{CNG}(\lambda, B)$, where

$$\lambda = \lambda^\star(0{\cdot}5, 0{\cdot}5, 0{\cdot}5, 0{\cdot}5), \quad B = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix}.$$

The prior implies that $\mathrm{corr}(\beta_i, \beta_{i+1}) = 1/2$ $(i = 1, 2)$, and equals zero otherwise, but a more general prior when $\mathrm{corr}(\beta_i, \beta_{i+1}) = \rho$ arises on taking

$$\lambda = \lambda^\star(1-\rho, 1-2\rho, \rho, 1-\rho, \rho), \quad B = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{pmatrix}.$$

An alternative prior specification would arise by setting $\eta_1 = \eta_2 = \eta_3 = \lambda^\star$, so that all the differences between regression effects are given the same prior as the regression effects. The potential solutions are $\lambda = \lambda^\star\{(3, 4, 4, 3, 4, 4, 3)/14 + v(1, 1, -1, 1, -1 - 1, 1)\}$. The solution for which all $\lambda_i \geqslant 0$ occurs when $-0{\cdot}567 < v < 0{\cdot}756$. There are three particularly interesting solutions: $\lambda^{(1)} = (\lambda^\star/2)(1, 1, 0, 1, 0, 0, 1)$, which has only idiosyncratic parts and a part shared by all three regression effects; $\lambda^{(2)} = (\lambda^\star/2)(0, 0, 1, 0, 1, 1, 0)$, which only has components shared by two regression effects; and $\lambda^{(3)} = (\lambda^\star/4)1_7$, which has nonzero λ_is for all combinations of regression effects. These solutions preserve the second-order correlations of $1/2$ for β but alter higher order dependencies. For example, the contrast $\beta_1 - 2\beta_2 + \beta_3$ has marginal density with a sparsity shape parameter $3\lambda^\star/2$ for all three cases. This will also be true for the similar contrasts $\beta_1 + \beta_2 + 2\beta_3$ and $-2\beta_1 + \beta_2 + \beta_3$. The sparsity shape parameter of the non-contrast $\beta_1 - 3\beta_2 + \beta_3$ depends on the choice of λ. The values for $\lambda^{(1)}$, $\lambda^{(2)}$ and $\lambda^{(3)}$ are $7\lambda^\star/4$, $3\lambda^\star/2$ and $2\lambda^\star$, respectively, and never smaller than $3\lambda^\star/2$.

3·4. *Examples for general p*

The priors in the previous section can easily be extended to $p > 3$. Suppose that we want serial dependence, so that β_i and β_{i+1} are correlated but β_i and β_{i+j} are uncorrelated for $j < -1$ and $j > 1$. We set $q = 2p - 1$ and define

$$\lambda_i = \begin{cases} \lambda^\star\rho & (i = 1, \ldots, p-1), \\ \lambda^\star(1-\rho) & (i = p, 2p-1), \\ \lambda^\star(1-2\rho) & (i = p+1, \ldots, 2p-2), \end{cases}$$

and

$$B_{ij} = \begin{cases} 1 & (i = 1;\ j = 1,\ j = p+2), \\ 1 & (i = 2, \ldots, p-1;\ j = i-1, i, p+1+i), \\ 1 & (i = p;\ j = p,\ j = 2p+1), \\ 0 & (\text{otherwise}), \end{cases}$$

where $\rho \leqslant 1/2$. It follows that $\beta_i \sim \mathrm{NG}\{\lambda^\star, 1/(2\gamma^2)\}$ marginally and that $\mathrm{corr}(\beta_i, \beta_{i+1}) = \rho$, $(i = 1, \ldots, p-1)$, and equals zero otherwise.

The priors with equal correlations between all regressors can also be extended. First, $\beta \sim \mathrm{CNG}(\lambda, \gamma^2 B)$ where $B = (I_p\ 1_p)$ and $\lambda = (\lambda^\star/2)1_{p+1}$ which has idiosyncratic components and one shared component. Secondly, $\beta \sim \mathrm{CNG}(\lambda, \gamma^2 B^{(p)})$ and $\lambda = (\lambda^\star/2^{p-1})1_{2^p-1}$.

3·5. *Regression with factors*

Suppose that the regressors are constructed from factors whose levels are unordered categorical variables, where the gth variable has p_g-levels. These variables may be included in a linear regression model using dummy variables and (1). We could overparameterize by including all p_g variables, but often

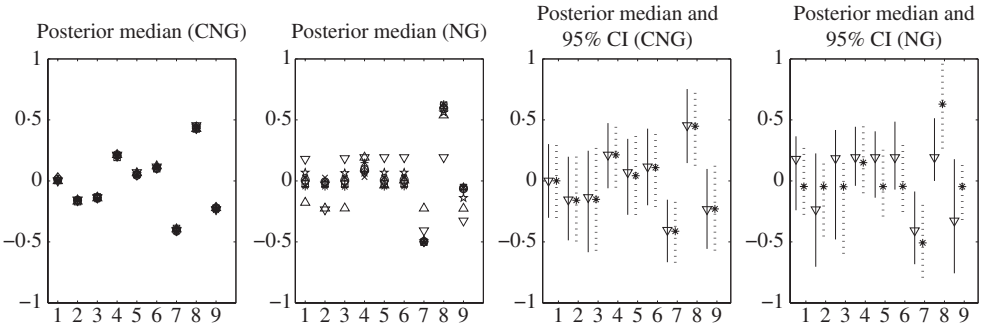

Fig. 2. Posterior medians and 95% credibility intervals, CI, for the regression coefficients using different base classes with the correlated normal-gamma, CNG, and independent normal-gamma, NG, priors. The base classes are coded 1 (circle), 2 (□), 3 (◇), 4 (pentagram), 5 (+), 6 (×), 7 (△), 8 (▽) and 9 (∗).

we omit one and parameterize relative to that one level. The linear regression is then written as

$$y = \alpha 1 + \sum_{g=1}^{G} X_g \beta_g + \epsilon,$$

where X_g is a $n \times (p_g - 1)$ matrix of dummy variables constructed by choosing the first level as a base and then defining $(X_g)_{ij} = 1$ only if the ith observation takes the $(j+1)$th level and 0 otherwise. The regression effect $\beta_{g,j}$ is interpreted as the difference between the effect of the $(j+1)$th level and the first level. Rearranging the levels allows any level to be the base level. We will assume that $\beta_1, \ldots, \beta_G$ are independent and that $\beta_i \sim \text{CNG}(\lambda_i, \gamma^2 B_i)$. The choice of B_i allows us to encourage shrinkage of the regression effects for some levels towards each other, which could be useful if the effects of level tended to cluster. For example, if this factor is categorical and there is no natural ordering so that the choice of base level is arbitrary then it would be natural to assume that the regression effects β_g would be invariant to permutations of the levels and also invariant to the choice of the base factor level. This can be achieved by assuming that β_g follows a $\text{CNG}(\lambda_g, \gamma^2 B^{(p)})$ prior where $\lambda_{g,i} = \lambda^\star / 2^{p_g - 1}$. It follows that the marginal prior for the regression effects are $\beta_{g,i} \sim \text{NG}\{\lambda^\star, 1/(2\gamma^2)\}$ and also that $\beta_{g,i} - \beta_{g,j} \sim \text{NG}\{\lambda^\star, 1/(2\gamma^2)\}$. The correlation between $\beta_{g,i}$ and $\beta_{g,j}$ equals 0·5 for $i, j = 1, \ldots, p_g$. If we were to change the base class to $j^\star$ then $\beta_{g,i} - \beta_{g,j^\star} = (\beta_{g,i} - \beta_{g,1}) - (\beta_{g,j^\star} - \beta_{g,1}) \sim \text{NG}\{\lambda^\star, 1/(2\gamma^2)\}$, as for base level j.

Direct application of shrinkage methods would potentially shrink the effect of some factor levels to zero whilst not shrinking the effect of other levels of that factor. This might be considered undesirable, since variable selection would usually be performed on the factor rather than the levels of the factor. The block structure avoids this problem.

4. EXAMPLE

To illustrate the use of the prior for grouped variables, we analyse data regressing the average score of graduating high-school students on the verbal component of SAT test in each state of the U.S.A. (Moore, 1995). The regressors are population of the state, percentage of graduating high-school students who took the SAT exam, state spending on public education per student, average teacher's salary in the state, with the U.S. Census region providing the grouping: East North Central (1), East South Central (2), Mid-Atlantic (3), Mountain (4), New England (5), Pacific (6), South Atlantic (7), West North Central (8) and West South Central (9). We found that it was necessary to take logs of the continuous regressors and the response, which were subsequently centred and scaled to unit variance. The results of fitting a linear regression model to the data with different base classes using the correlated normal-gamma prior and an independent normal-gamma prior are shown in Fig. 2. The effects of the continuous regressors are excluded but are similar across all priors and all choices of base class.

The results for different base classes are very similar for the correlated normal-gamma prior but are much more varied for independent normal-gamma priors. This illustrates the effect that the arbitrary choice of base level can have with shrinkage priors. Figure 2 shows that these differences can be quite substantial. For example, under independent normal-gamma priors, the 95% credible interval of the effect of regions 8 with 9 as the base class does not contain the posterior median of the effect with region 8 as the base class. These results would suggest that there is either little difference between regions 8 and 1 or that there is evidence of a difference. The correlated normal-gamma prior gives very similar credibility intervals, as we would expect.

5. Discussion

The priors discussed in this paper have the potential to shrink marginal regression effects and differences between them. Raman et al. (2009) and Kyung et al. (2010) discuss how the group lasso (Yuan & Lin, 2006) can be interpreted as a hierarchical prior for the regression effects where groups of effects are given a multivariate normal prior whose common scale has a gamma prior. This construction is markedly different from that described in this paper for grouped effects. The Bayesian interpretation of the group lasso shrinks groups of variables to zero through shrinking the scale of the conditional normal. On the other hand, the correlated normal-gamma prior simultaneously shrinks marginal effects and differences and so allows for the clustering of regression effects in a group.

Acknowledgement

We wish to thank two referees, an associate editor and the editor for searching comments that lead to substantial improvement.

References

Carvalho, C., Polson, N. & Scott, J. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–80.

Fan, J. & Li, R. Z. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *J. Am. Statist. Assoc.* **96**, 1348–60.

Griffin, J. E. & Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.* **5**, 171–88.

Hans, C. (2009). Bayesian lasso regression. *Biometrika* **96**, 835–45.

Kyung, M., Gill, J., Ghosh, M. & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal.* **5**, 369–412.

Mitchell, T. J. & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression (with discussion). *J. Am. Statist. Assoc.* **83**, 1023–36.

Moore, D. (1995). *The Basic Practice of Statistics*. New York: Freeman.

Moschopoulos, P. G. (1985). The distribution of the sums of independent gamma random variables. *Ann. Inst. Statist. Math.* **37**, 541–4.

Park, T. & Casella, G. (2008). The Bayesian lasso. *J. Am. Statist. Assoc.* **103**, 672–80.

Polson, N. G. & Scott, J. G. (2011). Shrink globally, act locally: sparse Bayesian regularization and prediction. In *Bayesian Statistics 9*, Ed. M. J. Bernardo J. M., Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith & M. West, pp. 501–38. Oxford: Clarendon Press.

Raman, S., Fuchs, T., Wild, P., Dahl, E. & Roth, V. (2009). The Bayesian group-lasso for analyzing contingency tables. In *Proc. 26th Int. Conf. Mach. Learn.*, 881–8. Ed. L. Bottou & M. Littman. Montreal: Omnipress.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

Tibshirani, R. J., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *J. R. Statist. Soc.* B **67**, 91–108.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B **68**, 49–67.

[*Received September* 2010. *Revised October* 2011]