Bayesian Analysis (2013) **8**, Number 3, pp. 691–702

# Some Priors for Sparse Regression Modelling

Jim E. Griffin[*] and Philip. J. Brown[†]

**Abstract.** A wide range of methods, Bayesian and others, tackle regression when there are many variables. In the Bayesian context, the prior is constructed to reflect ideas of variable selection and to encourage appropriate shrinkage. The prior needs to be reasonably robust to different signal to noise structures. Two simple evergreen prior constructions stem from ridge regression on the one hand and g-priors on the other. We seek to embed recent ideas about sparsity of the regression coefficients and robustness into these priors. We also explore the gains that can be expected from these differing approaches.

**Keywords:** Correlated priors; Canonical reduction; Multiple regression; g-priors; Markov chain Monte Carlo; Normal-Gamma prior; $p > n$; Ridge regression; Robust priors; Sparsity

## 1 Introduction

High dimensional problems are the norm rather than the exception with modern measuring instruments in many branches of science, technology and social science. There are many and varied approaches to dealing with the problems that arise with the multitude of parameters in commonly used models for such data, whether it be through forms of regularisation, penalisation or variable selection, in machine learning, classical or Bayesian statistics.

In the Bayesian literature, ideas of sparsity and robustness of the prior distribution have emerged, see for example Polson and Scott (2012). Sparsity is the assumption that many parameters can be set to values very close to zero without substantially affecting the fit of the model. Robustness is a property of the fatness of the tails of the prior and it is an open question in high dimensional problems with very flat likelihoods as to how heavy these tails should be. In the full Bayes context, we generally prefer modestly fat exponential tails of the normal mixed by a gamma distribution on the variance (Griffin and Brown 2010) to polynomial tails. We will use the computationally convenient scale mixture of normals family for each regression coefficients $\beta_i$, so that $\beta_i \sim \mathrm{N}(0, \psi_i)$, $\psi_i \sim G$, $i = 1, \ldots, p$, with $G$ a general mixing distribution. In this class, sparsity can be defined through the degree of spikiness of the prior of $\beta_i$, or equivalently $\psi_i$, at zero. In general, if the prior density of $\psi_i$ is $\pi(\psi_i) \propto \psi_i^{v_i - 1} f(\psi_i)$ close to zero (where $f(\psi_i) \to m$ as $\psi_i \to 0$) then we will refer to $v_i$ as the sparsity shape parameter, which in the case of Gamma$(\lambda, \theta)$ mixing is just $\lambda$, the shape parameter of the gamma (scale $\theta$, expectation $\lambda/\theta$). Smaller values of $v_i$ favour values of the

[*] School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, UK, J.E.Griffin-28@kent.ac.uk

[†] School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, UK, Philip.J.Brown@kent.ac.uk

DOI:10.1214/13-BA827

regression coefficients with greater sparsity. The idea is made precise by Carvalho et al. (2010) and Polson and Scott (2011) in terms of super-efficiency. Priors for parameters have generally assumed that parameters of interest are *a priori* independent. Recently however in the context of the normal-gamma mixtures Griffin and Brown (2012) have defined a $p$-variate multivariate correlated normal–gamma distribution for the $(p \times 1)$-dimensional vector $\beta$ of regression coefficients. This is referred to as $\mathrm{CNG}(\lambda, C)$ which relies on $q$ linear combinations of latent iid normal–gamma$(\lambda_j, 1/2)$ random variables, $j = 1, \dots, q$. The hyperparameter $C$ is a $(p \times q)$-dimensional matrix and the resultant covariance matrix of $\beta$ is the $(p \times p)$-dimensional matrix $2C\mathrm{Diag}(\lambda)C^T$, where $\mathrm{Diag}(\lambda)$ is a $(q \times q)$-dimensional diagonal matrix with $\lambda_1, \dots, \lambda_q$ on the diagonal. This was primarily defined to cope with grouping problems and factors in regression and analysis of variance, but we will employ it for a different purpose to robustify and combine two strands of early popular prior formulations, ridge and g-prior. This is the main purpose of our paper. A secondary but very practical purpose is to demonstrate that for many datasets simple prior structures, such as that appropriate to ridge regression are still very effective, despite, as we demonstrate, showing occasional deficiency, see also Fearn (1983).

## 2 Two old standards: ridge and g-priors

The multiple regression model assumes that

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

where $X$ is an $(n \times p)$-dimensional matrix of centred regressors and $\epsilon$ is normally distributed with mean zero and $\mathrm{COV}(\epsilon) = \sigma^2 I_n$. A common practice in practical regression problems is to scale the regressors so that each column of $X$ has unit sample variance. Scaling is important as it can adjust for different scales of measurement for the regressors, but it can also be important not to scale for co-measureable explanatory variables (that is variables measured on the same scale). If the explanatory variables are scaled, there is the danger that noisy low signals become overly influential. When it comes to analysing co-measureable explanatory variables later we will merely centre each of the explanatory variables by subtracting their mean.

Ridge regression was originally proposed on the basis of evidence of mean squared error improvements in ill-conditioned problems. Such improvements can be illusory if the informative directions of $\beta$ coincide with the ill-conditioned directions, see Casella (1980) for comprehensive minimax results. The Bayesian formulation of ridge regression corresponds to a spherical normal prior on the regression coefficients. The ridge prior is taken to be $\beta \sim N_p[0, (\sigma^2/\tau)I_p]$, where $\tau$ would then be the ridge constant which is added to the diagonal of the $X^TX$ matrix in standard ridge regression, Hoerl and Kennard (1970). The hyperparameter $\tau$ can then in turn be assigned a prior distribution or estimated by some form of cross-validation in a more classical context. It has often proved to be hard to beat for prediction in practice, whether because of or despite its simple roots. For example, several studies have found that ridge regression is competitive to other penalized maximum likelihood procedures such as the Lasso, elastic net, and adaptive Lasso (*e.g.* Waldron et al. 2011; Bøvelstad et al. 2007; Ogutu et al. 2012)

A radically different prior in terms of its relationship between the regression coefficients is provided by the conjugate *g-prior* of Zellner (1986). This is largely used with a different purpose in mind, that of hypothesis testing via Bayes factors. The corresponding elliptical prior is $\beta \sim N_p(0, g\sigma^2(X^TX)^{-1})$. This is data dependent in the ancillary sense that it depends on the design matrix $X$.

Reducing the regression model to a canonical form more clearly reveals the differences between the two priors and their resulting posterior expectation. In general we consider $p > n$ problems where the rank of $X$ is $r \leq \min(n-1, p)$ with columns orthogonal to the unit vector (by construction through mean subtraction). In this case, we can write the singular value decomposition (SVD) of $X = TDV^T$ where $T$ is an $(n \times n)$-dimensional orthonormal matrix, $D$ is an $n \times r$ matrix of zeros apart from $r$ singular values $s = (s_1, \ldots, s_r)$ on the principal diagonal, and $V$ is a $(p \times r)$-dimensional matrix for which $V^TV = I_r$ and which is completed by a $(p \times (p-r))$-dimensional matrix $\bar{V}$ to form an orthonormal $p \times p$ matrix. The squares of the singular values are eigenvalues of $X^TX$.

The model (1) becomes the canonical model

$$U_i = \begin{cases} \sqrt{n}\mu + \epsilon_i & i = 0 \\ s_i\alpha_i + \epsilon_i, & i = 1, \ldots, r \\ \epsilon_i, & i = r+1, \ldots, n-1. \end{cases}$$

with the $(n \times 1)$-dimensional vector $U = T^Ty$ and the $(r \times 1)$-dimensional vector $\alpha = V^T\beta$. The remaining $((p-r) \times 1)$-dimensional parameter vector $\bar{\alpha} = \bar{V}^T\beta$ is unidentified. The corresponding prior distributions transform to

- *Ridge prior* $\alpha_j \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2/\tau), j = 1, \ldots, r$; the $p-r$ remaining canonical parameters $\bar{\alpha}$ having the same zero mean spherical normal prior.
- *g-prior* $\alpha_j \overset{\text{ind}}{\sim} \text{N}(0, \sigma^2 s_j^{-2} g), j = 1, \ldots, r$. Strictly in this *g-prior*, the complementary $(p-r) \times 1$ parameters $\bar{\alpha}$ in the null space of $X$ are undefined although the form of model dependence in the identified canonical model would suggest infinite variance.

In both cases the intercept is given the vague prior $p(\mu) \propto$ const which leads to the posterior distribution of $\mu$ being $N(\bar{y}, \sigma^2/n)$. The least squares estimate of $\alpha_j$ is $\hat{\alpha}_j^{LS} = \frac{U_j}{s_j}$ for $j = 1, \ldots, r$. The posterior distributions of $\alpha$ are:

- *Ridge posterior* $\alpha|y$ has a normal distribution with mean $\hat{\alpha}_j^{ridge} = \frac{s_j^2}{s_j^2+\tau}\hat{\alpha}_j^{LS}$, and variance $\sigma^2/(s_j^2+\tau)$, $j = 1, \ldots, r$. The posterior for parameters $\bar{\alpha}$ which relate to the null space of $X$ are unchanged from the prior.
- *g-prior posterior* $\alpha|y$ has a normal distribution with mean $\hat{\alpha}_j^{gprior} = \frac{g}{1+g}\hat{\alpha}_j^{LS}$ and variance $\frac{g}{1+g}\sigma^2/s_j^2$, $j = 1, \ldots, r$. Again the parameter $\bar{\alpha}$ is unchanged from its

corresponding prior specification.

Estimation under the g-prior via the posterior mean may give poorly estimated regression coefficients when $X$ is nearly singular - the posterior mean shrinks $\hat{\alpha}_j^{LS}$ by the same amount whether data information, $s_j^2$, is great or small, and does not shrink to zero in the ill-conditioned case (as $s_j^2 \to 0$) when the least squares estimator uncertainty gets very large in those same near singular directions. On the other hand the ridge prior has no difficulty shrinking to zero as eigenvalues, $s_j^2$, become small. Thus as an estimator the g-prior provides little regularisation and has highly inflated variance with well-known implications for mean squared error.

A variety of priors have been suggested for the hyperparameter $g$, see Liang et al. (2008). More recently Maruyama and George (2011), influenced by Casella (1980), note that shrinkage of regression coefficients should not be great when least squares variance is small, rather when variance is large. This is the opposite to the effect induced by the standard $g-$prior as we have noted above. Maruyama and George have modified the $g-$prior in this direction, allowing more shrinkage of coefficients corresponding to small eigenvalues. They have also chosen a fat-tailed prior for $g$, and considered modifications for singular $X$. For the issue of singular $X$ they propose a completion of the singular prior which is spherical with the last and smallest non-zero eigenvalue. These beneficial moves retain the appeal of the g-prior in giving explicit formulae for Bayes factors used in hypothesis testing in variable selection. There remains the issue of robustness of the normal prior and the potential for instability due to the decision of what constitutes a small eigenvalue for the purpose of setting the rank.

Our approach, however, is different. Casella's observation was within the framework of normal priors when variance is wholly meaningful and his worries very apt. Putting a prior on $g$, however fat-tailed, still retains this normal tail behaviour *between* the regression coefficients, that is conditional on $g$. Our priors rather involve independent fattish-tailed normal-gamma priors, one for each estimable canonical regression coefficient. For the normal-gamma prior there is more than just variance to consider. We will show that even though the prior variance on coefficients tends to infinity corresponding to small eigenvalues, the prior becomes increasingly concentrated at zero, and hence can still provide total shrinkage to zero of the corresponding regression coefficient.

# 3 Eigen-CNG prior

We will use the normal–gamma mixture distribution as our basic ingredient in forming a robustified version of both the standard ridge prior and g-prior which incorporates both sparsity and shrinkage. Its robustness stems from its exponential tails. To do this we will exploit the correlated normal gamma (CNG) of Griffin and Brown (2012). The correlated normal-gamma prior for a $p$-dimensional parameter $\beta$ is defined by setting $\beta = C\phi$ where $C$ is a fixed $(p \times q)$-dimensional matrix and $\phi = (\phi_1, \phi_2, \dots, \phi_q)$ are independent normal-gamma random variables with shapes $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_q)$ and common scale $\frac{1}{2}$. This prior will be written $\beta \sim \mathrm{CNG}(\lambda, C)$. Griffin and Brown (2012) show that the sparsity shape parameter for $\beta_i$ is $\sum_{\{j|C_{ij}\neq 0\}} \lambda_j$. This controls the ability of the prior to encourage the posterior distribution to be shrunk very close to zero if the data allows. Generalizations of the ridge prior and $g$-prior arise from choosing $C = VA$ where $V$ arises from the SVD of $X$ and $A$ is a diagonal $(r \times r)$-dimensional matrix. It follows that $\alpha = V^T\beta$.

An analogue of the ridge prior uses $\beta \sim \text{CNG}(\lambda 1_r, \gamma V)$. This choice implies that the covariance matrix of $\beta$ is $2\lambda\gamma^2\text{Diag}(e)$, with $e^T = (1_r^T, 0_{p-r}^T)$, which is singular when the rank of $X$ is less that $p$, the number of regression coefficients. A simple and natural extension to a full non-singular $p$-dimensional prior can be defined by replacing zeros by ones in Diag(.) but is not needed for our purposes. To match the parametrization of the normal ridge prior we would take $2\lambda\gamma^2 = \sigma^2/\tau$.

The $g$-prior in the canonical model introduces dependence on the singular values in the prior. A very general prior with dependence on the singular values is

$$\beta \sim \text{CNG}(\lambda s_b, \gamma V S_{k+b/2}) \tag{2}$$

where $s_k = (s_1^k, \ldots, s_r^k)$, $(r \times r)$ matrix $S_k = \text{Diag}(s_{-k})$ and $b \geq 0$ is a hyperparameter. Here the singular values appear in the both arguments of the CNG prior. The covariance matrix of $\beta$ is $2\lambda\gamma^2 V S_k^2 V^T$. The dependence on the singular values of the sparsity shape parameter of the underlying normal-gamma random variables allows us to more aggressively regularise latent parameter directions corresponding to very small values of $s_j$. Indeed these same latent parameters will be set to zero for small singular value directions unless the likelihood suggests otherwise, as is shown at the end of this section.

Equivalently in the canonical reduction, $\alpha = V^T\beta$, the $\alpha_1, \ldots, \alpha_r$ are independent and

$$\alpha_i \sim \text{NG}(\lambda s_i^b, 1/(2\gamma^2 s_i^{-b-2k}))$$

or, expressed hierarchically,

$$\alpha_i \sim \text{N}(0, \phi_i)$$

and

$$\phi_i \sim \text{Gamma }(\lambda s_i^b, 1/(2\gamma^2 s_i^{-b-2k})). \tag{3}$$

The expectation of $\phi_i$ (which is the variance of $\alpha_i$) is $2\lambda\gamma^2 s_i^{-2k}$ which is unaffected by the $b$-terms. The $b$ term in the shape gives sparsity in the prior when needed.

The general prior in (2) has some interesting special cases:

- *Ridge-CNG.* The analogue to the ridge prior which was defined previously arises when $k = 0$ and $b = 0$, and is identified as a robust ridge which we will call ridge-CNG.

- *g-prior-CNG.* A natural analogue to the $g$-prior arises when $k = 1$ and $b > 0$ for $p > n - 1$ and is identified as a robust g-prior which we will call g-prior-CNG.

- *Combined full prior.* A compromise between these two priors arises when $k \sim \text{U}(0, 1)$ and $b \sim \text{U}(0, 1)$ and is identified as a compromise between our robust ridge and robust g-prior which will be called eigen-CNG.

Singular values in the shape of NG imply very small singular values will have their coefficients shrunk to zero adapting the g-prior to $p > n-1$. As we have said robustified ridge does not need this compensating adjustment. The use of singular values, $s_j^b$, in the shape of the gamma is neutral as far as covariance of $\beta$ is concerned as the $s_j^b$ of the shape is cancelled by the same term in the scale of the gamma random variables. It does however give high sparsity when $s_j$ is small so as to shrink coefficients to zero should the likelihood not deem otherwise. It thus enables division by singular values to be accomplished smoothly as $s_j \to 0$, with resulting coefficients set to zero.

We now show formally that the vanishing shape parameter acts as an antidote to the variance blowing-up in the canonically reduced model (3). If $b \neq 0$

$$p(\phi_i < \epsilon) = \int_0^{\epsilon} p(\phi_i) = \int_0^{2\gamma^2 s_i^{-(b+k)}\epsilon} \frac{1}{\Gamma(\lambda s_i^b)} \phi_i^{\lambda s_i^b} \exp\left\{-\phi_i\right\} \, d\phi_i$$

using the incomplete gamma function series expansion, see the result 6.5.29 of Abramowitz and Stegun (1964). This can be expressed as

$$\sum_{n=0}^{\infty} a_n$$

where

$$a_n = \frac{\left(1/2\gamma^{-2} s_i^{b+k}\epsilon\right)^{n+\lambda s_i^b}}{\Gamma(\lambda s_i^b + n + 1)} \exp\left\{-1/2\gamma^{-2} s_i^{b+k}\epsilon\right\}$$

and

$$\begin{aligned} \log a_n &= \left(n + \lambda s_i^b\right) \log\left(1/2\gamma^{-2}\epsilon\right) + n \log\left(s_i^{b+k}\right) \\ &+ \lambda s_i^b \log\left(s_i^{b+k}\right) - \log\Gamma(\lambda s_i^b + n + 1) - 1/2\gamma^{-2} s_i^{b+k}\epsilon. \end{aligned} \tag{4}$$

Clearly, as $s_i \to 0$, $\left(n + \lambda s_i^b\right) \to n$, $\log\left(s_i^{b+k}\right) \to -\infty$, $\lambda s_i^b \log\left(s_i^{b+k}\right) \to 0$, $\log\Gamma(\lambda s_i^b + n + 1) \to \log\Gamma(n+1)$ and $s_i^{b+k}\epsilon \to 0$.

It follows that

$$a_n \to \begin{cases} 1 & \text{if } n = 0 \\ c\, a^n & \text{if } n > 0, \end{cases}$$

with $0 < a < 1$ for suitably small $s_i$ and so

$$p(\phi_i < \epsilon) \to 1 \text{ as } s_i \to 0.$$

Clearly, $\mathrm{E}[\phi_i] = 2\lambda\gamma^2 s_i^{-2k}$ and so $\mathrm{E}[\phi_i] \to \infty$ as $s_i \to 0$. Consequently, $\phi_i$ has a distribution with a mean which goes to infinity as $s_i \to 0$ but whose mass is increasingly concentrated around 0 (with that mass limiting to 1). This is, in turn, a consequence of making the shape parameter $\lambda s_i^b$ become smaller as $s_i$ decreases to 0. This expresses a prior belief that as $s_i$ decreases, $\alpha_i$ is increasingly badly estimated, and so should be increasingly penalized. The construction of this prior imposes this penalization in a particular way through the shape parameter rather than the scale parameter of the normal-gamma prior.

## 4 Setting hyperparameters and computation

Firstly the intercept $\mu$ is given a vague (uniform on the real line) prior. The regression variance $\sigma^2$ is given the scale-invariant prior $p(\sigma^2) \propto \sigma^{-2}$.

There are two other hyperparameters, $\lambda, \gamma$, of the normal-gamma prior related by the prior mean of the normal-gamma being $\nu = 2\lambda\gamma^2$. As in Griffin and Brown (2010) we first give a prior for the shape parameter $\lambda$ and then for the derived parameter $\nu$ conditional on $\lambda$. The shape parameter $\lambda$ is given an inverted Beta prior distribution with density

$$p(\lambda) = 2(\sqrt{2}-1)(1+\lambda(\sqrt{2}-1))^{-3},$$

see for example equation (7-28) of Raiffa and Schlaifer (1961) with $p = 1, q = 2, b = 1/(\sqrt{2}-1)$, which has a prior median of 1, a mean of $(1/(\sqrt{2}-1)$, with higher moments not existing, and is positively skewed with a long right-hand tail. This prior offers more diffuseness than the chosen exponential prior in Griffin and Brown (2010) but is similarly anchored around the lasso prior of $\lambda = 1$. Again in similar fashion to Griffin and Brown (2010) we anchor the prior for $\nu$ empirically. The least squares estimate $\hat{\alpha}_j^{LS}$ of $\alpha_j$ is $U_j/s_j$ and its variance is $\sigma^2/s_j^2$. Under the prior $\mathrm{E}[\alpha_j^2] = \nu s_j^{-2k}$ and so a simple estimate of $\nu$ would be $(\hat{\alpha}_j^{LS})^2 s_j^{2k} = U_j^2 s_j^{2k-2}$. These estimates are combined as a weighted sum where the weights are inversely proportional to the variance of $\hat{\alpha}_j^{LS} s_j^k$. It is straightforward to show that the $j$-th weight is $w_j \propto \sigma^{-2} s_j^{2-2k}$ and so the weighted sum is $M = \frac{\sum_{j=1}^r w_j U_j^2 s_j^{2k-2}}{\sum_{j=1}^r w_j} = \frac{\sum_{j=1}^r U_j^2}{\sum_{j=1}^r s_j^{2-2k}}$. It is then assumed that $\nu$ has the prior density

$$p(\nu) = \nu^{-3} \exp\{-M/\nu\}$$

which is an inverted Gamma distribution with mean $M$. The parameters $k$ and $b$, when they appear, are given uniform distributions on $(0, 1)$. Inference in the linear regression models with these priors can be easily performed using Markov chain Monte Carlo (MCMC) methods. The model can be written in terms of $U_i|\alpha_i$ where $\alpha_i$ are given independent normal-gamma prior distributions and the appropriate MCMC methods are described in Griffin and Brown (2010).

## 5 Examples

We repeated the simulation study of Polson and Scott (2012), who give full details of implementation and availability of their six real data sets with more parameters than observations, $3 < p/n < 10$. We have also added 3 more data sets where more conventionally $p < n$. They are:

*CPS* - A cross-section of the May 1985 Current Population Survey by the US Census Bureau. Available in the AER package in R. Here there are $n = 334$ observations on $p = 16$ variables.

*Crime* - See Vandaele (1978) as used in Raftery et al. (1997), relating to crime rates in 1960 in $n = 47$ states with $p = 13$ explanatory variables.

*Ozone* - as used in Breiman and Friedman (1985) with $n = 330$, and $p = 8$ meteorlogical variables. It is available in 'gclus' in R.

## 5.1 Out-of-sample prediction

The performance of each algorithm is compared using Log Predictive Score (LPS) (Good 1952) which is less focussed on individual outliers than mean square error and gives a more balanced view of the predictive distribution. We randomly split the data into 50 pairs of testing and training samples where the training sample is approximately 75% of the full training set. Let $(X_i^{train}, y_i^{train})$ and $(X_i^{test}, y_i^{test})$ be the $i$-th training and testing samples respectively. The LPS is

$$\mathrm{LPS} = -\frac{1}{50}\frac{1}{N^{test}}\sum_{i=1}^{50}\sum_{j=1}^{N^{test}} \log p\left(y_{i,j}^{test} \,\middle|\, X_{i,j}^{test}, y_{i,j}^{train}, X_{i,j}^{train}\right)$$

where $N^{test}$ is the number of observations in each testing sample. Table 1 shows the LPS for various data sets. As well as the CNG priors, we used a standard normal ridge and g-prior and the ridge-based horseshoe (Ridge-HS) of (Carvalho et al. 2010) which gives independent horseshoe priors to the $\alpha_i$'s. The results indicate that the distributional assumptions in the prior play an important role. In the case where $p > n$, the sparse prior (CNG and horseshoe) outperform the normal versions with the exception of the cereal data. There are differences between the sparse methods but no clear order in terms of performance. The normal-based ridge prior performs much better when there are more observations than regressors, although the differences are generally small in this case. From a practical viewpoint one can judge differences of log predictive scores in a similar way to log Bayes factors: doubling them and referring to the sort of guidelines in Kass and Raftery (1995) with natural logs. From this standpoint ridge is generally competitive (twice differences less than 3) in all cases apart from the Yarn data, which we look at in more detail in the next section. A similar general conclusion might be adduced from Polson and Scott (2012) when comparing the use of the horseshoe prior

| Data Set | ridge-CNG | g-prior-CNG | eigen-CNG | Ridge-HS | Ridge | g-prior |
|---|---|---|---|---|---|---|
| Gasoline | 0.0046 | 0.0173 | -0.0209 | **-0.0273** | -0.0136 | – |
| Yarn | 2.19 | 0.39 | 0.80 | **0.25** | 3.23 | – |
| Cereal | 3.16 | 3.25 | 3.10 | 3.14 | **2.93** | – |
| N-mouse | **3.52** | 3.61 | 3.52 | 4.01 | 3.78 | – |
| M-drug | 2.76 | **2.76** | 2.76 | 2.92 | 3.90 | – |
| Liver | 3.69 | 3.73 | **3.68** | **3.68** | 4.87 | – |
| CPS | 0.62 | 0.66 | 0.62 | 0.78 | **0.61** | $\approx 10^{16}$ |
| Crime | 7.20 | 7.22 | 7.17 | 7.12 | **7.04** | 7.07 |
| Ozone | 2.84 | 2.83 | 2.83 | **2.82** | **2.82** | 2.87 |

Table 1: Log predictive score (LPS) for five competing methods on nine different data sets. The lowest LPS for each data set are bold.

with ridge regression in the case of mean square errors of prediction. The straight g-prior can come seriously unstuck though as with the CPS data when some variables are highly correlated. The g-prior is strictly undefined when $X$ is singular as with the six datasets where $p > n$ when at least $p - n + 1$ singular values are exactly zero. A generalised inverse could be used but would depend on the implementation and the numerical treatment of 'zero' and would anyway still leave some very small eigenvalues: we have therefore omitted any results for this case.

### 5.2 Yarn data

We focus on the *Yarn* dataset comprising 28 samples of Positive Emission Tomography (PET) yarns where density of yarn is regressed on the Near Infrared (NIR) spectrum of 268 wavelengths. The data set is available as part of the chemometrics package in R. This is the data set which showed the worst standard ridge regression (and other chemometric methods) as shown in the simulation study of Polson and Scott (2012), and backed up by our study.

A typical training sample fraction has 20 singular values $n-1 = 20$ for the 75% training fraction and non-zero eigenvalues of $X^TX$ 11.4543, 10.8525, 1.3581, $\ldots$, 0.0144, 0.0122, a condition number of almost 1000.

The 50 random splits give twice the LPS difference for Ridge versus both the eigen-CNG and g-prior-CNG as greater than 3, a positive improvement on figures for standard ridge regression. To explore why this might be we plotted both estimated regression components $\beta$ and canonical components $\alpha$ in Figure 1. The pictures for both eigen-CNG and ridge-CNG also display 95% credibility intervals. The canonical graphs on the right are particularly revealing and show that there are very real effects at the 8th smallest eigenvalue and even at smaller eigenvalues where the x-axis is ordered by these eigenvalues from largest to smallest. Standard ridge on the other hand, despite having a leave-one-out cross-validatory choice of ridge constant $\tau$, evidently shrinks out components corresponding to small eigenvalues far too much. A related point was made by Fearn (1983).

## 6 Discussion

We have developed a robust Bayesian prior for regression which incorporates features of both ridge and g-prior regression, flexibly weighting between these. In the $p > n - 1$ case the prior over the full parameter space is singular and thus shrinks out components where there is no information. This is essential if we wish to construct priors with dependence on the design matrix and so have similar structure to the g-prior. A simple non-singular extension of the Ridge-CNG is easy to make since the prior is spherical and can be extended to unidentified components, even though this would have no effect on prediction mean square error. A full non-singular extension incorporating the g-prior is possible but perhaps less natural.

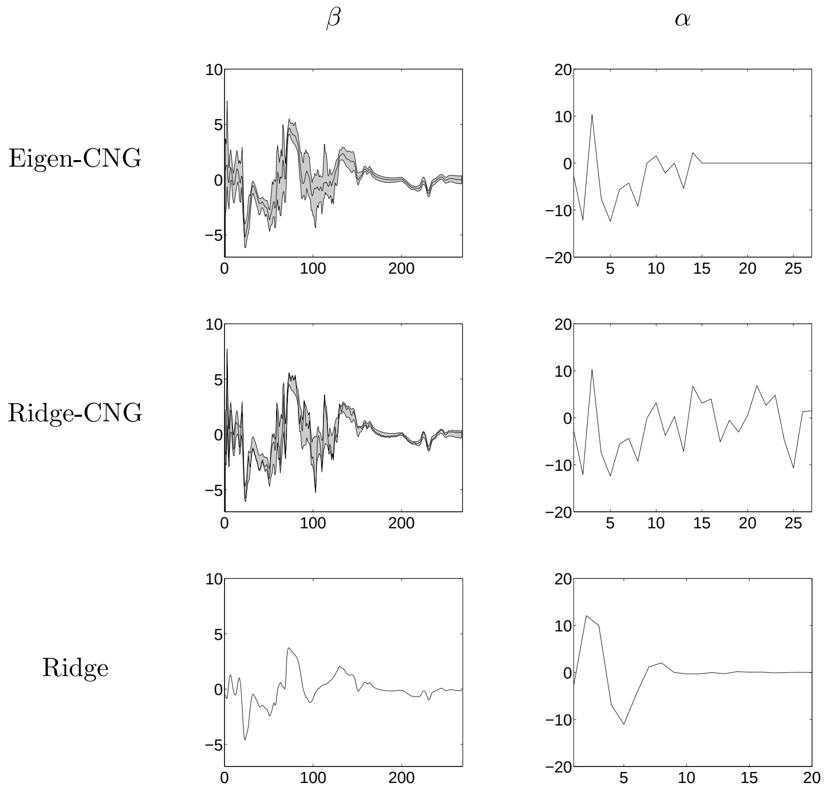

Figure 1: Yarn data: The posterior median of $\beta$ (with 95% credible interval) and $\alpha$.

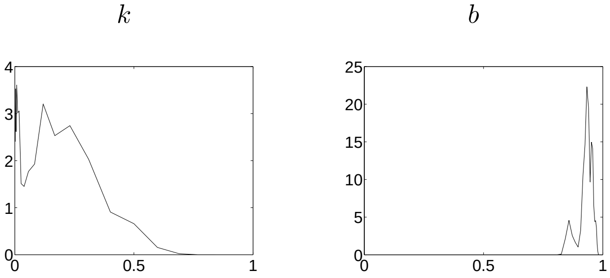

Figure 2: Yarn data: posterior densities of $k$ and $b$ with the full prior.

The Eigen-CNG embeds NG random effects giving semi-heavy tails. As noted small sparsity shapes offer strong potential shrinkage since there is heavy concentration in the prior on the parameter being zero. With semi-heavy tails despite this concentration at zero it will remain robust to strong real effects in the likelihood.

Our aim here was not to promote any one prior over any other and we have not tried to bring in other prior distributions, it is rather to show how the g-prior can be modified continuously to meet sparse settings and in so doing link with a robust version of ridge regression.

One aspect of our parameterisation is worth further comment and relates to the sparsity of the CNG. The matrices weighting the latent components in the CNG induce sparsity that is the aggregate of the normal gamma sparsities included in the linear compound. This has the unfortunate aspect of decreasing overall sparsity even though component sparsities may be getting very small. Within the flexible formulation this is automatically allowed for through the prior on overall gamma shape $\lambda$. However a more explicit allowance for this could be achieved by modified parameterisation dividing shape and multiplying scale by $\min(p, n-1)$ in the general Eigen-CNG prior. This preserves the covariance structure of the CNG whilst explicitly neutralising the accumulative nature of its sparsity, knowing that small values of sparsity are crucial for effective robust shrinkage. What evidence we have suggests that such a modification would make little difference though.

Standard ridge regression remains an effective method for prediction with many variables which is hard to beat in many circumstances. We have demonstrated however that it has an Achilles heel which implies it may do badly in certain adverse conditions which are not so unusual, namely sometimes in the more parameters than observations setting.

## References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. Dover. 696

Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigressi, A., and Lingjærde, O. C. (2007). "Predictive survival from microarray data – a comparative study." *Bioinformatics*, 23: 2080–2087. 693

Breiman, L. and Friedman, J. H. (1985). "Estimating optimal transformations for multiple regression and correlation." *Journal of the American Statistical Association*, 80: 580–598. 698

Carvalho, C., Polson, N., and Scott, J. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97: 465–480. 692, 698

Casella, G. (1980). "Minimax ridge regression estimation." *Annals of Statistics*, 8: 1036–1056. 692, 694

Fearn, T. (1983). "A misuse of ridge regression in the calibration of a near infrared reflectance instrument." *Journal of the Royal Statistical Society C: Applied Statistics*, 32: 73–79. 692, 699

Good, I. J. (1952). "Rational Decisions." *Journal of the Royal Statistical Society B*, 14: 107–114. 698

Griffin, J. E. and Brown, P. J. (2010). "Inference with Normal-Gamma prior distributions in regression problems." *Bayesian Analysis*, 5: 171–188. 691, 697

— (2012). "Structuring shrinkage: some correlated priors for regression." *Biometrika*, 99: 481–487. 692, 694

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge regression: biased estimation for nonorthogonal problems." *Technometrics*, 12: 55–67. 692

Kass, R. F. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American Statistical Association*, 90: 773–795. 698

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g Priors for Bayesian Variable Selection." *Journal of the American Statistical Association*, 103: 410–423. 694

Maruyama, Y. and George, E. I. (2011). "Fully Bayes Factors with a generalised g-prior." *Annals of Statistics*, 39: 2740–2765. 694

Ogutu, J. O., Schulz-Streeck, T., and Piepho, H.-P. (2012). "Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions." *BMC Proceedings*, 6 (Suppl 2): S10. 693

Polson, N. G. and Scott, J. G. (2011). "Shrink globally, act locally: sparse Bayesian regularization and prediction." In Bernardo J. M., M. J., Bayarri, Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics 9*, 501–538. Oxford: Clarendon Press. 692

— (2012). "Local shrinkage rules, Lévy processes, and regularized regression." *Journal of the Royal Statistical Society Series B*, 74: 287–311. 691, 697, 698, 699

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). "Bayesian model averaging for linear regression models." *Journal of the American Statistical Association*, 92: 179–191. 697

Raiffa, H. and Schlaifer, R. (1961). *Applied statistical decision theory*. M.I.T. Press. 697

Vandaele, W. (1978). "Participation in illigimate activities: Ehrlich revisited." In *Deterrence and Incapacitation*, 270–335. Washington, D. C.: National Academy of Sciences. 697

Waldron, L., Pintilie, M., Taso, M.-S., Shepherd, F. A., Huttenhower, C., and Jurisica, I. (2011). "Optimized application of penalized regression methods to diverse genomic data." *Bioinformatics*, 27: 3399–3406. 693

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233–243. Amsterdam: North Holland/Elsevier. 693