Bayesian Analysis (2010) **5**, Number 1, pp. 171–188

Inference with normal-gamma prior distributions in regression problems

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

Abstract. This paper considers the effects of placing an absolutely continuous prior distribution on the regression coefficients of a linear model. We show that the posterior expectation is a matrix-shrunken version of the least squares estimate where the shrinkage matrix depends on the derivatives of the prior predictive density of the least squares estimate. The special case of the normal-gamma prior, which generalizes the Bayesian Lasso (Park and Casella 2008), is studied in depth. We discuss the prior interpretation and the posterior effects of hyperparameter choice and suggest a data-dependent default prior. Simulations and a chemometric example are used to compare the performance of the normal-gamma and the Bayesian Lasso in terms of out-of-sample predictive performance.

Keywords: Multiple regression, $p > n$, Normal-Gamma prior, "Spike-and-slab" prior, Bayesian Lasso, Posterior moments, Shrinkage, Scale mixture of normals, Markov chain Monte Carlo

1 Introduction

The standard multiple linear regression model assumes that a vector of responses $y = (y_1, y_2, \ldots, y_n)$ can be represented as

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

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ are independent, $p(\epsilon_i) = \mathrm{N}(\epsilon_i|0, \sigma^2)$ and X is an $n \times p$ matrix of explanatory variables. Here, $\mathrm{N}(x|\mu, \sigma^2)$ denotes the density of a normal distribution with mean μ and variance σ^2. The scalar α is the intercept and $\mathbf{1}$ a $n \times 1$ unit vector. This paper is concerned with the Bayesian analysis of this model and, in particular, the choice of the prior distribution of the $(p \times 1)$-dimensional vector of regression coefficients β. A zero mean normal prior leads to the ridge estimator as posterior mean. This estimator performs poorly if there are large differences in the size of regression coefficients. Alternatively, we could perform variable selection and assume that only a subset of the variables have non-zero regression coefficients which mitigate the problems associated with the normal prior. The standard approach is the "spike-and-slab" prior (Mitchell and Beauchamp 1988). An indicator variable z_i is introduced to identify whether the i-th variable is included in the model ($z_i = 1$) or excluded ($z_i = 0$). The prior for β_i can be written as

$$\pi(\beta_i) = z_i \mathrm{N}(\beta_i|0, \sigma_\beta^2) + (1 - z_i)\delta_{\beta_i = 0}\,, \qquad p(z_i = 1) = w,$$

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

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

 DOI:10.1214/10-BA507

where $\delta_{\beta_i=0}$ represents the Dirac delta measure which places all its mass on zero and Mitchell and Beauchamp's uniform slab is replaced by the now traditional Gaussian "slab". The independent Bernoulli variables z_i have mean w, and so the hyperparameter w can be interpreted as the prior proportion of non-zero regressors. Alternatively, a prior distribution can be used for w and its value inferred from the data. The scale σ_β controls the variance of the prior. As we shall show in Section 2, this prior induces adaptive matrix shrinkage (Chamberlain and Leamer 1976) where the norm $|\mathrm{E}[\beta|y]|$ is a decreasing function of $|\hat{\beta}|$ and larger $|\hat{\beta}|$ are shrunk less than smaller $|\hat{\beta}|$. However, for fixed design X, very large $|\hat{\beta}|$ shrinkage asymptotes to some non-zero value determined by σ_β^2, (and the eigenvalues of $X^T X$).

Absolutely continuous prior distributions can also induce adaptive shrinkage and represent an interesting alternative to "spike-and-slab" priors. Park and Casella (2008) describe a Bayesian analysis where the regression coefficients are given independent double exponential prior distributions. The idea is motivated by the equivalence of the maximum a posteriori (MAP) estimator under this prior distribution and the Lasso estimator (Tibshirani 1996). The prior distribution can also be motivated as a member of the scale mixture of normals family. If we assume that $\beta_i|\Psi_i \sim \mathrm{N}(\beta_i|0, \Psi_i)$ then a double exponential prior for β_i arises when Ψ_i follows an exponential distribution (Andrews and Mallows (1974). The prior can be seen as a ridge prior where the variance of the normal is allowed to change from regression coefficient to regression coefficient, allowing differences in their scales. Bayesian estimation of regression models with other absolutely continuous priors has been investigated by several authors (MacKay 1996; Tipping 2001; Figueiredo and Jain 2001; Figueiredo 2003; Kiiveri 2008; Caron and Doucet 2008). However their work was restricted to MAP estimation rather than fully considering the posterior distribution.

The performance of estimators based on absolutely continuous priors depends on the form of the prior. For example, the double exponential prior used in the Bayesian Lasso has a single hyperparameter. In many regression problems this inflexibility can have a considerable effect on the inference since it restricts the prior beliefs that can be expressed. As we will show this choice fixes the rate of decay of the ordered regression coefficients. In the extreme case where some regression coefficients are zero then the mean of the exponential needs to be set to a small value to shrink those regression coefficients close to zero. However, this also leads to substantial shrinkage of the regression coefficients that are truly non-zero. In other words, the shrinkage is not adaptive enough for the problem. This can lead to both poor prediction accuracy and inference about the regression coefficients. This is an extreme case but the problem will also effect analyses where all the regression coefficients are truly non-zero but many are close to zero. The problem is particularly acute if n is small and p is large when prior assumptions play an important role in posterior inference.

This paper looks at the effect of the prior for β on the posterior expectation and variance of β. We show that the posterior expectation is a shrinkage estimator where the level of shrinkage is directly related to the form of the prior predictive distribution of the least squares estimator. We develop a generalisation of the double exponential

prior distribution for regression problems. The normal-gamma prior provides a natural extension and we show that it defines estimators that induce a wide-range of shrinkage behaviour, including effective selection (by shrinking the posterior expectation of many regression coefficients to values very close to zero).

The paper is organised in the following way. Section 2 discusses a very general expression for the posterior expectation of the regression coefficients showing that the shape of the prior predictive distribution plays a key role (and acts as an analogue to some of the results of Fan and Li (2001) for penalized maximum likelihood estimation). Section 3 considers the use of the normal-gamma distribution as a prior for regression coefficients and the interpretation of its hyperparameters in a Bayesian context. We finish the section by suggesting a simple, data-dependent default prior for the hyperparameters. Section 4 introduces a Gibbs sampling scheme for posterior inference including the use of the singular value decomposition for reduced dimension matrix inversion and computational speed. Section 5 applies the method to two simulation examples and a real problem in chemometrics and compares with results for estimation using the double exponential prior distribution. A discussion follows in section 6.

2 The posterior expectation of regression coefficients in linear regression

Suppose that the regression error variance σ^2 is known in the linear regression model given by equation (1), then the posterior expectation and variance of the $(p \times 1)$-dimensional vector β can be expressed in a useful form by extending a univariate result of Pericchi and Smith (1992).

Proposition 1. *Suppose that we have the linear regression model given by equation (1) where $n \geq p+1$, the design matrix X is non-singular and its columns have been centred, and the intercept α is independent of the $p \times 1$ vector β a priori. Let $\hat{\beta} = (X^T X)^{-1} X^T y$, the standard least squares estimator, and $h(\hat{\beta}) = \int N(\hat{\beta}|\beta, \sigma^2 (X^T X)^{-1}) \pi(\beta)\, d\beta$ where $\pi(\beta)$ is the prior distribution of β then*

$$E[\beta|\hat{\beta}] = \left(I - S(\hat{\beta})\right)\hat{\beta} \text{ and } V[\beta|\hat{\beta}] = \sigma^2 (X^T X)^{-1} - \sigma^4 (X^T X)^{-1} W(\hat{\beta}) (X^T X)^{-1}$$

where

$$S(\hat{\beta}) = \sigma^2 (X^T X)^{-1} R(\hat{\beta})$$

and $R(x)$ is a diagonal matrix with

$$R_{ii}(x) = -\frac{1}{x_i} \frac{\partial}{\partial x_i} \log h(x)$$

and $W(x) = -\frac{\partial}{\partial x} \frac{\partial}{\partial x^T} \log h(x)$.

The sampling density of $\hat{\beta}$ is $N(\beta, \sigma^2 (X^T X)^{-1})$ and so $h(\hat{\beta})$ is the prior predictive density of $\hat{\beta}$. The result may be extended to singular X and $p > n - 1$ using the

singular value decomposition of X and exploiting the scale mixture of normals, as illustrated in section 4. The result in Proposition 1 has several implications. The vector posterior expectation is always a matrix-shrunken version of the least squares estimator. The amount of shrinkage is controlled by the shape of h and the standard error of $\hat{\beta}$. Interestingly, the penalized maximum likelihood estimator can also be expressed as a shrinkage estimator (Fan and Li 2001). In that case the shrinkage is controlled by the derivative of the penalty function (which is related to the log prior density if the posterior mode is used as our estimator). In contrast, the posterior expectation depends on the derivative of the log predictive distribution. The predictive distribution can be considered a "smoothed" version of the prior distribution and so, in this case, the distinction between absolutely continuous and discrete prior distributions is blurred. Therefore, the posterior mean is a continuous function of $\hat{\beta}$, whereas the MAP estimate is not.

If the design matrix is orthogonal, the result can be simply expressed in terms of each regression coefficient:

$$\mathrm{E}[\beta_j|\hat{\beta}_j] = \hat{\beta}_j \left(1 - S^{(j)}(\hat{\beta}_j)\right)$$

and

$$\mathrm{V}[\beta_j|\hat{\beta}_j] = \frac{\sigma^2}{\sum_{i=1}^n x_{ij}^2} - \frac{\sigma^4}{(\sum_{i=1}^n x_{ij}^2)^2} W^{(j)}(\hat{\beta}_j),$$

where $W^{(j)}(\hat{\beta}_j) = -\frac{d^2}{d\hat{\beta}_j^2} \log h(\hat{\beta}_j)$ and $S^{(j)}(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^n x_{ij}^2} R_{(jj)}(\hat{\beta}_j)$ with $R_{(jj)}(\hat{\beta}_j) = -\frac{1}{\hat{\beta}_j} \frac{d}{d\hat{\beta}_j} \log\, h(\hat{\beta}_j)$. In this case the shrinkage factor is simply $S^{(j)}(\hat{\beta}_j)$. The shrinkage of β_j only depends on the univariate predictive density of $\hat{\beta}_j$ and its observed value. Clearly the shape of the prior distribution of β_j will directly affect the shrinkage and heavier tails will lead to less shrinkage. Conversely densities which are more peaked at zero will lead to larger shrinkage of small estimated values. A desirable property for fixed design X is that $\mathrm{E}[\beta_j|\hat{\beta}_j] \to \hat{\beta}_j$ as $\hat{\beta}_j \to \infty$. If $h(x) \approx \exp\left\{-\frac{1}{2}cx^2\right\}$ (i.e. the predictive distribution has normal tails), then $-\frac{1}{x}\frac{d}{dx} \log h(x) \to c$ and so $\mathrm{E}[\beta_j|\hat{\beta}_j]$ does not limit to $\hat{\beta}_j$ as $\hat{\beta}_j \to \infty$. The predictive distribution will have these tails if the prior on β is normal or if a normal "slab" is chosen in a "spike-and-slab" prior. If the tails of the prior distribution of β are heavier than normal, then the $\mathrm{E}[\beta_j|\hat{\beta}_j] \to \hat{\beta}_j$ as $\hat{\beta}_j \to \infty$, (see also Dawid (1973)).

3 The normal-gamma prior

A wide and natural class of prior densities for regression coefficients is the scale mixtures of normals (SMN) (see *e.g.* West (1987)), which we write as

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

where G is a mixing distribution. The prior can be expressed in a hierarchical form as

$$\beta_i|\Psi_i \sim \mathrm{N}(0, \Psi_i), \qquad \Psi_i \sim G. \tag{2}$$

This hierarchical form for the model shows that the i-th regression coefficient has a normal prior distribution conditional on an idiosyncratic variance (or scale), Ψ_i. This allows for larger differences in the sizes of the regression coefficients than would be possible under a normal prior. The marginal prior distribution for $\hat{\beta}_i$ has heavier than normal tails (apart from the degenerate case where G places all its mass at a single point). The "spike-and-slab" prior can be represented in this way by choosing

$$G(\Psi_i) = z_i \delta_{\Psi_i = \sigma^2_{\hat{\beta}}} + (1 - z_i)\delta_{\Psi_i = 0}.$$

The double exponential prior of the Bayesian Lasso arises if G is an exponential distribution.

An interesting choice of absolutely continuous prior is the normal-gamma distribution, which includes the double exponential prior as a special case. Let $\text{Ga}(x|c,d)$ represent the density of a gamma distribution with shape c and rate d so that

$$\text{Ga}(x|c,d) = \frac{d^c}{\Gamma(c)} x^{c-1} \exp\{-dx\}.$$

We refer to the distribution as $\text{Ga}(c,d)$. The normal-gamma distribution arises by assuming that the mixing distribution in a SMN has the density $g(x) = \text{Ga}(x|\lambda, 1/(2\gamma^2))$. The density function is expressible as

$$\pi(\beta_i) = \frac{1}{\sqrt{\pi} 2^{\lambda - 1/2} \gamma^{\lambda + 1/2} \Gamma(\lambda)} |\beta_i|^{\lambda - 1/2} K_{\lambda - 1/2}(|\beta_i|/\gamma), \tag{3}$$

where K is the modified Bessel function of the third kind. The variance of β_i is $v_\beta = 2\lambda\gamma^2$ and the excess kurtosis is $\frac{3}{\lambda}$. The gamma distribution can represent a wide-range of shapes. As the shape parameter λ decreases these include distributions that place a lot of mass close to zero but at the same time have heavy tails. Figure 1 shows the

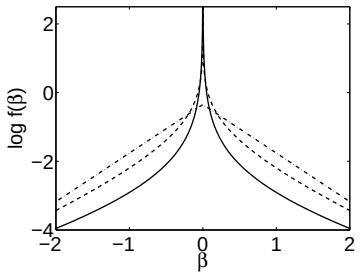

Figure 1: The log density of the normal-gamma prior with a variance of 2 and different values of λ. $\lambda = 0.1$ (solid line), $\lambda = 0.333$ (dot-dashed line) and $\lambda = 1$ (dashed line).

effect of shape parameter λ on the marginal prior distribution of β_i. The marginal distribution becomes more peaked at zero which places increasing mass close to zero as λ decreases. The distribution has proved a popular choice for modelling fat tails in finance (Bibby and Sorensen 2003), and is a member of the generalized hyperbolic

family (Barndorff-Nielsen and Blaesild 1981). The prior was considered by Griffin and Brown (2007), but the shape of the density made it difficult to obtain MAP estimates. More recently, Caron and Doucet (2008) have looked at MAP estimation and drawn a link to Lévy processes. The density has exponential tails, but the heaviness of the tail is controlled by λ, which can take any positive value.

The choice of λ and γ plays an important role in estimation. Park and Casella (2008) discuss an empirical Bayes estimation strategy for the hyperparameter of the Bayesian Lasso. However, with the normal-gamma prior, the posterior distribution of λ and γ can be highly multimodal and an empirical Bayes approach is very difficult to implement. Therefore we take a fully Bayesian approach and concentrate on choosing a prior distribution for the hyperparameters of the normal-gamma prior. In order to make an informed choice, we consider the effect of λ on both the prior and the posterior.

3.1 The effect of λ

It follows from the definition of the model in equations (1) and (2) that

$$\mathrm{V}[y_i|\Psi,\sigma^2] = \mathrm{V}[\alpha] + \sum_{j=1}^{p} \Psi_j + \sigma^2,$$

if the regressors have been standardized so that the sample mean and variance of each regressor is 0 and 1 respectively. The regression total variability is $(\sum_{k=1}^{p} \Psi_k)$. Thus $\zeta_j = \frac{\Psi_j}{\sum_{k=1}^{p} \Psi_k}$ can be interpreted as the proportion of total variability attributable to the j-th regressor. If the mixing distribution is gamma, then ζ follows a $\mathrm{Di}(\lambda,\lambda,\ldots,\lambda)$, a Dirichlet distribution with all parameters equal to λ. Consequently, the distribution of ζ is controlled by λ only and γ has no effect. Increasing λ will lead to more evenly distributed values of $\zeta_1,\zeta_2,\ldots,\zeta_p$ and small values of λ will be associated with large differences between the proportions. We look at the strength of this effect by considering $\zeta_{(1)} > \zeta_{(2)} > \cdots > \zeta_{(p)}$, which is the ordered version of ζ, and define $r_j = \log\zeta_{(j)} - \log\zeta_{(j+1)}$. Plotting $r_1, r_2, \ldots, r_{p-1}$ gives an indication of the rate at which the ordered proportions decay for a given prior distribution. These plots for various values of λ are shown in Figure 2. The shape of the curve defined by the values $r_1, r_2, \ldots, r_{p-1}$ is similar for all values of λ. The rate is fairly constant for small values of j but increases for larger values of j. The level is determined by both p and λ. Smaller values of λ and smaller values of p are associated with larger values of r_j (*i.e.* a faster decay). This illustrates a limitation of the Bayesian Lasso prior, $(\lambda = 1)$, which implies a particular set of values for r_j for a given p. Extending the prior to the normal-gamma distribution leads to a wider choice of decay rates.

The posterior properties of the regression coefficients can be studied using Proposition 1. The shrinkage factor, $S(\hat{\beta})$, for the posterior expectation of a single regression coefficient is plotted against the standard error (SE) of $\hat{\beta}$ in Figure 3 for different choices of λ. The shrinkage for small values of $\hat{\beta}$ changes markedly with λ (smaller values of λ are associated with larger amounts of shrinkage). When SE $< 1/5$ the graphs show a fast transition from a high level shrinkage to a low level of shrinkage (*e.g.* if SE $= 1/5$

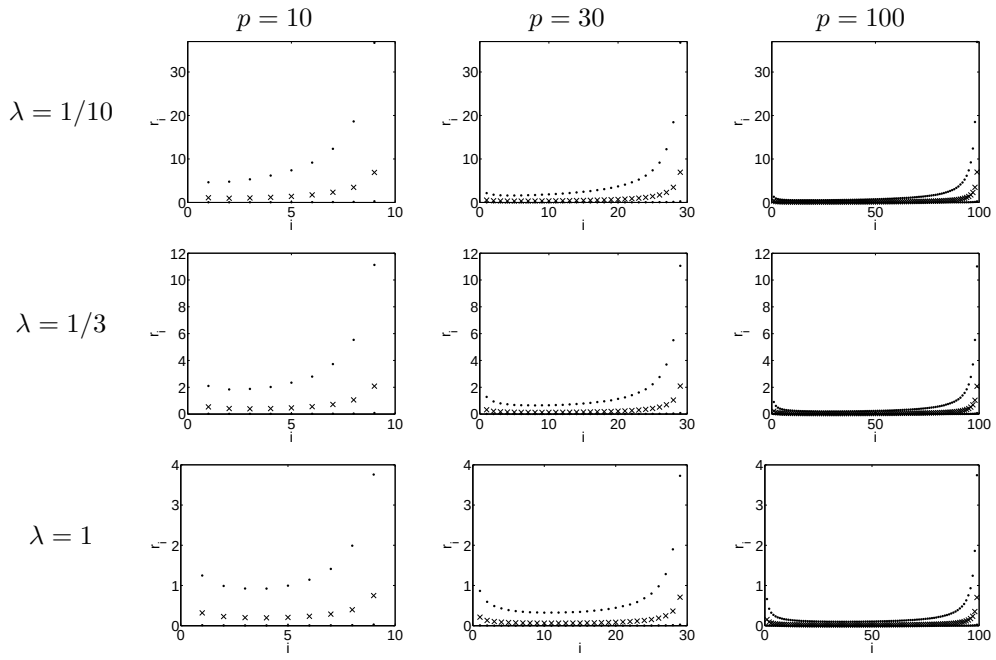

Figure 2: The prior distribution of $r_i = \log \zeta_{(i+1)} - \log \zeta_{(i)}$ represented by the median (cross) and 95% central region (dots).

this transition occurs at around 1). The positions of these transitions are controlled by the hyperparameter γ.

3.2 Similarity to "spike-and-slab"

The above shrinkage factor results for the normal-gamma prior show an adaptive pattern which can take a wide range of shapes according to the choice of hyperparameters. It is therefore interesting to ask whether the "spike-and-slab" prior leads to forms of shrinkage different from the normal-gamma. The hyperparameters of the normal-gamma prior can be matched to the "spike-and-slab" prior using the following argument. The prior proportion of non-zero coefficients, w, of the "spike-and-slab" can be elicited by choosing a prior guess of the number of non-zero regression coefficients, $p^\star$, and setting $w = p^\star/p$. For the normal-gamma prior, we could mimic the "spike and slab" prior by choosing λ so that most of the variation in the prior is contained in a small number of regressors. To make this idea operational, we chose λ to solve

$$\text{median}\left(\sum_{i=1}^{p^\star} \zeta_{(i)}\right) = 1 - \epsilon$$

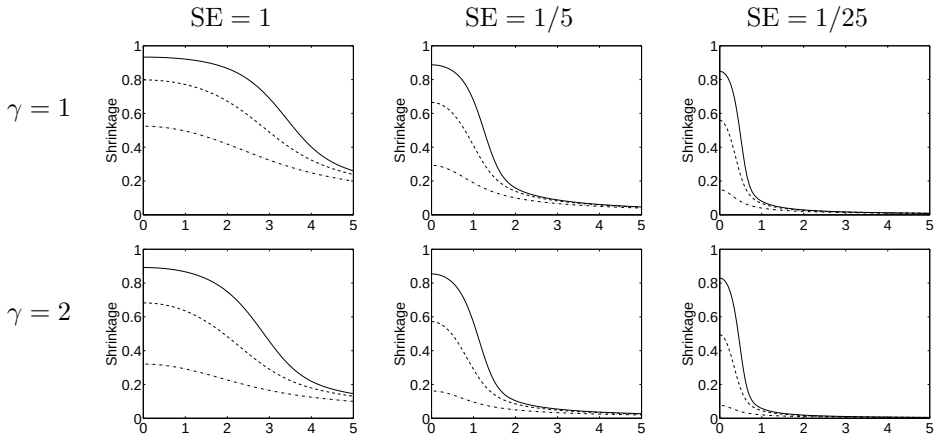

Figure 3: Shrinkage factor for different values of $\hat{\beta}$ for different priors: $\lambda = 0.1$ (solid line), $\lambda = 0.333$ (dashed line) and $\lambda = 1$ (dot-dashed line).

for a pre-specified "small" value of ϵ. Choosing the variance of the normal component $\sigma_\beta^2 = 2\lambda\gamma^2 p/p^\star$ guarantees that $\mathrm{V}[\beta_i]$ is the same under the two priors.

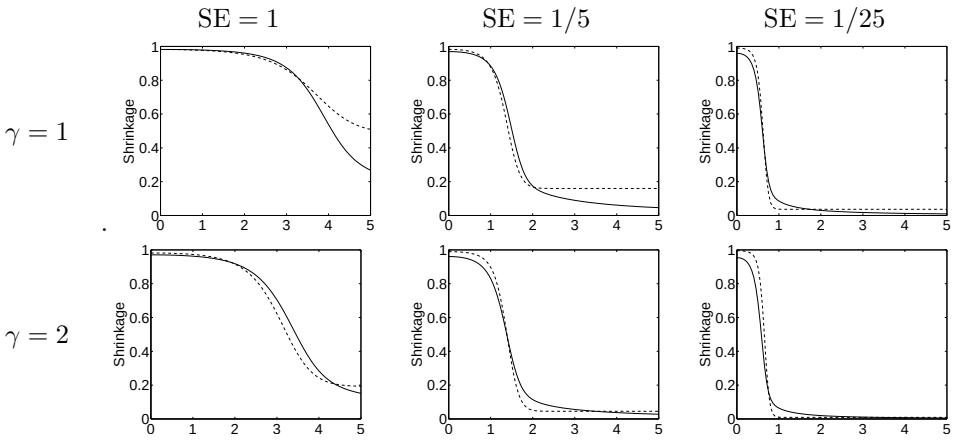

Figure 4: Shrinkage factor for different values of $\hat{\beta}$ for matched normal-gamma (solid line) and "spike-and-slab" priors (dashed line) with $p = 100$ and $p^\star = 5$.

Figure 4, with $\epsilon = 0.1$, shows the shrinkage factor $S(\hat{\beta})$ for a single regressor with $p = 100$ and $p^\star = 5$ for several values of the standard error (SE). The two priors lead to remarkably similar shrinkage for many values of the least squares estimate (only positive values are shown since this function will be symmetric around 0). The very sharp transition associated with the "spike-and-slab" prior is mimicked by the matching normal-gamma distribution. The main differences occur for very small values and very

large values of $\hat{\beta}$. When $\hat{\beta}$ is large, the "spike-and-slab" prior reverts to the constant shrinkage factor induced by a ridge prior, whereas the shrinkage for the normal-gamma prior tends to zero. If β is small, the shrinkage associated with the "spike-and-slab" prior tends to be larger than for the matching normal-gamma prior.

3.3 Prior hyperparameter settings

The hyperparameters of the normal-gamma distribution could be chosen to match the "spike-and- slab" prior as discussed in the Section 3.2. However, we choose a simpler route by directly specifying priors for λ and γ. A prior for λ which seems to work well in the simulations and our example is obtained by taking the prior of λ to be an exponential distribution with mean 1. This offers variability around the Bayesian Lasso prior ($\lambda = 1$). The prior for the scale parameter γ conditional on λ is given by $v_\beta = 2\lambda\gamma^2 \sim \text{IG}(2, M)$, where IG denotes the inverted gamma distribution, the inverse of a gamma distribution, so that $\text{IG}(2, M)$ has expectation M. When X is non-singular, $M = \frac{1}{p}\sum_{i=1}^{p} \hat{\beta}_i^2$ where $\hat{\beta}$ is the least squares estimate. When X is singular, as when $p > n - 1$, $M = \frac{1}{n}\sum_{i=1}^{p} \hat{\beta}_i^2$ where $\hat{\beta}$ is the Minimum Length Least Squares (MLLS) estimate. In the nonsingular X case, this is the same as that used in the Hoerl-Kennard-Baldwin estimate of the variance of β for the constant in ridge regression (apart from a Stein-type dimension correction) (see Hoerl et al. (1975) or Brown (1993), section 4.4). This completes the prior specification for β.

The prior for the intercept α in (1) is $\pi(\alpha) \propto 1$. Lastly, we choose a vague prior for the error variance so that $\sigma^{-2} \propto 1$.

4 Computational method

The posterior distribution of the parameters can be simulated using a Gibbs sampler with an additional Metropolis-Hastings update. The convergence of the method is improved by augmenting the model with the latent scale parameters $\Psi_1, \Psi_2, \ldots, \Psi_p$. The full conditionals used in the updating steps are given below.

Updating α and β

Let $\phi = (\alpha, \beta)^T$. The full conditional distribution of ϕ follows a joint normal distribution with mean $\left(X^{\star T}X^\star + \sigma^2\Lambda\right)^{-1} X^{\star T}y$ and variance $\sigma^2\left(X^{\star T}X^\star + \sigma^2\Lambda\right)^{-1}$, where

$$\Lambda = \text{diag}\left(0, \frac{1}{\Psi_1}, \frac{1}{\Psi_2}, \ldots, \frac{1}{\Psi_p}\right),$$

and $X^\star = [1 : X]$. It is computationally convenient in problems with $p > n - 1$ to express the mean and variance of this distribution using the following form which only involves the inversion of an $n \times n$ matrix rather than a larger $(p+1)\times(p+1)$ matrix. The standard MLE estimator will not be defined if $p > n - 1$. Consequently, the problem is

re-expressed in terms of an n-dimensional parameter, θ, for which the MLE exists. As in West (2003), the singular value decomposition of $X^\star$ can be written as $X^\star = F^T D A^T$ where A is a $((p+1) \times n)$-dimensional matrix such that $A^T A = I_n$, D is an $(n \times n)$-dimensional diagonal matrix and F is a $(n \times n)$-dimensional matrix for which $F^T F = I_n$ and $FF^T = I_n$. Clearly, we can write

$$X^\star \phi = (F^T D)\theta.$$

The MLE, $\hat{\theta}$, of θ is well-defined and has the form

$$\hat{\theta} = D^{-1} F y.$$

Let $\Lambda^\star = D^{-2}$ and $\Psi_0 = A^T \Psi A$. After some simplification we can express the posterior mean and covariance in terms of the inverse of an $n \times n$ matrix:

$$\mathrm{E}\left(\phi \,\middle|\, \Psi, \hat{\theta}\right) = \Psi A (\Psi_0 + \sigma^2 \Lambda^\star)^{-1} \hat{\theta}$$

and

$$\mathrm{V}\left(\phi \,\middle|\, \Psi, \hat{\theta}\right) = \Psi - \Psi A (\Psi_0 + \sigma^2 \Lambda^\star)^{-1} A^T \Psi.$$

Updating σ^2

The full conditional distribution of σ^{-2} is $\mathrm{Ga}\,(c^\star, d^\star)$ where $c^\star = n/2$ and $d^\star = (y - \alpha - X\beta)^T (y - \alpha - X\beta)/2$.

Updating Ψ

The parameter Ψ can effectively be updated in a block since the full conditional distributions of $\Psi_1, \Psi_2, \ldots, \Psi_p$ are independent. The full conditional distribution of Ψ_i follows a Generalized Inverse Gaussian distribution $\mathrm{GIG}\left(\lambda - \frac{1}{2}, 1/\gamma^2, \beta_i^2\right)$ where $\mathrm{GIG}(m, c, d)$ has the density

$$\frac{(c/d)^{m/2}}{2K_m(\sqrt{cd})} x^{m-1} \exp\left\{-\frac{1}{2}(cx + d/x)\right\}.$$

An algorithm for simulation of this distribution is described by Devroye (1986). A Matlab implementation is available in the "randraw" toolbox which is available from Matlab Download Central.

Updating hyperparameters of the normal-gamma prior

In section 3.3 we assigned priors for λ and γ. If we denote the prior for λ by $\pi(\lambda)$ then the full conditional of λ is

$$\pi(\lambda) \frac{1}{(2\gamma^2)^{p\lambda} (\Gamma(\lambda))^p} \left(\prod_{i=1}^{p} \Psi_i\right)^{\lambda},$$

which can be updated using a Metropolis-Hastings random walk update on $\log \lambda$. We propose $\lambda' = \exp\{\sigma_\lambda^2 z\}\lambda$, where z is standard normal then λ' is accepted with probability

$$\min\left\{1, \frac{\pi(\lambda')}{\pi(\lambda)}\left(\frac{\Gamma(\lambda)}{\Gamma(\lambda')}\right)^p \left((2\gamma^2)^{-p}\prod_{i=1}^p \Psi_i\right)^{\lambda'-\lambda}\right\}.$$

The tuning parameter σ_λ^2 is chosen to set the average acceptance rate at around 20-30%.

Now turning to the scale parameter γ, with $\gamma^{-2} \sim \text{Ga}(2, M/2\lambda)$ from section 3.3, γ^{-2} can be updated directly from its full conditional distribution, which is $\gamma^{-2} \sim \text{Ga}(e^\star, f^\star)$ where $e^\star = 2 + p\lambda$ and $f^\star = M/2\lambda + \frac{1}{2}\sum_{i=1}^p \Psi_i$.

5 Examples

We consider two simulation examples and a real data example from chemometrics. In each example, the MCMC algorithm was run for 40,000 iterations with the first 5,000 discarded as a burn-in.

5.1 Simulation 1

In the first simulation we generate data where the design matrix has independent elements which are standard normally distributed. The regression coefficients are drawn from normal-gamma priors with three possible shape parameters, (0.1, 1 and 3). In each case, the prior mean of the regression coefficients was chosen to be 1. Sample sizes considered were $n = 50$ and 200 with $p = 100$. In the first case, $n > p$ and $p > n$ in the second case. The performance of the double exponential prior is compared to the normal-gamma prior using root mean squared error of the estimation of β by the posterior mean. The double exponential was favored when $\lambda = 1$ and the normal-gamma in the other two cases. The posterior distributions for λ, γ are summarised in Table 1

			Normal-Gamma		RMSE	
λ	n	p	λ	$v_\beta = 2\lambda\gamma^2$	NG	DE
1	200	100	1.02 (0.54, 2.08)	0.59 (0.35, 1.02)	0.107	0.107
1	50	100	0.92 (0.31, 2.40)	0.25 (0.11, 0.89)	0.683	0.677
3	200	100	1.87 (0.98, 7.62)	0.64 (0.38, 1.08)	0.100	0.101
3	50	100	1.09 (0.30, 5.02)	0.19 (0.09, 0.68)	0.739	0.740
0.1	200	100	0.12 (0.07, 0.19)	0.55 (0.26, 1.36)	0.069	0.091
0.1	50	100	0.08 (0.05, 0.16)	0.94 (0.29, 5.16)	0.203	0.452

Table 1: Simulation 1: Posterior mean and 95% credibility interval of the parameters of the Normal-Gamma model and the root mean squared errors (RMSE) of estimating β by its posterior mean with the NG prior and double exponential (DE)

in terms λ, v_β, where variance $v_\beta = 2\lambda\gamma^2$. The values of λ and v_β seem well estimated

in each case. The credibility interval for λ increases as the true value of λ increases.

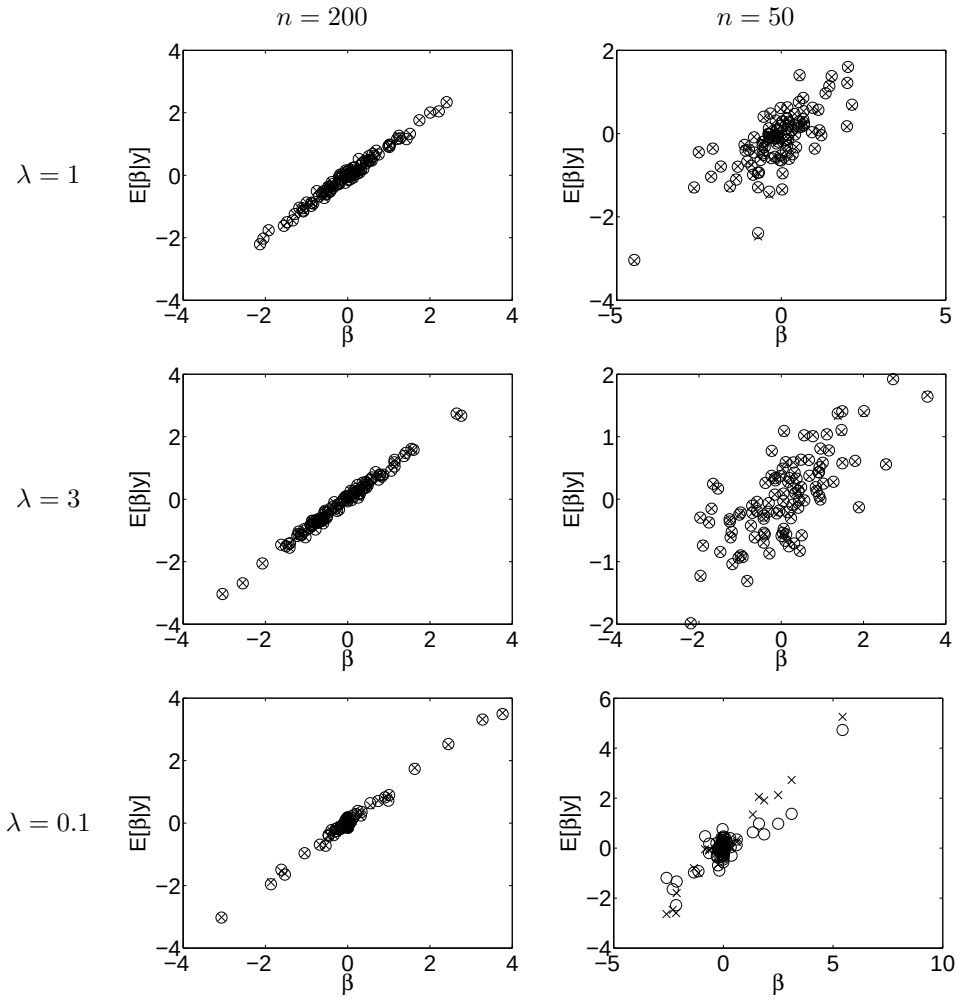

Figure 5: Plots of the true regression coefficients against their posterior means under the NG (crosses) and Lasso priors (circles), simulation 1.

The root mean square errors for estimation of the regression coefficients are given in Table 1. The performance of the double exponential and normal-gamma priors are roughly similar when $\lambda = 1$ (when the double exponential is the true distribution of the regression coefficients) and $\lambda = 3$. This is reassuring since it demonstrates that there is little lost in inference about the regression coefficients if the λ is considered unknown. However, there are substantial differences when $\lambda = 0.1$ and, as we would expect, the differences become larger as the sample size, n, becomes smaller. The RMSE is more than doubled by the double exponential prior relative to the normal-gamma prior when

$n = 50$ and $\lambda = 0.1$.

Figure 5 shows the estimated coefficient versus their true values. Clearly, the regression coefficients are well estimated when $n = 200$ with both prior distributions but less well when $n = 50$. If $\lambda = 0.1$, there is a substantial number of regression coefficients whose value is close to zero, which is better captured by estimates using the normal-gamma prior rather than the double exponential prior.

5.2 Simulation 2

In this example, the design matrix is generated in the same way as Simulation 1 but the regression coefficients have a different structure. In this case, only 10 regression coefficients are non-zero which are placed evenly throughout the vector of regression coefficients which can be written as

$$\beta_i = \begin{cases} \beta^\star & \text{mod}((i-1), p/10) = 0 \\ 0 & \text{otherwise.} \end{cases}$$

In these simulation $\beta^\star = 1, 5$, $n = 50$ and $p = 100, 200$. This is a challenging example since p is larger than n and our prior places no mass on the regression coefficients being zero. However, it would be re-assuring if the prior performs well for this type of data. Table 2 displays the means and 95% credibility intervals for the hyperparameters of

$\beta^\star$	n	p	λ	$v_\beta = 2\lambda\gamma^2$
1	50	100	0.14 (0.05, 1.22)	0.16 (0.03, 0.84)
5	50	100	0.030 (0.015, 0.054)	4.00 (1.19, 19.65)
1	50	200	0.34 (0.08, 1.56)	0.014 (0.004, 0.075)
5	50	200	0.018 (0.012, 0.0231)	1.95 (0.59, 8.55)

Table 2: Estimates of the hyperparameters of the normal-gamma distribution for simulation 2.

the NG model. The posterior mean value of λ is much smaller than the Lasso value of 1. The posterior means of the regression coefficients with the NG prior and DE prior are displayed in Figure 6. The coefficient estimates based on the NG prior out-perform those based on the DE prior. For large signal (i.e. both $\beta^\star$ and n large), the NG estimates identified all the correct $\beta^\star$ with very little attenuation.

5.3 Example: NIR spectroscopy data

The data consists of 215 near-infrared absorbance spectra of meat samples, recorded on a Tecator Infratec Food Analyzer (represented as a 100-channel absorbance spectrum in the wavelength range 850-1050nm) and the composition of each sample in terms of water, fat and protein content. We consider predicting fat content on the basis of its infrared spectrum using the 100 channels. The data is split in a training/monitoring/testing set

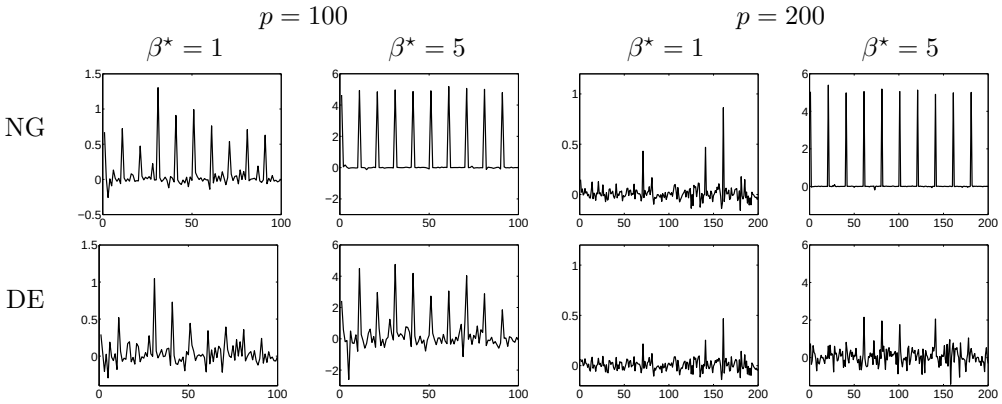

Figure 6: The posterior mean of β for simulation 2 with normal-gamma (NG) prior and double exponential prior (DE).

of 129/43/43 samples. The data, originally used by Borggaard and Thodberg (1992), is available at http://lib.stat.cmu.edu/datasets/tecator. More recently it was analysed in Eilers et al. (2009). We used the training and monitoring data comprising $n = 172 = 129 + 43$ samples (*all data*) as our main data set and also took a random subset of 60 of the training data to create a p larger than n data set (*small*). The RMSEs for

	All data	Small
normal-gamma	1.94	2.59
Lasso	3.54	3.09

Table 3: RMSEs for fat prediction

prediction of the 43-sample test set, using the normal-gamma and Lasso priors, are given in Table 3. The difference in predictive performance is not surprising when one looks at the posterior median of λ in the normal-gamma prior which is 0.020 with a 95% credibility interval of $(0.016, 0.026)$ on the full data.

The differences in the estimate of β under the two priors are shown in Figure 7, which shows the posterior means for the two datasets. Clearly, in the large data sets the RMSE is very different for the two methods as are the posterior means for β. Only a few regression coefficients are estimated to be far from zero with the normal-gamma prior unlike the Lasso prior which substantially overestimates many of the regression coefficients. The results for the smaller data set are similar. The posterior median of λ is 0.019 with a 95% credible interval of $(0.013, 0.032)$. In the normal-gamma prior there is more selection with only a few regression coefficients estimated to be far from zero. However, the estimates of these regression coefficients are similar to those with a larger sample size, unlike the Lasso prior whose estimates are much smaller.

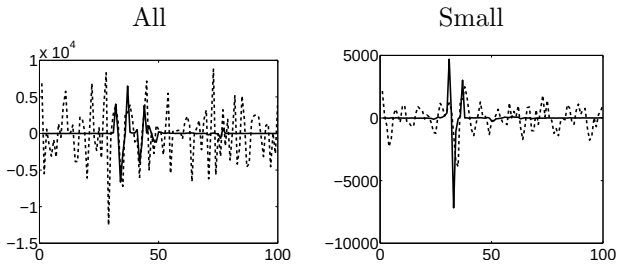

Figure 7: The posterior means of the β for the normal-gamma (solid) and Lasso (dashed) for two datasets.

6 Discussion

This paper considers the performance of absolutely continuous distributions as priors for regression coefficients. We demonstrate how the posterior expectation of the regression coefficients depends on the derivative of the prior predictive density of the least squares estimate of the regression coefficients. This allows us to compare the shrinkage induced by different absolutely continuous distributions.

A natural class of prior distributions are scale mixtures of normal distributions and we consider, in detail, the particular choice of a normal-gamma distribution. This distribution allows us to control prior beliefs about the decay of the absolute values of the ordered regression coefficients. At one extreme we have normal priors which promote a similarity between the magnitudes of the regression coefficients, and on the other hand prior distributions that can promote effective variable selection by extreme shrinkage of "small" regression coefficients to a value close to zero. A gamma mixing distribution can be interpreted through the well-known links between the gamma distribution and the Dirichlet distribution. We specify a default prior for the hyperparameters, shape λ and scale γ, of the normal-gamma and adopt a full Bayesian analysis.

A Gibbs sampler is proposed to fit the model with a normal-gamma prior, augmented by a Metropolis-Hastings step for updating the gamma shape parameter λ. Recently, there has been interest in regression problems where n is small and p is large. Naive application of Gibbs sampling would lead to the inversion of $(p \times p)$-dimensional matrix. We exploit the singular value decomposition of the design matrix to avoid this computationally expensive inversion, replacing it by the inversion of an $(n \times n)$-dimensional matrix. We have shown that the normal-gamma prior can better represent heterogeneity in regression effects relative to standard "spike-and-slab" priors in terms of predictive performance. The differences in the regression estimates under the priors are linked to the normal distribution commonly used as a "slab", which sets an upper limit on the shrinkage factors. The normal-gamma could also be used as the "slab" in a "spike-and-slab" prior. The use of such a distribution with heavy tails is an interesting area for future research. The computational methods developed in this paper could be easily extended to other generalized inverse Gaussian mixing distributions, which would

generate generalized hyperbolic priors for β.

Appendix A: Proof of Proposition 1

The proof follows closely that found in Pericchi and Smith (1992) generalised to multiparameter β and multiple regression. The centring of the columns of X implies that the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ are independent under the sampling distribution. The assumed prior independence implies that α and β are independent *a posteriori* and we can work with β only. From the definition of $h(x)$ as the predictive density of the $p \times 1$ sufficient statistic $\hat{\beta}$, it follows that

$$\begin{aligned}\frac{\partial}{\partial x}\log h(x) &= \frac{\int \frac{\partial}{\partial x}\mathrm{N}(x|\beta,\sigma^2(X^TX)^{-1})\pi(\beta)d\beta}{\int \mathrm{N}(x|\beta,\sigma^2(X^TX)^{-1})\pi(\beta)d\beta} \\ &= -\frac{\int \sigma^{-2}(X^TX)(x-\beta)\mathrm{N}(x|\beta,\sigma^2(X^TX)^{-1})\pi(\beta)d\beta}{\int \mathrm{N}(x|\beta,\sigma^2(X^TX)^{-1})\pi(\beta)d\beta}.\end{aligned}$$

Re-arranging we have

$$x + \sigma^2(X^TX)^{-1}\frac{\partial}{\partial x}\log h(x) = \frac{\int \beta\mathrm{N}(x|\beta,\sigma^2(X^TX)^{-1})\pi(\beta)d\beta}{\int \mathrm{N}(x|\beta,\sigma^2(X^TX)^{-1})\pi(\beta)d\beta}$$

and so the posterior expectation of $(p \times 1)$-dimensional vector β is given by

$$\mathrm{E}[\beta|\hat{\beta}] = \hat{\beta} - \sigma^2(X^TX)^{-1}s(\hat{\beta}) \tag{4}$$

where

$$s(x) = -\frac{\partial}{\partial x}\log h(x).$$

Clearly, this can can be written as

$$\mathrm{E}[\beta|\hat{\beta}] = \left(I - S(\hat{\beta})\right)\hat{\beta}$$

where

$$S(\hat{\beta}) = \sigma^2(X^TX)^{-1}R(\hat{\beta}),$$

with R as defined in the Proposition.

The form of the variance follows from observing that the mean square error is the 'variance plus squared bias' that is in matrix form

$$\mathrm{Var}(\beta|\hat{\beta}) = \mathrm{E}[(\beta-\hat{\beta})(\beta-\hat{\beta})^T|\hat{\beta}] - [\mathrm{E}[\beta|\hat{\beta}] - \hat{\beta}][\mathrm{E}[\beta|\hat{\beta}] - \hat{\beta}]^T.$$

From the result (4) we have

$$\mathrm{E}[\beta|\hat{\beta}] - \hat{\beta} = -\sigma^2(X^TX)^{-1}s(\hat{\beta}).$$

Letting $p(x|\beta)$ denote the sampling density of $\hat{\beta}$, differentiating it twice, we can write

$$(\beta - x)(\beta - x)^T p(x|\beta) = \sigma^4 (X^T X)^{-1} \frac{\partial^2}{\partial x \partial x^T} p(x|\beta)(X^T X)^{-1} + \sigma^2 (X^T X)^{-1} p(x|\beta).$$

Multiplying by the prior for β and integrating shows that

$$\mathrm{E}[(\beta - x)(\beta - x)^T|\hat{\beta}] = \sigma^4 (X^T X)^{-1} \left[\frac{\partial^2}{\partial x \partial x^T} h(x) \right] (X^T X)^{-1} \frac{1}{h(x)} + \sigma^2 (X^T X)^{-1}.$$

and so

$$\mathrm{Var}(\beta|\hat{\beta}) = \sigma^2 (X^T X)^{-1} - \sigma^4 (X^T X)^{-1} \left[s(\hat{\beta}) s(\hat{\beta})^T - \frac{1}{h(\hat{\beta})} \frac{\partial^2}{\partial x \partial x^T} h(x) \Big|_{x=\hat{\beta}} \right] (X^T X)^{-1}.$$

The result follows from noting that

$$W(x) = -\frac{\partial^2 \log h(x)}{\partial x \partial x^T} = \left[\left\{ \frac{\partial h(x)}{\partial x} \frac{\partial h(x)}{\partial x^T} \right\} - h(x) \frac{\partial^2 h(x)}{\partial x \partial x^T} \right] \Big/ h^2(x).$$

References

Andrews, D. F. and Mallows, C. L. ((1974). "Scale mixtures of normal distributions." *Journal of the Royal Statistical Society B*, 36: 99–102. 172

Barndorff-Nielsen, O. E. and Blaesild, P. (1981). "Hyperbolic distributions and ramifications: contributions to the theory and applications." In *Statistical Distributions in Scientific Work, Vol. 4*, 19–44. Dorderecht: Reidal. 176

Bibby, B. M. and Sorensen, M. (2003). "Hyperbolic Processes in Finance." In *Handbook of Heavy Tailed Distributions in Finance*, 211–248. Elsevier Science. 175

Borggaard, C. and Thodberg, H. H. (1992). "Optimal minimum neural interpretation of spectra." *Analytical Chemistry*, 64: 545–551. 184

Brown, P. J. (1993). *Measurement, Regression and Calibration*. Oxford: Clarendon Press. 179

Caron, F. and Doucet, A. (2008). "Sparse Bayesian nonparametric regression." In *Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland*. 172

Chamberlain, G. and Leamer, E. E. (1976). "Matrix weighted averages and posterior bounds." *Journal of the Royal Statistical Society B*, 38: 73–84. 172

Dawid, A. P. (1973). "Posterior expectations for large observations." *Biometrika*, 60: 664–667. 174

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York: Springer. 180

Eilers, P. H. C., Li, B., and Marx, B. D. (2009). "Multivariate calibration with single index regression." *Chemometrics and Intelligent Laboratory Systems*, 96: 196–202. 184

Fan, J. and Li, R. Z. (2001). "Variable selection via non-concave penalized likelihood and its oracle properties." *Journal of the American Statistical Association*, 96: 1348–1360. 173, 174

Figueiredo, M. A. T. (2003). "Adaptive sparseness for supervised learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25: 1150–1159. 172

Figueiredo, M. A. T. and Jain, A. K. (2001). "Bayesian learning of sparse classifiers." In *Proceedings IEEE Computer Society Conference in Computer Vision and Pattern Recognition*, volume 1, 35–41. 172

Griffin, J. E. and Brown, P. J. (2007). "Bayesian adaptive lassos with non-convex penalization." Technical report, IMSAS, University of Kent. 176

Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975). "Ridge regression: some simulations." *Communications in Statistics*, 4: 105–123. 179

Kiiveri, H. (2008). "A general approach to simultaneous model fitting and variable elimination in response models for biological data with many more variables than observations." *BMC Bioinformatics*, 9:195. 172

MacKay, D. J. C. (1996). "Bayesian methods for back-propagation networks." In *Models of Neural Networks III*, chapter 6, 211–254. New York: Springer. 172

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression (with Discussion)." *Journal of the American Statistical Association*, 83: 1023–1036. 171

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103: 672–680. 171, 172, 176

Pericchi, L. R. and Smith, A. F. M. (1992). "Exact and Approximate Posterior Moments for a Normal Location Parameter." *Journal of the Royal Statistical Society B*, 54: 793–804. 173, 186

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society B*, 58: 267–288. 172

Tipping, M. E. (2001). "Sparse Bayesian learning and the relevance vector machine." *Journal of Machine Learning Research*, 1: 211–244. 172

West, M. (1987). "On scale mixtures of normal distributions." *Biometrika*, 74: 646–648. 174

— (2003). "Bayesian Factor regression models in the large p, small n paradigm." In *Bayesian Statistics 7*, 733–742. Oxford: Clarendon Press. 180