Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **53**(4), 2011, 423–442 doi: 10.1111/j.1467-842X.2011.00641.x

BAYESIAN HYPER-LASSOS WITH NON-CONVEX PENALIZATION

JIM E. GRIFFIN AND PHILIP J. BROWN*

University of Kent

Summary

The Lasso has sparked interest in the use of penalization of the log-likelihood for variable selection, as well as for shrinkage. We are particularly interested in the more-variables-than-observations case of characteristic importance for modern data. The Bayesian interpretation of the Lasso as the maximum *a posteriori* estimate of the regression coefficients, which have been given independent, double exponential prior distributions, is adopted. Generalizing this prior provides a family of hyper-Lasso penalty functions, which includes the quasi-Cauchy distribution of Johnstone and Silverman as a special case. The properties of this approach, including the oracle property, are explored, and an EM algorithm for inference in regression problems is described. The posterior is multi-modal, and we suggest a strategy of using a set of perfectly fitting random starting values to explore modes in different regions of the parameter space. Simulations show that our procedure provides significant improvements on a range of established procedures, and we provide an example from chemometrics.

Key words: Bayesian variable selection; hyper-Lasso; non-convexity; normal-exponential-gamma; oracle property; penalized likelihood.

1. Introduction

Variable selection in regression has a number of purposes: to provide regularization for good estimation of effects, to provide good prediction and to identify clearly important variables. With the advent of modern instrumentation, very many variables are provided routinely. Often, though, there will be relatively few observations. For example, in functional genomics, microarray chips typically have as many as tens of thousands of genes spotted on their surface, and their behaviour may be investigated over perhaps 100 or so samples. Curve-fitting in proteomics and other application areas may involve an arbitrarily large number of variables, being limited only by the resolution of the instrument. In such circumstances, it is often desirable to be able to restrict attention to the few most important variables by some form of adaptive variable selection. Consequently, there is renewed interest in providing fast and effective algorithms for sifting through these many variables.

Classical subset selection procedures are usually computationally too time-consuming and, perhaps more importantly, suffer from inherent instability (Breiman 1996). Bayesian stochastic search variable selection methods have become increasingly popular, often adopting the 'spike and slab' prior formulation of Mitchell & Beauchamp (1988). George & McCulloch (1997), West (2003), Ishwaran & Rao (2005) and Fahrmeir, Kneib & Konrath (2010) explore their use for univariate responses, and Brown, Vannucci & Fearn (1998) consider multivariate

*Author to whom correspondence should be addressed.
School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury CT2 7NF, UK.
e-mail: J.E.Griffin-28@kent.ac.uk; Philip.J.Brown@kent.ac.uk

Acknowledgments. Our thanks go to Dr H. Kiiveri for suggesting the idea of perfectly fitting starting values. We would also like to thank two referees and the editors for suggestions that led to an improved presentation.

extensions. Model averaging to induce stability in the more-variables-than-observations case ($k \gg n$) has been developed by Brown, Vannucci & Fearn (2002), simultaneously allowing variables to be discarded and averaging only over selected variables. Despite the careful use of algorithms to speed up computations, these approaches are still too slow to deal with the vast numbers of variables of order 10 000 or even 100 000 with SNPs in genomics, and some form of pre-filtering is necessary.

One Bayesian-inspired approach that does offer the potential for much faster computation takes a continuous form of prior and looks merely for modes of the posterior distribution (which is often termed Maximum *a posteriori* or MAP estimation) rather than relying on Markov Chain Monte Carlo, as would be required by a full Bayesian analysis. Such formulations lead to penalized log-likelihood approaches, where the additive penalization of the log-likelihood is the log of the prior distribution. In this formulation, the Lasso estimator of Tibshirani (1996) is equivalent to a double exponential prior distribution, proposed in Bayesian wavelet analysis by Vidakovic (1998). A more extreme form of penalty is the normal-Jeffreys prior suggested by Figueiredo & Jain (2001) and Figueiredo (2003) and adopted in an extended generalized linear model setting by Kiiveri (2003). A power variant of the normal-Jeffreys prior is adopted by ter Braak (2006), leading to a proper posterior distribution in a fully Bayesian analysis. A fully Bayesian analysis of the Lasso-implied prior is given by Park & Casella (2008). We will propose a Bayesian model that extends the double exponential (Lasso) prior and uses MAP estimation. This allows us to use standard results about penalized maximum likelihood estimators to understand the properties of our estimator.

The Lasso has been a popular choice of penalty function whose frequentist properties have been extensively studied. It provides estimates that are sparse, as some regression coefficients will be estimated as zero. There has been interest in its performance for prediction and variable selection. Zou (2006) shows that in some circumstances the Lasso may be inconsistent for variable selection. Meinshausen & Bühlmann (2006) also discuss the conflict of optimal prediction and consistent variable selection in the Lasso. They prove that the optimal Lasso shrinkage parameter gives inconsistent variable selection results, with many noise features included in the predictive model. Zhao & Yu (2006) discuss conditions on the regressor correlation structure for consistent Lasso estimation of the true model. Wainwright (2009) establishes conditions on the sample size for accurate estimation of the sparsity structure with a given number of regressors and non-zero regression coefficients. The problem of potential inconsistency has been addressed by a number of authors. Fan & Li (2001) modified the Lasso's L_1 penalty to offer less shrinkage for large effects by defining their Smoothly Clipped Absolute Deviation penalty (SCAD). They show that the Lasso property of having exactly zero coefficients at a posterior mode requires the penalty to be singular at the origin. They discuss the 'oracle' property, whereby knowing beforehand which coefficients should be set to zero does not improve estimation asymptotically, and provide conditions on the penalty function (in their theorem 2). To address the problems of the Lasso, Zou (2006) proposes an adaptive Lasso whereby the penalties on coefficients are weighted differently according to consistent estimators of the regression coefficients. This places great demands on the data in relatively small samples, which can lead to poor predictive and model selection performance. We develop an alternative form of adaptive penalty in Section 3, taking into account the uncertainty in the estimated scales. The adaptive aspect is achieved by a further layer of the hierarchy using a prior distribution and averaging over the uncertainty. In a textbook

statistical setting, this is the general argument in favour of random effects models as opposed to multiple fixed effects. Furthermore, because the common hyperparameters are estimated by cross-validation from the data, it *borrows strength* between the regression coefficients whilst retaining sparsity and good asymptotic properties. The asymptotic properties of this estimator can be studied using the results of Fan & Li (2001) and Fan & Peng (2004) for penalized maximum likelihood estimators, and we show that our estimator can have the oracle property.

As noted by Zou & Hastie (2005), the Lasso will select at most n non-zero parameters. These authors also distinguish strictly convex penalties and the non-strictly convex Lasso penalty, which may consequently lead to a continuum of solutions. The literature has concentrated on convex penalized likelihoods, but our Bayesian priors involve non-convex penalties and penalized likelihoods and their consequent multiple solutions. We will explore these by randomly generating perfectly fitting starting values. It may be argued that it is artificial to demand a single solution to a problem that is inherently indeterminate, although we find it is often easy to find one very good estimator, avoiding the need to utilize the multiple solutions, and we generally do this. Because there are no probabilistic relationships between nodes, it is hardly natural to form a single estimator by any form of model averaging.

In Section 2 we show the connection between Bayesian MAP estimation and penalized maximum likelihood estimation, and review some results about penalized maximum likelihood estimators. In Section 3.1, we consider using scale mixtures of normal prior distributions for the regression coefficients and describe the normal-exponential-gamma prior and associated penalty. In Section 3.2, we compare shrinkage and selection of our preferred choice with more standard alternatives, and we verify its oracle property in Section 3.3. In Section 4 we implement MAP estimation for this class of priors through an EM algorithm for exploring the posterior modes using the singular value decomposition for reduced dimension computation. We show how alternative subsets can be fitted through multiple random perfectly fitting starting points, when k, the number of variables, is greater than n, the number of observations. Section 5 shows through a large-parameter-space systematic simulation study that our normal-exponential-gamma penalty has an excellent small-sample mean square error and false discovery rate performance compared with its competitors. This emphasizes that asymptotic features such as the oracle property are shared by many estimators and, whilst important, are not sufficient for good small-sample behaviour. Our normal-exponential-gamma penalty is oracle-compliant but in addition has its Bayesian adaptive roots to promote good small-sample behaviour. A simple spline function simulation example and a real chemometric large-k example follow. Some concluding remarks are made in Section 6.

2. Bayesian penalization

Throughout this paper we will be concerned with standard multiple regression with Gaussian errors, although it will become clear that generalization to other exponential family models is straightforward.

We assume that the explanatory variables have been centred and that any desired scaling of these variables has been undertaken. It should be noted that automatic scaling, by subtracting the mean and dividing by the standard deviation, may not be desirable when the variables are measured on the same scale, as it tends to inflate the relative importance of variables that

change little over the data. We assume that

$$\boldsymbol{y} = \theta \mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{y}$ is a $(n \times 1)$-dimensional response vector, $\mathbf{1}$ is a $(n \times 1)$-dimensional vector of 1s, $\boldsymbol{X}$ is a $(n \times k)$-dimensional matrix of regressors, $\boldsymbol{\beta}$ is a $(k \times 1)$-dimensional vector of regression coefficients and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ is a vector of independent $\mathrm{N}(0, \sigma^2)$ random variables. We do not restrict k to be less than n. The values in $\boldsymbol{y}$ are mean-corrected throughout, and we implicitly assume a vague prior for θ so that in effect we replace it by the sample mean $\bar{y}$ and remove θ from model (1) for estimation purposes. At first we will assume that σ^2 is known; later its estimation is finessed by absorbing it into the scale for the prior for $\boldsymbol{\beta}$ and using cross-validatory estimation.

The Bayesian approach takes a prior distribution $\pi(\boldsymbol{\beta})$ and uses the posterior distribution

$$\pi(\boldsymbol{\beta}|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})$$

for inference. If the mode is our summary of the posterior distribution and the $\beta_1, \beta_2, \ldots, \beta_k$ are i.i.d. *a priori* then the problem reduces to finding $\hat{\boldsymbol{\beta}}$ that minimizes

$$-\log \pi(\boldsymbol{y}|\boldsymbol{\beta}) - \sum_{i=1}^{k} \log \pi(\beta_i),$$

and $\hat{\boldsymbol{\beta}}$ will be termed the MAP estimator. This estimator coincides with the penalized maximum likelihood estimator (PMLE) with penalty function $p(\beta_i) = -\log \pi(\beta_i)$, and in the linear regression case we have

$$L = \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \sum_{i=1}^{k} p(|\beta_i|).$$

The form of penalty function will have implications for the shrinkage of the regression coefficient towards zero. If we have one regressor it is straightforward to show that the relationship between the PMLE $\tilde{\beta}$ and the MLE $\hat{\beta}$ is given by

$$\frac{\hat{\beta} - \tilde{\beta}}{\sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}} = \mathrm{sign}(\tilde{\beta})p'(|\tilde{\beta}|), \tag{2}$$

where $p'(\cdot)$ is the derivative of the penalty function and $\sigma(\boldsymbol{X}^\top\boldsymbol{X})^{-1/2}$ is the standard error of $\hat{\beta}$. The amount of shrinkage is directly controlled by the derivative of the penalty function. An extreme form of shrinkage results in $\tilde{\beta} = 0$, a form of variable selection. Fan & Li (2001) use equation (2) to show that the PMLE is zero if

$$|\hat{\beta}| < \min_{\zeta \neq 0}(|\zeta| + \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}p'(|\zeta|)),$$

which defines the threshold for the MLE below which the PMLE is zero. The dependence on the derivative of the penalty function (i.e. $-\log(\text{prior})$) also arises from robustness considerations in Li & Goel (2006). In contrast, with a fully Bayesian analysis rather than a Bayesian modal analysis, the shrinkage dependence is on the derivative of the log predictive probability density function (see the general result of Griffin & Brown 2010).

3. Bayesian hyper-Lasso

3.1. Derivation of the normal-exponential-gamma (NEG) penalty

The Lasso penalty $p(\beta) = \lambda|\beta|$ coincides with the mean-zero double exponential distribution, DE(0, λ), with probability density function

$$\frac{\lambda}{2}\exp(-\lambda|\beta|), \qquad -\infty < \beta < \infty,\ 0 < \lambda < \infty.$$

We will generalize the double exponential distribution to provide a MAP estimator and study its properties. The double exponential distribution can be simply generalized using its representation as a scale mixture of normals (see Andrews & Mallows 1974),

$$\pi(\beta_i) = \int \mathrm{N}(\beta_i|0, \psi_i)\, G(d\psi_i), \tag{3}$$

where $\mathrm{N}(Y|\theta, \sigma^2)$ denotes the probability density function of a random variable Y having a normal distribution with mean θ and variance σ^2, and G is a mixing distribution whose density, if it is defined, will be referred to as $g(\cdot)$. The double exponential prior distribution occurs when we have an exponential mixing distribution with probability density function

$$g(\psi_i) = \frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2\psi_i}{2}\right).$$

One possible generalization replaces the exponential mixing distribution by the gamma distribution $\mathrm{Ga}(\nu, \lambda^2/2)$, which has density

$$\frac{\lambda^{2\nu}}{2^\nu\Gamma(\nu)}\psi_i^{\nu-1}\exp\left(-\frac{\lambda^2\psi_i}{2}\right),$$

leading to the normal-gamma (NG) or alternatively named variance gamma (VG) of Bibby & Sorensen (2003). This has been used in a fully Bayesian analysis for regression problems by Griffin & Brown (2010).

Our preferred generalization of the Lasso prior is the normal-exponential-gamma (NEG) distribution, formed by allowing the parameter of the exponential mixing distribution to vary from coefficient to coefficient according to a gamma distribution. The prior can be written hierarchically as

$$\beta_i \sim \mathrm{DE}\big(0, (2\phi_i)^{1/2}\big), \qquad \phi_i \sim \mathrm{Ga}\left(\nu, \frac{1}{\lambda^2}\right).$$

The NEG distribution can be expressed in the form of (3) by using

$$g(\psi_i) = \nu\lambda^2\big(1 + \lambda^2\psi_i\big)^{-(\nu+1)} \qquad 0 < \nu, \lambda < \infty.$$

The mixing distribution is a subclass of the gamma-gamma distribution (Bernardo & Smith 1994, p. 120) or compound gamma distribution (Johnson, Kotz & Balakrishnan 1994, p. 381), which we will refer to as the exponential-gamma (EG) distribution. The density of the marginal distribution of β_i, the NEG probability density function, can be expressed (Gradshteyn &

Ryzik 1980, p. 319) as

$$\pi(\beta_i) = \frac{\nu}{\pi^{1/2}} 2^{\nu} \lambda \Gamma(\nu + 1/2) \exp\left(\frac{1}{4}\lambda^2 \beta_i^2\right) D_{-2(\nu+1/2)}(\lambda|\beta_i|),$$

as derived in the appendix of Griffin & Brown (2005). Computation of this function is described in Zhang & Jin (1996, section 13.5.1, p. 439). Coded versions are available from http://jin.ece.uiuc.edu/routines/routines.html for Fortran 77 and http://ceta.mit.edu/comp_spec_func/for Matlab. If ν is small, the computation of $\exp(z)D_{\nu}(z)$ is much more stable than computation of $D_{\nu}(z)$. This involves a simple modification of the method described in Zhang & Jin (1996).

The parameters λ and ν control the scale and the heaviness or shape of the tails, respectively. From Abramowitz & Stegun (1964, p. 689, equation 19.8.1) we see that for large $\lambda|\beta_i|$,

$$\pi(\beta_i) \approx c(\lambda|\beta_i|)^{-(2\nu+1)},$$

where c is a constant depending on λ and ν. Thus if $\nu = 0.5$ the distribution has the same tail behaviour as a Cauchy distribution. Furthermore, if $\nu > 1$, the expectation of ψ_i, that is the variance of β_i, exists and has the form $(\nu - 1)^{-1}\lambda^{-2}$. The excess kurtosis is $3\nu(\nu - 2)^{-1}$ if $\nu > 2$. This class of distributions can define distributions for which the variance is undefined ($\nu \leq 1$) and thus has a tail-to-spike balance that can be concentrated around zero and yet have fat tails. The distribution of β_i is singular at zero with a mode that is finite for all parameter values.

Although the emergence of parabolic cylinder functions may seem unappetizing, the distribution has precedents in the literature when $\nu = 0.5$. In fact, Johnstone & Silverman (2005) define a *quasi-Cauchy* that is exactly the NEG distribution with $\nu = 0.5$, although their derivation in terms of beta random variables is distinct but actually equivalent. Furthermore, Berger (1985, section 4.7.10) defines a robustness prior for which the NEG distribution with $\nu = 0.5$ again exactly corresponds to the univariate case of his multivariate prior. The Cauchy form of tail behaviour was also derived by Jeffreys (1961, section 5.2) in connection with hypothesis testing for a normal mean, with the requirement that one observation should give an indecisive result. The marginal distribution of β_i for the quasi-Cauchy special case also avoids the need for parabolic cylinder functions. Using integration by parts and equation 3.362 (2) from Gradshteyn & Ryzik (1980, p. 315), we obtain the probability density function of the NEG distribution for $\nu = 1/2$ as

$$\pi(\beta_i) = \lambda(2\pi)^{1/2}\left[1 - \frac{(\lambda|\beta_i|)\{1 - \Phi(\lambda|\beta_i|)\}}{\phi(\lambda|\beta_i|)}\right], \tag{4}$$

where $\phi(.)$ and $\Phi(.)$ are the probability density function and cumulative density function of the standard normal distribution. This form is also given as (13) in Johnstone & Silverman (2005).

Another (improper) prior that has been used in MAP estimation in regression is the normal-Jeffreys (NJ) prior of Figueiredo & Jain (2001), which has been used by Kiiveri (2003), and, in a power variant, by ter Braak (2006). It arises from the choice of an improper hyper-prior $g(\psi_i) \propto 1/\psi_i$ in (3), which in turn induces an improper prior for β_i of the form $\pi(\beta_i) \propto 1/|\beta_i|$ and thus the penalty $\sum \log|\beta_i|$. This prior and consequent penalty is the

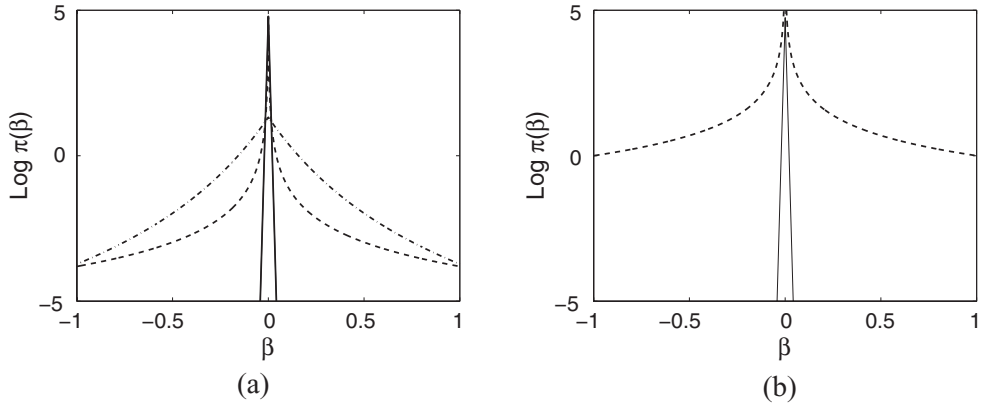

Figure 1. Log prior densities setting the central region $(-0.01, 0.01)$ to have probability $\eta = 0.9$ for (a) a double exponential distribution (solid line), NEG ($\nu = 1$) (dashed line) and NEG($\nu = 0.1$) (dotted line), and (b) a double exponential (solid line) and improper normal-Jeffreys (dashed line).

limiting case of the NEG when $\nu \to 0$, and, as we shall see later, the extreme form of penalization leads to very sparse solutions. However, it has some theoretical limitations, and the NEG can obtain better properties by including a shape and scale parameter.

Our reason for choosing the NEG is two-fold: it has a finite spike at zero for all parameter values and it has increasingly heavy tails as ν becomes smaller. A spike is useful for finding sparse solutions, but a finite spike at zero is necessary to have a well-defined MAP estimator. If the spike at zero is infinite, as with the NJ and NG ($\nu \leq 1/2$), the MAP estimator is always the origin, irrespective of the data. Heavy tails are important to avoid over-shrinkage of regression coefficients with large absolute values. The tail-to-spike behaviour is illustrated in Figure 1 for NEG, DE and NJ. For purposes of comparison, we specify one scale parameter by fixing the probability mass on the central region $(-\epsilon, \epsilon)$ to be η (except for the NJ, which has no scale parameter). The figure illustrates the effect of fixing $\eta = 0.9$ on the region $(-0.01, 0.01)$ for the two comparisons with the Lasso: (a) DE versus NEG and (b) DE versus NJ. The NEG distribution is able to maintain flat tails with a large preponderance of density around zero. It seems that the DE and NJ are at opposite extremes from the NEG, preserving good features of the NJ without the drawback of the extreme spike at zero.

In the next section we characterize the threshold properties of the NEG and some of its competitors in the special case of one parameter, or equivalently in general regression when the $\boldsymbol{X}^\top \boldsymbol{X}$ matrix is diagonal, i.e. the columns of $\boldsymbol{X}$ are orthogonal.

3.2. Properties

Various prior distributions, expressed in terms of their implied penalty function and its derivatives, are listed in Table 1. In the case of the NG distribution, $K_\nu(\cdot)$ is a Bessel function of the third kind.

It is illuminating to compare the set of $|\hat{\beta}|$ that will give PMLE estimates of exactly zero for various choices of the prior distribution. This threshold is defined by equation (2) when there is one regression parameter, $k = 1$. For the DE prior distribution, the threshold is $|\hat{\beta}| < \lambda\sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, which depends on the square of the standard error, and so the threshold

TABLE 1

Penalty functions and their derivatives induced by various choice for the prior distribution.

	$p(\beta)$	$p'(\lvert\beta\rvert)$
Double exponential$(0, \lambda)$	$\lambda\lvert\beta\rvert$	λ
NJ	$\log\lvert\beta\rvert$	$\frac{1}{\lvert\beta\rvert}$
Normal-gamma	$(1/2-\nu)\log\lvert\beta\rvert - \log K_{\nu-1/2}(\lambda\lvert\beta\rvert)$	$\lambda\frac{K_{\nu-3/2}(\lambda\lvert\beta\rvert)}{K_{\nu-1/2}(\lambda\lvert\beta\rvert)}$
NEG	$-\frac{\lambda^2\beta^2}{4} - \log D_{-2(\nu+1/2)}(\lambda\lvert\beta\rvert)$	$\lambda(2\nu+1)\frac{D_{-2(\nu+1)}(\lambda\lvert\beta\rvert)}{D_{-2(\nu+1/2)}(\lambda\lvert\beta\rvert)}$

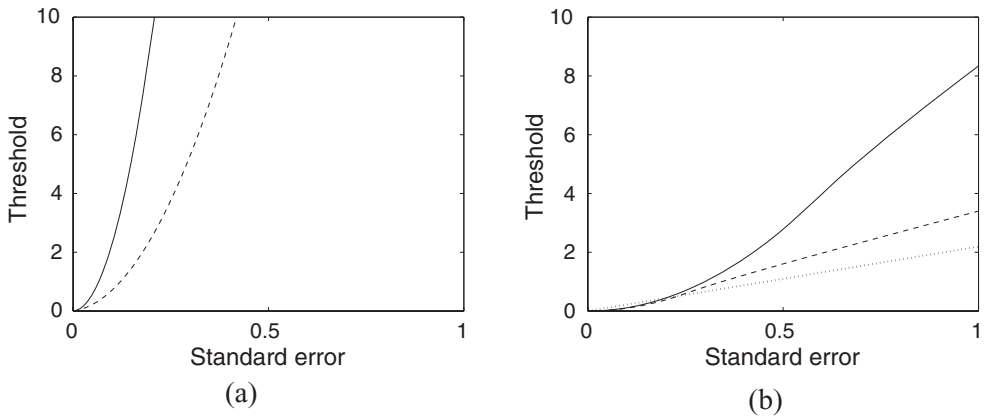

Figure 2. Threshold values as a function of the standard error under various prior choices with $\eta = 0.9$ and $\epsilon = 0.01$: (a) DE (solid line) and NG ($\nu = 0.1$) (dashed line), and (b) NEG distributions with $\nu = 10$ (solid line), $\nu = 1$ (dashed line) and $\nu = 0.1$ (dotted line).

interval shrinks at an uncomfortably fast rate of n^{-1}. In contrast, the NJ prior thresholds according to the rule $|\hat{\beta}| < 2\sigma(\boldsymbol{X}^\top\boldsymbol{X})^{-1/2}$ and the threshold depends linearly on the standard error, that is, at rate $n^{-1/2}$. Remarkably, the 2-multiplier that pops out is rather close to 1.96 for a single 5% normal test value. Figure 2 compares the threshold rules for the NG penalty and the NEG penalty. The latter has linear behaviour where the slope depends on ν, generalizing the NJ rule, and is thus more appealing. The NG case has substantially different behaviour and defines a much more conservative criterion. Much smaller values of λ would induce a linear threshold rule, but this contradicts our imposed prior property of a large mass close to zero.

3.3. The oracle property

The oracle property is important for penalized maximum likelihood estimation and says that the PMLE has the same asymptotic properties as the MLE calculated using only the regressors with non-zero regression coefficients under the true model. Only estimators that are sparse, in the sense that some regression coefficients are estimated to be zero, can have the oracle property. Fan & Li (2001) establish conditions on the penalty function that imply both the oracle property and sparsity for the PMLE if $k < n$. Fan & Peng (2004) extended these results to the case where the number of parameters increases with the sample size n

and is written k_n. Suppose that the penalty function is parametrized explicitly in terms of a scale parameter λ that has the property that $p_\lambda(\beta) \to 0$ as $\lambda \to 0$ for all values of β. Fan & Peng (2004) allow the scale parameter to depend on sample size, written λ_n, and establish conditions for the penalty function and likelihood that ensure that the PMLE has the oracle property, as well as asymptotic normality of the non-zero estimates with a covariance matrix that can be estimated using a sandwich formula and the validity of likelihood ratio test. We restrict attention to checking their conditions on the penalty function. Suppose that β_{n0j} are the true values of the parameters when the sample size is n. Let

$$a_n = \max_{1 \le j \le k_n} (p'_{\lambda_n}(|\beta_{n0j}|), \beta_{n0j} \neq 0)$$

and

$$b_n = \max_{1 \le j \le k_n} (p''_{\lambda_n}(|\beta_{n0j}|), \beta_{n0j} \neq 0),$$

then the conditions are

(A) $\liminf_{n \to +\infty} \liminf_{x \to 0+} \left(p'_{\lambda_n}(x)/\lambda_n \right) > 0$
(B) $a_n = O(n^{-1/2})$
(B*) $a_n = o((nk_n)^{-1/2})$
(C) $b_n \to 0$ as $n \to +\infty$
(C*) $b_n = o_p(k_n^{-1/2})$
(D) there are constants C and D such that, when $x_1, x_2 > C\lambda_n$, $|p''_{\lambda_n}(x_1) - p''_{\lambda_n}(x_2)| \le D|x_1 - x_2|$.

The oracle property, asymptotic normality of the non-zero regression coefficient estimates and validity of likelihood tests follow if conditions (A)–(D) are met and if $\lambda_n \to 0$, $(nk_n^{-1})^{1/2}\lambda_n \to \infty$ and $(k_n^5 n^{-1})^{1/2} \to 0$ as $n \to \infty$. The final conditions imply that k_n grows with n but not as fast as n.

The Lasso does not possess the oracle property as it does not satisfy condition (A). However, the NEG penalty can be shown to meet conditions (A)–(D) under the further condition (H) of Fan & Peng (2004), which states that if $\beta_{n01}, \ldots, \beta_{n0s_n}$ are the only non-zero regressors then

$$\min_{1 \le j \le s_n} \left(\frac{|\beta_{n0j}|}{\lambda_n} \right) \to \infty \text{ as } n \to \infty.$$

The first derivative of the NEG penalty function is given in Table 1, and the second derivative is

$$(p'(\lambda|\beta|))^2 - 2\lambda^2(2\nu + 1)(\nu + 1)\frac{D_{2(\nu+3/2)}(\lambda|\beta|)}{D_{2(\nu+1/2)}(\lambda|\beta|)}.$$

The first and second derivatives are monotone, implying that $a_n = p'_{\lambda_n}(\beta_n^{\min})$ and $b_n = p''_{\lambda_n}(\beta_n^{\max})$, where

$$\beta_n^{\min} = \min_{1 \le j \le k_n} (|\beta_{n0j}|, \beta_{n0j} \neq 0)$$

and

$$\beta_n^{\max} = \max_{1 \le j \le k_n} (|\beta_{n0j}|, \beta_{n0j} \neq 0).$$

Then, clearly under condition (H), $a_n \to 0$, $b_n \to 0$ as $n \to 0$. Condition (D) follows from noticing that $p''_\lambda(x) \to 0$ as $x \to \infty$. These conditions also imply that we can if we wish form a 'sandwich' estimator of the covariance matrix of the estimator to give standard errors in much the same way as Fan & Peng (2004).

Unfortunately, the NJ penalty does not have the oracle property, since $a_n = 1/|\beta_n^{\min}|$. Theorem 1 of Fan & Peng (2004) shows that this implies that, if β_{n0} are the true regression coefficients, then $\| \hat{\beta}_n - \beta_{n0} \| = O_p(k_n^{1/2}(n^{-1/2} + a_n))$ and the normal-Jeffreys' penalty can lead to a slow rate of convergence. The NG penalty can also have the oracle property. Whilst the oracle property is a notable asymptotic property, very many estimators can achieve it, so small-sample properties become crucial in choosing between them.

4. Implementing regression

In order to explore both inference aspects and algorithms for inference when the number of parameters may exceed the number of observations, we develop an EM algorithm and show how to create multiple starting values that fit the data perfectly. These can be used to explore alternative modes in the multi-modal posterior. We would usually find a set of local modes by starting the EM algorithm at a fixed number of random starting points and report the local mode with the highest posterior density as the MAP.

4.1. An EM algorithm to find a posterior mode of $\boldsymbol{\beta}$

Local posterior modes can be found using the EM algorithm (Dempster, Laird & Rubin 1977; Meng & van Dyk 1997), which has been suggested by both Figueiredo (2003) and Kiiveri (2003) as a means of fitting models using scale mixture of normal priors. In general, we use the EM algorithm to find a promising and small subset of variables with non-zero regression coefficients. In our case, the prior variances of the regression coefficients $\psi_1, \ldots, \psi_k$ are treated as missing data and imputed in the *E-step*, and the *M-step* maximizes over $\boldsymbol{\beta}$ conditional on these imputed variances. The algorithm could be extended to estimate an unknown σ^2 by including it in either the E- or the M-step. In the generalized linear model setting, Kiiveri (2003) applies the EM algorithm directly to the 'likelihood times prior'. We avoid the term 'posterior' here as this does not exist for the NJ prior used. The M-step is approximated by a Newton–Raphson line search for the MLE of $\boldsymbol{\beta}$ conditional on imputed variances, and the algorithm is started from a ridge regression estimate. In the normal linear regression case no approximations are necessary.

The standard EM algorithm outputs a sequence of estimates $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots$ that under regularity conditions converge to a local maximum of the posterior density of $\boldsymbol{\beta}$. The sequence is defined by iterating between an E-step, which for us averages over $\boldsymbol{\psi}$ for given $\boldsymbol{\beta}$, and an M-step, which maximizes over $\boldsymbol{\beta}$ for given $\boldsymbol{\psi}$. If the rank of $\boldsymbol{X}$ is r, the singular value decomposition of the centred design matrix is given by

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{D}\boldsymbol{A}^\top, \tag{5}$$

where $\boldsymbol{A}$ is a $(k \times r)$-dimensional matrix such that $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}_r$, with the columns of $\boldsymbol{A}$ being the r eigenvectors of $\boldsymbol{X}^\top \boldsymbol{X}$ corresponding to non-zero eigenvalues, $\boldsymbol{D}$ is an $(r \times r)$-dimensional diagonal matrix, and $\boldsymbol{F}$ is $(n \times r)$-dimensional matrix whose columns are the r eigenvectors of $\boldsymbol{X}\boldsymbol{X}^\top$ corresponding to non-zero eigenvalues and for which $\boldsymbol{F}^\top \boldsymbol{F} = \boldsymbol{I}_r$. The decomposition is computed before starting the EM algorithm and can be used to define an algorithm that only involves inversion of matrices of dimension at most $\min(n-1, k) \geq r$.

(i) **E-step**: Let

$$\psi_j^{(i)} = \frac{1}{\mathrm{E}\left(\psi_j^{-1} | \boldsymbol{\beta}^{(i-1)}\right)} = \frac{p'\left(\left|\beta_j^{(i-1)}\right|\right)}{\left|\beta_j^{(i-1)}\right|}, \quad j = 1, \ldots, k.$$

The derivatives $p'(|\beta|)$ are given in Table 1.

(ii) **M-step**: Set $\boldsymbol{\beta}^{(i)} = \boldsymbol{\Psi}^{(i-1)} \boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{\Psi}^{(i-1)} \boldsymbol{A} + \sigma^2 \boldsymbol{D}^{-2})^{-1} \hat{\boldsymbol{\alpha}}$, where $\boldsymbol{\Psi}^{(i-1)}$ is a diagonal matrix with elements $\psi_1^{(i-1)}, \ldots, \psi_k^{(i-1)}$ and $\hat{\boldsymbol{\alpha}} = \boldsymbol{D}^{-1} \boldsymbol{F}^\top \boldsymbol{y}$. This form involves the inversion of $(r \times r)$-dimensional matrices. When $k \gg n$ these matrices will be very much smaller than the $(k \times k)$-dimensional matrices that would be needed using standard results.

Work in linear or generalized linear models using NJ prior distributions (Kiiveri 2003; Figueiredo 2003) usually focuses on finding a single 'good' mode. Bae & Mallick (2004) and Mallick, Ghosh & Ghosh (2005), however, aim for full posterior simulation using Markov Chain Monte Carlo, but in favouring the NJ overlook the fact that the likelihood times prior for this remains improper as the likelihood for β at zero is bounded away from zero and hence the behaviour in the region of zero is still proportional to β^{-1} and integrates to $\log(\beta)$, which blows up at $zero$. This precludes full Bayesian posterior analysis using the NJ prior but does formally allow it to act as a device for generating modes from the 'likelihood times prior' in the spirit of penalized likelihood. It is yet another reason for our preference for the NEG, which retains some of the attractions of NJ but without the dominating spike at zero.

In the next section we explore where we might start the algorithm to find well-fitting local modes that have sparse solutions in the sense of involving few variables.

4.2. Perfectly fitting random starting values

When the number of parameters k exceeds the rank r of X, minimum length least squares (MLLS) gives the unique estimator $\hat{\boldsymbol{\beta}}_{\mathrm{MLLS}} = (\boldsymbol{X}^\top \boldsymbol{X})^+ \boldsymbol{X}^\top \boldsymbol{y}$, where '+' denotes the Moore–Penrose generalized inverse. This will provide a perfectly fitting solution with typically all coefficient estimates non-zero. In fact there will be a $(k-r)$-dimensional null space in which we can start our EM algorithm, with all least squares solutions fitting perfectly. These serve as different perfectly fitting starting points. They enable us to generate a set of well-fitting solutions, some of which will correspond to different modes. If we want a single solution for mean square error calculations our strategy is to choose the best of these in terms of penalized likelihood.

The orthogonal projection matrix is $\boldsymbol{I} - \boldsymbol{P} = \boldsymbol{I}_k - \boldsymbol{A}\boldsymbol{A}^\top$, a matrix of rank $(k-r)$ derived from the singular value decomposition (5). Consider generating a random k-vector z and take $\boldsymbol{w} = (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^\top)z$, calculated as $z - \boldsymbol{A}(\boldsymbol{A}^\top z)$. If we add this projected random vector to $\hat{\boldsymbol{\beta}}_{\mathrm{MLLS}}$ then we will have the same minimum length least squares 'perfectly' fitting solution,

since $\boldsymbol{X}\boldsymbol{w} = \boldsymbol{0}$, as verified by

$$\begin{aligned}\boldsymbol{X}\boldsymbol{w} &= \boldsymbol{F}\boldsymbol{D}\boldsymbol{A}^{\top}(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^{\top})\boldsymbol{z} \\ &= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{A}^{\top} - \boldsymbol{A}^{\top})\boldsymbol{z} = \boldsymbol{0}.\end{aligned}$$

Thus we can add $\boldsymbol{w}$ to $\hat{\boldsymbol{\beta}}_{\mathrm{MLLS}}$ and obtain a 'perfectly' fitting starting point. We can repeat this as often as we like, or design the $\boldsymbol{z}$ to span the space. Typically the seed $\boldsymbol{z}$ would be generated as independent normal elements with zero means, and we choose a common variance that reflects the typical or near-largest of the variances in the sampling distribution of least squares $\hat{\boldsymbol{\beta}}$, as given by the Moore–Penrose generalized inverse. To this end we ordered the k components of $\hat{\boldsymbol{\beta}}$, $\hat{\beta}_{(1)} \leq \hat{\beta}_{(2)}, \ldots, \leq \hat{\beta}_{(k)}$, and the average of the largest from $\hat{\beta}_{([0.9k])}$ upwards. Other more graphical strategies could be sensible if, for example, there is distinct jump in size of the larger elements.

The approach above is mildly inefficient in the sense that it requires the generation of k random values when only $(k - r)$ are required to cover the space orthogonal to the least squares fit. A potential way around this is to calculate $\bar{\boldsymbol{A}}$, the $k \times (k - r)$ set of eigenvectors completing the set $\boldsymbol{A}$. Now suppose that we a have a random $(k - r) \times 1$ vector $\boldsymbol{u}$, then $\bar{\boldsymbol{A}}\boldsymbol{u}$ may be added to $\hat{\boldsymbol{\beta}}_{\mathrm{MLLS}}$ to achieve a 'perfectly' fitting starting point. This is easily seen since $\boldsymbol{X}\bar{\boldsymbol{A}}\boldsymbol{u} = \boldsymbol{F}\boldsymbol{D}\boldsymbol{A}^{\top}\bar{\boldsymbol{A}}\boldsymbol{u} = \boldsymbol{0}$, as the eigenvectors in $\boldsymbol{A}$ and $\bar{\boldsymbol{A}}$ are orthogonal. The lack of quick algorithms to generate $\bar{\boldsymbol{A}}$ may make this modified approach unattractive though, and it is not an approach that we have used.

5. A simulation study and two examples

A small-sample simulation study of some alternative methods is here systematically compared with the NEG in the $k \gg n$ setting. We then apply the NEG prior to an example of fitting to simulated data from a sine function with added error using a spline basis. This is followed finally by a real example involving prediction of the composition of biscuits.

5.1. Multiple regression simulation study

We have conducted a simulation study to compare a variety of estimation methods, including that using our preferred NEG prior. The Gaussian error regression model is simulated with error variance $\sigma^2 = 1$, and each of the explanatory variables for the ith observations, X_i, following an AR(1) process with lag 1 correlation $\rho = 0.5, 0.8$. The simulation has $n = 100$ observations, $k = 500, 2000, 10\,000$ variables, $k^* = 10$ non-zero coefficients of $\boldsymbol{\beta}$, with either all the non-zero coefficients $\beta = 1$, or $\beta = 5$, with extra points of $\beta = 10$ for the high dimensional case of $k = 10\,000$. All non-zero values were equally spaced in the k variable design. This design is similar in spirit to that of Zou & Hastie (2005), although more ambitious as regards dimensions. Hyper-parameters, for example ν, μ with $\mu = \nu\lambda^2$, closely related to the prior precision of the regression coefficients, were chosen by five-fold cross-validation.

The methods compared are as follows.

(i) The NEG prior, in versions with both parameters (ν, μ) chosen by cross-validation (ν is chosen from 0.1, 0.5, 1 or 2), or with ν fixed at 0.5 (leading to the quasi-Cauchy) and μ chosen by cross-validation.

TABLE 2

Mean squared errors for regression simulations with $n = 100$ observations and $k^* = 10$ non-zero coefficients β, error variance 1.0, best method in bold.

β	ρ	k	Lasso	SCAD	AL_{MLLS}	AL_{Ridge}	NEG	$NEG_{\nu=0.5}$	NJ
1	0.5	500	1.91	**1.08**	1.88	1.84	1.13	1.14	1.23
1	0.5	2000	2.01	3.38	2.06	2.03	**1.31**	1.59	2.05
1	0.5	10 000	2.85	3.17	2.82	3.44	**2.62**	2.70	3.33
1	0.8	500	1.84	**1.20**	1.91	1.92	1.27	1.27	**1.20**
1	0.8	2000	1.96	3.33	1.98	1.95	**1.51**	1.66	2.10
1	0.8	10 000	2.74	3.47	2.87	2.72	**2.61**	2.70	3.29
5	0.5	500	2.20	**1.07**	2.33	2.30	1.09	1.16	1.11
5	0.5	2000	2.32	15.97	2.41	2.32	**1.16**	1.25	3.10
5	0.5	10 000	7.96	16.28	8.10	8.36	7.18	**6.19**	10.94
5	0.8	500	2.34	**1.06**	2.49	2.65	1.11	1.12	1.13
5	0.8	2000	2.19	15.82	2.28	2.31	**1.20**	1.27	1.36
5	0.8	10 000	5.53	15.54	5.60	5.72	**2.87**	3.14	9.32
10	0.5	10 000	16.76	30.96	18.29	18.02	**9.52**	11.73	24.20
10	0.8	10 000	11.98	31.00	12.23	12.40	**7.33**	7.40	19.53

(ii) The Lasso with the penalty parameter estimated by cross-validatory choice.
(iii) The SCAD penalty with the penalty parameter estimated by cross-validatory choice.
(iv) The adaptive Lasso (AL) using either the MLLS with a Moore–Penrose generalized inverse or ridge (from a separate cross-validatory choice) for the estimate β_* in the construction of their adaptive weight function, $w = 1/|\beta_*|^\eta$, with η chosen as either 0.5, 1 or 2 (by cross-validation).
(v) The normal-Jeffreys (NJ) prior, which has no free parameters to estimate.

The Lasso and adaptive Lasso estimates are found using the 'lars' package in R, and the SCAD estimates are found using the 'ncvreg' package in R. We use two measures to compare the various penalty functions. The first concentrates on prediction performance by estimating the mean squared error (MSE) of the prediction of a further 100 observations. The second looks at variable selection performance. If we rate a non-zero regression coefficient as a 'discovery' then we can define the false discovery rate (FDR) as the number of false discoveries divided by the total number of discoveries. The expected values of these quantities were estimated by generating 20 independent datasets and taking the mean of the measures on each dataset.

The MSE results are given in Table 2. The best possible performance is close to 1 (the value of the error variance). The SCAD penalty is generally best for $k = 500$, and the NEG penalty is generally the best for larger numbers of variables, with NEG ($\nu = 0.5$), the quasi-Cauchy, not far behind. The adaptive Lasso performs generally no better than the Lasso. We surmise that this is due to consistent but otherwise inaccurate estimates of the differing Lasso parameters caused by the large number of regressors degrading the accuracy of MLLS or ridge regression. The NJ is surprisingly good given that it lacks adaptive flexibility with no hyper-parameters to estimate, but tends to break down when there are very many parameters ($k = 10\,000$).

Table 3 shows these false discovery rates averaged over the datasets generated for each simulation experiment. It is well known that the Lasso penalty generally makes a large number of discoveries, which leads to a poor false discovery rate. This can be seen in Table 4, where

TABLE 3

False discovery rate (percentage) for regression simulations with $n = 100$ observations and $k^* = 10$ non-zero coefficients β, error variance 1.0, best method in bold.

β	ρ	k	Lasso	SCAD	AL_{MLLS}	AL_{Ridge}	NEG	$NEG_{\nu=0.5}$	NJ
1	0.5	500	48	31	52	51	18	20	**17**
1	0.5	2000	68	47	67	67	39	47	**24**
1	0.5	10 000	76	69	75	68	**46**	46	NA
1	0.8	500	54	42	50	49	19	23	**14**
1	0.8	2000	74	60	72	75	49	59	**30**
1	0.8	10 000	81	72	80	80	**49**	53	NA
5	0.5	500	41	**0**	35	37	9	25	11
5	0.5	2000	67	**3**	63	67	18	45	9
5	0.5	10 000	76	**23**	82	75	34	43	55
5	0.8	500	47	**1**	45	45	12	15	12
5	0.8	2000	72	**0**	70	70	24	43	3
5	0.8	10 000	81	**7**	81	80	34	31	51
10	0.5	10 000	79	79	74	82	**33**	48	73
10	0.8	10 000	86	86	85	84	**26**	40	61

TABLE 4

Number of discoveries for regression simulations with $n = 100$ observations and $k^* = 10$ non-zero coefficients β, error variance 1.0.

β	ρ	k	Lasso	SCAD	AL_{MLLS}	AL_{Ridge}	NEG	$NEG_{\nu=0.5}$	NJ
1	0.5	500	23	17	29	24	13	13	13
1	0.5	2000	39	22	40	37	18	21	9
1	0.5	10 000	41	32	44	33	20	20	0
1	0.8	500	25	18	20	19	14	15	15
1	0.8	2000	44	26	38	47	20	26	9
1	0.8	10 000	45	29	44	50	19	22	0
5	0.5	500	21	10	16	18	11	16	16
5	0.5	2000	38	10	34	40	14	19	13
5	0.5	10 000	54	18	60	51	22	27	21
5	0.8	500	20	10	19	21	12	12	12
5	0.8	2000	42	10	39	41	15	19	10
5	0.8	10 000	60	11	57	55	20	20	19
10	0.5	10 000	56	56	51	63	19	30	38
10	0.8	10 000	74	74	68	64	21	25	33

it usually makes twice as many 'discoveries' as the NEG. The SCAD penalty function tends to perform better in simulations with $\beta = 5$, and the NJ and NEG when $\beta = 1$. The NEG typically performs best in the challenging example where $k = 10\,000$. We have found that the NEG usually makes more discoveries and may be preferred to the NJ, which is too stringent and breaks down with very high dimensions, getting lost in such low-information settings, without the ability to adapt through an extra scale parameter. This is backed up by the better mean square error performance of NEG in Table 2 for those entries where NJ has very low false discovery rates. In two places NJ makes no discoveries, and in these cases the FDR is designated as NA. The NEG does better or almost as well on FDR, but has better mean square error performance in corresponding instances where it fares worse.

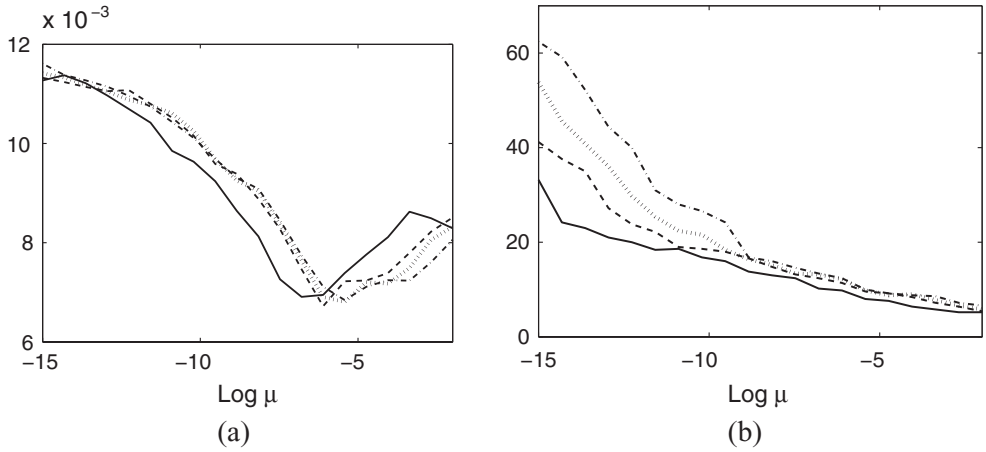

Figure 3. Spline fitting five-fold cross-validation with the NEG penalty for various values of ν and $\mu = \nu\lambda^2$. Panel (a) shows the average MSE, and (b) shows the average number of included variables. In both cases $\nu = 0.1$ (solid line), $\nu = 0.5$ (dashed line), $\nu = 1$ (dotted line) and $\nu = 2$ (dot–dashed line).

5.2. Spline simulation

Our second example applies regularization to the problem of fitting a curve using piecewise linear splines. This allows the visualization of regularizing effects using the estimated curve. We assume that the function can be well expressed in the form

$$f(x) = \sum_{i=1}^{k} \beta_i \max\{0, x - K_i\},$$

where $K_1, K_2, \ldots, K_k$ are knots points that are equally spaced in the interval (a, b) and so $K_i = a + \{(i-1)(k-1)\}/(b-a)$. We observe pairs (x_i, y_i), which is a noisy version of a function $f(x_i)$, and the problem of estimating β_i is a linear regression problem in a non-linear basis. Osborne, Presnell & Turlach (1998) applied a Lasso penalty to this problem, and we compare this approach with NJ and NEG penalization. If k is large, there will be substantial correlation between subsequent regressors, owing to the closeness of the knot points, which makes inference by regression methods a challenging problem. We fit $n = 30$ observations: x_i are uniformly distributed on (0, 1) and $y_i = \sin(2\pi x_i) + \epsilon_i$, where ϵ_i is drawn from a normal distribution with mean 0 and variance 0.01. We have $k = 500$ knot points between $a = -0.3$ and $b = 1.3$. The hyper-parameters of the Lasso and the NEG are estimated using five-fold cross-validation.

The results for the NEG penalization are illustrated in Figure 3 by the average MSE error for the test set (panel a) and the average number of non-zero estimates (panel b). The average MSE is mainly determined by the choice of $\mu = \nu\lambda^2$, and the average non-zero regressor decreases as μ is increased. For any value of μ, larger values of ν give more non-zero estimates.

The summary graphs for 20 perfectly fitting random starts are given in Figure 4 for the three penalty functions using parameters chosen by cross-validation. The number of regressors chosen by the three methods follows the order that we would expect. The Lasso finds a single

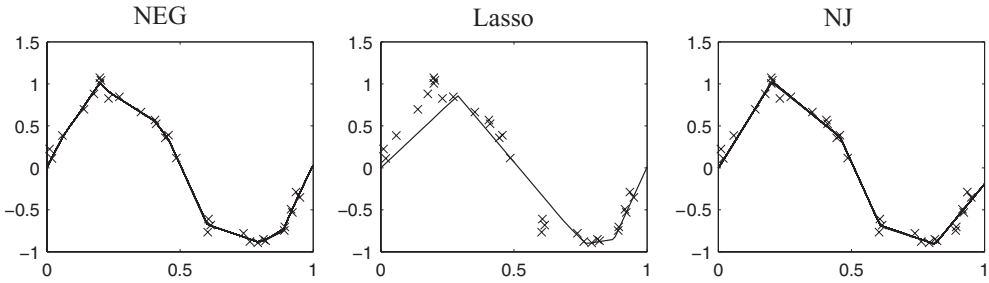

Figure 4. Spline fitting at 20 perfectly fitting random starts for the NJ, Lasso and NEG penalties using the hyper-parameters chosen by cross-validation. The fitted models for each of the 20 modes are shown overlaid by the observations.

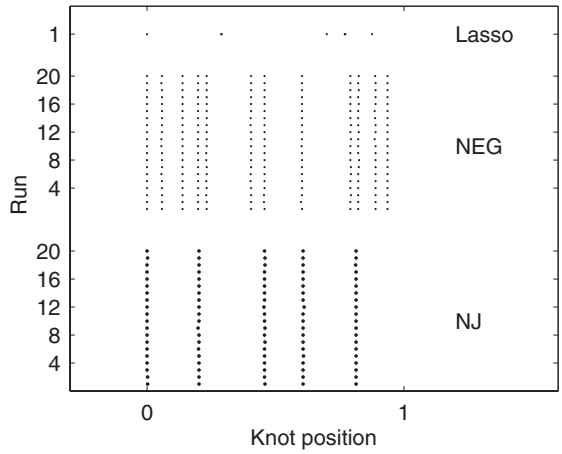

Figure 5. Spline-fitting knots for 20 perfectly fitting random starts for the NJ, Lasso and NEG penalties using the hyper-parameters chosen by cross-validation. The figures show the regression estimates (the area of the dots is proportional to the absolute value of the regression estimate).

set of 10 non-zero regression estimates, whereas the NEG finds around 8 and the NJ finds around 6.

Figure 5 shows the differences in the chosen knot points. The NEG and NJ find similar sets of knots, with the NJ choosing only one knot at around 0.45 compared with the NEG's usual two. Moreover, the NEG chooses two knots between 0.8 and 0.9 to the NJ's one. The Lasso estimate uses fewer knots at small values (below 0.1) and more knots between 0.1 and 0.5. This manifests itself by a poor fit between 0 and about 0.25 (Fig. 4b). The fits can be compared using the MSE of predicting a further 100 observations randomly generated in the same way as the data. The poor fit of the Lasso leads to a MSE of 0.0608, which is much larger than that for the NEG (0.0140) and the NJ (0.0163).

5.3. Biscuit NIR data

The data are taken from Osborne *et al.* (1984) and were used in Brown, Fearn & Vannucci (2001), where the data set-up is described in some detail. The predictor variables

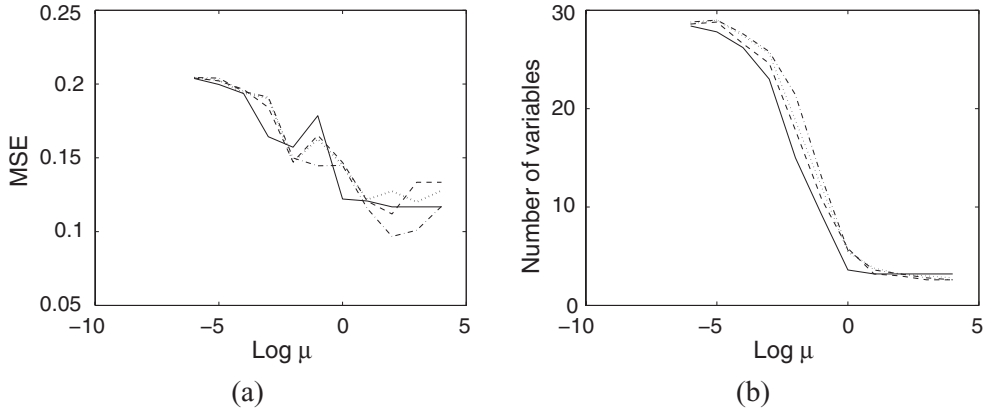

Figure 6. Biscuit data five-fold cross-validation with the NEG penalty for various values of ν and $\mu = \nu\lambda^2$. Panel (a) shows the average MSE, and (b) shows the average number of included variables. In both cases $\nu = 0.1$ (solid line), $\nu = 0.5$ (dashed line), $\nu = 1$ (dotted line) and $\nu = 2$ (dot–dashed line).

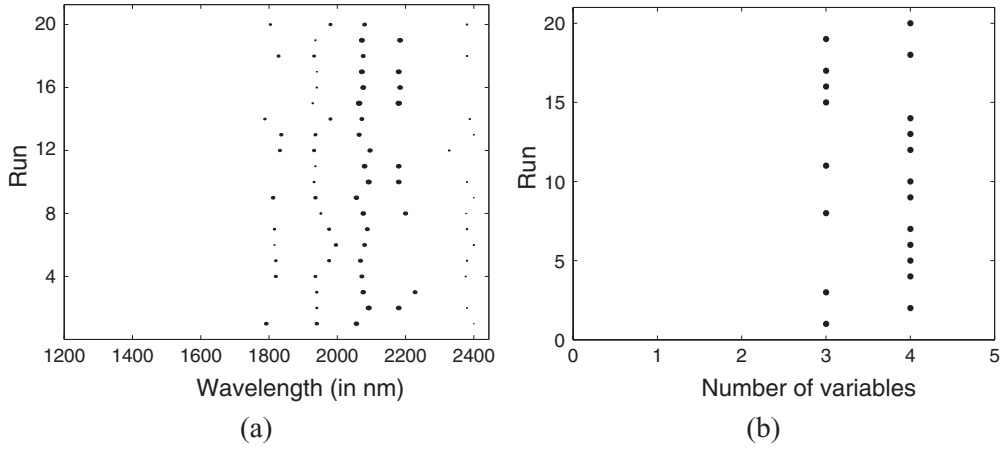

Figure 7. Biscuit data: 20 perfectly fitting random starts for the NEG penalties using the hyper-parameters chosen by cross-validation. Panel (a) shows the regression estimates (the area of the dots is proportional to the absolute value of the regression estimate) and (b) shows the number of wavelengths selected in each run.

are measurements of the NIR reflectance spectrum of biscuit dough pieces and the amount of fat, flour, sugar and water that each piece contains. There are 39 samples in the training data and 31 in the final validation set. We have reduced and thinned the reflectance spectra to 300 wavelengths from 1202 to 2400 nm in steps of 4 nm. The hyper-parameter values of the NEG penalty are chosen using five-fold cross-validation on the training sample. The variables are not scaled, as it is important not to change the relative scales of reflectance at different wavelengths as this would promote reflectances that are very small and may be largely noise; see also comments in Section 2. The response Y chosen was the flour content, which was also centred and scaled by its standard deviation over the 20 samples. These standardizations help numerical stability and allow easy interpretation of fit.

Here the real data has the error variance σ^2 to contend with, but we are able to finesse its estimation by absorbing it into the prior scale λ of the regression coefficients and hence into hyper-parameter μ. The hyper-parameters ν, μ are selected by cross-validation averaging over five splits, and the results are shown in Figure 6(a). Figure 6(b) gives the parallel effect on the number of wavelengths chosen. The hyper-parameters values chosen were $\nu = 1$, $\mu = 100\,000$. The results of finding estimates using the NEG penalized likelihood with these hyper-parameters over 20 perfect random starts are depicted in Figure 7. Each mode found has three or four wavelengths with non-zero regression coefficients. Most modes include a wavelength around positions 1920 and 2080 nm. Three further regions are identified by some of the modes around 1800, 2200 and 2400 nm.

The average MSE on the validation set (31 observations) is 0.0565 (94% explained), which is competitive with that achieved in Brown *et al.* (2001) via full MCMC and a 'slab-and-spike' prior.

6. Conclusions

We have developed an adaptive hyper-Lasso motivated by a hierarchical Bayesian framework. The Lasso itself is unable to do well simultaneously in (a) prediction and (b) identification of significant variables. This can be viewed as a problem of its inflexibility in 'tail-to-spike' behaviour with one parameter (a scale parameter) fits all. Our normal-exponential-gamma prior has two parameters for flexibility. While estimation is more sensitive to mis-specification of the scale than of the shape, extreme mis-specification of the shape (with the Lasso) can lead to serious consequences for estimation accuracy. The normal-Jeffreys works surprisingly well for some sparse problems. However, it tends to break down for very large problems when its lack of scale means it loses the ability to adapt. An effective subclass, which seems to lose little on the two-parameter NEG, is provided by the quasi-Cauchy with $\nu = 1/2$. It has almost optimal performance across the NEG class with one less parameter, is more robust to different signal-to-noise ratios (different values of beta) than the NJ, and is more accurate than the Lasso. Furthermore, its density, given by (4), is a function of simple normal probability functions and can be quickly computed. The scale parameter is chosen using cross-validation. Recent work has suggested that choosing the scale parameter using the BIC gives better results (Wang, Li & Tsai 2007; Wang, Li & Leng 2009), and this seems an interesting direction for future work.

Our NEG approach is non-convex and allows one to explore alternative selections that also fit well. The EM algorithm, exploiting the scale mixture of normals characterization of the NEG prior, is able quickly and successfully to find very predictive small subsets. In future work we will explore the use of the NEG prior for modal generalized linear modelling. Based on an earlier version of our paper, Hoggart *et al.* (2008) successfully applied the NEG for the analysis of many SNPs (100Ks+) in genomics, with logistic regression and a different optimization algorithm. This implementation has been shown by Ayers & Cordell (2010) to give better power to detect causal SNPs when compared with the Lasso, elastic net and other methods. Kiiveri (2008) has used the normal-gamma and developed software for the generalized linear model with the EM algorithm and is incorporating the NEG into his GENERAVE software. We have shown in the simulation study that our NEG succeeds in its aims, namely

(i) it provides an adaptive method for variable selection in regression,
(ii) it avoids undue attenuation of large effects,
(iii) classically it can possess the *oracle* property of not needing to know *a priori* the zero regression coefficients,
(iv) algorithms exist to enable very high-dimensional problems to be tackled,
(v) it can easily cope with generalized linear models.

References

ABRAMOWITZ, M. & STEGUN, I.A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*. New York: Dover.

ANDREWS, D.F. & MALLOWS, C.L. (1974). Scale mixtures of normal distributions. *J. Roy. Statist. Soc. Ser. B* **36**, 99–102.

AYERS, K.L. & CORDELL, H.J. (2010). SNP selection in genome wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology* **34**, 879–891.

BAE, K. & MALLICK, B.K. (2004). Gene selection using two-level hierarchical Bayesian model. *Bioinformatics* **20**, 3423–3430.

BERGER, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

BERNARDO, J.M. & SMITH, A.F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

BIBBY, B.M. & SORENSEN, M. (2003). Hyperbolic processes in finance. In *Handbook of Heavy Tailed Distributions in Finance*, ed. S. Rachev, pp. 211–248. Amsterdam: Elsevier Science.

TER BRAAK, C.J. F. (2006). Bayesian sigmoid shrinkage with improper variance priors and an application to wavelet de noising. *Comput. Statist. Data Anal.* **51**, 1232–1242.

BREIMAN, L. (1996). Heuristics of instability and stabilization in model selection. *Annals Statist.* **24**, 2350–2383.

BROWN, P.J., VANNUCCI, M. & FEARN, T. (1998). Multivariate Bayesian variable selection and prediction. *J. Roy. Statist. Soc. Ser. B* **60**, 627–641.

BROWN, P.J., FEARN, T. & VANNUCCI, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *J. Amer. Statist. Assoc.* **96**, 398–408.

BROWN, P.J., VANNUCCI, M. & FEARN, T. (2002). Bayes model averaging with selection of regressors. *J. Roy. Statist. Soc. Ser. B* **64**, 519–536.

DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. (1977). Maximum-likelihood from incompleted data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1–38.

FAHRMEIR, L., KNEIB, T. & KONRATH, S. (2010). Bayesian regularisation in stuctured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Stat. Comput.* **20**, 203–219.

FAN, J. & LI, R.Z. (2001). Variable selection via non-concave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348–1360.

FAN, J. & PENG, H. (2004). Non-concave penalized likelihood with a diverging number of parameters. *Annals Statist.* **32**, 928–961.

FIGUEIREDO, M.A.T. (2003). Adaptive sparseness for supervised learning. *IEEE Trans. Pattern Analysis and Machine Intelligence* **25**, 1150–1159.

FIGUEIREDO, M.A.T. & JAIN, A.K. (2001). Bayesian learning of sparse classifiers. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Volume 1*, pp. 35–41. Los Alamitos: IEEE Computer Society.

GEORGE, E.I. & MCCULLOCH, R.E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7**, 339–373.

GRADSHTEYN, I.S. & RYZIK, I.M. (1980). *Tables of Integrals, Series and Products: Corrected and enlarged edition*. Burlington: Academic Press.

GRIFFIN, J.E. & BROWN, P.J. (2005). Alternative prior distributions for variable selection with very many more variables than observations. Available from URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.126.3430 &rep=rep1 &type=pdf

GRIFFIN, J.E. & BROWN, P.J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Anal.* **5**, 171–188.

HOGGART, C.J., WHITTAKER, J.C., DE IORIO, M. & BALDING, D.J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies, *PLoS Genetics* **4**: e1000130.

ISHWARAN, H. & RAO, J.S. (2005). Spike and slab variable selection: frequentist and Bayesian strategies. *Annals Statist.* **33**, 730–773.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd edn. Oxford: Oxford University Press.

JOHNSON, N.L., KOTZ, S. & BALAKRISHNAN, N. (1994). *Continuous Univariate Distributions, Volume I*, 2nd edn. New York: Wiley.

JOHNSTONE, I.M. & SILVERMAN, B.W. (2005). Empirical Bayes selection of wavelet thresholds. *Annals Statist.* **33**, 1700–1752.

KIIVERI, H. (2003). A Bayesian approach to variable selection when the number of variables is very large. *IMS Lecture Notes Monogr. Ser.* **40**, 127–143.

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.

LI, B. & GOEL, P.K. (2006). Regularized optimization in statistical learning: A Bayesian perspective. *Statist. Sinica* **16**, 411–424.

MALLICK, B.K., GHOSH, D. & GHOSH, M. (2005). Bayesian classification of tumours by using gene expression data. *J. Roy. Statist. Soc. Ser. B* **67**, 219–234.

MEINSHAUSEN, N. & BÜHLMANN, P. (2006). High dimensional graphs and variable selection with the Lasso. *Annals Statist.* **34**, 1436–1462.

MENG, X.L. & VAN DYK, D.A. (1997). The EM algorithm – an old folk song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 511–567.

MITCHELL, T.J. & BEAUCHAMP, J.J. (1988). Bayesian variable selection in linear regression (with discussion). *J. Amer. Statist. Assoc.* **83**, 1023–1036.

OSBORNE, B.G., FEARN, T., MILLER, A.R. & DOUGLAS, S. (1984). Application of near infrared reflectance spectroscopy to compositional analysis of biscuits and biscuit doughs. *J. Sci. Food Agric.* **35**, 99–105.

OSBORNE, M.R., PRESNELL, B. & TURLACH, B.A. (1998). Knot selection for regression splines via the LASSO. In *Computing Science and Statistics. Dimension Reduction, Computational Complexity and Information. Proceedings of the 30th Symposium on the Interface*, ed. S. Weisberg, pp. 44–49. Fairfax Station: Interface Foundation of North America.

PARK, T. & CASELLA, G. (2008). The Bayesian Lasso. *J. Amer. Statist. Assoc.* **103**, 672–680.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267–288.

VIDAKOVIC, B. (1998). Wavelet-based nonparametric bayes methods. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Müller and D. Sinha, pp. 133–156. New York: Springer-Verlag.

WAINWRIGHT, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using L1-constrained quadratic programming (Lasso). *IEEE Trans. Inform. Theory* **55**, 2183–2202.

WANG, H., LI, R. & TSAI, C.L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.

WANG, H., LI, B. & LENG, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. Roy. Statist. Soc. Ser. B* **71**, 671–683.

WEST, M. (2003). Bayesian factor regression models in the large p small n paradigm. In *Bayesian Statistics 7*, eds. J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West, pp. 733–742. Oxford: Clarendon Press.

ZHANG, S. & JIN, J. (1996). *Computation of Special Functions*. New York: Wiley.

ZHAO, P. & YU, B. (2006). On model selection consistency of LASSO. *J. Mach. Learn. Res.* **7**, 2541–2563.

ZOU, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418–1429.

ZOU, H. & HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67**, 301–320.