

Bayesian adaptive lassos with non-convex penalization

James E. GRIFFIN and Philip J. BROWN *

July 19, 2007

Abstract

The lasso (Tibshirani,1996) has sparked interest in the use of penalization of the log-likelihood for variable selection, as well as shrinkage. Recently, there have been attempts to propose penalty functions which improve upon the Lassos properties for variable selection and prediction, such as SCAD (Fan and Li, 2001) and the Adaptive Lasso (Zou, 2006). We adopt the Bayesian interpretation of the Lasso as the maximum *a posteriori* (MAP) estimate of the regression coefficients, which have been given independent, double exponential prior distributions. Generalizing this prior provides a family of adaptive lasso penalty functions, which includes the quasi-cauchy distribution (Johnstone and Silverman, 2005) as a special case.

The properties of this approach are explored. Our suggested penalization can also

*Jim Griffin is Lecturer, and Phil Brown is Pfizer Professor of Medical Statistics, Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, CT2 7NF, U.K; Tel.:+44-1227-823794; Fax: +44-1227-827932;E-mail *J.E.Griffin-28@kent.ac.uk*; *Philip.J.Brown@kent.ac.uk*.

Our thanks go to Dr. H. Kiiveri for suggesting the idea of perfectly fitting starting values

have the oracle property whereby it is not advantageous asymptotically to know the subset of variables that are zero. We are particularly interested in the more-variables-than-observations case of characteristic importance for data arising in chemometrics, genomics and proteomics - to name but three. Our methodology can give rise to multiple modes of the posterior distribution and we show how this may occur even with the convex lasso. These multiple modes do no more than reflect the indeterminacy of the model. We give fast algorithms and suggest a strategy of using a set of perfectly fitting random starting values to explore different regions of the parameter space with substantial posterior support. Simulations show that our procedure provides significant improvements on a range of established procedures and we provide an example from chemometrics.

KEYWORDS: Bayesian variable selection in regression, scale mixtures of normals, Normal Exponential Gamma, adaptive lasso, oracle property, penalized likelihood, non-convexity.

1 INTRODUCTION

Variable selection in regression has several 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, often vastly more than the number of observations, are provided routinely. For example in functional genomics microarray chips typically have as many as tens of thousand genes spotted on their surface and their behavior may be investigated over perhaps one hundred 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 often it

is 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. Here we do not attempt to inject subject-matter prior knowledge, rather to give generic procedures that will be successful across a wide range of applications.

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 (SSVS) methods have become increasingly popular often adopting the ‘spike and slab’ prior formulation of Mitchell and Beauchamp (1988), see also George and McCulloch (1997), and Brown *et al* (1998) for multivariate extensions and more recently in the more-variables- than- observations case, ($k \gg n$), by Brown *et al* (2002), West (2003). In these approaches Bayesian averaging helps to induce stability. Despite 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 form of Bayesian approach which does offer the potential for much faster computation takes a continuous form of prior and looks merely for modes of the posterior distribution rather than relying on full MCMC. Such formulations lead to penalized log likelihood approaches where the additive penalization of the log likelihood is the log of the prior distribution. Tibshirani’s (1996) lasso 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 (Figueiredo and Jain 2001, Figueiredo 2003), adopted in an extended generalized linear model setting by Kiiveri (2003). Ter Braak (2006) adopts a power variant of the Jeffreys prior for propriety of the posterior, concentrating

more on MCMC and the full posterior distribution.

Within the penalized likelihood literature Fan and Li (2001) have modified the lasso's L_1 penalty so as to offer less shrinkage for large effects, their Smoothly Clipped Absolute Deviation penalty (SCAD). They show that the lasso property of giving a mode of exactly zero requires the penalty to be singular at the origin. They also discuss (their Theorem 2) the 'oracle' property whereby knowing beforehand which coefficients should be set to zero does not improve estimation asymptotically. Zou (2006) takes up the oracle theme, showing that in some circumstances the lasso may be inconsistent for variable selection. Meinshausen and 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. Consequently Zou (2006) proposes an adaptive lasso whereby coefficients are weighted differently. Our preferred Bayesian alternative developed in section 2 is automatically adaptive, effectively achieved by continuously varying the lasso parameter. It will also adapt to providing negligible shrinkage for large effects in the spirit of SCAD.

As noted by Zou and Hastie (2005) the lasso needs to select at most n non-zero parameters. They also draw the line between 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 infer non-convex penalties and penalized likelihoods and their consequent multiple solutions. We will explore these by means of random perfectly fitting starting values. We argue that it is artificial to demand a single solution to a problem that is inherently indeterminate, although it is often easy to find one very good estimator avoiding the need to form a single estimator by

averaging.

In section 2 we set consider using scale mixture of normal prior distributions for the regression coefficients and develop the particular normal-exponential-gamma, showing its connections with existing approaches and the critical tail to spike weighting of various competitors. In section 3 we compare shrinkage and selection of our preferred choice with more standard alternatives. In section 4 we implement the class of priors through an EM algorithm for exploring the posterior modes and 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 gives a counter example to the uniqueness of the lasso when the number of variables is greater than the number of observations $k > n$. Section 6 illustrates the ideas via a simple simulation and a more systematic simulation study together with an real example. Some concluding remarks are made in Section 7.

2 BAYESIAN PENALIZATION

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

We assume that the explanatory variables have been centered and any scaling of these variables has been undertaken if desired. It may be noted that automatic scaling to ‘correlation form’ may not be desirable when the variables are on the same scale as it will just tend to inflate the relative importance of variables that change little over the data. We assume

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

where $Y(n \times 1)$, is the response vector and $X(n \times k)$ is the matrix of regressors, $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ and these are independent $N(0, \sigma^2)$. We do not restrict k to be less than n . Implicitly throughout we are assuming a vague prior for μ so that in effect we replace it by the sample mean of the \bar{Y} . At least initially we will assume that σ^2 is known.

For reasons of convenience and flexibility we will concentrate on priors for $\beta_j, j = 1, \dots, k$ which are scale mixtures of normals, see for example West (1987).

Here

$$\pi(\beta_i) = \int \mathbf{N}(\beta_i|0, \psi_i) G(d\psi_i) \quad (2)$$

where $\mathbf{N}(Y|\mu, \sigma^2)$ denotes the probability density function of a random variable Y having a normal distribution with mean μ and variance σ^2 . Here G is the mixing distribution and its density, if it is defined, will be referred to as $g(\cdot)$.

Taking the negative log prior gives a direct analogue to classically penalizing the negative log-likelihood and then minimizing. The lasso is a member of this class. The mean-zero double exponential distribution, $\text{DE}(0, 1/\gamma)$ with probability density function

$$\frac{1}{2\gamma} \exp\{-|\beta|/\gamma\}, \quad -\infty < \beta < \infty, \quad 0 < \gamma < \infty$$

is defined by an exponential mixing distribution, $\text{Ex}\left(\frac{1}{2\gamma^2}\right)$, with probability density function

$$g(\psi_i) = \frac{1}{2\gamma^2} \exp\{-\psi_i/[2\gamma^2]\}. \quad (3)$$

2.1 The normal-exponential gamma (NEG)

Our preferred generalization of the lasso prior is formed by allowing the scale parameter to vary from coefficient to coefficient. Specifically if we write (3) as $z \exp(-z\psi_i)$ and assume Z has a gamma mixing distribution with parameters

λ, γ^2 and density proportional to $z^{\lambda-1}\exp(-\gamma^2 z)$ then the density for ψ_i is a subclass of the gamma-gamma distribution (Bernardo and Smith, 1994, p120), the exponential-gamma (EG). The density of the marginal distribution of β_i can be expressed using Gradshteyn & Ryzik (1980, p319) as

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

where $D_\nu(z)$ is the parabolic cylinder function. Computation of this functions is described in Zhang and Jin (1996, section 13.5.1, p439), 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 λ 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 and Jin (1996).

The parameter γ and λ control the scale and the heaviness of the tails respectively. From Abramowitz and Stegun (1964, p689 eqn 19.8.1) we see that for large $\frac{|\beta_i|}{\gamma}$

$$\pi(\beta_i) \approx c \left(\frac{|\beta_i|}{\gamma}\right)^{-(2\lambda+1)}.$$

Thus if $\lambda = 0.5$ the distribution has the same tail behavior as a Cauchy. Also if $\lambda > 1$, the expectation of ψ_i and the variance of β_i exist and have the form $\frac{\gamma^2}{(\lambda-1)}$. The excess kurtosis is $3\frac{\lambda}{\lambda-2}$ if $\lambda > 2$. This class of distributions can define distributions for which the variance is undefined ($\lambda \leq 1$) and thus has a tail-to-spike balance which can be concentrated around zero and yet have fat tails. The distribution of β is singular at zero with a mode that is finite for all parameter values. We will refer to the marginal distribution of β_i with density (4) as the normal-exponential-gamma (NEG) distribution.

Although the emergence of parabolic cylinder functions may seem unappetiz-

ing, the distribution has precedents in the literature when $\lambda = 0.5$. The precedents arise because when convolved with an equal variance normal the result is a convenient explicit form. In fact Johnstone and Silverman (2005) define a *quasi-Cauchy* which is exactly the NEG when $\lambda = 0.5$. Also Berger (1985, section 4.7.10) defines a robustness prior which again exactly corresponds in the univariate case of his multivariate prior. The Cauchy form of tail behavior 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 Gradshteyn and Ryzik (1980, p315, 3.362 eqn 2) we obtain for $\lambda = 1/2$

$$\pi(\beta_i) = \frac{\sqrt{2\pi}}{\gamma} \left\{ 1 - \frac{[\frac{|\beta_i|}{\gamma}][1 - \Phi(\frac{|\beta_i|}{\gamma})]}{\phi(\frac{|\beta_i|}{\gamma})} \right\}, \quad (5)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal. This form is also given as (13) of Johnstone and Silverman (2005).

Before going onto properties of the general NEG prior we list several alternatives. The normal-Jeffreys (NJ) prior distribution arise from the improper hyperprior $g(\psi_i) \propto \frac{1}{\psi_i}$ which in turn induces an improper prior for β_i of the form $\pi(\beta_i) \propto \frac{1}{|\beta_i|}$. This has been used by Figueiredo & Jain (2001), Kiiveri(2003) and in a power variant by ter Braak (2006). Another alternative which is proper is the Normal Gamma (NG) or Variance Gamma of Bibby and Sorenson (2003).

Our reason for choosing the NEG is two-fold: it has a finite spike at zero for all parameter values (not so NJ or NG) and it has fat tails for λ small. We will see that these properties are important if we want to find sparse solutions. The tail to spike behavior is illustrated in figure 1 for NEG, DE and NJ. For comparison we specify one scale parameter by fixing probability mass on the central region $(-\epsilon, \epsilon)$ to be η (except the Normal-Jeffreys 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 v NEG and (b) DE v 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 with the NEG preserving good features of the NJ without the drawback of the extreme spike at zero.

Figure 1 here

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 $X'X$ matrix is diagonal.

3 SELECTION AND SHRINKAGE

It is natural to regard the negative prior utility as a penalty function given as $p(\beta)$, where $p(\beta) = -\log \pi(\beta)$. The problem of finding a maximum *a posteriori* (MAP) estimate of β can be expressed as a penalized likelihood problem where β is chosen to find a minimum of the function

$$L = \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) + \sum_{i=1}^k p(|\beta_i|). \quad (6)$$

In one dimension typically for spiked priors there may be a posterior mode at zero as well as the data driven mode away from zero. With weak evidence the mode at zero will be the only mode. As evidence of an effect strengthens so a turning point appears away from zero. With more evidence still this mode will dominate. Thus there may be one or two modes. In higher dimensions we may have a highly multi-modal posterior distribution. The turning point of the posterior distribution with the largest density will be called the *penalized MLE* (PMLE) and reserve the term *Maximum A posteriori Probability* (MAP) for the overall mode (which may

be zero).

The choice of penalty function will have implications for the shrinkage of the regression coefficient. 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

$$\frac{\hat{\beta} - \tilde{\beta}}{\sigma^2/X^T X} = \text{sign}(\tilde{\beta})p'(|\tilde{\beta}|). \quad (7)$$

where $p'(\cdot)$ is the derivative of the penalty function and $\frac{\sigma}{\sqrt{X^T X}}$ 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 sets $\tilde{\beta} = 0$ which will be useful for variable selection. Fan and Li (2001) use equation (7) to show that the PMLE is zero if

$$|\hat{\beta}| < \min_{\theta \neq 0} \left\{ |\theta| + \frac{\sigma^2}{X^T X} p'(|\theta|) \right\}.$$

We shall refer to the quantity on the right-hand side of the inequality as the turning point threshold. The dependence on the derivative of the penalty function also arises from robustness considerations in Li and Goel (2006). Various penalty functions together with their derivatives are listed in Table 1.

Table 1 here

The oracle property is important for penalized maximum likelihood estimation (PMLE) and says that the PMLE has the same asymptotic properties as the MLE knowing which regression coefficients are zero. Fan and Li (2001) establish conditions for the oracle property in their Theorem 2.

The penalized maximum likelihood estimate defined by the NEG penalty will have the oracle property. The analogous scale parameter to λ of their paper is $1/\gamma$.

The important quantity is $a_n = \max\{p'_{\gamma_n}/n : \beta_j \neq 0\}$ which is $\frac{2\lambda+1}{n\gamma_n} \frac{D_{2(\lambda+1)}(\min(\frac{|\beta_j|}{\gamma_n}))}{D_{2(\lambda+1/2)}(\min(\frac{|\beta_j|}{\gamma_n}))}$ for the NEG penalty since the derivative is monotone decreasing. As $\frac{1}{\gamma_n} \rightarrow 0$

$a_n \rightarrow 0$. Then the oracle property and root- n consistency occur when $\frac{\sqrt{n}}{\gamma_n} \rightarrow \infty$ and $\frac{1}{\gamma_n} \rightarrow 0$.

It is illuminating to compare the turning point threshold for various choices of the prior distribution. For the double exponential prior distribution, the threshold is $|\hat{\beta}| < \frac{1}{\gamma} \frac{\sigma^2}{X^T X}$ which depends on the square of the standard error and so shrinks at an uncomfortably fast rate of $1/n$. In contrast, the normal-Jeffreys prior thresholds according to the rule $|\hat{\beta}| < 2 \frac{\sigma}{\sqrt{X^T X}}$ and the threshold depends linearly on the standard error, with $1/\sqrt{n}$. 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 normal-gamma penalty and the normal-exponential-gamma penalty. The latter has linear behavior where the slope depends on λ , generalizing the normal-Jeffreys rule and is thus more appealing. The normal-gamma case has substantially different behavior and defines a much more conservative criterion. Much larger values of γ would induce a linear threshold rule but this contradicts our imposed prior property of a large mass close to zero.

Figure 2 here

We have earlier noted that the global mode may not be the data driven non-zero mode. An exception is the double exponential prior for which it can be shown that the PMLE and the MAP estimate coincide. In fact the infinite spike at zero for the Jeffreys prior and the normal-gamma with $\lambda < 0.5$ always appears in the posterior and can dominate the search for a mode away from zero (a turning point). However, the NEG prior distribution always renders a finite mode at zero in the posterior.

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 and can explore alternative modes in the multi-modal posterior.

4.1 An EM algorithm to find a mode of β

Local posterior modes can be found using the EM algorithm (Dempster *et al* 1977, Meng and van Dyk 1997) which has been suggested by both Kiiveri (2003) and Figueiredo (2003) as a means for 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 ψ_1, \dots, ψ_k are treated as missing data. Alternatively Kiiveri (2003) suggests applying the EM algorithm directly to the ‘likelihood times prior’ in the generalized linear model setting. The M-step is approximated by a Newton-Raphson line search for the MLE of β 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 $\beta^{(1)}, \beta^{(2)}, \dots$ that under regularity conditions converge to a local maximum of $\beta|y$. The sequence is defined by iterating between an E step which for us averages over ψ for given β and an M step which maximizes over β for given ψ .

1. **E-step:** Let $\psi_j^{(i)} = \frac{1}{\mathbb{E}[\frac{1}{\psi_j} | \beta^{(i-1)}]} = \frac{p'(|\beta_j^{(i-1)}|)}{|\beta_j^{(i-1)}|}$ for $j = 1, \dots, k$. The derivatives $p'(|\beta|)$ are given in table 1.
2. **M-step:** Set $\beta^{(i)} = \Psi^{(i-1)} A (A^T \Psi^{(i-1)} A + \sigma^2 D^{-2})^{-1} \hat{\alpha}$ where we calculate

the singular value decomposition of $X = FDA^T$. The matrix A is $(k \times r)$ -dimension matrix such that $A^T A = I_r$ with columns of A the r eigenvectors of $X^T X$ corresponding to non-zero eigenvalues, D is an $(r \times r)$ -dimension diagonal matrix and F is $(n \times r)$ -dimension matrix whose columns are the r eigenvectors of XX^T corresponding to non-zero eigenvalues and for which $F^T F = I_r$. We also define $\Psi^{(i-1)} = \text{Diag}(\psi_1^{(i-1)}, \dots, \psi_k^{(i-1)})$ and $\hat{\alpha} = D^{-1} F^T y$. This form allows involves the inversion of $r \times r$ matrices where $r \leq \min(n - 1, k)$ is the rank of X . When $k \gg n$ these matrices will be very much smaller than the $k \times k$ matrices that would be needed using standard results.

Much of the work in linear or generalized linear models using normal-Jeffreys penalty functions, see Kiiveri (2003), Figueiredo (2003)) who try to find a single mode. Bae and Mallick (2004) and Mallick *et al* (2005) on the other hand go for full posterior simulation using MCMC, but in favoring 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 behavior in the region of zero is still proportional to $1/\beta$ 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

The Minimum Length Least Squares (MLLS) (also ridge for small ridge constant) fit to the data for $k > r$ is $\hat{\beta}_{MLLS} = (X^T X)^+ X^T y$ 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.

The singular value decomposition of the centered design matrix is

$$X = F \text{Diag}(d_1, d_2, \dots, d_r) A^T$$

. The orthogonal projection matrix is $I - P = I_k - AA^T$ a matrix of rank $(k - r)$. Consider generating a random k -vector z and take $w = (I - AA^T)z$, calculated as $z - A(A^T z)$. If we add this projected random vector to $\hat{\beta}_{MLLS}$ then we will have the same Minimum length least squares ‘perfectly’ fitting solution since $Xw = 0$, as verified by

$$\begin{aligned} Xw &= F D A^T (I - AA^T) z \\ &= U D (A^T - A^T) z = 0 \end{aligned}$$

Thus we can add w to $\hat{\beta}_{MLLS}$ and get a ‘perfectly’ fitting starting point. We can repeat this as often as we like or design the z to span the space. Typically the seed 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 $\hat{\beta}$, as given by the Moore-Penrose generalized inverse. To this end we ordered the the p components of $\hat{\beta}$, $\hat{\beta}_{(1)} \leq \hat{\beta}_{(2)}, \dots, \leq \hat{\beta}_{(p)}$ and the average of the largest from $\hat{\beta}_{([0.9p])}$ 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 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 calculate \bar{A} , the $k \times (k - r)$ set of eigenvectors completing the set A . Now suppose we have a random $(k - r) \times 1$ vector u , then $\bar{A}u$ may be added to $\hat{\beta}_{MLLS}$ to achieve a ‘perfectly’ fitting starting point. This is easily seen since $X\bar{A}u = FDA^T\bar{A}u = 0$ since the eigenvectors in A and \bar{A} are orthogonal. Lack of quick algorithms to generate \bar{A} may make this modified approach unattractive though.

In the next section we note that regions of indeterminacy can affect even the convex penalization of the lasso when $k > \text{rank}(X)$, essentially because although convex it is not then strictly convex.

5 NON-UNIQUENESS OF LASSO

With L_1 componentwise loss in Gaussian multiple regression the penalized negative log-likelihood is convex but not strictly convexity. Thus in under-determined contexts the lasso may give an interval of maxima rather than a unique maximum. We give a counterexample to uniqueness which will bring out when this will occur. The examples illustrate cases where the sufficient conditions for uniqueness of Appendix B1, Theorem 5 of Rosset *et al* (2004) do not hold. The examples begin by being quite specialized but move on to much more plausible settings. The message is that although the symmetry required may not be exactly present, near symmetry often is, and this may lead to a lack of robustness and an interchangeability of variables.

5.1 A general counter example

We first derive a result for an example essentially considered by Zou and Hastie (2005).

This very specialized example will then be generalized to a much richer context from which we will be able to draw insights even when the conditions only approximately hold. For the example in its original form there are k variables and n observations but the variables are repeated for each observation, that is the n -vectors satisfy $x_i = x_j$, $i, j \in (1, 2, \dots, k)$. Denote the common n -vector as x . Symmetry is crucial. This is an example where all the variables are perfectly correlated and there is additional symmetry in the variables, that is rather than the general $x_{li} = b_{ij}x_{lj}$, $l = 1, \dots, n$ it is necessary that $|b_{ij}| = |b|$, independent of variable labels, w.l.o.g take $b = 1$. The linear model becomes

$$Y_l = x_l \sum_{j=1}^k \beta_j + \epsilon_l, \quad l = 1, \dots, n.$$

The function to minimize becomes

$$L = \sum_{l=1}^n \left\{ y_l - x_l \left(\sum_{j=1}^k \beta_j \right) \right\}^2 + c \left(\sum_{j=1}^k |\beta_j| \right) \quad (8)$$

Let $\theta = \sum_{j=1}^k \beta_j$ then the least squares estimate of θ is $\hat{\theta} = y^T x / (x^T x)$, w.l.o.g. assumed to be non negative. Apart from an additive constant (8) becomes

$$L^* = x^T x (\theta - \hat{\theta})^2 + c \left(\sum_{j=1}^k |\beta_j| \right). \quad (9)$$

The MLE of β does not exist: the likelihood is constant on the plane $\sum_{j=1}^k \beta_j = \hat{\theta}$. Parallel to this plane and above it both terms in (9) are increased, whereas below the least squares plane the first term increases but the second decreases provided we are in the positive quadrant. Differentiating (9) on the simplex for turning points, the minimum or lasso is attained at

$$\theta = \hat{\theta} - c / \left[2 \sum_1^n x_l^2 \right]$$

or zero if this changes the sign from $\hat{\theta}$.

Now we can strengthen and extend this example by considering just a subset of the variables being perfectly symmetrically correlated. Suppose there are $k_1 > 1$ of these and their labels are the subset $j \in S_1$ with the complementary set $j \in S_2$.

Then

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon$$

with $X_1, n \times k_1$ and $X_2, n \times k_2$. Here we assume $X_1 = x1^T$ with n -vector x and 1 a k_1 -tuple of ones. The model becomes

$$y_l = x_l \sum_{j \in S_1} \beta_j + \sum_{j \in S_2} x_{lj} \beta_j + \epsilon_l$$

Let $\theta = \sum_{j \in S_1} \beta_j$, now we can see that if the lasso is applied to the reduced problem in which the design matrix is $X_0 = \{x: X_2\}$ a $n \times (k_2 + 1)$ matrix and the lasso solution does not set $\hat{\theta} = 0$ then there will be multiple solutions on the simplex $|\theta| = \sum_{j \in S_1} |\beta_j|$. Perversely the usual practice of standardizing x -variables will promote such symmetry and if correlations are near unity for any subset of at least two variables then there will be flat sections in the penalized likelihood space and near indeterminacy. This will be highly likely by chance in high dimensional problems where $k \gg n$.

6 EXAMPLES

We will first 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 by a simulation study of some alternative methods systematically compared with the NEG in the $n \ll k$ setting. Finally we give a real example involving prediction of the composition of biscuits.

6.1 Spline simulation

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

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

where K_1, K_2, \dots, K_p are knots points which are equally spaced in the interval (a, b) and so $K_i = a + \frac{i-1}{p-1}(b - a)$. We observe pairs (x_i, y_i) , which is a noisy versions of $f(x_i)$, and the problem of estimating β_i is a linear regression problem in a non-linear basis. Osborne *et al* (1998) have applied a Lasso penalty to this problem and we compare this approach to Normal-Jeffreys and NEG penalisation. If p is large, there will be substantial correlation between subsequent regressors, due to the closeness of the knots points, which makes inference by regression methods a challenging problem. We fit $n = 30$ observations: x_i are uniformly distributed on $(0, 1)$ and $y_i = f(x_i) + \epsilon_i$ where ϵ_i is drawn from a normal distribution with mean 0 and variance 0.01. We have $p = 500$ knot points between $a = -0.3$ and $b = 1.3$. The hyperparameters of the Lasso and the NEG are estimated using 5-fold cross-validation.

Figure 3 here

The results for the NEG penalisation 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 $\mu = \frac{\lambda}{\gamma^2}$ and the average non-zero regressors falls as μ is increased. For fixed μ larger values of λ (leading to fatter tails) gives more non-zero estimates. It is worth not-

ing that although the number of non-zero regressors should be less than n that, in practice, the number of included regressors can be larger than n due to the high correlation of the regressors.

Figure 4 here

The summary figures for 20 perfectly fitting random starts are given in figures 4 (a) and 4(b) for the three penalty functions using parameters chosen by cross-validation. In each case, the fitted curves follow the data well (figure 4(b)). The variety of fitted curve for a given method is the main difference with the Lasso showing the least differences and the Normal-Jeffreys the most. This results is also illustrated by the position of the knot-point with non-zero regression parameter estimates (figure 5). The NEG and Normal-Jeffreys prior distributions show a spread of knots points have non-zero estimates in different modes whereas the Lasso will typically pick a single point across all modes. There is also substantial differences between the number of non-zero regressors found using the Normal-Jeffreys and NEG priors, which are typically 7 or 8, compared to the Lasso fits which use many more (figure 4(a)). The absolute values of the estimated regression coefficient, shown in figure 5 where the area of the dots is proportional to the absolute size of the regression estimate, give some indications about the different levels of sparsity.

Figure 5 here

6.2 Multiple regression simulation

We have conducted a simulation study to compare a variety of estimation methods including that corresponding to our preferred NEG prior. The Gaussian error regression model is simulated with error variance $\sigma^2 = 1$. The $n \times k$ design matrix X is simulated with an autoregressive order (AR(1)) structure with lag 1 correlation $\rho = 0.5, 0.8$. The simulation has $n = 100$ observations, $k = 500, 2000$ variables, $k^* = 10$ nonzero coefficients of β , with either all the non-zero coefficients $\beta = 1$ or $\beta = 5$, equally spaced in the k variable design. Hyperparameters, eg λ, μ in the NEG, were chosen by 5-fold cross validation and further tested on 10 datasets of 100 observations.

The methods compared are:

1. The normal exponential gamma (NEG) prior, in versions with both parameters free to be chosen from $\lambda = 0.1, 0.5, 1$ and 2 by cross validation and with $\lambda = 0.1, 0.5$ fixed
2. The Lasso, with one parameter estimated by -validatory choice
3. The Adaptive Lasso (AL) using either the Minimum Length Least Squares (MLLS) with a Moore-Penrose generalized inverse or Ridge (from separate cross-validatory choice) for the estimate β_* in the construction of their adaptive weight function, $w = \frac{1}{|\beta_*|^\gamma}$ with γ chosen as either $0.5, 1$ or 2 .
4. The normal Jeffreys (NJ) prior which has no free parameters to estimate
5. The L_2 penalisation (Ridge regression) with its constant estimated by cross-validation.

The Mean squared error results for the $2^3 = 8$ cases are given in Table 2.

Table 2 here

To summarise these results:

- NEG is generally the best with a MSE close to the oracle unity of the error variance.
- The adaptive lasso is generally no better than the lasso.
- The NJ is surprisingly good given that it lacks adaptive flexibility with no hyperparameters to estimate
- Ridge is generally bad, which is hardly surprising in that its prior assumption of a exchangeable normal distribution would expect a good balance of non zero β 's, not such a small number relative to the number of parameters. It will come into its own with a higher proportion of non-zero β .

6.3 Biscuits NIR data

Figure 6 here

The data is taken from Osborne *et al* (1984) and was used again in Brown *et al* (2001), where the data set-up are described in some detail. The predictor variables 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 1202nm to 2400nm in steps of 4nm. The hyperparameter values of the NEG penalty are chosen using 5-fold cross-validation. For each split of the training sample into a training and testing subsample the EM algorithm is run once the training data has been centered and standardized by the median of SDs of the X -variables and the same mean and standard deviation used to adjust the 19 test spectra. We've avoided scaling

to ‘correlation form’ since it is important not to change the relative scales of reflectance at different wavelengths as this would promote reflectances which are very small and may be largely noise. The response Y chosen was the flour content which was also centered and scaled by its standard deviation over the 20 samples. These standardizations help numerical stability and allow easy interpretation of fit.

The hyperparameters μ, λ are selected by cross-validation averaging over 5 splits and the results are shown in figure 6 (a). Figure 6 (b) gives the parallel effect on number of wavelengths chosen. The hyperparameters values chosen were $\lambda = 1, \mu = 100000$. The results of finding estimates using the NEG penalized likelihood with these hyperparameters over 20 perfect random starts are depicted in figure 7. Each mode found has 3 or 4 wavelengths with non-zero regression coefficients. Most modes include a wavelength around position 1920nm and 2080nm. Three further regions are identified by some of the modes around 1800nm, 2200nm and 2400nm.

Figure 7 here

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

7 CONCLUSIONS

We have developed a wholly adaptive lasso motivated by a Bayesian framework. The lasso itself is unable to simultaneously do well in (a) prediction and (b) identification of significant variables. This can be viewed as a problem of its inflex-

ibility in ‘tail to spike’ behavior with one parameter (a scale parameter) fits all. Our Normal-Exponential-Gamma prior has two parameters for flexibility, one for the shape and one for the scale although within this class the shape parameter seems far less important in terms of our cross-validation studies. An effective subclass which seems to lose little on the 2-parameter NEG is provided by the quasi-Cauchy with $\lambda = 1/2$. Also its density, given by equation (5), is a function of simple normal probability functions and can be quickly computed.

We have shown in the simulation study that our NEG succeeds in its aims. We have also shown how the absence of strict convexity in the lasso leads to multiple solutions and indeterminacy when the number of variables is larger than the number of observations ($k > n$). Our NEG approach is non-convex and can allow one to explore alternative selections which also fit well. Our 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.

REFERENCES

- Abramowitz, M. and Stegun, I. A. (Eds.) (1964) “Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables,” Dover: New York.
- Bae, K. and 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,” Berlin: Springer.
- Bernardo, J. M. and Smith, A. F. M. (1994): “Bayesian Theory,” Wiley : Chich-

ester.

- Bibby, B. M. and Sorensen, M. (2003): "Hyperbolic Processes in Finance, in *Handbook of Heavy Tailed Distributions in Finance* S. Rachev (ed.): , Elsevier Science, 211-248.
- Breiman, L.(1996): "Heuristics of instability and stabilization in model selection," *Annals of Statistics*, 24, 2350-238 .
- Brown, P. J., Vannucci, M. and Fearn, T. (1998): "Multivariate Bayesian variable selection and prediction," *Journal of the Royal Statistical Society B*, 60, 627-641.
- Brown, P. J., Fearn, T. and Vannucci, M. (2001): "Bayesian wavelet regression on curves with application to a spectroscopic calibration problem," *Journal of the American Statistical Association*, 96, 398-408.
- Brown, P. J., Vannucci, M. and Fearn, T. (2002): "Bayes model averaging with selection of regressors," *Journal of the Royal Statistical Society B*, 64, 519-536.
- Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977): "Maximum-likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society B*, 39, 1-38.
- Fan, J. and Li, R.Z. (2001): "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348-1360.
- Figueiredo, M. A. T. and Jain, A. K. (2001): "Bayesian learning of sparse classifiers," *Proceedings IEEE Computer Society Conference in Computer Vision and Pattern Recognition*, Vol 1, 35-41.
- Figueiredo, M. A. T. (2003): "Adaptive sparseness for supervised learning,"

- IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 1150-1159.
- George, E. I. and McCulloch, R. E. (1997): "Approaches for Bayesian variable selection," *Statistica Sinica* 7, 339-373.
- Gradshteyn, I. S. and Ryzik, I. M. (1980) "Tables of Integrals, Series and Products: Corrected and Enlarged Edition," (A. Jeffrey, Ed.) Academic Press: New York.
- Jeffreys, H. (1939/1961) "Theory of Probability", 3rd Edition 1961, Oxford: Clarendon Press
- Johnstone, I. M. and Silverman, B. W. (2005): "Empirical Bayes selection of wavelet thresholds," *Annals of Statistics*, 33, 1700-1752.
- Kiiveri, H. (2003): "A Bayesian approach to variable selection when the number of variables is very large," In Goldstein, D.R. (Ed) "Science and Statistics: Festschrift for Terry Speed" *Institute of Mathematical Statistics Lecture Notes-Monograph Series*, Vol 40, 127-143.
- Li, B. and Goel, P. K. (2006): "Regularized optimization in statistical learning: A Bayesian perspective," *Statistica Sinica*, 16, 411-424.
- Mallick, B. K., Ghosh, D. and Ghosh, M. (2005): "Bayesian classification of tumours by using gene expression data," *Journal of the Royal Statistical Society B*, 67, 219-234.
- Meinshausen, N. and Bühlmann, P. (2006) "High dimensional graphs and variable selection with the lasso", *Annals of Statistics*, 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)," *Journal of the Royal Statistical Society B*, 59, 511-567.

- 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.
- 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. and Turlach, B. A. (1998): “Knot selection for regression splines via the LASSO,” in *Dimension Reduction, Computational Complexity, and Information, Proceedings of the 30'th Symposium on the Interface, Interface 98* (Editor S. Weisberg), Interface Foundation of North America, 44-49.
- Rosset, S., Zhu, J. and Hastie, T. (2004) “Boosting as a Regularized Path to a Maximum Margin Classifier”, *Journal of Machine Learning Research*, 5, 941-973.
- ter Braak, C. J. F. (2006) “Bayesian sigmoid shrinkage with improper variance priors and an application to wavelet denoising”, *Computational Statistics and Data Analysis*, 51, 1232-1242.
- Tibshirani, R. (1996): “Regression shrinkage and selection via the lasso,” *Journal of the Royal Statistical Society B*, 58, 267-288.
- Vidakovic, B. (1998): “Wavelet-Based Nonparametric Bayes Methods,” in *Practical Nonparametric and Semiparametric Bayesian Statistics* D. Dey, P. Muller and D. Sinha (eds.):, New York : Springer-Verlag, 133-156.
- West, M. (2003): “Bayesian Factor regression models in the large p , small n paradigm,” In Bernardo J. M. *et al* (Eds), “Bayesian Statistics 7”, 733-742: Clarendon Press: Oxford.

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

Zhang, S. and Jin, J. (1996): "Computation of Special Functions," Wiley : New York.

Zou, H. (2006) "The adaptive lasso and its oracle properties", *Journal of the American Statistical Association*, 101, 1418-1429.

Zou, H. and Hastie, T. (2005) "Regularization and variable selection via the elastic net", *Journal of the Royal Statistical Society, B*, 67, 301-320.