Alternative prior distributions for variable selection with very many more variables than observations

J.E. Griffin*
Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K.

P.J. Brown
Institute of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, CT2 7NF, U.K.

15th May 2005

Abstract

The problem of variable selection in regression and the generalised linear model is addressed. We adopt a Bayesian approach with priors for the regression coefficients that are scale mixtures of normal distributions and embody a high prior probability of proximity to zero. By seeking modal estimates we generalise the lasso. Properties of the priors and their resultant posteriors are explored in the context of the linear and generalised linear model especially when there are more variables than observations. We develop EM algorithms that embrace the need to explore the multiple modes of the non log-concave posterior distributions. Finally we apply the technique to microarray data using a probit model to find the genetic predictors of osteo- versus rheumatoid arthritis.

Keywords: Bayesian modal analysis, Variable selection in regression, Scale mixtures of normals, Improper Jeffreys prior, lasso, Penalised likelihood, EM algorithm, Multiple modes, More variables than observations, Singular value decomposition, Latent variables, Probit regression.

*Jim Griffin was a member of the Institute of Mathematics, Statistics and Actuarial Science, University of Kent at the start of this research. Corresponding author: Jim Griffin, Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. Tel.: +44-1227-82 3865; Fax: +44-24-7652 4532; Email: J.E.Griffin@warwick.ac.uk.

1 Introduction

It is common nowadays to be able to investigate very many variables simultaneously with data collected on relatively few samples. For example in functional genomics microarray chips typically have as many as ten thousand genes spotted on their surface and their behaviour 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.

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), Wolfe *et al* (2004), Brown *et al* (1998) for multivariate extensions and more recently in the more- variables- than- observations case by $(k >> n)$, 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) of some applications 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 MCMC to fully investigate the posterior distribution. Such formulations lead to penalised log likelihood approaches where the additive penalisation 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 generalised linear model setting by Kiiveri (2003). From a different viewpoint Fan and Li (2001) have modified the lasso's L_1 penalty so as to offer less shrinkage for large effects, see also Fan and Peng (2004).

Early examples of parallel approaches in the machine learning literature are Automatic Relevance Determination of Mackay (1994) and the Relevance Vector Machine of Tipping and Faul (2003).

In this paper we concentrate on priors for the effects which are scale mixtures of normal distributions in a broad sense. These bridge the full range from the lasso to the extreme Jeffreys-based prior. We explore thresholding properties and multimodality. In the context of multiple regression and later probit regression, we develop estimation procedures and fast EM style algorithms for estimation utilising the inherent dimensionality $\{\min(n, k)\}$ of in-

formation. In the probit case, our method of hyperparameter choice is geared to prediction characteristics of some canonical models and although data dependent helps to avoid over-shrinkage. We finish by analysing microarray data on two forms of arthritis earlier analysed by Sha *et al* (2003).We embrace the multimodality through plots of genes included in modes as ranked either by posterior or log-likelihood value. We are able to reveal subsets of highly discriminating genes.

2 Generalising lasso estimation

There are at least two ways of generalising the lasso in a Bayesian setting. One is to use an exponential power prior for β, see Box and Tiao (1973, p157); the other is to use a scale mixture of normals, see West (1987). Non Bayesian analogues and adaptations of the former are to be found in Knight and Fu (2000), Fan and Li (2001). We will rather devote our attention to scale mixtures of normals as these are easier to deal with analytically and are richer in form.

2.1 Scale mixture of normal prior distributions

If we wish to construct distributions that bridge the gap between the normal-Jeffreys prior and the double exponential distribution, a natural class of prior distributions to consider for each regression coefficient, β_i, would be scale mixtures of normal distributions where

$$\pi(\beta_i) = \int \mathrm{N}(\beta_i|0,\psi_i)\, G(d\psi_i) \tag{1}$$

where $\mathrm{N}(Y|\mu,\sigma^2)$ 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 $g(\cdot)$. The prior variance of the regression coefficients, if it exists, can be simply expressed in terms of the mean of the mixing distribution since

$$\begin{aligned}\mathrm{V}(\beta_i) &= \mathrm{E}_{\psi_i}(\mathrm{V}(\beta_i|\psi_i)) + \mathrm{V}_{\psi_i}(\mathrm{E}(\beta_i|\psi_i)) \\ &= \mathrm{E}_{\psi_i}(\mathrm{V}(\beta_i|\psi_i)) \\ &= \mathrm{E}_{\psi_i}(\psi_i).\end{aligned}$$

If we assume that domain knowledge will not be included in the prior, the mixing distribution seems a natural place to include the belief that only a few regressors will be important to give a good fit to the data. Most Bayesian approaches to variable selection make use of the form $G(\cdot)$ to aid inference. A traditional approach to variable selection, (Mitchell and Beauchamp,

1988, George and McCulloch, 1997), expresses the prior distribution for β_i as a mixture distribution

$$i.e. \qquad \pi(\beta_i) = \theta\, \mathrm{N}(\beta_i|0, \sigma^2_\beta) + (1-\theta)\, \delta_{\beta_i=0} \tag{2}$$

where $\delta_{x=a}$ is the Dirac measure which places measure 1 on $\{x = a\}$. The parameter $0 < \theta < 1$ can be interpreted as the probability that a variable is included in the model and σ^2_β is the prior variance of the regression coefficients included in the model. If we make use of the obvious extension of the normal distribution by defining $\mathrm{N}(x|\mu, 0) = \delta_{x_i=\mu}$, the mixing distribution can be expressed as

$$g(\psi_i) = \theta\, \delta_{\psi_i=\sigma^2_\beta} + (1-\theta)\, \delta_{\psi_i=0}. \tag{3}$$

Other particular mixture distributions of interest can also be represented in this scale mixture form.

1. The mean-zero double exponential distribution, $\mathrm{DE}(0, 1/\gamma)$ with probability density function

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

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

   $$g(\psi_i) = \frac{1}{2\gamma^2} \exp\left\{-\psi_i/[2\gamma^2]\right\}. \tag{4}$$

2. The normal-Jeffreys (NJ) prior distribution arise from the improper hyperprior

   $$g(\psi_i) \propto \frac{1}{\psi_i}, \tag{5}$$

   which in turn induces an improper prior for β_i of the form $\pi(\beta_i) \propto \frac{1}{|\beta_i|}$.

3. A well-known result shows that the Student t distribution on $\lambda > 0$ degrees of freedom, scale parameter $\gamma > 0$, can be expressed using an inverse-gamma mixing distribution

   $$g(\psi_i) = \mathrm{IG}\left(\frac{\lambda}{2}, \frac{\gamma^2\lambda}{2}\right), \tag{6}$$

   where $\mathrm{IG}(a, b)$ is the inverse of a gamma with shape a and natural parameter b.

4. One possible extension to the exponential mixing distribution is the gamma distribution

   $$g(\psi_i) = \mathrm{Ga}\left(\psi_i \,\middle|\, \lambda, \frac{1}{2\gamma^2}\right),\ 0 < \lambda, \gamma < \infty. \tag{7}$$

   The double exponential distribution is regained if $\lambda = 1$ and as λ becomes smaller the mixing distribution can put more mass close to zero. The corresponding marginal

distribution of β is often called a normal-gamma (NG) or variance-gamma distribution which has proved a popular choice for modelling fat tails in finance (*e.g.* Bibby and Sorensen, 2003) and is a member of the generalized hyperbolic family (see *e.g.* Barndorff-Nielsen and Blaesild 1981). The marginal distribution of β_i has the density

$$\pi(\beta_i) = \frac{1}{\sqrt{\pi}2^{\lambda-1/2}\gamma^{\lambda+1/2}\Gamma(\lambda)}|\beta_i|^{\lambda-1/2}K_{\lambda-1/2}(|\beta_i|/\gamma) \tag{8}$$

where K is the modified Bessel function of the third kind. The variance of β_i is $2\lambda\gamma^2$ and the excess kurtosis is $\frac{3}{\lambda}$.

5. Another extension arises from placing a further mixing distribution on the scale parameter of the exponential mixing distribution. A gamma mixing distribution with parameters λ, γ^2 on the natural parameter of the exponential leads to a subclass of the gamma-gamma distribution (Bernardo and Smith, 1994, p120). The density of the mixing distribution on ψ_i has the form

$$g(\psi_i) = \frac{\lambda}{\gamma^2}(1 + \psi_i/\gamma^2)^{-(\lambda+1)} \; 0 < \lambda, \gamma < \infty. \tag{9}$$

The density of the marginal distribution of β_i can be expressed as

$$\pi(\beta_i) = \frac{\lambda}{\sqrt{\pi}}\frac{2^\lambda}{\gamma}\Gamma(\lambda + 1/2)\exp\left\{\frac{1}{4}\frac{\beta^2}{\gamma^2}\right\}D_{-2(\lambda+1/2)}\left(\frac{|\beta|}{\gamma}\right) \tag{10}$$

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 $\frac{|\beta_i|}{\gamma}$

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

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, unlike the normal-gamma class, can define distributions for which the variance is undefined and thus has a rather different tail-to-spike balance. The distribution function of ψ_i is also available in closed form as

$$G(\psi_i) = 1 - \left(1 + \frac{\psi_i}{\gamma^2}\right)^{-\lambda}.$$

We will refer to the marginal distribution of β_i with density (10) as the normal-exponential-gamma (NEG) distribution and the marginal distribution of ψ_i as the exponential-gamma (EG) distribution.

We would expect all of these methods to improve upon a normal prior distribution with fixed variance, which would have the mixing distribution

$$g(\psi_i) = \delta_{\psi_i = \sigma^2_\beta},$$

since moving some mass in the mixing distribution either to zero in the case of (3) or close to zero in (4), (5), (6), (7) and (9) is consistent with our prior belief that many of the regression coefficient are close to zero and hence their values will be drawn from distributions with small variances. A natural starting point would be to re-consider equation (4) and question whether it accurately reflects our prior beliefs. If not, a wider class of prior distribution can be generated by elaborating the exponential mixing distribution, leading to the Student t, NG or NEG above. The relative merits of these are discussed in what follows.

2.2 Shapes and Limits

Some of the mixing distributions described above and their corresponding densities for β are displayed in Figure 1. Generally the expectation of the normal variance ψ is fixed at unity by appropriate setting of the hyperparameters, except when this does not exist as in the last pair of figures when for the NEG $\lambda = 0.1$, for which the expectation of $1/\psi$ is fixed at unity.

Aside from incorporating the density of the 'lasso' as a special case many of the scale mixture of normals will have the normal-Jeffreys as a limiting density form. For example the normal-gamma (NG) given by (8) goes to this improper limit when $\lambda \downarrow 0$ and $\gamma \uparrow \infty$. This degenerate limiting form has infinite mass, an infinite spike at zero and flatness for large values of $|\beta|$, and as a consequence does not penalise such large values. The spike at zero has strong consequences for the modal behaviour of the posterior, not all of them welcome as we shall see. Whereas the normal-gamma does have an infinite spike at zero for $\lambda \leq 1/2$, the normal-exponential-gamma distribution has the advantage of a finite limit at zero for all parameters values in its range and incorporates as limiting cases the double exponential prior (as $\lambda, \gamma \uparrow \infty$) and the normal-Jeffreys case (as $\lambda, \gamma \downarrow 0$).

In the distribution of β, we now compare the relative weights centrally versus in the tails of NG, NEG, DE, NJ and Student *t*. For all choices of prior (except the normal-Jeffreys), at least one scale parameter must be chosen. For comparison we simply specify one scale parameter by fixing probability mass on the central region $(-\epsilon, \epsilon)$ to be η. Figure 2 illustrates the effect of fixing $\eta = 0.9$ on the region $(-0.01, 0.01)$ for the four comparisons with the lasso, (a) DE v NEG, (b) DE v NG, (c) DE v t and (d) DE v NJ. The normal-gamma

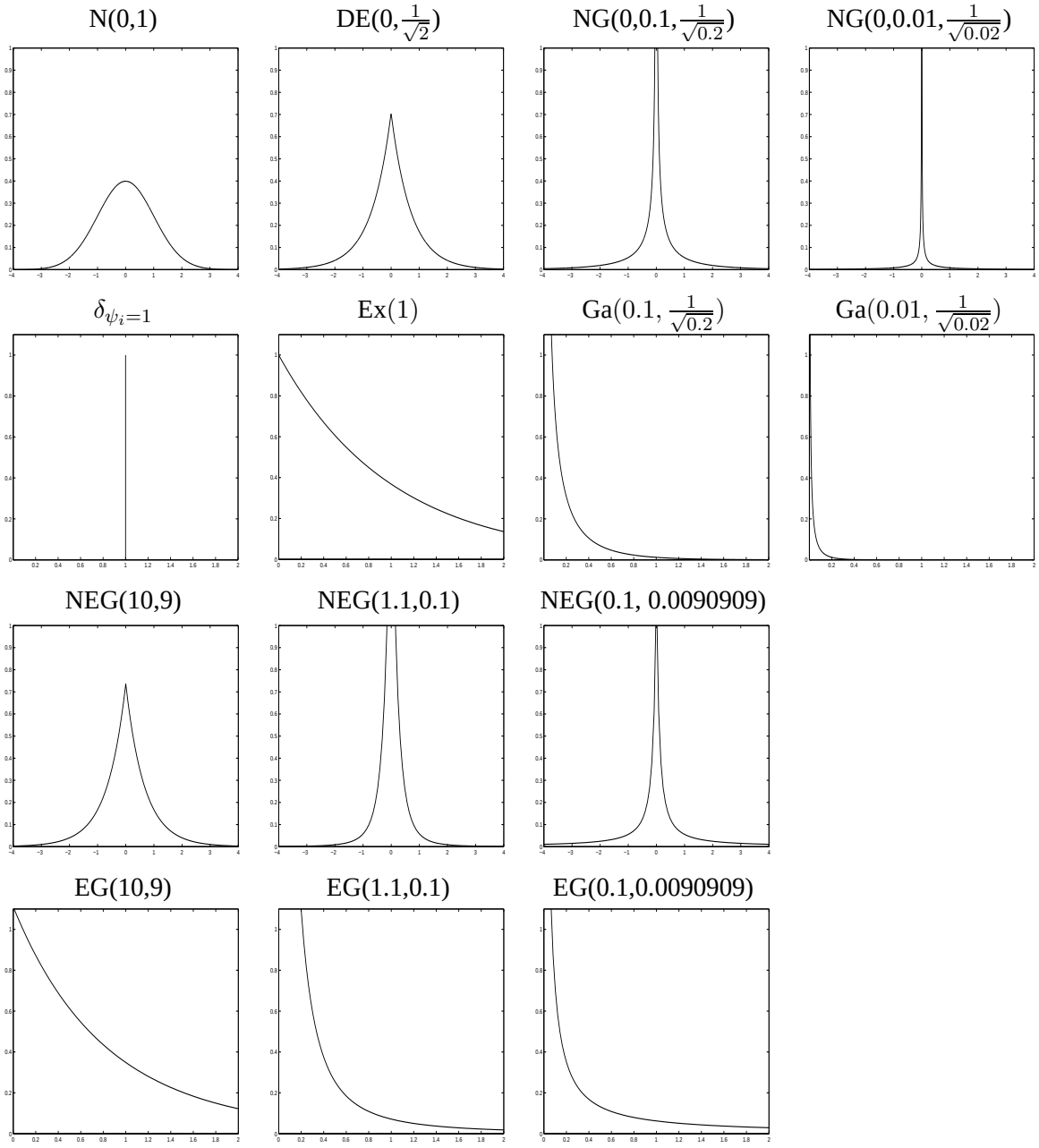


Figure 1: Various forms considered for the prior distribution for β with their associated mixing distribution

choice (panel (b)) is markedly different in tail behavior to the other three choices. The NEG distribution is able to maintain flat tails with a much larger value of the density of zero than the t-distribution and captures the main features of the normal-Jeffreys prior. In summary the DE and NJ are at opposite extremes with the NEG preserving good features of the NJ

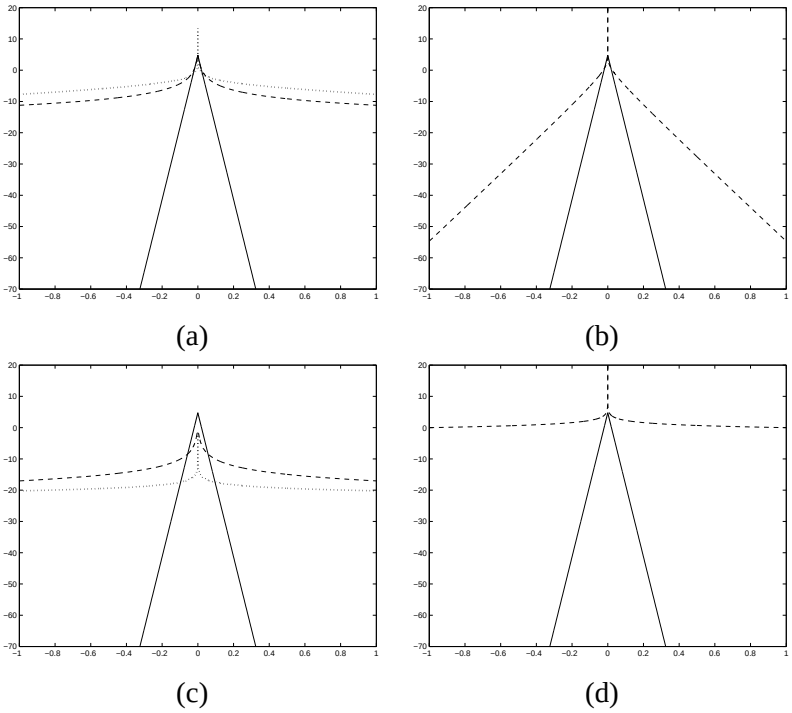


Figure 2: Log prior densities setting the central region (-0.01,0.01) to have probability $\eta = 0.9$ for: (a) double exponential distribution (solid line), NEG ($\lambda = 1$) (dashed line) and NEG($\lambda = 0.1$) (dotted line), (b) double exponential distribution (solid line) and NG ($\lambda = 0.1$) (dashed line), (c) double exponential (solid line), t-distribution ($\lambda = 2$) (dashed line) and t-distribution ($\lambda = 0.2$), and (d) double exponential (solid line) and improper normal-Jeffreys (dashed line)

without the drawback of the extreme spike at zero.

2.3 Thresholding for variable selection

The five distributions can express our belief that a small number of regressors can fit the data well but also allow a wide-range of other properties. It is important to choose appropriate forms that lead to a useful variable selection procedure.

A standard interpretation of Bayes theorem, is that the log posterior distribution is additive in data and prior information as given by

$$\log \pi(\beta|y) = \log f(y|\beta) + \log \pi(\beta), \tag{11}$$

where log probability is a measure of utility (Bernardo and Smith, 1994). It is natural to regard the negative prior utility as a penalty function given as $p(\beta)$, where

$$p(\beta) = -\log \pi(\beta).$$

It is the relative contribution of the two components on the right hand side of (11) that determines the posterior. Turning points of the posterior are then obtained by setting to zero the derivative of (11) and hence depend on the sum of the classical *efficient score function*, $-\partial \log f(y|\beta)/\partial\beta$ and the derivative of the penalty function. In the case of a single parameter, we will generally assume that turning-point (TP) thresholding, that is setting the penalized estimator $\tilde{\beta} = 0$, will occur iff there is no turning point. In which case with the class of penalty functions considered, the posterior is monotone decreasing in $|\beta|$ that is the only mode is at $\beta = 0$. Strictly if there is a turning point and the posterior function is non monotone then there may also be a mode at zero. A preference for a turning point follows the approach of Fan and Li (2001) and could be more formally computed by consideration of probability mass in the neighbourhood of zero, even when there is a spike at zero. An alternative choice, more simply computed with many regressors, is the true posterior mode which will be called the Bayesian threshold, that is the mode with the highest posterior mass. If there is one regressor, the lasso case, where the prior distribution is double exponential, is the only one of our chosen distributions where these thresholds are identical (see Appendix 1). Various penalty functions together with their derivatives are listed in Table 1.

	$p(\beta)$	$p'(\|\beta\|)$
double exponential$(0, \frac{1}{\gamma})$	$\frac{\|\beta\|}{\gamma}$	$\frac{1}{\gamma}$
normal-Jeffreys	$\log\|\beta\|$	$\frac{1}{\|\beta\|}$
IG$\left(\frac{\lambda}{2}\frac{\lambda\gamma^2}{2}\right)$	$\frac{\lambda+1}{2}\log(1+\beta^2/\lambda\gamma^2)$	$\frac{\lambda+1}{\lambda\gamma^2+\beta^2}\|\beta\|$
normal-gamma	$\left(\frac{1}{2}-\lambda\right)\log\|\beta\| - \log K_{\lambda-1/2}\left(\frac{\|\beta\|}{\gamma}\right)$	$\frac{1}{\gamma}\frac{K_{\lambda-3/2}\left(\frac{\|\beta\|}{\gamma}\right)}{K_{\lambda-1/2}\left(\frac{\|\beta\|}{\gamma}\right)}$
NEG	$-\frac{\beta^2}{4\gamma^2} - \log D_{-2(\lambda+\frac{1}{2})}\left(\frac{\|\beta\|}{\gamma}\right)$	$\frac{(\lambda+1/2)}{\gamma}\frac{D_{-2(\lambda+1)}\left(\frac{\|\beta\|}{\gamma}\right)}{D_{-2(\lambda+\frac{1}{2})}\left(\frac{\|\beta\|}{\gamma}\right)}$

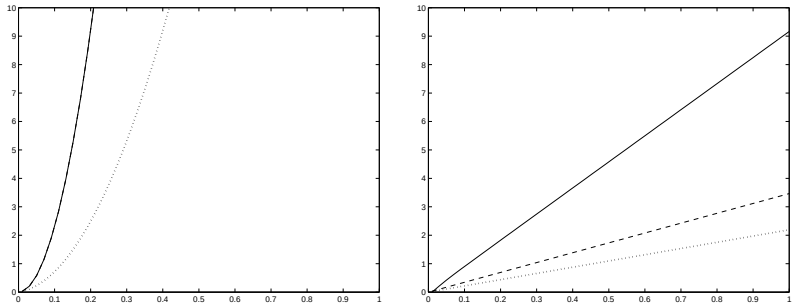
Table 1: Penalty functions and their derivatives induced by various choice for the hyperprior

Our approach will be applied to the generic problem of multiple regression, with the generalised linear model as a possible extension. It is assumed that we observe an $(n \times k)$-dimensional data matrix, X, and an $(n \times 1)$-dimensional response, y. The relationship between the responses and the data is modelled by a linear regression

$$\pi(y|\beta, \sigma^2, X) = \mathrm{N}(y|X\beta, \sigma^2 I)$$

where $\mathrm{N}(x|\mu, \Sigma)$ denotes a multivariate normal distribution with mean μ and variance Σ. The problem of finding a maximum *a posteriori* (MAP) estimate of β can be expressed as a penalised likelihood problem where β is chosen to find a minimum of the function

$$L = \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) + \sum_{i=1}^{k} p(|\beta_i|) \tag{12}$$



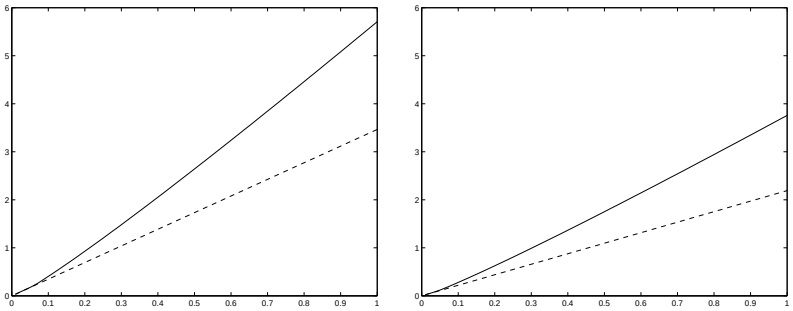
(a) (b)

Figure 3: TP thresholding rule for $\hat{\beta}$ as a function of the standard error under different prior choices with $\eta = 0.9$ and $\epsilon = 0.01$: (a) double exponential distribution (solid line) and normal-gamma ($\lambda = 0.1$) (dotted line), and (b) normal-exponential-gamma distributions with $\lambda = 10$ (solid line), $\lambda = 1$ (dashed line) and $\lambda = 0.1$ (dotted line)

where $p(x) = -\log \pi(x)$ is the penalty function. In generalised linear models the negative log-likelihood or deviance replaces the first term of (12). A particular case is probit regression as applied in section 4 where the information content of the likelihood is somewhat less than in the normal linear model.

Fan and Li (2001) consider the link between the choice of penalty function (or prior distribution in our case) and the TP thresholding value. The MAP estimate will be zero only if the maximum likelihood estimate (MLE) is smaller than this threshold value. In a univariate regression problem, for the maximum likelihood estimator $\hat{\beta}$, the parameter is set to zero if $|\hat{\beta}| < \min_{\theta \neq 0}\{|\theta| + \frac{\sigma^2}{X^T X} p'(|\theta|)\}$ where $p'(\cdot)$ is the derivative of the penalty function and $\frac{\sigma}{\sqrt{X^T X}}$ is the standard error of $\hat{\beta}$. A comparison with some of the prior distributions described above is illuminating. For the double exponential prior distribution, thresholding occurs if $|\hat{\beta}| < \frac{1}{\gamma}\frac{\sigma^2}{X^T X}$ which depends on the square of the standard error. In contrast, the normal-Jeffreys prior thresholds according to the rule $|\hat{\beta}| < 2\frac{\sigma}{\sqrt{X^T X}}$ and the thresholding depends linearly on the standard error. Figure 3 compares the thresholding rules for the normal-gamma penalty and the normal-exponential-gamma penalty. The latter has linear behaviour where the slope depends on λ, generalising the normal-Jeffreys rule and is thus more appealing. The normal-gamma case has substantially different behaviour and defines a much more conservative criterion. Much larger values of γ would induce a linear thresholding rule but this contradicts our imposed prior property of a large mass close to zero.

The Bayesian threshold for the normal-Jeffreys and normal-gamma choices with $\lambda < 0.5$ are undefined because the prior density value at 0 is infinite and the posterior mode is consequently zero for any set of observations. However, the NEG prior distribution always has a finite mode at zero. Figure 4 compares the TP and Bayesian thresholding rules. The



$\lambda = 0.1$ $\lambda = 1$

Figure 4: Bayesian threshold (solid line) and TP threshold (dashed line) for the NEG prior distribution with $\eta = 0.9$ and $\epsilon = 0.01$

Bayesian threshold is more conservative and almost doubles the thresholding value.

The discussion so far has centred around thresholding but the choice of penalty function will also have implications for the shrinkage of non-zero estimates. For example, Johnstone and Silverman (2005) suggest that overshrinkage of non-zero estimates can lead to better predictive performance in wavelet regression. Differentiating (12), the relationship between the penalised MLE $\tilde{\beta}$ and the MLE $\hat{\beta}$

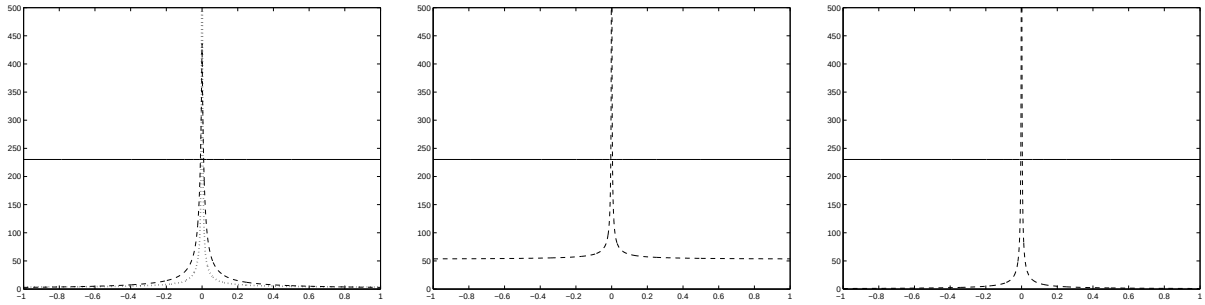
$$\frac{\hat{\beta} - \tilde{\beta}}{\sigma^2 / \sum_{i=1}^{n} x_i^2} = \text{sign}(\tilde{\beta}) p'(|\tilde{\beta}|)$$

shows that the amount of shrinkage is directly controlled by the derivative of the penalty function. Figure 5 illustrates various choice of penalty function with a chosen value of the probability mass η on the interval $(-\epsilon, \epsilon)$. The flat tails of the normal-Jeffreys and normal-exponential-gamma distributions lead to small derivative for large values of $\hat{\beta}$ and $\tilde{\beta} \approx \hat{\beta}$, which implies the so-called oracle property of Fan and Li (2001). The normal-gamma choice maintains a substantial derivative in the tails (which is approximately $\frac{1}{\gamma}$).

2.4 Modal estimates with multiple parameters

The following section extends the univariate results to problems with two regressors. First, for k parameters, returning to the penalised likelihood function, L, the derivative can be expressed as

$$\begin{aligned} \frac{dL}{d\beta} &= X^T X\beta - X^T y + \text{sign}(\beta) p'(|\beta|) \\ (X^T X)^{-1} \frac{dL}{d\beta} &= \beta - \hat{\beta} + (X^T X)^{-1} \text{sign}(\beta) p'(|\beta|) \end{aligned} \tag{13}$$



(a) (b) (c)

Figure 5: Penalty functions if $\eta = 0.9$ and $\epsilon = 0.01$ for: (a) double exponential distribution (solid line), NEG ($\lambda = 1$) and NEG($\lambda = 0.1$), (b) double exponential distribution (solid line) and NG ($\lambda = 0.1$) (dashed line), and (c) double exponential (solid line) and normal-Jeffreys (dashed line)

where

$$\text{sign}(\beta) = \begin{pmatrix} \text{sign}(\beta_1) & & 0 \\ & \ddots & \\ 0 & & \text{sign}(\beta_2) \end{pmatrix}, |\beta| = \begin{pmatrix} |\beta_1| \\ \vdots \\ |\beta_2| \end{pmatrix}$$

Turning points away from zero can only occur if there exists a value of β for which some elements of $\frac{dL}{d\beta}$ are zero. The mode with the largest number of non-zero parameter estimates will be preferred. In the bivariate case, we assume that

$$X^T X = \begin{pmatrix} c & -\rho\sqrt{cd} \\ -\rho\sqrt{cd} & d \end{pmatrix}$$

where c and d are the sum of squares for the first and second variable respectively and ρ is the correlation between the maximum likelihood estimators $\hat{\beta}_1$ and $\hat{\beta}_2$, which has the opposite sign to the correlation between the two independent variables

2.4.1 Lasso Regions

The relationship between thresholding and the values of $\hat{\beta}_1$ and $\hat{\beta}_2$ can be studied analytically for the lasso penalty. There are five regions into which $\hat{\beta}_1$ and $\hat{\beta}_2$ can fall which are shown in figure 6 (only positive correlation is considered; the relationship between $\hat{\beta}_1$ and $-\hat{\beta}_2$ shows the effect of negative correlation) and derived in the Appendix 2. Four of these regions arise when there is a single posterior mode. Each region is defined by a combination of thresholding or not thresholding either estimate. However, a bimodal posterior distribution is also possible and figure 6 shows the values of $\hat{\beta}_1$ and $\hat{\beta}_2$ which lead to it as the lighest of the three grey shades. The five regions are colour coded, moving from white to black, as: no thresholding; bimodal; β_2 only; β_1 only; or both variables thresholded. In the following

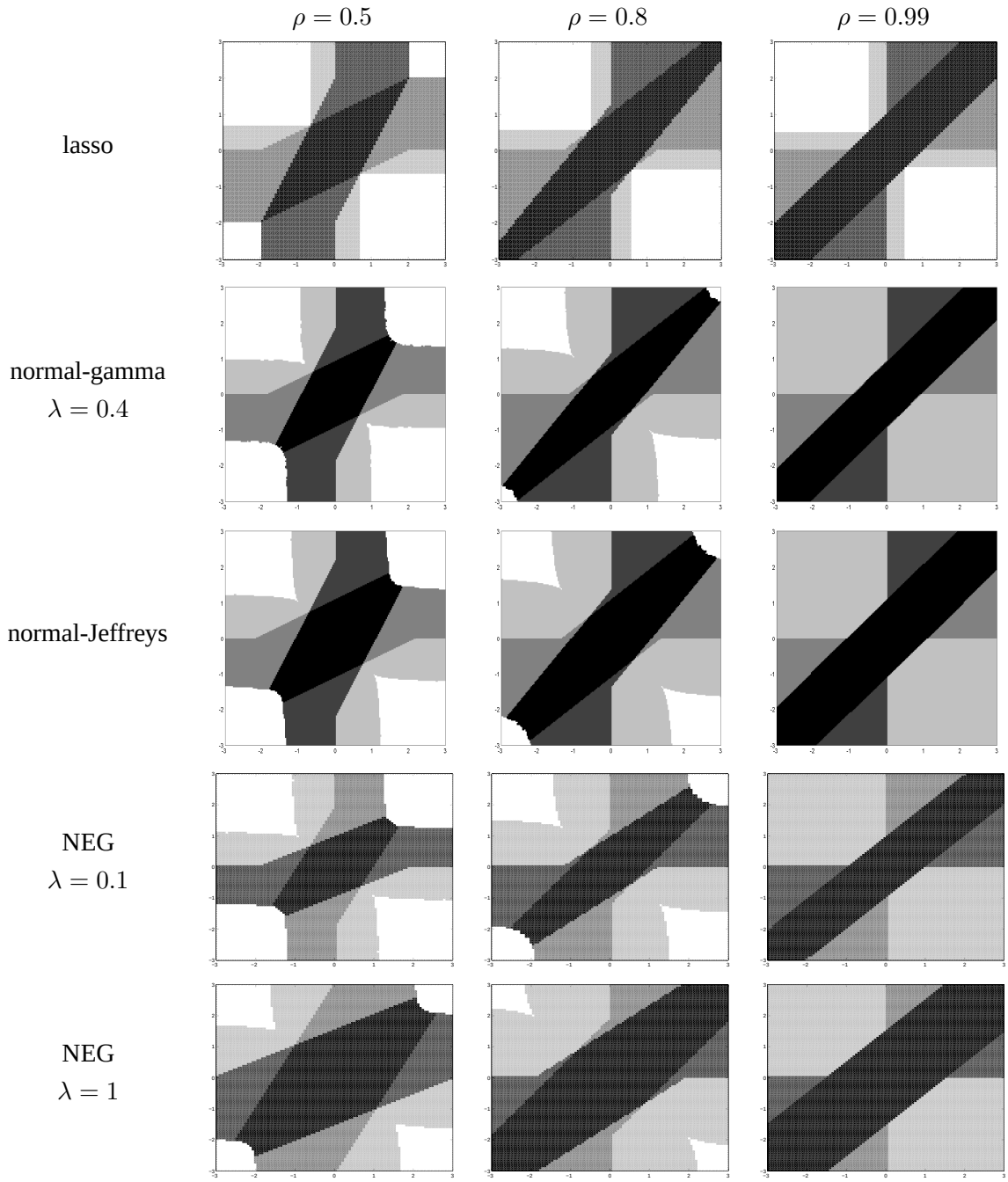


Figure 6: The regions where different types of thresholding occur either moving through shades of grey from black to white: only mode at 0 (black); β_1 set to zero only; β_2 set to zero only; two local modes; internal mode (white) for $c = 1$, $b = 1$

section, we will discuss resolving the bimodality by using the global posterior mode as the estimate. Each graph is symmetric in the lines $\hat{\beta}_1 = \hat{\beta}_2$ and $\hat{\beta}_1 = -\hat{\beta}_2$. The values of

$\hat{\beta}_1$ and $\hat{\beta}_2$ where no thresholding occurs clearly define four disjoint squares. This property is independent of correlation but the region where both regressors are thresholded forms a rhomboid whose shape changes with the value of ρ. This agrees with the observation that if there is high correlation between the regressors there is a tendency for the MLEs to produce spurious relationships. In those situations, $\hat{\beta}_1$ and $\hat{\beta}_2$ have similar absolute values and the predicted values will be near constant. The volume of the region will be determined by the ratios $\frac{1}{\gamma\sqrt{c}}$ and $\frac{1}{\gamma\sqrt{d}}$.

2.4.2 General Regions

In contrast with the lasso regions, the shapes of non-thresholded regions (white in figure) depend on the correlation for the normal-gamma and normal-Jeffreys penalty functions (Figure 6). These relationships are less amenable to analytical work and the regions are drawn by finding the type of thresholding on a grid of values. Both penalty functions lead to similar regions which are substantially different to those defined by the lasso penalty. Two striking differences are the shape of the region where both variables are thresholded and the shape of the region with a bimodal posterior. If both ML estimators have the same sign the no thresholding region becomes larger whereas if the signs are different the no thresholding region becomes smaller. The gap is filled by an expansion of the region with a bimodal posterior. These regions are intermediate between full thresholding (black) and no thresholding (white). This region is small and close to all axes with the double exponential prior but the shape depends on the correlation in the NEG case. In fact, the largest value of the correlation leads to this region filling almost all of the two quadrants where $\hat{\beta}_1$ and $\hat{\beta}_2$ have opposing signs. In other words, the thresholding depends on the difference of $\hat{\beta}_1$ and $\hat{\beta}_2$ and for correlations close to -1, the thresholding depends on the sum of $\hat{\beta}_1$ and $\hat{\beta}_2$.

The lasso and NEG penalties also define Bayesian thresholding regions (Figure 7). Unlike the one-dimensional case, the Bayesian and TP thresholding regions differ with a lasso penalty. The bi-modal region is divided into regions where one variable is thresholded. In contrast, the NEG penalty defines a substantially larger region where both estimates are shrunk to zero. Otherwise one of the regressors is set to zero and the line $\hat{\beta}_1 = -\hat{\beta}_2$ acts as a dividing line between these two cases. The difference in thresholding between the lasso and NEG penalty suggest that the latter will shrink more variables from the model.

It is hard to make any general comments about thresholding in higher dimensions, suffice that there are $\min(n, k)$ non-zero estimates. In the case of infinite spikes at zero (NJ, NG for $\lambda \leq 1/2$) then this infinite spike will persist for all subsets of at most $\min(n, k)$ genes.

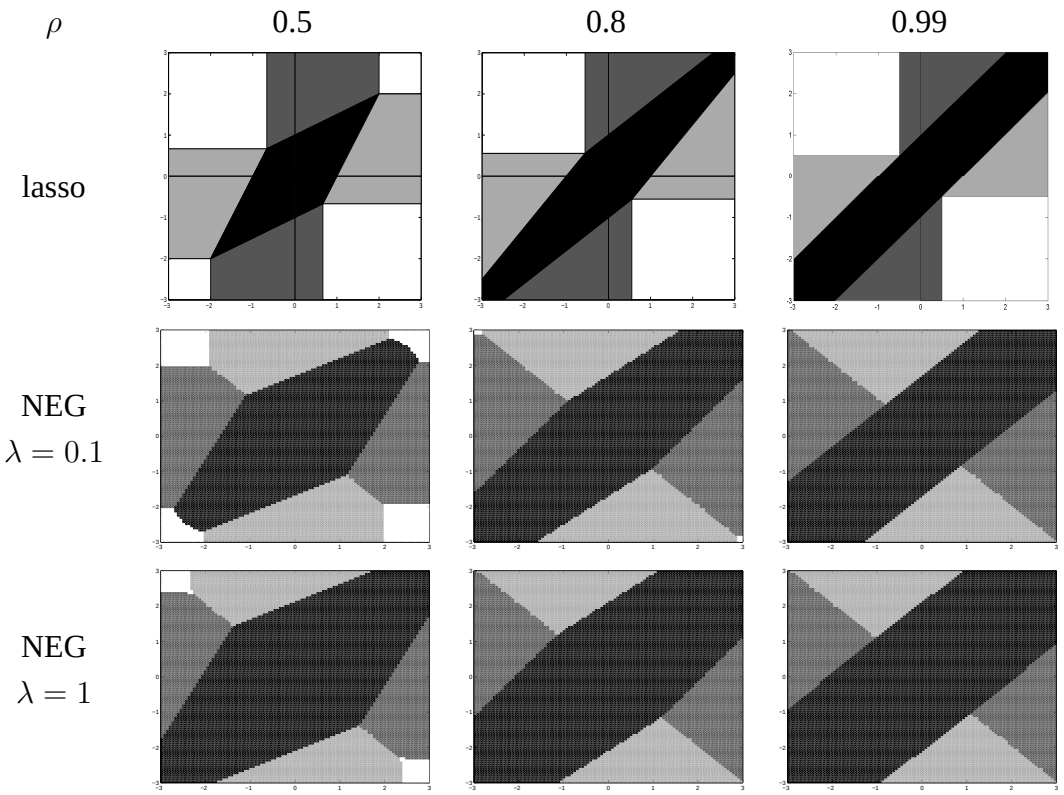


Figure 7: The Bayesian thresholding region with a NEG distribution. The parameters are chosen such that $\pi(\beta \in [-0.01, 0.01]) = 0.25$ for various values of λ.

2.5 Relationship to model choice

Heuristically, we can think of the posterior mode as a variable selection method since setting a regression coefficient to zero removes a variable from the model. It is useful to define an indicator variable s_i that takes the value 0 if the i-th regressor is excluded from the model (when $\hat{\beta}_i = 0$) and 0 otherwise (when $\hat{\beta}_i \neq 0$). For fixed $s = (s_1, \ldots, s_k)$, local posterior modes obey the condition

$$0 = \beta^\star - \hat{\beta}^\star + (X^{\star T} X^\star)^{-1} \text{sign}(\beta^\star) p'(|\beta^\star|)$$

where $X^\star$ is the submatrix of X constructed using the columns for which $s_i = 1$ and $\beta^\star = \{\beta_i | s_i = 1\}$. If such a posterior mode $\tilde{\beta}^\star$ exists then

$$\hat{\beta}^\star = \tilde{\beta}^\star + (X^{\star T} X^\star)^{-1} \text{sign}(\tilde{\beta}^\star) p'(|\tilde{\beta}^\star|)$$

where $\hat{\beta}^\star$ is the ML estimate of $\beta^\star$. The value of s that minimises

$$L = \frac{1}{2\sigma^2}(y - X^\star \tilde{\beta}^\star)^T (y - X^\star \tilde{\beta}^\star) + \sum_{i|s_i=1} p(|\tilde{\beta}_i^\star|) + \sum_{i|s_i=0} p(0).$$

corresponds to the global posterior mode of β. The normal-Jeffreys and NEG with small λ define a penalty that is almost constant for a range of suitably large values of $\tilde{\beta}_i^\star$. This penalty is represented by p_1 and L simplifies to

$$\begin{aligned} L &= \frac{1}{2\sigma^2}(y - X^\star\tilde{\beta}^\star)^T(y - X^\star\tilde{\beta}^\star) + rp_1 + (k-r)p_2 \\ &= \frac{1}{2\sigma^2}(y - X^\star\tilde{\beta}^\star)^T(y - X^\star\tilde{\beta}^\star) + k\,p_2 + r(p_1 - p_2) \end{aligned}$$

where $p_2 = p(0)$ and r is the number of non-zero estimates. The term $k\,p_2$ is constant across all s and can be dropped which leaves the criterion

$$\frac{1}{\sigma^2}(y - X^\star\tilde{\beta}^\star)^T(y - X^\star\tilde{\beta}^\star) + 2r(p_1 - p_2),$$

where the first term is more generally the deviance.

The indicator variables that correspond to the posterior mode defines a model selection criterion that is a trade-off between goodness-of-fit and a penalty for each included parameter. This form has been a recurring idea in the model selection literature. Standard choices for the penalty are Akaike's information criteria (AIC) (Akaike, 1974) where $p_1 - p_2 = -1$ and a Bayesian variant (BIC) (Schwarz, 1978) $p_1 - p_2 = -\frac{1}{2}\log n$. A typical choice for NEG of $\lambda = 0.1$, $\eta = 0.9$ and $\epsilon = 0.01$ would lead to values of $p_1 - p_2$ around -15, which is substantial larger than the penalties under the AIC and BIC for values of n which are of the order of hundreds of observations. The penalty is much closer to the Risk Inflation Criterion (RIC) (Foster and George 1994) who choose $p_1 - p_2 = -\log k$ for large k.

A further decomposition shows the relationship between the residual sum of squares calculated using the least squares estimates,

$$\frac{1}{\sigma^2}(y - X^\star\hat{\beta}^\star)^T(y - X^\star\hat{\beta}^\star) + \frac{1}{\sigma^2}(p'(|\tilde{\beta}^\star|))^T(X^{\star T}X^\star)^{-1}p'(|\tilde{\beta}^\star|) + 2r(p_1 - p_2).$$

3 Inference for regression and probit regression

This section discusses posterior inference, in particular methods for finding local posterior modes, for probit regression models using the classes of prior distributions already described. Initially we concentrate on estimation for a normal prior distribution which will be an important component of our analysis.

3.1 Estimation with normal prior distributions

The prior distribution for $\beta, (k \times 1)$ is assumed to have the form

$$\pi(\beta) = \mathrm{N}(\beta|0, \Psi)$$

where Ψ is a $(k \times k)$-matrix. Typically this matrix will be a diagonal matrix although the derivations in this section do not assume this special form. The standard MLE estimator will not be defined if k is larger than n. Consequently, the problem is re-expressed in terms of an n-dimensional parameter, γ, for which the MLE exists. As in West (2003), the singular value decomposition of X can be written as $X = F^T D A^T$ where A is $(k \times n)$-dimension matrix such that $A^T A = I_n$, D is an $(n \times n)$-dimension diagonal matrix and F is $(n \times n)$-dimension matrix for which $F^T F = I_n$ and $F F^T = I_n$. Clearly, we can write

$$X\beta = (F^T D)\gamma$$

where $\gamma = A^T \beta$ and the MLE, $\hat{\gamma}$, of γ is well-defined and has the form

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

The sampling distribution $\hat{\gamma}$ and the prior distribution of the n-dimensional parameter γ which is estimated by $\hat{\gamma}$ can be represented as

$$\pi(\hat{\gamma}|\gamma, \Psi, X) = \mathrm{N}(\hat{\gamma}|\gamma, \sigma^2 D^{-2} = \Lambda^\star),$$

$$\pi(\gamma|\Psi, X) = \mathrm{N}(0, A^T \Psi A = \Psi_0)$$

and the posterior distribution of γ is

$$\pi(\gamma|\hat{\gamma}, \Psi, X) = \mathrm{N}(\gamma|\Psi_0(\Psi_0 + \Lambda^\star)^{-1}\hat{\gamma}, (\Lambda^{\star\,-1} + \Psi_0^{-1})^{-1}). \tag{14}$$

In order to calculate the posterior distribution of the regression parameters, β, we consider the full singular value decomposition which represents X as $F^T D^* K^T$ where the first n columns of $K, (k \times k)$ are $A, (k \times n)$, the last $(n - k)$ columns as C given as $K = (A\ , C)$, and $D^*, (n \times k)$ with

$$D^\star = \begin{pmatrix} D & 0 \end{pmatrix}.$$

In this case, $K^T K = I_k$ and $K K^T = I_k$ and K is invertible with $K^{-1} = K^T$. If $\gamma^\star = K^T \beta$, the first n elements of $\gamma^\star$ are γ and we define the last $(k - n)$ elements to be τ. In this parametrization τ are exactly those dimensions that are independent of the data. Using this re-parametrization, the posterior distribution of β is simply related to the posterior distribution for $\gamma^\star$ which can be expressed as

$$\pi(\gamma^\star|\hat{\gamma}, \Psi, X) = \pi(\tau|\gamma, \Psi)\, \pi(\gamma|\hat{\gamma}, \Psi, X)$$

where

$$\pi(\tau|\gamma, \Psi, X) = \mathrm{N}(\tau|C^T \Psi A (A^T \Psi A)^{-1}\gamma, C^T \Psi C - C^T \Psi A (A^T \Psi A)^{-1} A^T \Psi C).$$

$$\mathrm{E}(\gamma|\hat{\gamma},\Psi) = \Psi_0(\Psi_0+\Lambda^\star)^{-1}\hat{\gamma}$$

and

$$\mathrm{E}(\tau|\hat{\gamma},\Psi) = C^T\Psi A(A^T\Psi A)^{-1}\mathrm{E}(\gamma|\hat{\gamma}).$$

The normality of both $\pi(\tau|\gamma,\Psi)$ and $\pi(\gamma|\hat{\gamma},\Psi,X)$ combined with the linear mean of τ in γ implies that $\gamma^\star$ has a normal posterior distribution. The transformation from β is well-defined and has the form $\beta = K\gamma^\star$ implying that β will also be normally distribution *a posteriori*. This distribution can be characterised by its posterior mean and variance. Computationally, we want to calculate these quantities whilst avoiding inversions of $(k\times k)$-dimensional matrices. After some simplification we can express the posterior mean and covariance in a form where only matrix that needs inverting is an $n\times n$-dimension matrix

$$\begin{aligned}\mathrm{E}(\beta|\Psi,\hat{\gamma}) &= \Psi A(A^T\Psi A)^{-1}\mathrm{E}(\gamma|\hat{\gamma},\Psi)\\ &= \Psi A(A^T\Psi A)^{-1}(\Psi_0^{-1}+\Lambda^{\star\,-1})^{-1}\Lambda^{\star\,-1}\hat{\gamma}\\ &= \Psi A(\Psi_0+\Lambda^\star)^{-1}\hat{\gamma} \qquad (15)\end{aligned}$$

and

$$\begin{aligned}\mathrm{V}(\beta|\Psi,\hat{\gamma}) &= \Psi - \Psi A(A^T\Psi A)^{-1}A^T\Psi + \Psi A(A^T\Psi A)^{-1}\mathrm{V}_{\gamma|\hat{\gamma},\Psi}(\gamma)(A^T\Psi A)^{-1}A^T\Psi\\ &= \Psi - \Psi A(A^T\Psi A)^{-1}A^T\Psi + \Psi A(A^T\Psi A)^{-1}(\Psi_0^{-1}+\Lambda^{\star\,-1})^{-1}(A^T\Psi A)^{-1}A^T\Psi\\ &= \Psi - \Psi A(\Psi_0+\Lambda^\star)^{-1}A^T\Psi. \qquad (16)\end{aligned}$$

Finally, we note that the marginal distribution of $\hat{\gamma}$ given Ψ can also be derived and has the form

$$\pi(\hat{\gamma}|\Psi) = \mathrm{N}(0, A^T\Psi A + \sigma^2 D^{-2}). \qquad (17)$$

3.2 Bayesian binary regression

The analysis of binary data arising from microarray experiments can exploit the normal theory developed thus far by introducing latent variables. There is also appeal in working directly with the log-likelihood as discussed earlier, see Kiiveri (2003). However here we focus on the method proposed by Albert and Chib (1993) which exploits a latent variable characterisation to reduce probit regression analysis to that of regression albeit at the expense of creating n latent variables. We assume that the response for the i-th individual is z_i and introduces latent parameters y_i such that

$$y_i|z_i,\beta \sim \mathrm{N}(X_i\beta, 1)$$

and $y_i > 0 \iff z_i = 1$. The model for z_i is a traditional probit regression analysis

$$\pi(z_i = 1|\beta) = \Phi(X_i\beta).$$

where Φ is the cumulative distribution function of a standard normal distribution. Importantly, if β has a normal prior distribution, the posterior distribution of $\beta|y_1, \ldots, y_n$ is also normal.

Much of the work using normal-Jeffreys penalty functions, Kiiveri (2003), Figueiredo (2003)) attempts 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 favouring the NJ overlooks 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 $1/\beta$ and integrates to $\log(\beta)$, which blows up at $zero$. See Gelfand and Sahu (1999) for more detailed analysis of such improprieties. 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 penalised 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.

3.3 Choosing hyperparameters

The standard subjectivist interpretation of the prior distribution is an expression of our beliefs about the likely values of β and, in this case, the number of non-zero regression coefficients needed to explain the variation in the responses. However, this approach can be problematic when combined with the MAP estimation procedure. Consider a probit regression model with a relatively diffuse prior distribution for β_0 (in the sense that its effect can be ignored when comparing local modes). The penalized likelihood function is

$$L = \sum_{i=1}^{n} z_i \log \Phi(\beta_0 + X_i\beta) + \sum_{i=1}^{n}(1 - z_i) \log(1 - \Phi(\beta_0 + X_i\beta)) - \sum_{i=1}^{k} p(|\beta_i|).$$

If only the j-th regressor takes a non-zero value, $\tilde{\beta}_j$, and the intercept is $\tilde{\beta}_0^{(j)}$ then

$$L = \sum_{i=1}^{n} z_i \log \Phi(\tilde{\beta}_0^{(j)} + X_{ij}\tilde{\beta}_j) + \sum_{i=1}^{n}(1-z_i) \log(1-\Phi(\tilde{\beta}_0^{(j)} + X_{ij}\tilde{\beta}_j)) - p(|\tilde{\beta}_j|) - (k-1)p(0).$$

Comparing this value to the penalized log likelihood for a "null model" for which all regression coefficients, except the intercept, are set to zero shows that the "null model" will be superior unless there is at least one regressor for which

$$\sum_{i=1}^{n} z_i \log \frac{\Phi(\tilde{\beta}_0^{(j)} + X_{ij}\tilde{\beta}_j)}{\Phi(\tilde{\beta}_0)} + \sum_{i=1}^{n}(1 - z_i) \log \frac{1 - \Phi(\beta_0^{(j)} + X_{ij}\tilde{\beta}_j)}{1 - \Phi(\tilde{\beta}_0)} > p(|\tilde{\beta}_j|) - p(0)$$

where $\tilde{\beta}_0$ is the estimated intercept in the null model. The improvement in the log likelihood, on the left-hand side of the equation, is bounded since a perfectly fitting model has log likelihood zero. If the difference between the penalty for a zero estimate and a typical non-zero estimate is large, we will define a penalty functions for which the "null model" is superior to all other model. However, we believe that a small number of genes will explain the differences between the classes. To avoid a problem of "over-penalisation", we first define L_{min}, the penalized log-likelihood for the null model,

$$\begin{aligned} L_{min} &= \log\hat{\theta}\sum_{i=1}^{n} z_i + \log(1-\hat{\theta})\sum_{i=1}^{n}(1-z_i) - kp(0) \\ &= n[\hat{\theta}\log\hat{\theta} + (1-\hat{\theta})\log(1-\hat{\theta})] - kp(0) \end{aligned}$$

where $\hat{\theta} = \frac{\sum_{i=1}^{n} z_i}{n}$. The log likelihood at any posterior mode lie must between L_{min} and 0. If we could find $\beta^\star$, a subset of β with k' elements which could perfectly fit the data, it would have penalized log likelihood

$$0 - \sum_{x\in\beta^\star} p(x) - (k-k')p(0).$$

The null model will not be the global mode if there is a subset $\beta^\star$ whose log posterior is greater than L_{min} or

$$\sum_{x\in\beta^\star} [p(|x|) - p(0)] < n[\hat{\theta}\log\hat{\theta} + (1-\hat{\theta})\log(1-\hat{\theta})].$$

The quantity on the left-hand side controls the level of thresholding and suggests a simple method for controlling its value relative to the log likelihood of the null model on the left-hand side. Decide on a value for k' and expected value for the estimate of a non-zero β, say φ, then

$$p(\varphi) - p(0) = \frac{n}{k'}[\hat{\theta}\log\hat{\theta} + (1-\hat{\theta})\log(1-\hat{\theta})].$$

where $\hat{\theta}$ is estimated from the data. Now we have a prior which enables us to fix the scale parameter γ, and being data dependent will tend to avoid overshrinkage and a mode at the origin. Although data dependent, the prior only depends on the data through design parameters, sample size, n, and proportion of observations in the disease group, $\hat{\theta}$.

3.4 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. The heavy tails of our

$g(\psi)$	$\mathrm{E}\left[\frac{1}{\psi_j}\middle\vert\beta_j\right]$
$\mathrm{Ga}(\nu,\frac{1}{2\gamma^2})$	$\frac{1}{\gamma\lvert\beta_j\rvert}\frac{K_{\nu-\frac{3}{2}}\left(\frac{\lvert\beta_j\rvert}{\gamma}\right)}{K_{\nu-\frac{1}{2}}\left(\frac{\lvert\beta_j\rvert}{\gamma}\right)}$
Jeffreys	$\frac{1}{\beta_j^2}$
$\mathrm{IG}(\lambda,\gamma^2/2)$	$\frac{1+2\lambda}{\beta_j^2+\gamma^2}$
EG	$\frac{(\lambda+1/2)}{\gamma\lvert\beta_j\rvert}\frac{D_{-2(\lambda+1)}\left(\frac{\lvert\beta_j\rvert}{\gamma}\right)}{D_{-2(\lambda+\frac{1}{2})}\left(\frac{\lvert\beta_j\rvert}{\gamma}\right)}$

Table 2: The forms of $\mathrm{E}[\frac{1}{\psi_j}|\beta_j]$ for some mixing distributions

prior distribution can lead to slow convergence. In general, we use the EM algorithm to find a promising and small subset of variables with non-zero regression coefficients. Once this subset has been found a standard optimization technique, such as conjugate gradient, can be used to find the posterior mode using the variables in the subset. In our case, the prior variances of the regression coefficients $\psi_1,\ldots,\psi_k$ and the unobserved values $y_1,\ldots,y_n$ are treated as missing data. Kiiveri (2003) suggests applying the EM algorithm directly to the 'likelihood times prior' in the generalised 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.

The standard EM algorithm outputs a sequence of estimates $\beta^{(1)},\beta^{(2)},\ldots$ that under regularity conditions converge to a local maximum of $\beta|z$. The sequence is defined by iterating between an E step and an M step

1. E-step: Let $\Lambda_{jj}^{(i)}=\frac{1}{\mathrm{E}[\frac{1}{\psi_j}|\beta^{(i-1)}]}$ for $j=1,\ldots,k$ and

$$y_j^{(i)}=\mathrm{E}\left[y\middle|\beta^{(i-1)}\right]=\begin{cases}\zeta_j-\frac{1}{\Phi(-\zeta_j)}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\zeta_j^2\right\} & \text{if } z_i=0\\ \zeta_j+\frac{1}{1-\Phi(-\zeta_j)}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\zeta_j^2\right\} & \text{if } z_i=1\end{cases}$$

   where $\zeta=X\beta^{(i-1)}$. The forms of $\mathrm{E}\left[\frac{1}{\psi_j}\middle|\beta^{(i-1)}\right]$ for various choices of penalty function are shown in table 2, with that for the Exponential Gamma prior derived in Appendix 3.

2. M-step: Set $\beta^{(i)}$ equal to the mode of $\pi(\beta|\Lambda^{(i-1)},y^{(i-1)})$, which will follow a normal distribution. The new value $\beta^{(i)}$ will be equal to the expectation of this distribution and a computationally efficient form is shown in equation (15).

The naive use of this EM algorithm can often lead to a sequence converging to the empty model where $\beta_j=0$ for all j. Several strategies lead to improved convergence of this EM

algorithm. A poorly chosen initial value $\beta^{(0)}$ can cause convergence problems. Before finding the posterior mode using these prior distributions, a posterior mode with a normal prior distribution with fixed variance $\Psi = I$ is found. A second problem we face is the lack of information from our data. If there are a large number of competing variables with similar, but useful, predictive properties, the algorithm will blindly remove all the variables because for any variable there are many other similar choices. A powered version of the likelihood is useful for counter-acting this problem. The idea is called Determinstic Annealing EM (DAEM) and was introduced by Udea and Nakano (1995) (see also McLachlan and Peel (2000), pp 58-60). They suggest multiplying the log-likelihood by a constant $\phi^{(i)}$ in the i-th iteration of the EM algorithm. The sequence should be chosen to converge to 1. We will assume that each observation occurs $q^{(i)}$ times in the dataset ($q^{(i)}$ and $\phi^{(i)}$ will have the same effect on the algorithm). The standard EM algorithm is run using this powered likelihood with a sequence of values for the power (a typical starting value would be 32) converging to 1. If both the likelihood and prior distribution were powered then this would be an annealing approach which should give better discrimination between competing posterior modes. Only powering the likelihood defines a pseudo-posterior distribution which gives more weight to the data than in the posterior distribution. We anticipate that this extra data information will guide the EM algorithm towards interesting areas of the parameter space. A powered likelihood will also lead to decreased standard errors for estimated parameters which should lead to less thresholding of variables. We expect that by smoothly changing the power, the thresholding also changes smoothly. Therefore, we hope to initially identify a promising subset of the variables associated with a large value of the power and shrink this set as the power decreases.

The second idea attempts to alleviate a practical problem that the algorithm can be overwhelmed by the large number of variables. In other words, the variables tend to be shrunk from the model at a uniform rate and with a large number of variables the data can often be fitted using relative small values of all regression coefficients. This will often lead to convergence to a mode at the origin. Updating subsets of the variables in the maximisation step of the EM algorithm allows us to vary the rates at which regression coefficients are shrunk to zero. In particular, only the k^* lowest $|\beta_i|$ are updated where k^* is uniformly distributed over the range $[0.5k, 0.8k]$. This step is initially alternated with a full update. After an initial period of alternating updates, only full updates are used. This will guarantee convergence of the algorithm. We would want to maximise the conditional distribution of the some parameter conditional on the other parameters. However, this exact updating is computationally expensive and we update β by maximising the marginal distribution of β and checking that this change leads to an increase in the posterior density value of β. A rather different promising strategy for improving convergence of the EM algorithm, not tried here, is that of parameter

expansion, see Liu *et al* (1998), or in the context of MCMC for the probit model Liu (2001, section 8.5).

4 Application to Arthritis Diagnosis

Bayesian thresholding using a NEG prior distribution is applied to a problem in immunology. The study measured gene expression level for 755 genes of known function for two groups of patients. A rheumatoid arthritis (RA) group with 24 subjects and a osteoarthritis (OA) group with 7 subjects. Three expression level measurements were taken for each sample and averaged to reduce noise-levels. The data was previously analysed in Sha *et al* (2003) and the interested reader is referred to this paper for a more detailed description of the experiments. The value of the λ parameter set to be 0.1 of the NEG distribution were chosen with the "typical" value of a non-zero regression coefficient) set to 2 and k' set to be 5 and 2.5. Here we will not attempt an exploration of sensitivity to hyperparameter choices λ and k' nor the model's application to other datasets.

Figure 8 shows the form of the penalty function and its derivative (which controls the amount of shrinkage) for the two choices of hyperparameter, $k' = 5$ solid line and $k' = 2.5$ dotted.

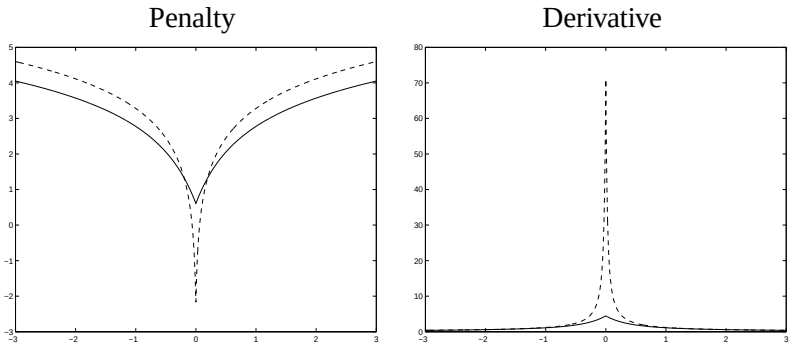


Figure 8: The penalty function $p(|\beta|)$, and its derivative for the two choices of hyperparameter, $k' = 5$ (solid), $k' = 2.5$ dotted

For both hyperparameter settings, the EM algorithm was started at 100 randomly chosen initial values of the regression coefficients. 89 and 60 distinct modes were found for $k' = 5$ and $k' = 2.5$ respectively. The posterior distribution of β is highly multi-modal with seemingly no overall dominating mode in both cases. Figure 9 shows the empirical cdf of $\log[\pi(y|\hat{\beta})\pi(\hat{\beta})]$ when $k' = 5, 2.5$ for all the local posterior modes found and also the log likelihood values as a function of number of genes selected. For comparison the lowest value

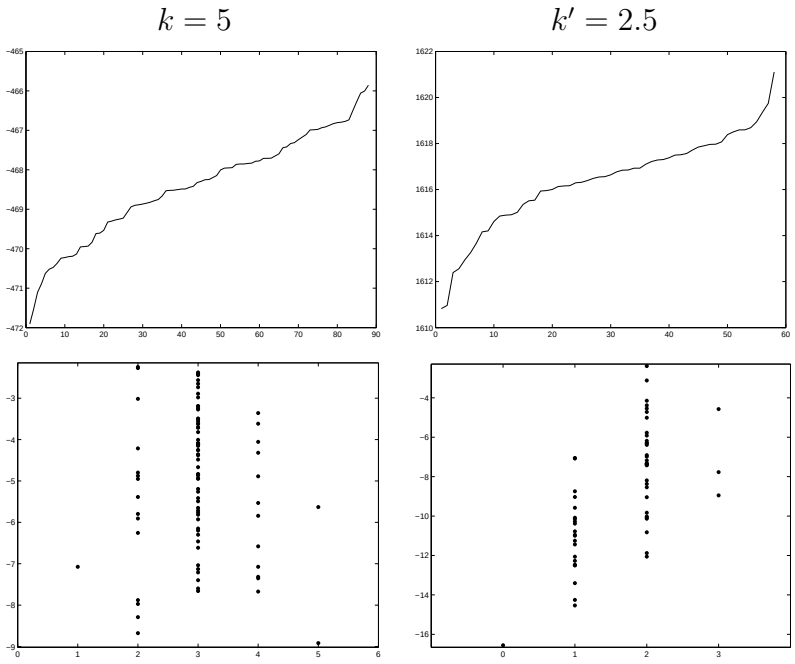
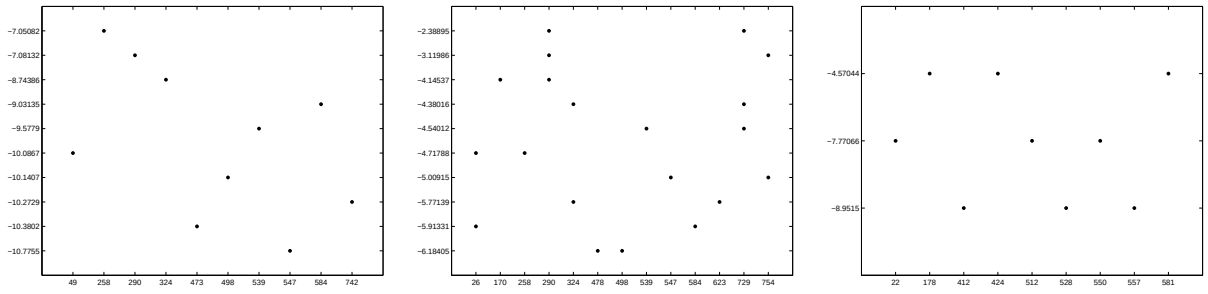


Figure 9: The top row shows plots of ranked values of $\log[\pi(y|\hat{\beta})\pi(\hat{\beta})]$ for the distinct values of the modes found and the bottom row show the values of the log likelihood for each mode (unsorted) as a function number of selected genes

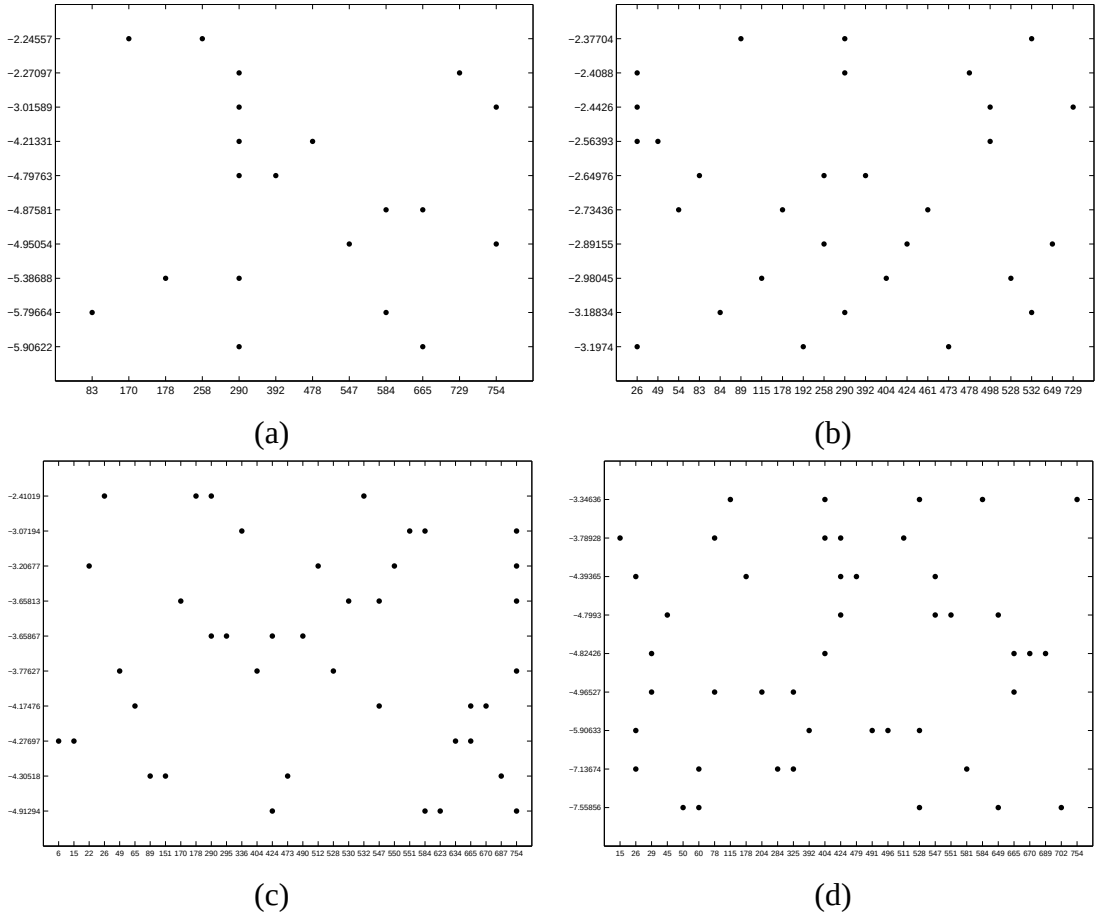
is associated with the model using no genes which has a log likelihood of -16.5589. Clearly many of the modes found have roughly similar posterior density values. Although, we would still like to express a preference for models including smaller numbers of regressors. Figures 10,11 show the actual genes chosen for $k' = 2.5, 5$. Figure 12 shows the regressors whose estimated coefficients are non-zero in the top ten modes ranked by their posterior density values with labels of selected genes on the x-axis. This posterior density includes the value of the penalty function and penalises less parsimonious models. Reducing the value of k' to 2.5 leads, unsurprisingly, to sparser models. Several of the genes appeared in both figures and in particular, the genes 290 appears in the top 3 modes. The data for the included variables with the dividing hyperplane (the locus of points for which the probability of membership to the two groups is equal with $k' = 2.5$) are shown in figure 13. This suggests that 290 has substantial power to distinguish between the two disease categories.

The effect of each gene on group membership can be gauged from figure 14 which plots the non-zero regression coefficient estimate for a selection of the genes. It shows that high expression levels of the genes 290 (Immunoglobulin Kappa heavy chain) and 754 (ZAP 70) are associated with a larger chance of belonging to the rheumatoid arthritis group than the osteoarthritis group, as does 729. In contrast, high expression levels of gene 170 are asso-



(a) (b) (c)

Figure 10: The log likelihood for the top modes with $k' = 2.5$ which include: (a) 1 variable, (b) 2 variables and (c) 3 variables



(a) (b) (c) (d)

Figure 11: The log likelihood for the top modes with $k' = 5$ which include: (a) 2 variable, (b) 3 variables, (c) 4 variables and (d) 5 variables

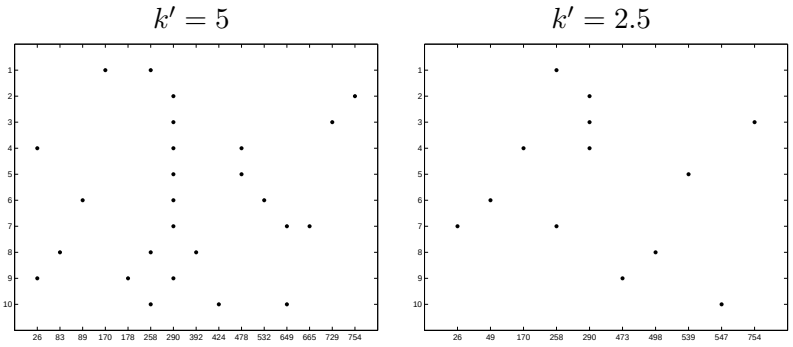


Figure 12: A summary of the ten modes with the highest posterior density values found by the algorithm

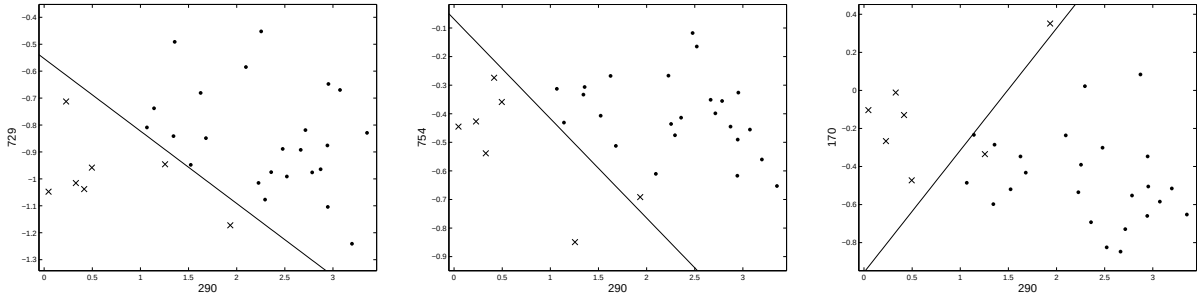


Figure 13: Plot of the gene expression levels for two genes for the rheumatoid arthritis group (dots) and the osteoarthritis group (crosses) with the fitted dividing hyperplane

ciated with a higher chance of membership of the osteoarthritis group. Gene 290 is coding for a B lymphocyte-specific gene and clearly is very strongly associated with disease class in this very small data set. It also came to the fore in the MCMC approach of Sha *et al* (2003).

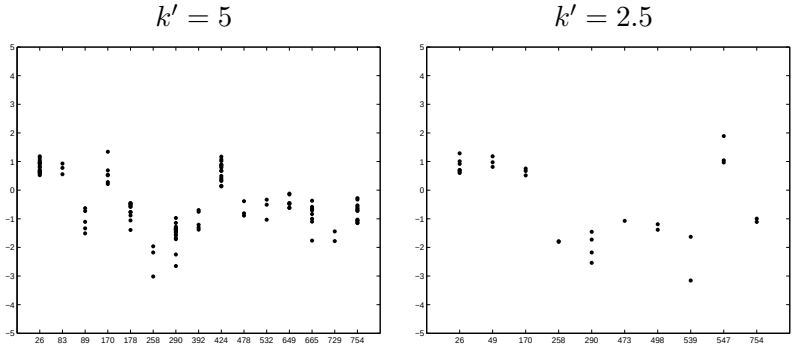


Figure 14: Modal regression parameter estimates for each gene

These results show important genes which should be included in a predictive model both

singly and in combination.

5 Discussion

In this paper we have proposed a class of prior distributions suggestive of a population of coefficients many of which are zero or near zero. By focusing on the modes of the posterior through an EM algorithm we are able to develop a method that does not search subsets but can produce coefficients that are exactly zero, much in the spirit of the original lasso. Our preferred prior is a member of the class of scale mixtures of normals, the normal-exponential-gamma (NEG), which allows a spike at zero which is not infinite and is proper over its full range. It retains some of the strong thresholding properties of the normal-Jeffreys with weak shrinkage for large coefficients without its evident overpowering drawback of impropriety of both prior and posterior.

We have compared the thresholding properties of several differing choices of prior in the scale mixture of normal class, illustrating the shape of regions in two dimensions. In higher dimensions these are harder to characterise although formulae are provided. In cases of more variables, k, than observations, n, as in the microarray example, then only $\min(n, k)$ coefficients can be non-zero.

We have developed an EM algorithm strategy to find modal estimates. Convergence is an issue with the latent variable probit model where information in the likelihood is weaker than in the linear regression model. One arm of our strategy powers up the likelihood whilst the other updates selectively within EM. Direct maximisation of the posterior utilising Newton-Raphson with EM as in Kiiveri (2003) would be an alternative worth exploring in the context of generalised linear modelling. Our algorithm uses the singular value decomposition to reduce the dimensions of coefficient space whilst retaining full information content and thresholding on the original coefficients rather than those derived.

The modal analysis quantifies the posterior probabilities of coefficients being outside of a near-zero region and uses these to select interesting variables. We are also able to look at variables in combination.

Acknowledgements The second author (PJB) is grateful to CSIRO for sponsoring a visit in 2002 and to H. Kiiveri for discussions on his method, which gave impetus to the present work. Our thanks also to G. Amphlett and GlaxoSmithKline for providing the data.

References

Abramowitz, M. and Stegun, I. A. (Eds.) (1964) "Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables," Dover: New York.

Akaike, H. (1974): "A new look at the statistical identification model," *IEEE Transactions on Automatic Control*, 19, 716-723.

Albert, J. and Chib, S. (1993): "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669-679.

Bae, K. and Mallick, B. K. (2004): "Gene selection using two-level hierarchical Bayesian model," *Bioinformatics*, 20, 3423-3430.

Barndorff-Nielsen, O. E. and Blaesild, P. (1981): "Hyperbolic distributions and ramifications: contributions to the theory and applications," in *Statistical Distributions in Scientific Work, Vol. 4* C. Taillie, G. Patil and B. Baldessari (ed.): , Reidal : Dorderecht.

Bernardo, J. M. and Smith, A. F. M. (1994): "Bayesian Theory," Wiley : Chichester.

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.

Box, G. E. P. and Tiao, G. C. (1973) "Bayesian Inference in Statistical Analysis," Wiley: New York.

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., 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.

Fan, J. and Peng, H. (2004): "Nonconcave penalized likelihood with diverging number of parameters," *Annals of Statistics*, 32, 928-961.

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.

Foster, D. P. and George, E. I. (1994): "The risk inflation criterion for multiple regression," *Annals of Statistics*, 22, 1947-75.

Gelfand, A. E. and Sahu, S. K. (1999): "Identifiability, improper priors, and Gibbs sampling for generalised linear models." *Journal of the American Statistical Association*, 94, 247-253.

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.

Johnstone, I. M. and Silverman, B. W. (2005): "Empirical Bayes selection of wavelet thresholds," *Annals of Statistics*, to appear.

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.

Knight, K. and Fu, W. (2000) "Asymptotics for lasso-type estimators", Annals of Statistics, 28, 1356-1378.

Liu, C. H., Rubin, D. B. and Wu, Y. N. (1998): "Parameter expansion to accelerate EM: The PX-EM algorithm," Biometrika, 85, 755-770.

Liu, J. S. (2001): "Monte Carlo Strategies in Scientific Computing," Springer: New York.

MacKay, D. J. C. (1994): "Bayesian methods for back-propagation networks," In Domany, E. *et al* (Eds) "Models of Neural Networks III" Chapter 6, 211-254.

McLachlan, G. J. and Peel, D. (2000): "Finite Mixture Models," Wiley: New York.

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.

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.

Schwarz, G. (1978): "Estimating the dimension of a model," *Annals of Statistics*, 6, 461-464.

Sha, N., Vannucci, M., Brown, P. J., Trower, M. K., Amphlett G., Falciani F. (2003): "Gene selection in arthritis classification with large-scale microarray expression profiles." *Comparative and Functional Genomics*, 4, 171-181.

Tibshirani, R. (1996): "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society B*, 58, 267-288.

Tipping, M. E. and Faul, A. (2003): "Fast marginal likelihood maximisation for sparse Bayesian models," In Frey, B. and Bishop, C. M. (Eds) *Proceedings 9th International Workshop on Artificial Intelligence and Statistics, Key West, Florida*.

Ueda, N. and Nakano, R. (1995): "Deterministic annealing variants of EM," In G. Tesauro, D. S. Tourestzky, T. K. Leen (Eds) *Advances in Neural Information Processing Systems 7*, 545-552, MIT Press.

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.

Wolfe, P. J., Godsill, S. J. and Ng, W. J. (2004): "Bayesian variable selection and regularisation for time-frequency surface estimation," *Journal of the Royal Statistical Society, B* , 66, 575-589.

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

Appendix 1

For the lasso and simple regression, equation (13) has a turning point if $|\hat{\beta}| > \frac{1}{\gamma}\frac{\sigma^2}{\sum_{i=1}^n x_i^2}$ at the point

$$\tilde{\beta} = \hat{\beta} - \text{sign}(\hat{\beta})\frac{1}{\gamma}\frac{\sigma^2}{\sum_{i=1}^n x_i^2} \tag{18}$$

which is also the Bayesian threshold since

$$\begin{aligned}\log \pi(\tilde{\beta}|y) - \log \pi(0|y) &= -\frac{1}{2\sigma^2}\left[\tilde{\beta}^2 \sum_{i=1}^n x_i^2 - 2\tilde{\beta}\sum_{i=1}^n x_i y_i\right] - \frac{1}{\gamma}|\tilde{\beta}| \\ &= -\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}\left[\tilde{\beta}^2 - 2\tilde{\beta}\hat{\beta} + 2\frac{\sigma^2}{\sum_{i=1}^n x_i^2}\frac{1}{\gamma}|\tilde{\beta}|\right]\end{aligned}$$

and substituting in equation (18) and noting that sign $\left(\tilde{\beta}\right) = \text{sign}\left(\hat{\beta}\right)$ gives

$$\begin{aligned}\log \pi(\tilde{\beta}|y) - \log \pi(0|y) &= -\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}\left[\tilde{\beta}^2 - 2\tilde{\beta}^2 - 2\frac{1}{\gamma}\frac{\sigma^2}{\sum_{i=1}^n x_i^2}\tilde{\beta}\text{sign}(\tilde{\beta}) + 2\frac{\sigma^2}{\sum_{i=1}^n x_i^2}\frac{1}{\gamma}|\tilde{\beta}|\right] \\ &= \frac{\sum_{i=1}^n x_i^2}{2\sigma^2}\tilde{\beta}^2 > 0.\end{aligned}$$

Appendix 2

The region of different types of thresholding with a lasso penalty and two variables can be derived in the following way. The assumed form of X^TX implies that

$$(X^TX)^{-1} = \begin{pmatrix} \frac{1}{c(1-\rho^2)} & \frac{\rho}{(1-\rho^2)\sqrt{cd}} \\ \frac{\rho}{(1-\rho^2)\sqrt{cd}} & \frac{1}{d(1-\rho^2)} \end{pmatrix}$$

and equation (13) can be re-arranged to give

$$\frac{1}{c}\frac{\partial L}{\partial \beta_1} = \beta_1 - \hat{\beta}_1 + \frac{1}{c}\text{sign}(\beta_1)\frac{1}{\gamma} + \sqrt{\frac{d}{c}}\rho(\hat{\beta}_2 - \beta_2)$$

and

$$\frac{1}{d}\frac{\partial L}{\partial \beta_2} = \beta_2 - \hat{\beta}_2 + \frac{1}{d}\text{sign}(\beta_2)\frac{1}{\gamma} + \sqrt{\frac{c}{d}}\rho(\hat{\beta}_1 - \beta_1).$$

If a mode exists with both parameters non-zero, the following conditions must hold

$$0 = \text{sign}(\beta_1)|\beta_1| - \hat{\beta}_1 + \frac{1}{c}\text{sign}(\beta_1)\frac{1}{\gamma} + \sqrt{\frac{d}{c}}\rho(\hat{\beta}_2 - \beta_2)$$

and

$$0 = \text{sign}(\beta_2)|\beta_2| - \hat{\beta}_2 + \frac{1}{d}\text{sign}(\beta_2)\frac{1}{\gamma} + \sqrt{\frac{c}{d}}\rho(\hat{\beta}_1 - \beta_1).$$

Some algebra gives the expression.

$$|\beta_1| = \text{sign}(\beta_1)\hat{\beta}_1 - \frac{1}{\gamma\sqrt{c}}\frac{1}{1-\rho^2}\left[\frac{1}{\sqrt{c}} + \frac{\rho}{\sqrt{d}}\text{sign}(\beta_1)\text{sign}(\beta_2)\right]$$

and

$$|\beta_2| = \text{sign}(\beta_2)\hat{\beta}_2 - \frac{1}{\gamma\sqrt{d}}\frac{1}{1-\rho^2}\left[\frac{1}{\sqrt{d}} + \frac{\rho}{\sqrt{c}}\text{sign}(\beta_1)\text{sign}(\beta_2)\right]$$

Since the right-hand side of both expression is greater than zero,

$$\text{sign}(\beta_1)\hat{\beta}_1 > \frac{1}{\gamma\sqrt{c}}\frac{1}{1-\rho^2}\left[\frac{1}{\sqrt{c}} + \frac{\rho}{\sqrt{d}}\text{sign}(\beta_1)\text{sign}(\beta_2)\right]$$

and

$$\text{sign}(\beta_2)\hat{\beta}_2 > \frac{1}{\gamma\sqrt{d}}\frac{1}{1-\rho^2}\left[\frac{1}{\sqrt{d}} + \frac{\rho}{\sqrt{c}}\text{sign}(\beta_1)\text{sign}(\beta_2)\right]$$

So, the regions where neither parameter is thresholded take the form of four squares. The regions where one parameter is thresholded are bounded by the four disjoint squares and the lines

$$\hat{\beta}_2 = \frac{1}{\sqrt{d}}\left[\frac{\rho}{\sqrt{c}}c\hat{\beta}_1 + \frac{\text{sign}(\beta_2)}{\sqrt{d}}\frac{1}{\gamma}\right]$$

for the region where β_1 is thresholded and

$$\hat{\beta}_1 = \frac{1}{\sqrt{c}}\left[\frac{\rho}{\sqrt{d}}d\hat{\beta}_2 + \frac{\text{sign}(\beta_1)}{\sqrt{c}}\frac{1}{\gamma}\right]$$

for the region where β_2 is thresholded. These lines cross at the corners of the region where both variable are thresholded. If c and d are equal, the graph is symmetric in the lines $y = x$ and $y = -x$. The region where both regressor is thresholded forms a rhomboid whose shape changes with the value of ρ. The volume of the region will be determined by the ratios $\frac{1}{\gamma\sqrt{c}}$ and $\frac{1}{\gamma\sqrt{d}}$.

If the sum of squares for the two variables (c and d) are not equal, the shape of the region where both regressors are thresholded can be less regular. The shape is a closed figure except if

$$\frac{1}{\gamma\sqrt{c}}\frac{1}{1-\rho^2}\left[\frac{1}{\sqrt{c}} + \frac{\rho}{\sqrt{d}}\text{sign}(\beta_1)\text{sign}(\beta_2)\right] < 0$$

or

$$\frac{1}{\gamma\sqrt{d}}\frac{1}{1-\rho^2}\left[\frac{1}{\sqrt{d}} + \frac{\rho}{\sqrt{c}}\text{sign}(\beta_1)\text{sign}(\beta_2)\right] < 0$$

$$\rho\text{sign}(\beta_1)\text{sign}(\beta_2) < -\sqrt{\frac{d}{c}}, \rho\text{sign}(\beta_1)\text{sign}(\beta_2) < -\sqrt{\frac{c}{d}}$$

This condition reduces to

$$\rho\text{sign}(\beta_1)\text{sign}(\beta_2) < -\sqrt{\frac{\min\{c,d\}}{\max\{c,d\}}}$$

Appendix 3

Here we derive results for normal exponential gamma distribution for (i) the marginal distribution of β (ii) the derivative of its form as a penalty function and (iii) the form of the E-step for it. ¿From Gradshtein Ryzik (1980, p319)

$$\int_0^\infty x^{\nu-1}(x+\beta^\star)^{-\nu+1/2}\exp\{-\mu x\}\,dx = 2^{\nu-1/2}\Gamma(\nu)\mu^{-1/2}\exp\{\beta^\star\mu/2\}D_{1-2\nu}(\sqrt{2\beta^\star\mu}) \tag{19}$$

$$\int_0^\infty x^{\nu-1}(x+\beta^\star)^{-\nu-1/2}\exp\{-\mu x\}dx = 2^{\nu}\Gamma(\nu)\beta^{\star-1/2}\exp\{\beta^\star\mu/2\}D_{-2\nu}(\sqrt{2\beta^\star\mu}) \tag{20}$$

(i) Marginal distribution of β

$$\begin{aligned}\pi(\beta) &= \int_0^\infty \frac{1}{\sqrt{2\pi\Psi}} \exp\left\{-\frac{1}{2}\frac{\beta^2}{\Psi}\right\} \frac{\lambda}{\gamma^2}(1+\Psi/\gamma^2)^{-(\lambda+1)}\, d\Psi \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}}\phi^{-3/2} \exp\left\{-\frac{1}{2}\phi\beta^2\right\} \frac{\lambda}{\gamma^2}\left(\frac{\phi+\frac{1}{\gamma^2}}{\phi}\right)^{-(\lambda+1)} d\phi \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi}}\phi^{\lambda-1/2} \exp\left\{-\frac{1}{2}\phi\beta^2\right\} \frac{\lambda}{\gamma^2}\left(\phi+\frac{1}{\gamma^2}\right)^{-(\lambda+1)} d\phi\end{aligned}$$

Let $\nu = \lambda + 1/2$, $\mu = \frac{1}{2}\beta^2$ and $\beta^\star = \frac{1}{\gamma^2}$ and substitute into equation (20)

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

(ii) The derivative of the penalty function

$$-\frac{d}{d\beta}\log\pi(\beta) = \beta\frac{\int_0^\infty \frac{1}{\sqrt{2\pi}}\phi^{\lambda+1/2} \exp\left\{-\frac{1}{2}\phi\beta^2\right\} \frac{\lambda}{\gamma^2}\left(\phi+\frac{1}{\gamma^2}\right)^{-(\lambda+1)} d\phi}{\int_0^\infty \frac{1}{\sqrt{2\pi}}\phi^{\lambda-1/2} \exp\left\{-\frac{1}{2}\phi\beta^2\right\} \frac{\lambda}{\gamma^2}\left(\phi+\frac{1}{\gamma^2}\right)^{-(\lambda+1)} d\phi}$$

Using the substitution $\nu = \lambda + 3/2$, $\mu = \frac{1}{2}\beta^2$ and $\beta^\star = \frac{1}{\gamma^2}$ and substitute into equation (19)

$$\begin{aligned}&\int_0^\infty \frac{1}{\sqrt{2\pi}}\phi^{\lambda+1/2} \exp\left\{-\frac{1}{2}\phi\beta^2\right\} \frac{\lambda}{\gamma^2}\left(\phi+\frac{1}{\gamma^2}\right)^{-(\lambda+1)} d\phi \\ &= \frac{1}{\sqrt{\pi}}\frac{\lambda}{\gamma^2}2^{\lambda+1}\Gamma(\lambda+3/2)\frac{1}{|\beta|} \exp\left\{\frac{1}{4}\frac{\beta^2}{\gamma^2}\right\} D_{1-2\nu}\left(\frac{|\beta|}{\gamma}\right)\end{aligned}$$

$$-\frac{d}{d\beta}\log\pi(\beta) = (\lambda+1/2)\frac{1}{\gamma}\frac{\beta}{|\beta|}\frac{D_{-(2\lambda+2)}\left(\frac{|\beta|}{\gamma}\right)}{D_{-(2\lambda+1)}\left(\frac{|\beta|}{\gamma}\right)}$$

(ii) The E-step of the E-M algorithm.

$$\begin{aligned}E\left(\frac{1}{\Psi}|\beta\right) &= \frac{1}{\pi(\beta)}\int_0^\infty \frac{1}{\sqrt{2\pi}\Psi^{3/2}} \exp\left\{-\frac{1}{2}\frac{\beta^2}{\Psi}\right\} \frac{\lambda}{\gamma^2}(1+\Psi/\gamma^2)^{-(\lambda+1)}\, d\Psi \\ &= \frac{1}{\pi(\beta)}\int_0^\infty \frac{1}{\sqrt{2\pi}}\phi^{-1/2} \exp\left\{-\frac{1}{2}\phi\beta^2\right\} \frac{\lambda}{\gamma^2}\left(\frac{\phi+\frac{1}{\gamma^2}}{\phi}\right)^{-(\lambda+1)} d\phi \\ &= \frac{1}{\pi(\beta)}\int_0^\infty \frac{1}{\sqrt{2\pi}}\phi^{\lambda+1/2} \exp\left\{-\frac{1}{2}\phi\beta^2\right\} \frac{\lambda}{\gamma^2}\left(\phi+\frac{1}{\gamma^2}\right)^{-(\lambda+1)} d\phi.\end{aligned}$$

For the integral let $\nu = \lambda + 3/2$, $\mu = \frac{1}{2}\beta^2$ and $\beta^\star = \frac{1}{\gamma^2}$ and substitute into equation (19), then using $\pi(\beta)$ derived above the E-step formula

$$= \frac{(\lambda + 1/2)}{\gamma|\beta|} \frac{D_{-(2\lambda+2)}\left(\frac{|\beta|}{\gamma}\right)}{D_{-(2\lambda+1)}\left(\frac{|\beta|}{\gamma}\right)}.$$