The Bayesian Lasso

Trevor Park and George Casella†

University of Florida, Gainesville, Florida, USA

Summary. The Lasso estimate for linear regression parameters can be interpreted as a Bayesian posterior mode estimate when the priors on the regression parameters are independent double-exponential (Laplace) distributions. This posterior can also be accessed through a Gibbs sampler using conjugate normal priors for the regression parameters, with independent exponential hyperpriors on their variances. This leads to tractable full conditional distributions through a connection with the inverse Gaussian distribution. Although the Bayesian Lasso does not automatically perform variable selection, it does provide standard errors and Bayesian credible intervals that can guide variable selection. Moreover, the structure of the hierarchical model provides both Bayesian and likelihood methods for selecting the Lasso parameter. The methods described here can also be extended to other Lasso-related estimation methods like bridge regression and robust variants.

Keywords: Gibbs sampler, inverse Gaussian, linear regression, empirical Bayes, penalised regression, hierarchical models, scale mixture of normals

1. Introduction

The Lasso of Tibshirani (1996) is a method for simultaneous shrinkage and model selection in regression problems. It is most commonly applied to the linear regression model

$$\boldsymbol{y} = \mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of responses, μ is the overall mean, $\boldsymbol{X}$ is the $n \times p$ matrix of *standardised* regressors, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\mathsf{T}$ is the vector of regression coefficients to be estimated, and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of independent and identically distributed normal errors with mean 0 and unknown variance σ^2. The estimate of μ is taken as the average $\bar{y}$ of the responses, and the Lasso estimate $\widehat{\boldsymbol{\beta}}$ minimises the sum of the squared residuals, subject to a given bound t on its L_1 norm. The entire path of Lasso estimates for all values of t can be efficiently computed via a modification of the related LARS algorithm of Efron et al. (2004). (See also Osborne et al. (2000).)

For values of t less than the L_1 norm of the ordinary least squares estimate of $\boldsymbol{\beta}$, Lasso estimates can be described as solutions to unconstrained optimisations of the form

$$\min_{\boldsymbol{\beta}} \quad (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})^\mathsf{T}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) \quad + \quad \lambda \sum_{j=1}^{p} |\beta_j|$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \bar{y}\mathbf{1}_n$ is the mean-centred response vector and the parameter $\lambda \geq 0$ relates implicitly to the bound t. The form of this expression suggests that the Lasso may be

†*Address for correspondence:* Department of Statistics, University of Florida, 103 Griffin/Floyd Hall, Gainesville, FL 32611, USA.
E-mail: tpark@stat.ufl.edu

interpreted as a Bayesian posterior mode estimate when the parameters β_i have independent and identical double exponential (Laplace) priors (Tibshirani, 1996; Hastie et al., 2001, Sec. 3.4.5). Indeed, with the prior

$$\pi(\boldsymbol{\beta}) \;=\; \prod_{j=1}^{p} \frac{\lambda}{2} e^{-\lambda|\beta_j|} \tag{1}$$

and an independent prior $\pi(\sigma^2)$ on $\sigma^2 > 0$, the posterior distribution, conditional on $\tilde{\boldsymbol{y}}$, can be expressed as

$$\pi(\boldsymbol{\beta}, \sigma^2 | \tilde{\boldsymbol{y}}) \;\propto\; \pi(\sigma^2)\; (\sigma^2)^{-(n-1)/2} \exp\left\{ -\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta})^{\mathsf{T}}(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta}) \;-\; \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

(This can alternatively be obtained as a posterior conditional on $\boldsymbol{y}$ if μ is given an independent flat prior and removed by marginalisation.) For any fixed value of $\sigma^2 > 0$, the maximising $\boldsymbol{\beta}$ is a Lasso estimate, and hence the posterior mode estimate, if it exists, will be a Lasso estimate. The particular choice of estimate will depend on λ and the choice of prior for σ^2.

Maximising the posterior, though sometimes convenient, is not a particularly natural Bayesian way to obtain point estimates. For instance, the posterior mode is not necessarily preserved under marginalisation. A fully Bayesian analysis would instead suggest using the mean or median of the posterior to estimate β. Though such estimates lack the model selection property of the Lasso, they do produce similar individualised shrinkage of the coefficients. The fully Bayesian approach also provides credible intervals for the estimates, and λ can be chosen by marginal (Type-II) maximum likelihood or hyperprior methods (Section 5).

For reasons explained in Section 4, we shall prefer to use conditional priors on $\boldsymbol{\beta}$ of the form

$$\pi(\boldsymbol{\beta}|\sigma^2) \;=\; \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}} \tag{2}$$

instead of prior (1). We can safely complete the prior specification with the (improper) scale invariant prior $\pi(\sigma^2) = 1/\sigma^2$ on σ^2 (Section 4).

Figure 1 compares Bayesian Lasso estimates with the ordinary Lasso and ridge regression estimates for the diabetes data of Efron et al. (2004), which has $n = 442$ and $p = 10$. The figure shows the paths of Lasso estimates, Bayesian Lasso posterior median estimates, and ridge regression estimates as their corresponding parameters are varied. (The vector of posterior medians minimises the L_1-norm loss averaged over the posterior. The Bayesian Lasso posterior mean estimates were almost indistinguishable from the medians.) For ease of comparison, all are plotted as a function of their L_1 norm relative to the L_1 norm of the least squares estimate. The Bayesian Lasso estimates were computed over a grid of λ values using the Gibbs sampler of Section 3 with the scale-invariant prior on σ^2. The estimates are medians from 10000 iterations of the Gibbs sampler after 1000 iterations of burn-in.

The Bayesian Lasso estimates seem to be a compromise between the Lasso and ridge regression estimates: The paths are smooth, like ridge regression, but are more similar in shape to the Lasso paths, particularly when the L_1 norm is relatively small. The vertical line in the Lasso panel represents the estimate chosen by n-fold (leave-one-out) cross validation

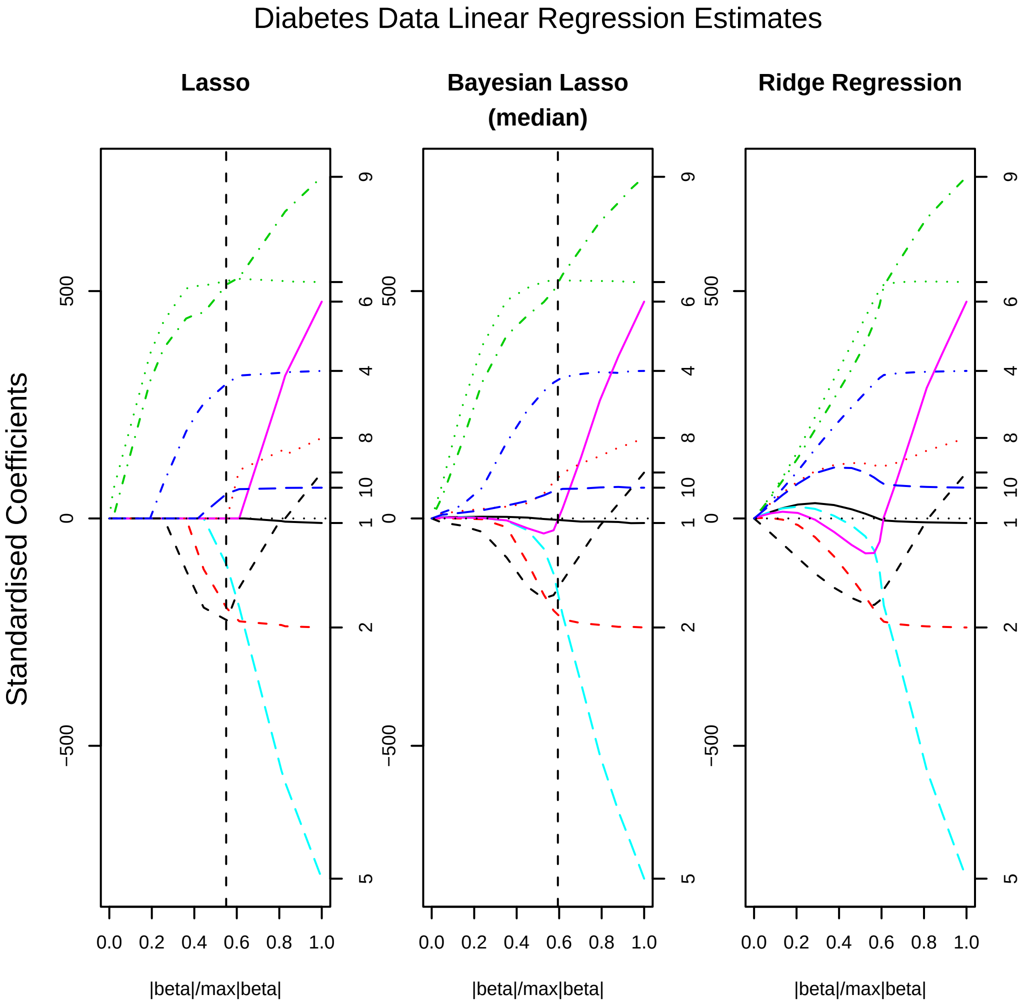

Fig. 1. Lasso, Bayesian Lasso, and Ridge Regression trace plots for estimates of the diabetes data regression parameters versus relative L_1 norm, with vertical lines for the Lasso and Bayesian Lasso indicating the estimates chosen by, respectively, n-fold cross validation and marginal maximum likelihood.

Table 1. Estimates of the linear regression parameters for the diabetes data.

Variable		Bayesian Lasso (marginal m.l.)	Bayesian Credible Interval (95%)	Lasso (n-fold c.v.)	Lasso ($t \approx 0.59$)	Least Squares
(1)	age	-3.73	(-112.02, 103.62)	0.00	0.00	-10.01
(2)	sex	-214.55	(-334.42, -94.24)	-195.13	-217.06	-239.82
(3)	bmi	522.62	(393.07, 653.82)	521.95	525.41	519.84
(4)	map	307.56	(180.26, 436.70)	295.79	308.88	324.39
(5)	tc	-173.16	(-579.33, 128.54)	-100.76	-165.94	-792.18
(6)	ldl	-1.50	(-274.62, 341.48)	0.00	0.00	476.75
(7)	hdl	-152.12	(-381.60, 69.75)	-223.07	-175.33	101.04
(8)	tch	90.43	(-129.48, 349.82)	0.00	72.33	177.06
(9)	ltg	523.26	(332.11, 732.75)	512.84	525.07	751.28
(10)	glu	62.47	(-51.22, 188.75)	53.46	61.38	67.63

(see e.g. Hastie et al., 2001), while the vertical line in the Bayesian Lasso panel represents the estimate chosen by marginal maximum likelihood (Section 5.1).

With λ selected by marginal maximum likelihood, medians and 95% credible intervals for the marginal posterior distributions of the Bayesian Lasso estimates for the diabetes data are shown in Figure 2. For comparison, the figure also shows the least squares and Lasso estimates (both the one chosen by cross-validation, and the one that has the same L_1 norm as the Bayesian posterior median to indicate how close the Lasso can be to the Bayesian Lasso posterior median). The cross-validation estimate for the Lasso has a relative L_1 norm of approximately 0.55 but is not especially well-defined. The norm-matching Lasso estimates (at relative L_1 norm of approximately 0.59) perform nearly as well. Corresponding numerical results are shown in Table 1. The Bayesian posterior medians are remarkably similar to the Lasso estimates. The Lasso estimates are well within the credible intervals for all variables, whereas the least squares estimates are outside for four of the variables, one of which is significant.

The Bayesian marginal posterior distributions for the elements of $\boldsymbol{\beta}$ all appear to be unimodal, but some have shapes that are distinctly non-Gaussian. For instance, kernel density estimates for variables 1 and 6 are shown in Figure 3. The peakedness of these densities is more suggestive of a double exponential than of a Gaussian density.

2. A Hierarchical Model Formulation

The Bayesian posterior median estimates shown in Figure 1 were obtained from a Gibbs sampler that exploits the following representation of the double exponential distribution as a scale mixture of normals:

$$\frac{a}{2}e^{-a|z|} \;=\; \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-z^2/(2s)}\; \frac{a^2}{2} e^{-a^2 s/2}\, ds, \qquad a > 0.$$

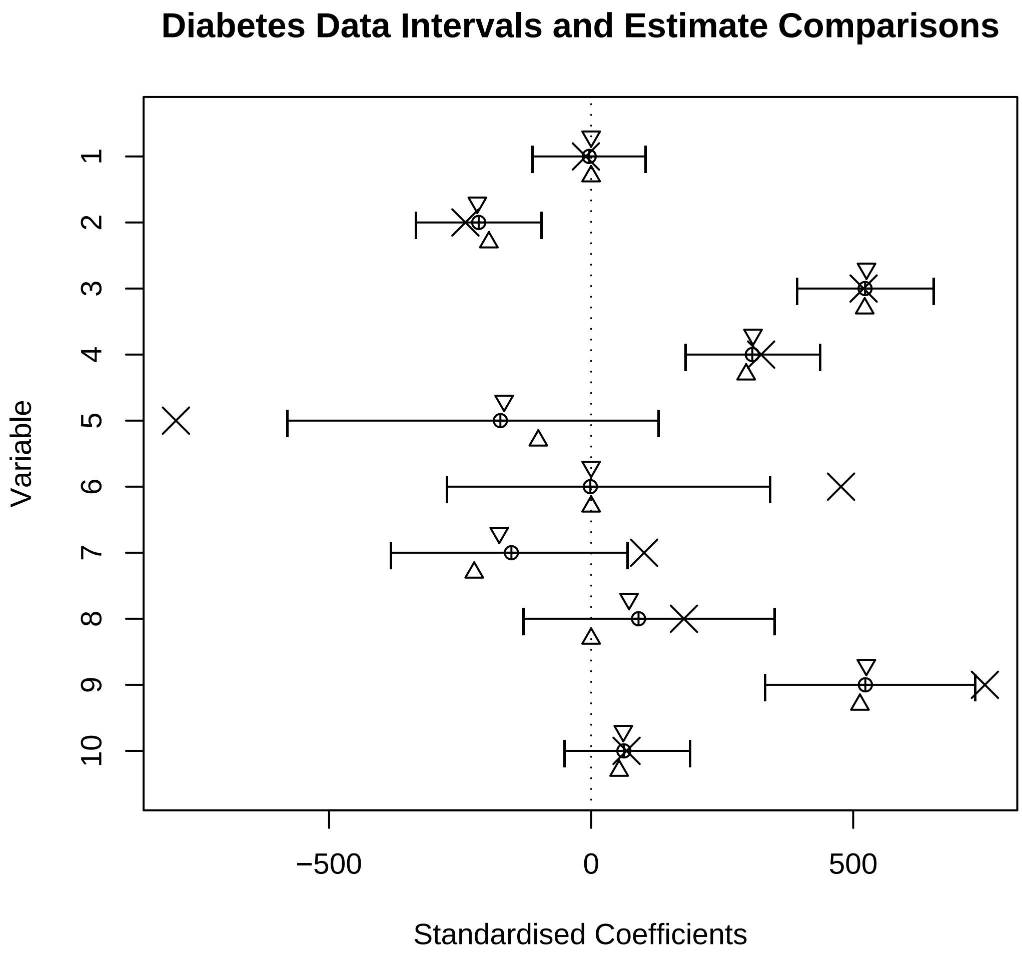

Fig. 2. Posterior median Bayesian Lasso estimates ($\oplus$) and corresponding 95% credible intervals (equal-tailed) with λ selected according to marginal maximum likelihood (Section 5.1). Overlaid are the least squares estimates ($\times$), Lasso estimates based on n-fold cross-validation ($\triangle$), and Lasso estimates chosen to match the L_1 norm of the Bayes estimates ($\bigtriangledown$).

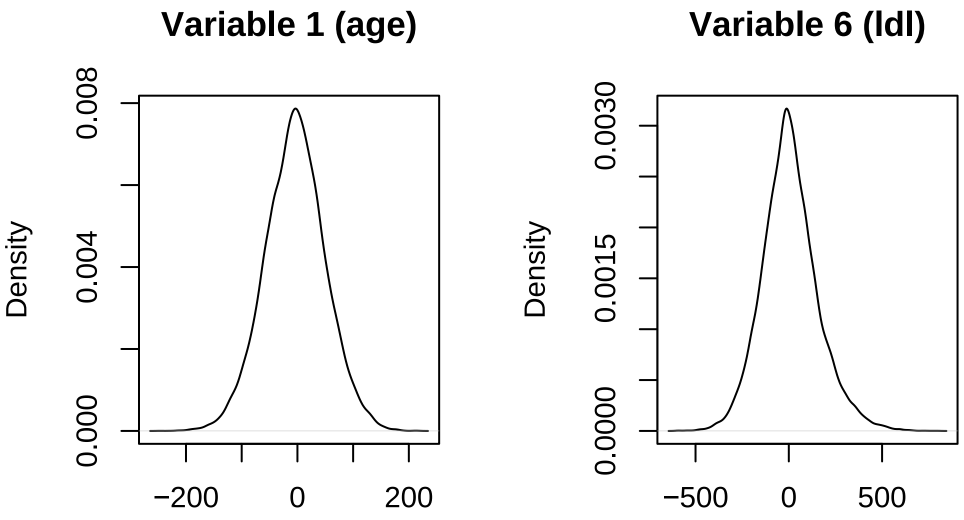

Fig. 3. Marginal posterior density function estimates for the Diabetes data variables 1 and 6. These are kernel density estimates based on 30000 Gibbs samples.

See, e.g. Andrews and Mallows (1974). This suggests the following hierarchical representation of the full model:

$$\begin{aligned} \boldsymbol{y} \mid \mu, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 &\sim N_n\big(\mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n\big) \\ \boldsymbol{\beta} \mid \tau_1^2, \ldots, \tau_p^2, \sigma^2 &\sim N_p(\mathbf{0}_p, \sigma^2 \boldsymbol{D}_\tau), \qquad \boldsymbol{D}_\tau = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2) \\ \tau_1^2, \ldots, \tau_p^2 &\sim \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2}\, d\tau_j^2, \qquad \tau_1^2, \ldots, \tau_p^2 > 0 \\ \sigma^2 &\sim \pi(\sigma^2)\, d\sigma^2 \end{aligned} \tag{3}$$

with $\tau_1^2, \ldots, \tau_p^2$ and σ^2 independent. (The parameter μ may be given an independent, flat prior.) After integrating out $\tau_1^2, \ldots, \tau_p^2$, the conditional prior on $\boldsymbol{\beta}$ has the form (2). Prior (1) can alternatively be obtained from this hierarchy if (3) is replaced by

$$\boldsymbol{\beta} \mid \tau_1^2, \ldots, \tau_p^2 \sim N_p(\mathbf{0}_p, \boldsymbol{D}_\tau), \qquad \boldsymbol{D}_\tau = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2) \tag{4}$$

so that $\boldsymbol{\beta}$ is (unconditionally) independent of σ^2. Section 3 details a Gibbs sampler implementation for the hierarchy of (3), which exploits a conjugacy involving the inverse Gaussian distribution. The hierarchy employing (4) could also be easily implemented in a Gibbs sampler, but Section 4 illustrates some difficulties posed by this prior due to the possibility of a non-unimodal posterior. Hierarchies that employ other distributions for $\tau_1^2, \ldots, \tau_p^2$ can be used to produce Bayesian versions of methods related to the Lasso, as discussed in Section 6.

Bayesian analysis using this general form of hierarchy predates widespread use of the Gibbs sampler (e.g. West, 1984). Figueiredo (2003) proposes the hierarchical representation using (4) for use in an EM algorithm to compute the ordinary Lasso estimates by regarding $\tau_1^2, \ldots, \tau_p^2$ as "missing data," although this is not as efficient as the LARS algorithm.

The hierarchy that employs (4) is an example of what Ishwaran and Rao (2005) refer to as "spike-and-slab" models, in generalisation of the terminology of Mitchell and Beauchamp (1988). But true spike-and-slab models tend to employ two-component mixtures for the elements of $\boldsymbol{\beta}$, one concentrated near zero (the spike) and the other spread away from zero (the slab). An early example of such a hierarchy is the Bayesian variable selection method of George and McCulloch (1993), in which $\tau_1^2, \ldots, \tau_p^2$ are given independent two-point priors with one point close to zero. George and McCulloch (1997) propose alternative versions of this method that condition on σ^2 and so are more akin to using (3). In the context of wavelet analysis (or orthogonal designs more generally), Clyde et al. (1998) use a prior similar to (3), but with independent Bernoulli priors on $\tau_1^2, \ldots, \tau_p^2$, yielding a degenerate spike exactly at zero. Clyde and George (2000) extended this by effectively using heavier-tailed distributions for the slab portion and for the error distribution through scale mixtures of normals, although they did not consider the double-exponential distribution.

Yuan and Lin (2005) propose a prior for the elements of $\boldsymbol{\beta}$ with a degenerate spike at zero and a double exponential slab, but instead of performing a Bayesian analysis choose to approximate the posterior. Their analysis leads to estimates chosen similarly to the original Lasso and lack any corresponding interval estimates.

3. Gibbs Sampler Implementation

We will use the typical inverse gamma prior distribution on σ^2,

$$\pi(\sigma^2) \;=\; \frac{\gamma^a}{\Gamma(a)}\left(\sigma^2\right)^{-a-1} e^{-\gamma/\sigma^2}, \qquad \sigma^2 > 0 \qquad (a > 0, \gamma > 0), \tag{5}$$

although other conjugate priors are available (see Athreya (1986)). We will also assume an independent, flat (shift-invariant) prior on μ. With the hierarchy of (3), which implicitly produces prior (2), the joint density becomes

$$\begin{aligned}
f(\boldsymbol{y}|\mu,\boldsymbol{\beta},\sigma^2)\;\pi(\sigma^2)\;\pi(\mu)\prod_{j=1}^{p}\pi(\beta_j|\tau_j^2,\sigma^2)\;\pi(\tau_j^2) =&\\
&\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\,e^{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\mu\mathbf{1}_n-\boldsymbol{X\beta})^{\mathsf{T}}(\boldsymbol{y}-\mu\mathbf{1}_n-\boldsymbol{X\beta})}\\
&\times\frac{\gamma^a}{\Gamma(a)}\left(\sigma^2\right)^{-a-1}e^{-\gamma/\sigma^2}\prod_{j=1}^{p}\frac{1}{\left(2\pi\sigma^2\tau_j^2\right)^{1/2}}\,e^{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2}\,\frac{\lambda^2}{2}\,e^{-\lambda^2\tau_j^2/2}.
\end{aligned}$$

Now, letting $\overline{y}$ be the average of the elements of $\boldsymbol{y}$,

$$\begin{aligned}
(\boldsymbol{y}-\mu\mathbf{1}_n-\boldsymbol{X\beta})^{\mathsf{T}}(\boldsymbol{y}-\mu\mathbf{1}_n-\boldsymbol{X\beta}) \;&=\; (\overline{y}\mathbf{1}_n-\mu\mathbf{1}_n)^{\mathsf{T}}(\overline{y}\mathbf{1}_n-\mu\mathbf{1}_n) \;+\; (\tilde{\boldsymbol{y}}-\boldsymbol{X\beta})^{\mathsf{T}}(\tilde{\boldsymbol{y}}-\boldsymbol{X\beta})\\
&=\; n\,(\overline{y}-\mu)^2 \;+\; (\tilde{\boldsymbol{y}}-\boldsymbol{X\beta})^{\mathsf{T}}(\tilde{\boldsymbol{y}}-\boldsymbol{X\beta})
\end{aligned}$$

because the columns of $\boldsymbol{X}$ are standardised. The full conditional distribution of μ is thus normal with mean $\overline{y}$ and variance σ^2/n. In the spirit of the Lasso, μ may be integrated out, leaving a joint density (marginal only over μ) proportional to

$$\frac{1}{\left(\sigma^2\right)^{(n-1)/2}}\,e^{-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}}-\boldsymbol{X\beta})^{\mathsf{T}}(\tilde{\boldsymbol{y}}-\boldsymbol{X\beta})}\,\left(\sigma^2\right)^{-a-1}e^{-\gamma/\sigma^2}\prod_{j=1}^{p}\frac{1}{\left(\sigma^2\tau_j^2\right)^{1/2}}\,e^{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2}\,e^{-\lambda^2\tau_j^2/2}.$$

Note that this expression depends on $\boldsymbol{y}$ only through $\tilde{\boldsymbol{y}}$. The conjugacy of the other parameters remains unaffected, and thus it is easy to form a Gibbs sampler for $\boldsymbol{\beta}$, σ^2 and $(\tau_1^2,\ldots,\tau_p^2)$ based on this density.

The full conditional for $\boldsymbol{\beta}$ is multivariate normal: The exponent terms involving $\boldsymbol{\beta}$ are

$$-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}}-\boldsymbol{X\beta})^{\mathsf{T}}(\tilde{\boldsymbol{y}}-\boldsymbol{X\beta})-\frac{1}{2\sigma^2}\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{D}_\tau^{-1}\boldsymbol{\beta} \;=\; -\frac{1}{2\sigma^2}\Big\{\boldsymbol{\beta}^{\mathsf{T}}\big(\boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}+\boldsymbol{D}_\tau^{-1}\big)\boldsymbol{\beta}-2\,\tilde{\boldsymbol{y}}^{\mathsf{T}}\boldsymbol{X\beta}+\tilde{\boldsymbol{y}}^{\mathsf{T}}\tilde{\boldsymbol{y}}\Big\}.$$

Letting $\boldsymbol{A} \;=\; \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X}+\boldsymbol{D}_\tau^{-1}$ and completing the square transforms the term in brackets to

$$\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{A\beta}\;-\;2\,\tilde{\boldsymbol{y}}^{\mathsf{T}}\boldsymbol{X\beta}\;+\;\tilde{\boldsymbol{y}}^{\mathsf{T}}\tilde{\boldsymbol{y}} \;=\; \big(\boldsymbol{\beta}-\boldsymbol{A}^{-1}\boldsymbol{X}^{\mathsf{T}}\tilde{\boldsymbol{y}}\big)^{\mathsf{T}}\boldsymbol{A}\,\big(\boldsymbol{\beta}-\boldsymbol{A}^{-1}\boldsymbol{X}^{\mathsf{T}}\tilde{\boldsymbol{y}}\big)\;+\;\tilde{\boldsymbol{y}}^{\mathsf{T}}\big(\boldsymbol{I}_n-\boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^{\mathsf{T}}\big)\tilde{\boldsymbol{y}},$$

so $\boldsymbol{\beta}$ is conditionally multivariate normal with mean $\boldsymbol{A}^{-1}\boldsymbol{X}^{\mathsf{T}}\tilde{\boldsymbol{y}}$ and variance $\sigma^2\,\boldsymbol{A}^{-1}$.

The full conditional distribution of σ^2 is inverse gamma: The terms in the joint distribution involving σ^2 are

$$\left(\sigma^2\right)^{-(n-1)/2-p/2-a-1}\,\exp\Big\{-\frac{1}{\sigma^2}\big((\tilde{\boldsymbol{y}}-\boldsymbol{X\beta})^{\mathsf{T}}(\tilde{\boldsymbol{y}}-\boldsymbol{X\beta})/2\;+\;\boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{D}_\tau^{-1}\boldsymbol{\beta}/2\;+\;\gamma\big)\Big\}$$

so σ^2 is conditionally inverse gamma with shape parameter $(n-1)/2 + p/2 + a$ and scale parameter $(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta})^{\mathsf{T}}(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta})/2 + \boldsymbol{\beta}^{\mathsf{T}}\boldsymbol{D}_{\tau}^{-1}\boldsymbol{\beta}/2 + \gamma$.

For each $j = 1, \ldots, p$ the portion of the joint distribution involving τ_j^2 is

$$\left(\tau_j^2\right)^{-1/2} \exp\left\{ -\frac{1}{2}\left(\frac{\beta_j^2/\sigma^2}{\tau_j^2} + \lambda^2\tau_j^2\right)\right\},$$

which happens to be proportional to the density of the reciprocal of an inverse Gaussian random variable. Indeed, the density of $\eta_j^2 = 1/\tau_j^2$ is proportional to

$$\left(\eta_j^2\right)^{-3/2} \exp\left\{ -\frac{1}{2}\left(\frac{\beta_j^2}{\sigma^2}\,\eta_j^2 + \frac{\lambda^2}{\eta_j^2}\right)\right\} \quad \propto \quad \left(\eta_j^2\right)^{-3/2} \exp\left\{ -\frac{\beta_j^2\left(\eta_j^2 - \sqrt{\lambda^2\sigma^2/\beta_j^2}\right)^2}{2\sigma^2\eta_j^2}\right\},$$

which compares with one popular parameterisation of the inverse Gaussian density (Chhikara and Folks, 1989):

$$f(x) \;=\; \sqrt{\frac{\lambda'}{2\pi}}\, x^{-3/2} \exp\left\{ -\frac{\lambda'(x-\mu')^2}{2(\mu')^2 x}\right\}, \qquad x > 0,$$

where $\mu' > 0$ is the mean parameter and $\lambda' > 0$ is a scale parameter. (The variance is $(\mu')^3/\lambda'$.) Thus the distribution of $1/\tau_j^2$ is inverse Gaussian with

$$\text{mean parameter } \mu' \;=\; \sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}} \quad \text{and} \quad \text{scale parameter } \lambda' \;=\; \lambda^2.$$

A relatively simple algorithm is available for simulating from the inverse Gaussian distribution (Chhikara and Folks, 1989, Sec. 4.5), and a numerically stable variant of the algorithm is implemented in the language `R`, in the contributed package `statmod` (Smyth, 2005).

The Gibbs sampler simply samples cyclically from the distributions of $\boldsymbol{\beta}$, σ^2, and $(\tau_1^2, \ldots, \tau_p^2)$ conditional on the current values of the other parameters. Note that the sampling of $\boldsymbol{\beta}$ is a block update, and the sampling of $(\tau_1^2, \ldots, \tau_p^2)$ is also effectively a block update since $\tau_1^2, \ldots, \tau_p^2$ are conditionally independent. Our experience suggests that convergence is reasonably fast.

Parameter μ is generally of secondary interest, but the Gibbs sample can be used to perform inference on it if desired. As noted previously, the posterior of μ conditional on the other parameters is normal with mean $\overline{y}$ and variance σ^2/n. It follows that the marginal mean and median are $\overline{y}$, and the variance and other properties of the marginal posterior may be obtained using the Gibbs sample of σ^2.

4. The Posterior Distribution

The joint posterior distribution of $\boldsymbol{\beta}$ and σ^2 under priors (2) and (5) is proportional to

$$\left(\sigma^2\right)^{-(n+p-1)/2-a-1} \exp\left\{ -\frac{1}{\sigma^2}\left((\tilde{\boldsymbol{y}} - \boldsymbol{X\beta})^{\mathsf{T}}(\tilde{\boldsymbol{y}} - \boldsymbol{X\beta})/2 + \gamma\right) - \frac{\lambda}{\sqrt{\sigma^2}}\sum_{j=1}^{p}|\beta_j|\right\}. \tag{6}$$

The form of this density indicates that we may safely let $a = 0$ and, assuming that the data do not admit a perfect linear fit (i.e. $\tilde{\boldsymbol{y}}$ is not in the column space of $\boldsymbol{X}$), also let $\gamma = 0$. This corresponds to using the non-informative scale-invariant prior $1/\sigma^2$ on σ^2. The posterior remains integrable for any $\lambda \geq 0$. Note also that λ is unitless: A change in the units of measurement for $\boldsymbol{y}$ does not require any change in λ to produce the equivalent Bayesian solution. (The $\boldsymbol{X}$ matrix is, of course, unitless because of its scaling.)

For comparison, the joint posterior distribution of $\boldsymbol{\beta}$ and σ^2 under prior (1), with some independent prior $\pi(\sigma^2)$ on σ^2, is proportional to

$$\pi(\sigma^2)\,(\sigma^2)^{-(n-1)/2}\,\exp\left\{-\frac{1}{2\sigma^2}(\tilde{\boldsymbol{y}}-\boldsymbol{X\beta})^\mathsf{T}(\tilde{\boldsymbol{y}}-\boldsymbol{X\beta})\;-\;\lambda\sum_{j=1}^{p}|\beta_j|\right\}. \tag{7}$$

In this case, λ has units that are the reciprocal of the units of the response, and any change in units will require a corresponding change in λ to produce the equivalent Bayesian solution.

It can be shown that posteriors of the form (6) generally do not have more than one mode for any $a \geq 0, \gamma \geq 0, \lambda \geq 0$ (see the Appendix). In contrast, posteriors of the form (7) may have more than one mode. For example, Figure 4 shows the contours of an bimodal joint density of β and $\log(\sigma^2)$ when $p = 1$ and $\pi(\sigma^2)$ is the scale-invariant prior $1/\sigma^2$. (Similar bimodality can occur even if $\pi(\sigma^2)$ is proper.) This particular example results from taking $p = 1$, $n = 10$, $\boldsymbol{X}^\mathsf{T}\boldsymbol{X} = 1$, $\boldsymbol{X}^\mathsf{T}\tilde{\boldsymbol{y}} = 5$, $\tilde{\boldsymbol{y}}^\mathsf{T}\tilde{\boldsymbol{y}} = 26$, $\lambda = 3$. The mode on the lower right is near the least-squares solution $\beta = 5, \sigma^2 = 1/8$, while the mode on the upper left is near the values $\beta = 0, \sigma^2 = 26/9$ that would be estimated for the selected model in which β is set to zero. The crease in the upper left mode along the line $\beta = 0$ is a feature produced by the “sharp corners” of the L_1 penalty. Not surprisingly, the marginal density of β is also bimodal (not shown). When $p > 1$, it may be possible to have more than two modes, though we have not investigated this.

Presence of multiple posterior modes causes both conceptual and computational problems. Conceptually, it is questionable whether a single posterior mean, median, or mode represents an appropriate summary of a bimodal posterior. A better summary would be separate measures of the centres of each mode, along with the approximate amount of probability associated with each, in the spirit of “spike and slab” models (Ishwaran and Rao, 2005), but this would require an entirely different methodology.

Computationally, posteriors having multiple offset modes are a notorious source of convergence problems in the Gibbs sampler. Although it is possible to implement a Gibbs sampler for posteriors of the form (7) when $\pi(\sigma^2)$ is chosen to be the conjugate inverse gamma distribution using a derivation similar to that of Section 3, we were able to construct examples that make the convergence of this Gibbs sampler much too slow for practical use. A Gibbs sampler can be alternated with non-Gibbs steps designed to facilitate mixing by allowing jumps between modes, but such methods are more complicated and generally rely upon either knowledge of the locations of all modes or access to an effective search strategy.

5. Choosing the Bayesian Lasso Parameter

The parameter of the ordinary Lasso can be chosen by cross-validation, generalised cross-validation, and ideas based on Stein’s unbiased risk estimate (Tibshirani, 1996). The Bayesian Lasso also offers some uniquely Bayesian alternatives: empirical Bayes via marginal (Type II) maximum likelihood, and use of an appropriate hyperprior.

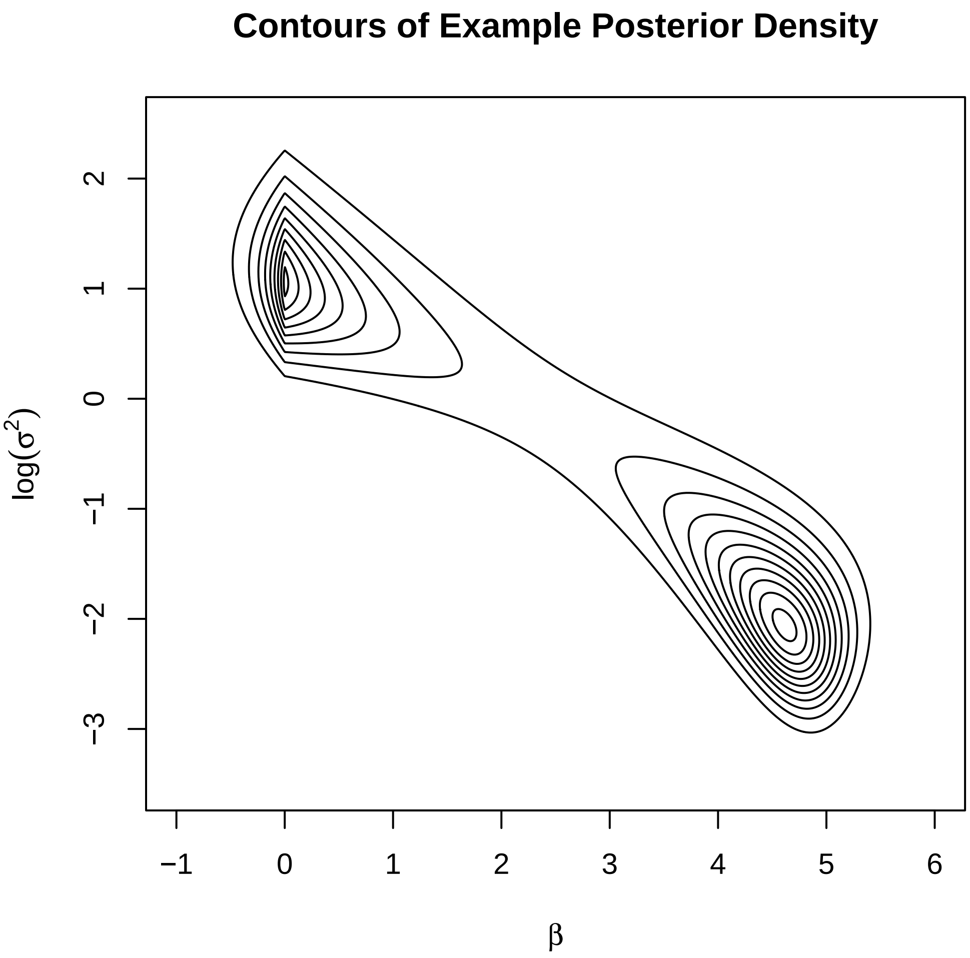

Fig. 4. Contour plot of an artificially-generated posterior density of $(\beta, \log(\sigma^2))$ of the form (7) that manifests bimodality.

5.1. Empirical Bayes by Marginal Maximum Likelihood

If the hierarchy of Section 2 is regarded as a parametric model, the parameter λ has a likelihood function that may be maximised to obtain an empirical Bayes estimate. Casella (2001) proposes a Monte Carlo EM algorithm that complements a Gibbs sampler implementation. For the Bayesian Lasso, the steps are

(a) Let $k = 0$ and choose initial $\lambda^{(0)}$.
(b) Generate a sample from the posterior distribution of $\boldsymbol{\beta}, \sigma^2, \tau_1^2, \ldots, \tau_p^2$ using the Gibbs sampler of Section 3 with λ set to $\lambda^{(k)}$.
(c) (E-Step:) Approximate the expected "complete-data" log likelihood for λ by substituting averages based on the Gibbs sample of the previous step for any terms involving expected values of $\boldsymbol{\beta}, \sigma^2$, or $\tau_1^2, \ldots, \tau_p^2$.
(d) (M-Step:) Let $\lambda^{(k+1)}$ be the value of λ that maximises the expected log likelihood of the previous step.
(e) Return to the second step, and iterate until desired level of convergence.

The "complete-data" log likelihood based on the hierarchy of Section 2 with the conjugate prior (5) is

$$-\left((n+p-1)/2+a+1\right)\ln\left(\sigma^2\right) - \frac{1}{\sigma^2}\left((\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}}(\tilde{\boldsymbol{y}}-\boldsymbol{X}\boldsymbol{\beta})/2 + \gamma\right)$$
$$- \frac{1}{2}\sum_{j=1}^{p}\ln\left(\tau_j^2\right) - \frac{1}{2}\sum_{j=1}^{p}\frac{\beta_j^2}{\sigma^2\tau_j^2} + p\ln\left(\lambda^2\right) - \frac{\lambda^2}{2}\sum_{j=1}^{p}\tau_j^2$$

(after neglecting some additive constant terms not involving λ). The ideal E-step of iteration k involves taking the expected value of this log likelihood conditional on $\tilde{\boldsymbol{y}}$ under the current iterate $\lambda^{(k)}$ to get

$$Q(\lambda|\lambda^{(k)}) = p\ln\left(\lambda^2\right) - \frac{\lambda^2}{2}\sum_{j=1}^{p}E_{\lambda^{(k)}}\left[\tau_j^2\middle|\tilde{\boldsymbol{y}}\right] + \text{terms not involving } \lambda$$

(in the usual notation associated with EM). The M-step admits a simple analytical solution: The λ maximising this expression becomes the next EM iterate

$$\lambda^{(k+1)} = \sqrt{\frac{2p}{\sum_{j=1}^{p}E_{\lambda^{(k)}}\left[\tau_j^2\middle|\tilde{\boldsymbol{y}}\right]}}.$$

Of course, the conditional expectations must be replaced with the sample averages from the Gibbs sampler run.

When applied to the diabetes data using the scale invariant prior for σ^2 $(a = 0, \gamma = 0)$, this algorithm yields an optimal λ of approximately 0.237. The corresponding vector of medians for $\boldsymbol{\beta}$ has L_1 norm of approximately 0.59 relative to least squares (as indicated in Figure 1). Table 1 lists these posterior median estimates along with two corresponding sets of Lasso estimates, one chosen by n-fold cross-validation and one chosen to match the L_1 norm of the Bayes estimate. The Bayes estimates are very similar to the Lasso estimates in both cases.

We have found that the convergence rate of the EM algorithm can be dramatically affected by choice of the initial value of λ. Particularly large choices of λ can cause convergence of the EM algorithm to be impractically slow. Even when it converges relatively quickly, the accuracy is ultimately limited by the level of approximation of the expected values. When each step uses the same fixed number of iterations in the Gibbs sampler, the iterates will not converge but instead drift randomly about the true value, with the degree of drift depending on the number of Gibbs sampler iterations. McCulloch (1997) and Booth and Hobert (1999) encountered similar problems when employing Monte Carlo maximum likelihood methods to fit generalised linear mixed models, and suggested ways to alleviate the problem. (These remedies typically involve increasing the Monte Carlo replications as the estimates near convergence.)

Monte Carlo techniques for likelihood function approximation are also easy to implement. For notational simplicity, let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \tau_1^2, \ldots, \tau_p^2)$. Then, for any λ and λ_0, the likelihood ratio can be written

$$\begin{aligned}\frac{L(\lambda|\tilde{\boldsymbol{y}})}{L(\lambda_0|\tilde{\boldsymbol{y}})} &= \int \frac{L(\lambda|\tilde{\boldsymbol{y}})}{L(\lambda_0|\tilde{\boldsymbol{y}})}\, \pi_\lambda(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})\, d\boldsymbol{\theta} \;=\; \int \frac{f_\lambda(\tilde{\boldsymbol{y}},\boldsymbol{\theta})\, \pi_{\lambda_0}(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})}{\pi_\lambda(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})\, f_{\lambda_0}(\tilde{\boldsymbol{y}},\boldsymbol{\theta})}\, \pi_\lambda(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})\, d\boldsymbol{\theta} \\ &= \int \frac{f_\lambda(\tilde{\boldsymbol{y}},\boldsymbol{\theta})}{f_{\lambda_0}(\tilde{\boldsymbol{y}},\boldsymbol{\theta})}\, \pi_{\lambda_0}(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})\, d\boldsymbol{\theta}\end{aligned}$$

where f_λ is the complete joint density for a particular λ and π_λ is the full posterior. Since f_λ is known explicitly for all λ, the final expression may be used to approximate the likelihood ratio as a function of λ from a single Gibbs sample taken at the fixed λ_0. In particular,

$$\frac{f_\lambda(\tilde{\boldsymbol{y}},\boldsymbol{\theta})}{f_{\lambda_0}(\tilde{\boldsymbol{y}},\boldsymbol{\theta})} = \left(\frac{\lambda^2}{\lambda_0^2}\right)^p \exp\left\{-(\lambda^2-\lambda_0^2)\sum_{j=1}^p \frac{\tau_j^2}{2}\right\}$$

and thus

$$\frac{L(\lambda|\tilde{\boldsymbol{y}})}{L(\lambda_0|\tilde{\boldsymbol{y}})} = \left(\frac{\lambda^2}{\lambda_0^2}\right)^p \int \exp\left\{-(\lambda^2-\lambda_0^2)\sum_{j=1}^p \frac{\tau_j^2}{2}\right\} \pi_{\lambda_0}(\tau_1^2,\ldots,\tau_p^2|\tilde{\boldsymbol{y}})\, d\tau_1^2 \cdots d\tau_p^2 .$$

(The approximation is best in the neighbourhood of λ_0.) As a by-product, this expression may also be used to establish conditions for existence and uniqueness of the maximum likelihood estimate through the apparent connection with the posterior moment generating function of $\sum_{j=1}^p \tau_j^2/2$.

Figure 5 shows an approximation to the logarithm of this likelihood ratio for the diabetes data, using the Gibbs sampler of Section 3 with the scale-invariant prior for σ^2 and with λ_0 taken to be the maximum likelihood estimate (approximately 0.237). The figure includes the nominal 95% reference line based on the usual chi-square approximation to the log likelihood ratio statistic. The associated confidence interval $(0.125, 0.430)$ corresponds to the approximate range $(0.563, 0.657)$ of relative L_1 norms for the vector of posterior medians (compare Figure 1).

Using the marginal maximum likelihood estimate for λ is an empirical Bayes approach that does not automatically account for uncertainty in the maximum likelihood estimate. However, the effect of this uncertainty can be evaluated by considering the range of values of λ contained in the approximate 95% confidence interval stated above. Informal investigation of the sensitivity to λ, by using values at the extremes of the approximate 95% confidence

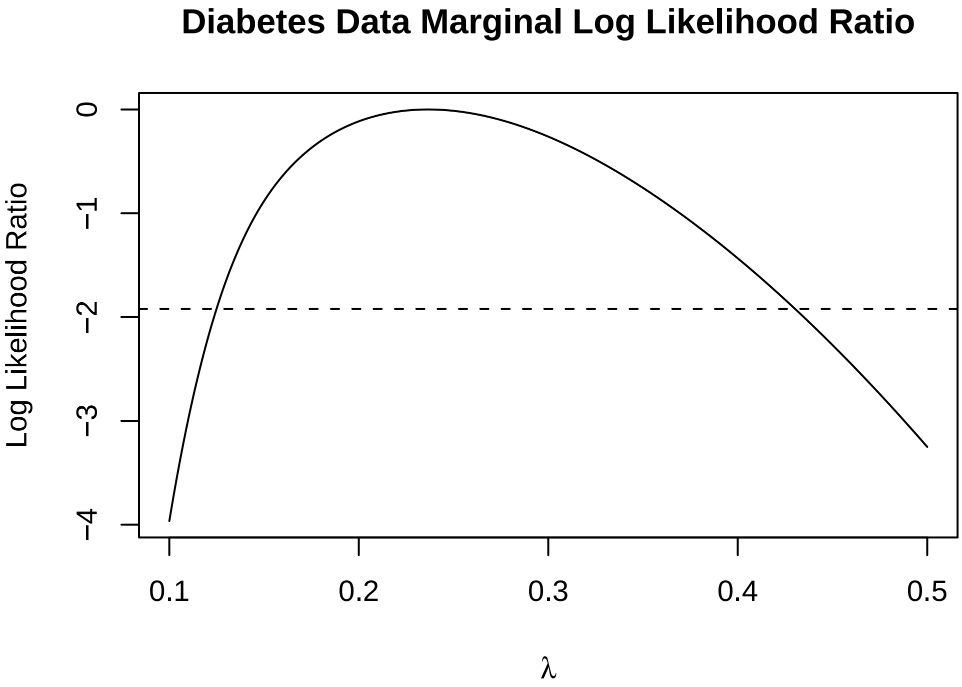

Fig. 5. The log likelihood ratio $\log\{L(\lambda|\tilde{\boldsymbol{y}})/L(\lambda_{\mathsf{MLE}}|\tilde{\boldsymbol{y}})\}$ for the diabetes data, as approximated by a Monte Carlo method described in the text. The horizontal reference line at $-\chi^2_{1,0.95}/2$ suggests the approximate 95% confidence interval $(0.125, 0.430)$.

interval, reveals that the posterior median estimates are not particularly sensitive to the uncertainty in λ, but that the range of the credible sets can be quite sensitive to λ. In particular, choosing λ near the low end of its confidence interval widens the 95% credible intervals enough to include the least squares estimates. An alternative to this approach is to adopt the fully Bayesian model that puts a hyperprior on λ. This is discussed in the next section.

5.2. Hyperpriors for the Lasso Parameter

Placing a hyperprior on λ is appealing because it both obviates the choice of λ and automatically accounts for the uncertainty in its selection that affects credible intervals for the parameters of interest. However, this hyperprior must be chosen carefully, as certain priors on λ may induce not only multiple modes and but also non-integrability of the posterior distribution.

For convenience, we will regard λ^2 as the parameter, rather than λ, throughout this section. We consider the class of gamma priors on λ^2 of the form

$$\pi(\lambda^2) \;=\; \frac{\delta^r}{\Gamma(r)}\left(\lambda^2\right)^{r-1} e^{-\delta\lambda^2}, \qquad \lambda^2 > 0 \qquad (r > 0, \delta > 0) \tag{8}$$

because conjugacy properties allow easy extension of the Gibbs sampler. The improper scale-invariant prior $1/\lambda^2$ for λ^2 (formally obtained by setting $r = 0$ and $\delta = 0$) is a tempting choice, but it leads to an improper posterior, as will be seen subsequently. Moreover, scale invariance is not a very compelling criterion for choice of prior in this case because λ is unitless when prior (2) is used for $\boldsymbol{\beta}$ (Section 4).

When prior (8) is used in the hierarchy of (3), the product of the factors in the joint density that involve λ is

$$\left(\lambda^2\right)^{p+r-1} \exp\left\{-\lambda^2\left(\frac{1}{2}\sum_{j=1}^{p}\tau_j^2 + \delta\right)\right\}$$

and thus the full conditional distribution of λ^2 is gamma with shape parameter $p + r$ and *rate* parameter $\sum_{i=1}^{p}\tau_i^2/2 \;+\; \delta$. With this specification, λ^2 can simply join the other parameters in the Gibbs sampler of Section 3, since the full conditional distributions of the other parameters do not change.

The parameter δ must be sufficiently larger than zero to avoid computational and conceptual problems. To illustrate why, suppose the improper prior $\pi(\lambda^2) \;=\; \left(\lambda^2\right)^{r-1}$ (formally the $\delta = 0$ case of (8)) is used in conjunction with priors (2) and (5). Then the joint density of $\tilde{\boldsymbol{y}}$, $\boldsymbol{\beta}$, σ^2, and λ^2 (marginal only over $\tau_1^2, \ldots, \tau_p^2$) is proportional to

$$\left(\lambda^2\right)^{p/2+r-1}\left(\sigma^2\right)^{-n/2-p/2-a-1} \exp\left\{-\frac{1}{\sigma^2}\left((\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})/2 \,+\, \gamma\right) \;-\; \frac{\sqrt{\lambda^2}}{\sqrt{\sigma^2}}\sum_{i=1}^{p}|\beta_i|\right\}$$

Marginalising over λ^2 (most easily done by making a transformation of variable back to λ) produces a joint density of $\tilde{\boldsymbol{y}}$, $\boldsymbol{\beta}$, and σ^2 proportional to

$$\left(\sigma^2\right)^{-n/2+r-a-1} \exp\left\{-\frac{1}{\sigma^2}\left((\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})/2 \,+\, \gamma\right)\right\}\left(\sum_{i=1}^{p}|\beta_i|\right)^{-p-2r}$$

and further marginalising over σ^2 gives

$$\left((\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})/2 \,+\, \gamma\right)^{-n/2+r-a} \left(\sum_{i=1}^{p} |\beta_i|\right)^{-p-2r} .$$

For fixed $\tilde{\boldsymbol{y}}$, both of these last two expressions are degenerate at $\boldsymbol{\beta} = \mathbf{0}$ and bimodal (unless the least squares estimate is exactly $\mathbf{0}$). The same computational and conceptual problems result as discussed in Section 4. Moreover, taking $r = 0$ produces a posterior that is not integrable due to the singularity at $\boldsymbol{\beta} = \mathbf{0}$.

It is thus necessary to choose a proper prior for λ^2, though to reduce bias it is desirable to make it relatively flat, at least near the maximum likelihood estimate. If, for the diabetes data, we take $r = 1$ and $\delta = 1.78$ (so that the prior on λ^2 is exponential with mean equal to about ten times the maximum likelihood estimate), then the posterior median for λ is approximately 0.279 and a 95% equal-tailed posterior credible interval for λ is approximately $(0.139, 0.486)$. Posterior medians and 95% credible intervals for the regression coefficients are shown in Figure 6, along with the intervals from Figure 2 for comparison. The two sets of intervals are practically identical in this case.

6. Extensions

Hierarchies based on various scale mixtures of normals have been used in Bayesian analysis both to produce priors with useful properties and to robustify error distributions (West, 1984). The hierarchy of Section 2 can be used to mimic or implement many other methods through modifications of the priors on $\tau_1^2, \ldots, \tau_p^2$ and σ^2. One trivial special case is ridge regression, in which the τ_j^2's are all taken to have degenerate distributions at the same constant value. We briefly list Bayesian alternatives corresponding to two other Lasso-related methods.

- Bridge Regression

  One direct generalisation of the Lasso (and ridge regression) is penalised regression by solving (Frank and Friedman, 1993)

  $$\min_{\boldsymbol{\beta}} \qquad (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})^{\mathsf{T}}(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) \quad + \quad \lambda \sum_{j=1}^{p} |\beta_j|^q$$

  for some $q \geq 0$ (the $q = 0$ case corresponding to best-subset regression). See also Hastie et al. (2001, Sec. 3.4.5) and Fu (1998), in which this is termed "bridge regression" in the case $q \geq 1$. Of course, $q = 1$ is the ordinary Lasso and $q = 2$ is ridge regression.

  The Bayesian analogue of this penalisation involves using a prior on $\boldsymbol{\beta}$ of the form

  $$\pi(\boldsymbol{\beta}) \,\propto\, \prod_{j=1}^{p} e^{-\lambda|\beta_j|^q}$$

  although, in parallel with (2), we would emend this to

  $$\pi(\boldsymbol{\beta}|\sigma^2) \,\propto\, \prod_{j=1}^{p} e^{-\lambda\left(|\beta_j|/\sqrt{\sigma^2}\right)^q} .$$

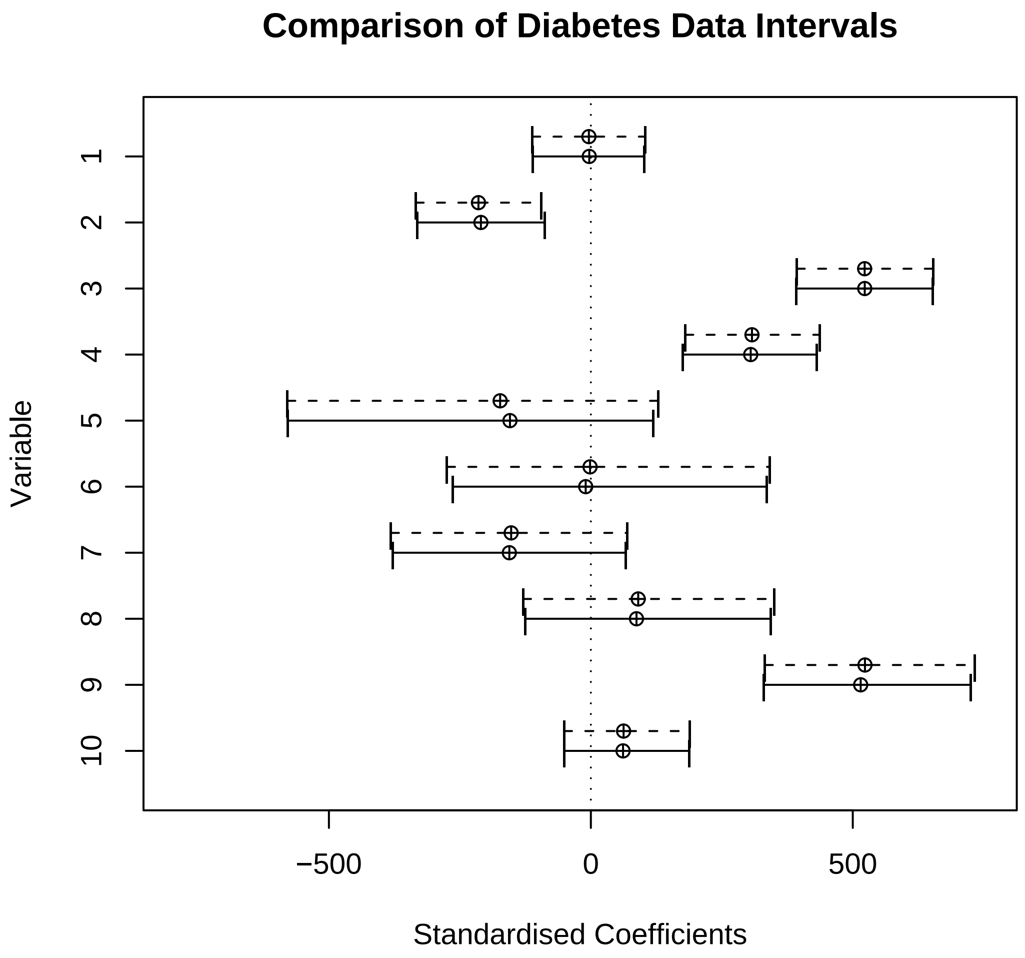

Fig. 6. Posterior median Bayesian Lasso estimates and corresponding 95% credible intervals (solid lines) from the fully hierarchical formulation with λ^2 having an exponential prior with mean $1/1.78$. The empirical Bayes estimates and intervals of Figure 2 are plotted in dashed lines above these for comparison.

Thus the elements of $\boldsymbol{\beta}$ have (conditionally) independent priors from the *exponential power* distribution (Box and Tiao, 1973) (also known as the "generalised Gaussian" distribution in electrical engineering literature), though technically this name is reserved for the case $q \geq 1$. Whenever $0 < q \leq 2$, this distribution may be represented by a scale mixture of normals. Indeed, for $0 < q < 2$,

$$e^{-|z|^q} \propto \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-z^2/(2s)} \, \frac{1}{s^{3/2}} \, g_{q/2}\left(\frac{1}{2s}\right) ds$$

where $g_{q/2}$ is the density of a positive stable random variable with index $q/2$ (West, 1987; Gneiting, 1997), which generally does not have a closed form expression. A hierarchy of the type in Section 2 is applicable by placing appropriate independent distributions on $\tau_1^2, \ldots, \tau_p^2$. Their resulting full conditional distributions are closely related to certain exponential dispersion models (Jørgensen, 1987). It is not clear whether an efficient Gibbs sampler can be based on this hierarchy, however.

- The "Huberized Lasso"

Rosset and Zhu (2004) illustrate that the Lasso may be made more robust by using loss functions that are less severe than the quadratic. They illustrate the result of solving

$$\min_{\boldsymbol{\beta}} \quad \sum_{i=1}^{n} L(\tilde{y}_i - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}) \quad + \quad \lambda \sum_{j=1}^{p} |\beta_j|,$$

where L is a once-differentiable piecewise quadratic Huber-type loss function that is quadratic in a neighbourhood of zero and linearly increases away from zero outside of that neighbourhood. It is not easily possible to implement an exact Bayesian analogue of this technique, but it is possible to implement a Bayesian analogue of the very similar hyperbolic loss

$$L(d) \;=\; \sqrt{\eta(\eta + d^2/\rho^2)}$$

for some parameters $\eta > 0$ and $\rho^2 > 0$. Note that this is almost quadratic near zero and asymptotically approaches linearity away from zero.

The key idea for robustification is to replace the usual linear regression model with

$$\boldsymbol{y} \sim N_n(\mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{D}_\sigma)$$

where $\boldsymbol{D}_\sigma = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. (Note the necessary re-introduction of the overall mean parameter μ, which can safely be given an independent, flat prior.) Then independent and identical priors are placed on $\sigma_1^2, \ldots, \sigma_n^2$. To mimic the hyperbolic loss, an appropriate prior for $(\sigma_1^2, \ldots, \sigma_n^2)$ is

$$\prod_{i=1}^{n} \frac{1}{2K_1(\eta)\rho^2} \exp\left(-\frac{\eta}{2}\left(\frac{\sigma_i^2}{\rho^2} + \frac{\rho^2}{\sigma_i^2}\right)\right)$$

where K_1 is the modified Bessel K function with index 1, $\eta > 0$ is a shape parameter, and $\rho^2 > 0$ is a scale parameter. The scale parameter ρ^2 can be given the non-informative scale-invariant prior $1/\rho^2$, and the prior (3) on $\boldsymbol{\beta}$ would use ρ^2 in place of σ^2. Upon applying this prior and integrating out $\sigma_1^2, \ldots, \sigma_n^2$, the conditional density of the observations given the remaining parameters is

$$\prod_{i=1}^{n} \frac{1}{2K_1(\eta)\sqrt{\eta\rho^2}} \exp\Big(- \sqrt{\eta\big(\eta + (y_i - \mu - \boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta})^2/\rho^2\big)}\Big)$$

(Gneiting, 1997), which has the desired hyperbolic form. The Gibbs sampler is easy to implement here because the full conditional distributions of the σ_i^2's are reciprocal inverse Gaussian, and the full conditional distribution of ρ^2 is in the class of *generalised* inverse Gaussian distributions, for which reasonably efficient simulation algorithms exist (Atkinson, 1982).

7. Discussion

For the diabetes data, results from the Bayesian Lasso are surprisingly similar to those from the ordinary Lasso. Although computationally more intensive, it is just as easy to implement and provides finite-sample interval estimates, which are not available for the ordinary Lasso. The asymptotics for Lasso-type estimators by Knight and Fu (2000) might be used to construct frequentist confidence sets, but it isn't clear what their small-sample properties might be.

Credible sets allow assessment of practical significance as well as statistical significance. If a parameter must meet a certain threshold to be considered significant, a credible set will indicate the degree of certainty that this requirement is met.

Correcting Bayesian credible sets for multiple comparisons is reasonably straightforward. For instance, simultaneous intervals for the elements of $\boldsymbol{\beta}$ can be obtained by expanding the hyper-rectangle defined by the uncorrected credible intervals until it includes 95% of the sampled points. The sides of this expanded credible set would then be nearly exact simultaneous credible intervals for the coefficients.

The ordinary Lasso, as computed using the LARS algorithm, has the property that at most $n-1$ variables may have nonzero coefficients, which is not necessarily desirable when $n-1 \ll p$. In contrast, the $n-1<p$ case poses no such problems for the Bayesian version. In informal simulations under the condition $n-1<p$, we have observed convergence to a solution that is nearly a legitimate solution to the normal equations and has all medians clearly nonzero, in contrast to the Lasso solution, which necessarily sets $p-n+1$ coefficients to zero.

Appendix: Derivation of Unimodality

We demonstrate that the joint posterior distribution of $\boldsymbol{\beta}$ and $\sigma^2 > 0$ under prior

$$\begin{aligned}\pi(\boldsymbol{\beta}, \sigma^2) &= \pi(\sigma^2) \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}}\, e^{-\lambda|\beta_j|/\sqrt{\sigma^2}} \\ &= \pi(\sigma^2)\, \frac{\lambda^p}{2^p\,(\sigma^2)^{p/2}}\, e^{-\lambda\|\boldsymbol{\beta}\|_1/\sqrt{\sigma^2}}\end{aligned}$$

is unimodal (for typical choices of π), in the sense that every upper level set $\{(\boldsymbol{\beta}, \sigma^2) \mid \pi(\boldsymbol{\beta}, \sigma^2) > x,\ \sigma^2 > 0\}, x > 0$, is connected.

Unimodality in this sense is immediate for densities that are log concave. Unfortunately, this isn't quite true in this case, but we can instead show that it is true under a continuous transformation with continuous inverse, which will prove unimodality just as effectively, since connected sets are the images of connected sets under such a transformation.

The log posterior is

$$\log\big(\pi(\sigma^2)\big) \;-\; \frac{n+p-1}{2}\log(\sigma^2) \;-\; \frac{1}{2\sigma^2}\,\|\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \;-\; \lambda\|\boldsymbol{\beta}\|_1/\sqrt{\sigma^2} \tag{9}$$

after dropping all additive terms that involve neither $\boldsymbol{\beta}$ nor σ^2. Consider the transformation defined by

$$\boldsymbol{\phi} \;\leftrightarrow\; \boldsymbol{\beta}/\sqrt{\sigma^2} \qquad \rho \;\leftrightarrow\; 1/\sqrt{\sigma^2},$$

which is continuous with a continuous inverse when $0 < \sigma^2 < \infty$. Note that this is simply intended as a coordinate transformation, not as a transformation of measure (i.e. no Jacobian), so that upper level sets for the new parameters correspond under the transformation to upper level sets for the original parameters. In the transformed parameters, (9) becomes

$$\log\big(\pi\big(1/\rho^2\big)\big) \;+\; (n+p-1)\,\log(\rho) \;-\; \frac{1}{2}\,\|\rho\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\phi}\|_2^2 \;-\; \lambda\|\boldsymbol{\phi}\|_1, \tag{10}$$

where π is the prior density for σ^2. The second and fourth terms are clearly concave in $(\rho, \boldsymbol{\phi})$, and the third term is a concave quadratic in $(\rho, \boldsymbol{\phi})$. Thus, the expression is concave if $\log\big(\pi(1/\rho^2)\big)$ is concave (because a sum of concave functions is concave). Function $\log\big(\pi(1/\rho^2)\big)$ is concave if, for instance, the prior on σ^2 is the inverse gamma prior (5) or the scale-invariant prior $1/\sigma^2$.

The log posterior is thus unimodal in the sense that every upper level set is connected, though this does not guarantee a unique maximiser. A sufficient condition for a unique maximiser is that the $\boldsymbol{X}$ matrix has full rank and $\tilde{\boldsymbol{y}}$ is not in the column space of $\boldsymbol{X}$, since this makes the third term of (10) strictly concave.

References

Andrews, D. F., and Mallows, C. L. (1974), "Scale Mixtures of Normal Distributions," *Journal of the Royal Statistical Society*, Ser. B, 36, 99–102.

Athreya, K. B. (1986), "Another Conjugate Family for the Normal Distribution," *Statistics & Probability Letters*, 4, 61–64.

Atkinson, A. C. (1982), "The Simulation of Generalized Inverse Gaussian and Hyperbolic Random Variables," *SIAM Journal on Scientific and Statistical Computing*, 3, 502–515.

Booth, J. G., and Hobert, J. P. (1999), "Maximizing Generalized Linear Mixed Model Likelihoods with an Automated Monte Carlo EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 61, 265–285.

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Addison-Wesley.

Casella, G. (2001), "Empirical Bayes Gibbs Sampling," *Biostatistics*, 2, 485–500.

Chhikara, R. S., and Folks, L. (1989), *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*, Marcel Dekker Inc.

Clyde, M., and George, E. I. (2000), "Flexible Empirical Bayes Estimation for Wavelets," *Journal of the Royal Statistical Society*, Ser. B, 62, 681–698.

Clyde, M., Parmigiani, G., and Vidakovic, B. (1998), "Multiple Shrinkage and Subset Selection in Wavelets," *Biometrika*, 85, 391–401.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499.

Figueiredo, M. A. T. (2003), "Adaptive Sparseness for Supervised Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25, 1150–1159.

Frank, I. E., and Friedman, J. H. (1993), "A Statistical View of Some Chemometrics Regression Tools (Disc: P136-148)," *Technometrics*, 35, 109–135.

Fu, W. J. (1998), "Penalized Regressions: The Bridge versus the Lasso," *Journal of Computational and Graphical Statistics*, 7, 397–416.

George, E. I., and McCulloch, R. E. (1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

— (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–374.

Gneiting, T. (1997), "Normal Scale Mixtures and Dual Probability Densities," *Journal of Statistical Computation and Simulation*, 59, 375–384.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag Inc.

Ishwaran, H., and Rao, J. S. (2005), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *Annals of Statistics*, 33, 730–773.

Jørgensen, B. (1987), "Exponential Dispersion Models," *Journal of the Royal Statistical Society*, Ser. B, 49, 127–162.

Knight, K., and Fu, W. (2000), "Asymptotics for Lasso-Type Estimators," *The Annals of Statistics*, 28, 1356–1378.

McCulloch, C. E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 92, 162–170.

Mitchell, T. J., and Beauchamp, J. J. (1988), "Bayesian Variable Selection in Linear Regression (C/R: p1033-1036)," *Journal of the American Statistical Association*, 83, 1023–1032.

Osborne, M. R., Presnell, B., and Turlach, B. A. (2000), "A New Approach to Variable Selection in Least Squares Problems," *IMA Journal of Numerical Analysis*, 20, 389–404.

Rosset, S., and Zhu, J. (2004), "Discussion of "Least Angle Regression" By Efron et. al." *Annals of Statistics*, 32, 469–475.

Smyth, G. (2005), *statmod: Statistical Modeling*; R package version 1.1.1.

Tibshirani, R. (1996), “Regression Shrinkage and Selection Via the Lasso,” *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

West, M. (1984), “Outlier Models and Prior Distributions in Bayesian Linear Regression,” *Journal of the Royal Statistical Society*, Ser. B, 46, 431–439.

— (1987), “On Scale Mixtures of Normal Distributions,” *Biometrika*, 74, 646–648.

Yuan, M., and Lin, Y. (2005), “Efficient Empirical Bayes Variable Selection and Esimation in Linear Models,” *Journal of the American Statistical Association*, 100, 1215–1225.