Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

The Bayesian Lasso

Trevor Park & George Casella

To cite this article: Trevor Park & George Casella (2008) The Bayesian Lasso, Journal of the American Statistical Association, 103:482, 681-686, DOI: 10.1198/016214508000000337

To link to this article: https://doi.org/10.1198/016214508000000337

The Bayesian Lasso

Trevor PARK and George CASELLA

The Lasso estimate for linear regression parameters can be interpreted as a Bayesian posterior mode estimate when the regression parameters have independent Laplace (i.e., double-exponential) priors. Gibbs sampling from this posterior is possible using an expanded hierarchy with conjugate normal priors for the regression parameters and independent exponential priors on their variances. A connection with the inverse-Gaussian distribution provides tractable full conditional distributions. The Bayesian Lasso provides interval estimates (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. Slight modifications lead to Bayesian versions of other Lasso-related estimation methods, including bridge regression and a robust variant.

KEY WORDS: Empirical Bayes; Gibbs sampler; Hierarchical model; Inverse Gaussian; Linear regression; Penalized regression; Scale mixture of normals.

1. INTRODUCTION

The Lasso of Tibshirani (1996) estimates linear regression coefficients through L_1-constrained least squares. The Lasso is usually used to estimate the regression parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ in the model

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

where $\mathbf{y}$ is the $n \times 1$ vector of responses, μ is the overall mean, $\mathbf{X}$ is the $n \times p$ matrix of *standardized* regressors, and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of independent and identically distributed normal errors with mean 0 and unknown variance σ^2. For convenience, Lasso estimates often are viewed as L_1-penalized least squares estimates. They achieve

$$\min_{\boldsymbol{\beta}} (\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^\top (\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

for some $\lambda \geq 0$, where $\tilde{\mathbf{y}} = \mathbf{y} - \bar{y}\mathbf{1}_n$. The entire path of Lasso estimates for all values of λ can be efficiently computed through a modification of the LARS algorithm of Efron, Hastie, Johnstone, and Tibshirani (2004) (see also Osborne, Presnell, and Turlach 2000a).

Noting the form of the penalty term in (2), Tibshirani (1996) suggested that Lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace (i.e., double-exponential) priors. Motivated by this connection, several other authors subsequently proposed using Laplace-like priors (e.g., Figueiredo 2003; Bae and Mallick 2004; Yuan and Lin 2005). We consider a fully Bayesian analysis using a conditional Laplace prior specification of the form

$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}} \tag{3}$$

and the noninformative scale-invariant marginal prior $\pi(\sigma^2) = 1/\sigma^2$ on σ^2. Conditioning on σ^2 is important, because it guarantees a unimodal full posterior (see App. A). Without this, the posterior may not be unimodal, as shown by example in Appendix B. Lack of unimodality slows convergence of the Gibbs sampler and makes point estimates less meaningful.

Figure 1 compares posterior median estimates for this Bayesian Lasso model 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 these estimates as their respective shrinkage parameters are varied. For ease of comparison, each is plotted as a function of its relative L_1 norm. The Bayesian Lasso estimates appear 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. Specifically, the Bayesian Lasso appears to pull the more weakly related parameters to 0 faster than ridge regression does, indicating a potential advantage of the Laplace prior over a Gaussian (or a Student-t) prior. The vertical line in the Lasso panel represents the estimate chosen by n-fold (leave-one-out) cross-validation (see, e.g., Hastie, Tibshirani, and Friedman 2001), whereas the vertical line in the Bayesian Lasso panel represents the estimate chosen by marginal maximum likelihood (Sec. 3.1).

With λ selected by marginal maximum likelihood, posterior medians and 95% credible intervals for the diabetes data regression parameters are shown in Figure 2. For comparison, the least squares estimates and the Lasso estimates for two different values of λ are also shown. Although they are not sparse in the exact sense, the Bayesian posterior medians are remarkably similar in value to the Lasso estimates. Moreover, all of the Lasso estimates are well within the credible intervals, whereas the least squares estimates are outside for four of the variables, one of which is significant.

The following sections outline a simple and practical Gibbs sampler implementation for the Bayesian Lasso and offer methods that address the choice of λ.

2. HIERARCHICAL MODEL AND GIBBS SAMPLER

The Gibbs sampler for the Bayesian Lasso exploits the following representation of the Laplace distribution as a scale mixture of normals (with an exponential mixing density):

$$\frac{a}{2} e^{-a|z|} = \int_0^\infty \frac{1}{\sqrt{2\pi s}} e^{-z^2/(2s)} \frac{a^2}{2} e^{-a^2 s/2}\, ds, \qquad a > 0 \tag{4}$$

Trevor Park is Assistant Professor (E-mail: *tpark@stat.ufl.edu*) and George Casella is Distinguished Professor (E-mail: *casella@stat.ufl.edu*), Department of Statistics, University of Florida, Gainesville, FL 32611. This work was supported by National Security Agency grant H98230-07-1-0031 and by National Science Foundation grants DMS-04-05543, DMS-0631632, and SES-0631588.

Journal of the American Statistical Association
June 2008, Vol. 103, No. 482, Theory and Methods
DOI 10.1198/016214508000000337

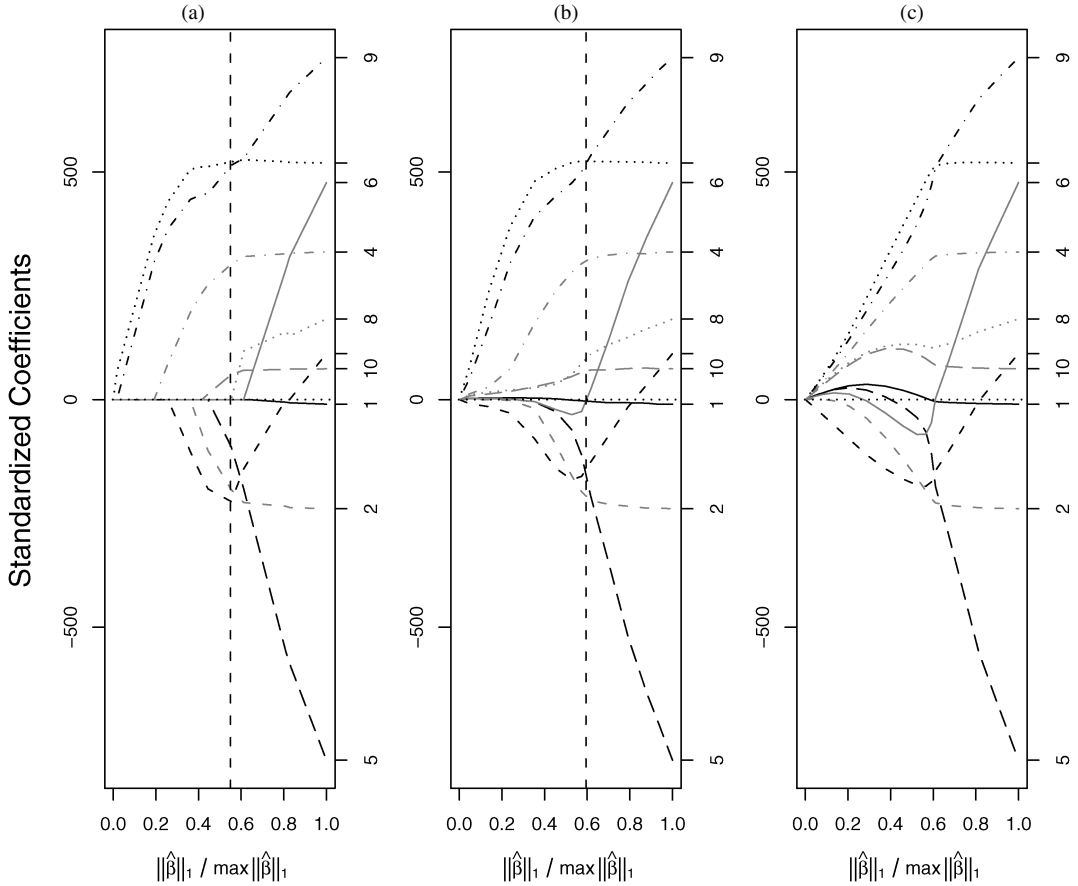

Figure 1. Lasso (a), Bayesian Lasso (b), and ridge regression (c) trace plots for estimates of the diabetes data regression parameters versus the relative L_1 norm, with vertical lines for the Lasso and Bayesian Lasso indicating the estimates chosen by n-fold cross-validation and marginal maximum likelihood. The Bayesian Lasso estimates were posterior medians computed over a grid of λ values, using 10,000 consecutive iterations of the Gibbs sampler of Section 2 (after 1,000 burn-in iterations) for each λ.

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

$$\begin{aligned} \mathbf{y}|\mu, \mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathrm{N}_n(\mu \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n), \\ \boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2 &\sim \mathrm{N}_p(\mathbf{0}_p, \sigma^2 \mathbf{D}_\tau), \\ \mathbf{D}_\tau &= \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2), \\ \sigma^2, \tau_1^2, \ldots, \tau_p^2 &\sim \pi(\sigma^2)\, d\sigma^2 \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2}\, d\tau_j^2, \\ \sigma^2, \tau_1^2, \ldots, \tau_p^2 &> 0. \end{aligned} \tag{5}$$

(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 desired form (3). We use the improper prior density $\pi(\sigma^2) = 1/\sigma^2$, but any inverse-gamma prior for σ^2 also would maintain conjugacy.

Similar hierarchies based on (4) have been used by other authors. Figueiredo (2003) used such a hierarchy in conjunction with an EM algorithm to compute a marginal posterior mode. Bae and Mallick (2004) proposed a variant of this hierarchy and a corresponding Gibbs sampler for probit binary regression (which does not need a separate variance parameter).

Because the columns of $\mathbf{X}$ are centered, it is easy to analytically integrate μ from the joint posterior under its independent, flat prior. Because μ is rarely of interest, we marginalize it out in the interest of simplicity and speed. If desired, it can be reintroduced with a full conditional distribution that is normal with mean $\overline{y}$ and variance σ^2/n.

Marginalizing over μ does not affect conjugacy. The full conditional distributions of $\boldsymbol{\beta}$, σ^2, and $\tau_1^2, \ldots, \tau_p^2$ are still easy to sample, and they depend on the centered response vector $\tilde{\mathbf{y}}$. The full conditional for $\boldsymbol{\beta}$ is multivariate normal with mean $\mathbf{A}^{-1}\mathbf{X}^\top \tilde{\mathbf{y}}$ and variance $\sigma^2 \mathbf{A}^{-1}$, where $\mathbf{A} = \mathbf{X}^\top \mathbf{X} + \mathbf{D}_\tau^{-1}$. The full conditional for σ^2 is inverse-gamma with shape parameter $(n-1)/2 + p/2$ and scale parameter $(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^\top (\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})/2 + \boldsymbol{\beta}^\top \mathbf{D}_\tau^{-1} \boldsymbol{\beta}/2$, and $\tau_1^2, \ldots, \tau_p^2$ are conditionally independent, with $1/\tau_j^2$ conditionally inverse-Gaussian with parameters

$$\mu' = \sqrt{\frac{\lambda^2 \sigma^2}{\beta_j^2}} \qquad \text{and} \qquad \lambda' = \lambda^2$$

in the parameterization of the inverse-Gaussian density given by

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

(Chhikara and Folks 1989). These full conditionals form the basis for an efficient Gibbs sampler, with block updating of $\boldsymbol{\beta}$

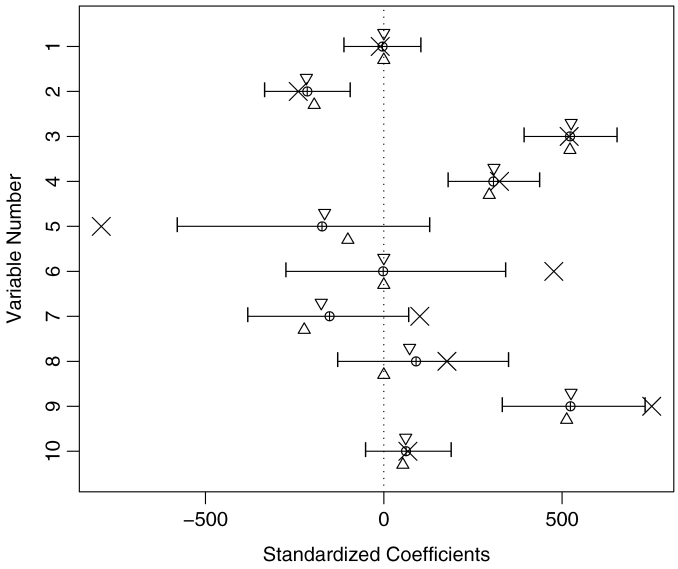

Figure 2. Posterior median Bayesian Lasso estimates ($\oplus$) and corresponding 95% credible intervals (equal-tailed) with λ selected according to marginal maximum likelihood (Sec. 3.1). Overlaid are the least squares estimates ($\times$), Lasso estimates based on n-fold cross-validation ($\triangle$), and Lasso estimates chosen to match the L_1 norm of the Bayes estimates ($\triangledown$). The variables were described by Efron et al. (2004): (1) age, (2) sex, (3) bmi, (4) map, (5) tc, (6) ldl, (7) hdl, (8) tch, (9) ltg, and (10) glu.

and $(\tau_1^2, \ldots, \tau_p^2)$. Our experience suggests that convergence is reasonably fast.

3. CHOOSING THE BAYESIAN LASSO PARAMETER

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

3.1 Empirical Bayes by Marginal Maximum Likelihood

Casella (2001) proposed a Monte Carlo EM algorithm that complements a Gibbs sampler and provides marginal maximum likelihood estimates of hyperparameters. For the Bayesian Lasso, each iteration of the algorithm involves running the Gibbs sampler using a λ value estimated from the sample of the previous iteration. Specifically, iteration k uses the Gibbs sampler of Section 2 with hyperparameter $\lambda^{(k-1)}$ (i.e., the estimate from iteration $k-1$) to approximate the ideal updated estimate,

$$\lambda^{(k)} = \sqrt{\frac{2p}{\sum_{j=1}^{p} E_{\lambda^{(k-1)}}[\tau_j^2 | \tilde{\mathbf{y}}]}},$$

by replacing the conditional expectations with averages from the Gibbs sample. (A derivation is provided in App. C.) We suggest the initial value

$$\lambda^{(0)} = p\sqrt{\hat{\sigma}^2_{\mathrm{LS}}} \Big/ \sum_{j=1}^{p} |\hat{\beta}_j^{\mathrm{LS}}|,$$

where $\hat{\sigma}^2_{\mathrm{LS}}$ and $\hat{\beta}_j^{\mathrm{LS}}$ are estimates from the usual least squares procedure. This empirical estimate tends to be smaller than the maximizing λ, but our experience suggests that only extreme initial overestimates of λ lead to slow convergence. Because the expectations are estimated from the Gibbs sampler, the successive λ estimates will not quite converge, but will eventually drift randomly about the true maximum likelihood estimate, with less drift if more Gibbs samples are taken in each iteration.

When applied to the diabetes data, this algorithm yields an optimal λ of approximately .237. The corresponding vector of medians for $\boldsymbol{\beta}$ has an L_1 norm of approximately .59 relative to least squares (as indicated in Fig. 1). Figure 2 shows that these posterior median estimates are very similar to certain Lasso estimates.

A Gibbs sample can be used with importance sampling methods for ratios of normalizing constants to approximate the likelihood ratio surface for λ near the maximizer. For the diabetes data, we used this method to obtain the approximate 95% likelihood ratio confidence interval (.125, .430) for λ, using the usual chi-squared approximation.

3.2 Hyperpriors for the Lasso Parameter

An alternative to choosing λ explicitly is to give it a diffuse hyperprior. We consider the class of gamma priors on λ^2 (not λ) of the form

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)}(\lambda^2)^{r-1}e^{-\delta\lambda^2}, \qquad \lambda^2 > 0 \ (r > 0, \delta > 0), \quad (6)$$

because the resulting conjugacy allows easy extension of the Gibbs sampler. The improper scale-invariant prior $1/\lambda^2$ for λ^2 ($r = 0, \delta = 0$) is tempting, but it leads to an improper posterior. Moreover, scale invariance is not a very compelling criterion, because λ is unitless.

When prior (6) is used in the hierarchy of (5), the full conditional distribution of λ^2 is gamma with shape parameter $p + r$ and *rate* parameter $\sum_{j=1}^{p} \tau_j^2/2 + \delta$. With this specification, λ^2 can simply join the other parameters in the Gibbs sampler, because the full conditional distributions of the other parameters do not change.

The prior density for λ^2 should approach 0 sufficiently fast as $\lambda^2 \to \infty$ (to avoid mixing problems) but should be relatively flat and place high probability near the maximum likelihood estimate. For the diabetes data, if we run this augmented Gibbs sampler with $r = 1$ and $\delta = 1.78$ (so that the prior on λ^2 is exponential with mean equal to about 10 times the maximum likelihood estimate), then the posterior median for λ is approximately .279, and a 95% equal-tailed posterior credible interval for λ is approximately (.139, .486). Posterior medians and 95% credible intervals for the regression coefficients are practically identical to those shown in Figure 2.

4. 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 specified in 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 all of the τ_j^2's are given degenerate distributions at the same constant value. We next briefly describe Bayesian alternatives to two other Lasso-related methods.

4.1 Bridge Regression

One direct generalization of the Lasso (and ridge regression) is penalized regression by solving (Frank and Friedman 1993)

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

for some $q \geq 0$ (with the $q = 0$ case corresponding to best-subset regression). (See also Hastie et al. 2001, sec. 3.4.5, Knight and Fu 2000, and Fu 1998, in which this is termed "bridge regression," at least in the case where $q \geq 1$.) Of course, $q = 1$ is the ordinary Lasso, and $q = 2$ is ridge regression.

The Bayesian analog of this penalization 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 keeping with (3), we would instead use

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

Thus the elements of $\boldsymbol{\beta}$ have (conditionally) independent priors from the *exponential power* distribution, although technically this term is reserved for the case where $q \geq 1$. 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 discussed 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). Whether an efficient Gibbs sampler can be based on this hierarchy is not clear, however.

4.2 The "Huberized Lasso"

Rosset and Zhu (2004) illustrated that the Lasso may be made more robust by using loss functions that are less severe than the quadratic. They studied estimates that solve

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

where L is a once-differentiable piecewise quadratic Huber-type loss function that is quadratic in a neighborhood of 0 and linearly increases away from 0 outside of that neighborhood. Implementing an exact Bayesian analog is not easy, but it is possible to implement a Bayesian analog of a 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 $d = 0$ and asymptotically approaches linearity as $|d| \to \infty$.

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

$$\mathbf{y}|\mu, \mathbf{X}, \boldsymbol{\beta}, \sigma_1^2, \ldots, \sigma_n^2 \sim \mathrm{N}_n(\mu\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \mathbf{D}_\sigma),$$

where $\mathbf{D}_\sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. (The parameter μ can be safely given an independent, flat prior as before, but marginalizing over μ is not as easy in this case.) 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 noninformative scale-invariant prior $1/\rho^2$, and the prior (5) on $\boldsymbol{\beta}$ would use ρ^2 in place of σ^2. On applying this prior and integrating out $\sigma_1^2, \ldots, \sigma_n^2$, the conditional density of the observations given the remaining parameters is

$$\prod_{i=1}^{n} \frac{1}{2K_1(\eta)\sqrt{\eta\rho^2}} \exp\left(-\sqrt{\eta\left(\eta + (y_i - \mu - \mathbf{x}_i^\top \boldsymbol{\beta})^2/\rho^2\right)}\right)$$

(Gneiting 1997), which has the desired hyperbolic form. The Gibbs sampler is easy to implement 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 *generalized* inverse-Gaussian distributions, for which reasonably efficient simulation algorithms exist (Atkinson 1982).

5. DISCUSSION

For the diabetes data, results from the Bayesian Lasso are strikingly similar to those from the ordinary Lasso. Although more computationally intensive, the Bayesian Lasso is similarly easy to implement and automatically provides interval estimates for all parameters, including the error variance.

The various proposed standard error estimators for the Lasso are not considered fully satisfactory. Approximate analytical methods proposed by Tibshirani (1996) and Fan and Li (2001), for example, fail to provide reasonable standard error estimates for the parameters estimated to be 0. The method of Osborne et al. (2000b) avoids this problem but is still based on a smooth approximation.

The methods of Section 3 for choosing λ are not intended for direct application to the ordinary Lasso, because λ serves the Bayesian Lasso in a somewhat different capacity. But they potentially could aid in choosing λ for the Lasso. The Lasso parameter is generally chosen using cross-validation, but the n-fold cross-validation choice is often unstable. For the diabetes data of Section 1, it was in fact rather poorly defined. (More general K-fold cross-validation may be more stable, but at the price of some bias.) Choosing the Lasso parameter so that the estimates resemble those of the Bayesian Lasso (by, e.g., choosing them to match in L_1 norm, as in Fig. 2) could result in more stable Lasso estimates.

The extension of the Bayesian Lasso to generalized linear models (GLMs) may yield further advantages. The Bayesian Lasso may be more computationally competitive relative to the Lasso in this context, because the fast special algorithms for the Lasso apply only to the linear case. Lasso algorithms for GLMs are generally much slower, making cross-validation computationally demanding. Extending the Bayesian Lasso will require methodological modifications, as demonstrated by Bae and Mallick (2004) for probit regression, as well as reintroduction of the intercept parameter μ (because it is unlikely to be analytically integrable from the posterior). But with careful implementation, the Bayesian Lasso need not require much more computation for GLMs than for linear models.

APPENDIX A: UNIMODALITY UNDER PRIOR (3)

Here we demonstrate that the joint posterior distribution $\pi(\boldsymbol{\beta}, \sigma^2|\tilde{\mathbf{y}})$ of $\boldsymbol{\beta}$ and $\sigma^2 > 0$ under the prior

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

is unimodal for typical choices of $\pi(\sigma^2)$ and any choice of $\lambda \geq 0$, in the sense that for every $c > 0$, the upper level set

$$\{(\boldsymbol{\beta}, \sigma^2)|\pi(\boldsymbol{\beta}, \sigma^2|\tilde{\mathbf{y}}) > c, \sigma^2 > 0\}$$

is connected. The log posterior is

$$\ln(\pi(\sigma^2)) - \frac{n+p-1}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\|\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}\|_2^2 - \lambda\|\boldsymbol{\beta}\|_1/\sqrt{\sigma^2} \tag{A.1}$$

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

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

is continuous with a continuous inverse when $0 < \sigma^2 < \infty$, and thus unimodality in the original coordinates is equivalent to unimodality in these transformed coordinates. In the transformed coordinates, (A.1) becomes

$$\ln(\pi(1/\rho^2)) + (n+p-1)\ln(\rho) - \frac{1}{2}\|\rho\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\phi}\|_2^2 - \lambda\|\boldsymbol{\phi}\|_1. \tag{A.2}$$

The second and fourth terms are clearly concave in $(\boldsymbol{\phi}, \rho)$, and the third term is a concave quadratic in $(\boldsymbol{\phi}, \rho)$. Thus (A.2) is concave, and hence the posterior is unimodal, provided that $\ln(\pi(1/\rho^2))$ is concave. The function $\ln(\pi(1/\rho^2))$ is concave if, for instance, σ^2 has the scale-invariant prior $1/\sigma^2$ or any inverse-gamma prior.

APPENDIX B: BIMODALITY UNDER THE UNCONDITIONAL PRIOR

If instead of (3), the unconditional Laplace prior

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

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

$$\pi(\sigma^2)(\sigma^2)^{-(n-1)/2} \times \exp\left\{-\frac{1}{2\sigma^2}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^\top(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) - \lambda\sum_{j=1}^{p}|\beta_j|\right\}. \tag{B.2}$$

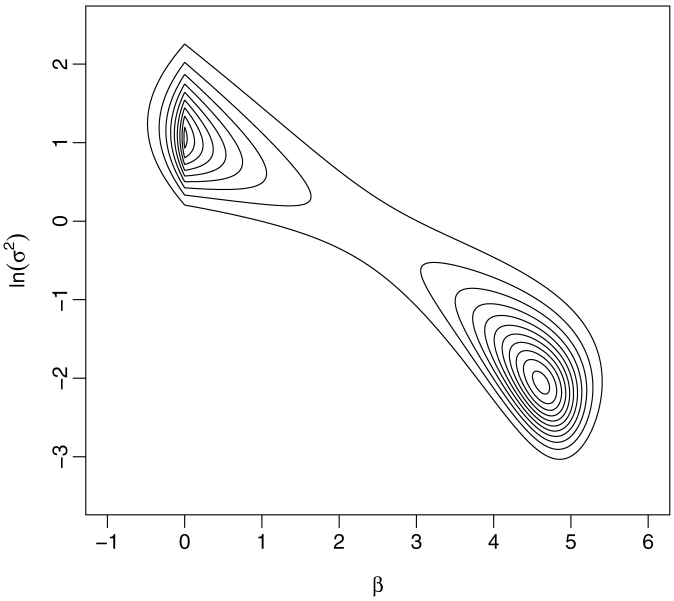

Figure B.1. Contour plot of an artificially generated posterior density of $(\beta, \ln(\sigma^2))$ of the form (B.2) that manifests bimodality. The logarithm of σ^2 is used only because it provides better visual scaling.

(Note that 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.)

In Appendix A it was shown that using the conditional prior (3) leads to a unimodal posterior for any choice of $\lambda \geq 0$, for many reasonable choices of $\pi(\sigma^2)$. In contrast, posteriors of the form (B.2) can easily have more than one mode. For example, Figure B.1 shows the contours of a bimodal joint density of β and $\ln(\sigma^2)$ 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 $p = 1$, $n = 10$, $\mathbf{X}^\top\mathbf{X} = 1$, $\mathbf{X}^\top\tilde{\mathbf{y}} = 5$, $\tilde{\mathbf{y}}^\top\tilde{\mathbf{y}} = 26$, and $\lambda = 3$. The mode on the lower right is near the least squares solution $\beta = 5, \sigma^2 = 1/8$, whereas that on the upper left is near the values $\beta = 0$ and $\sigma^2 = 26/9$, which would be estimated for the selected model in which β is set to 0. 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 posterior density of β alone is also bimodal (results not shown). We have not thoroughly investigated the case where $p > 1$, but perhaps an example of a joint posterior with more than two modes might be possible in that case, with the modes corresponding to setting various subsets of the elements of $\boldsymbol{\beta}$ to 0.

APPENDIX C: EMPIRICAL BAYES UPDATE

The Monte Carlo EM method for empirical Bayes estimation of hyperparameters proposed by Casella (2001) essentially treats the parameters as "missing data" and then uses the EM algorithm to iteratively approximate the hyperparameters, substituting Monte Carlo estimates for any expected values that cannot be computed explicitly. For the Bayesian Lasso, the Gibbs sampler is used to estimate the expected values.

The hierarchy of Section 2 with the conjugate inverse-gamma prior

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

yields the "complete-data" log-likelihood

$$-((n+p-1)/2+a+1)\ln(\sigma^2)-\frac{1}{\sigma^2}(\|\tilde{\mathbf{y}}-\mathbf{X}\boldsymbol{\beta}\|_2^2/2+\gamma)$$
$$-\frac{1}{2}\sum_{j=1}^{p}\ln(\tau_j^2)-\frac{1}{2}\sum_{j=1}^{p}\frac{\beta_j^2}{\sigma^2\tau_j^2}+p\ln(\lambda^2)-\frac{\lambda^2}{2}\sum_{j=1}^{p}\tau_j^2$$

(after dropping constant terms not involving λ). The scale-invariant prior $1/\sigma^2$, formally obtained by taking $a=0$ and $\gamma=0$, alternatively can be used without invalidating any of the following discussion.

Iteration k makes use of the estimate $\lambda^{(k-1)}$ from the previous iteration (or the initial value, if $k=1$). Ideally, the E-step involves taking the expected value of the log-likelihood, conditional on $\tilde{\mathbf{y}}$ and under $\lambda^{(k-1)}$, to get

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

(in the usual notation associated with EM). The M-step maximizes this expression over λ to produce the next estimate, $\lambda^{(k)}$. In this case there is a simple analytical solution,

$$\lambda^{(k)}=\sqrt{\frac{2p}{\sum_{j=1}^{p}E_{\lambda^{(k-1)}}[\tau_j^2|\tilde{\mathbf{y}}]}}.$$

Of course, the conditional expectations are just the posterior expectations under the hyperparameter $\lambda^{(k-1)}$, and thus they can be estimated using the sample averages from a run of the Gibbs sampler described in Section 2.

[Received October 2007. Revised January 2008.]

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.

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

Bae, K., and Mallick, B. K. (2004), "Gene Selection Using a Two-Level Hierarchical Bayesian Model," *Bioinformatics*, 20, 3423–3430.

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*, New York: Marcel Dekker.

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

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

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," *Technometrics*, 35, 109–135.

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

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*, New York: Springer-Verlag.

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.

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

——— (2000b), "On the LASSO and Its Dual," *Journal of Computational and Graphical Statistics*, 9, 319–337.

Rosset, S., and Zhu, J. (2004), Discussion of "Least Angle Regression," by B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *The Annals of Statistics*, 32, 469–475.

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.