Bayesian Analysis (2010) **5**, Number 2, pp. 369–412

Penalized Regression, Standard Errors, and Bayesian Lassos

Minjung Kyung*, Jeff Gill†, Malay Ghosh‡ and George Casella§

Abstract. Penalized regression methods for simultaneous variable selection and coefficient estimation, especially those based on the lasso of Tibshirani (1996), have received a great deal of attention in recent years, mostly through frequentist models. Properties such as consistency have been studied, and are achieved by different lasso variations. Here we look at a fully Bayesian formulation of the problem, which is flexible enough to encompass most versions of the lasso that have been previously considered. The advantages of the hierarchical Bayesian formulations are many. In addition to the usual ease-of-interpretation of hierarchical models, the Bayesian formulation produces valid standard errors (which can be problematic for the frequentist lasso), and is based on a geometrically ergodic Markov chain. We compare the performance of the Bayesian lassos to their frequentist counterparts using simulations, data sets that previous lasso papers have used, and a difficult modeling problem for predicting the collapse of governments around the world. In terms of prediction mean squared error, the Bayesian lasso performance is similar to and, in some cases, better than, the frequentist lasso.

Keywords: Hierarchical Models, Gibbs Sampling, Geometric Ergodicity, Variable Selection

1 Introduction

A large amount of effort has gone into the development of penalized regression methods for simultaneous variable selection and coefficient estimation. In practice, even if the sample size is small, a large number of predictors is typically included to mitigate modeling biases. With such a large number of predictors, there might exist problems among explanatory variables, in particular, there could be a problem with multicollinearity. Also, with a large number of predictors there is often a desire to select a smaller subset that not only fits as well as the full set of variables, but also contains the more important predictors. Such concerns have led to prominent development of least squares (LS) regression methods with various penalties to discover relevant explanatory factors and to get higher prediction accuracy in linear regression.

We consider a linear regression model with n observations on a dependent variable

*Department of Statistics, University of Florida, Gainesville, FL, http://www.stat.ufl.edu/~kyung/

†Department of Political Science, Washington University in St. Louis, St Louis, MO, mailto:jgill@wustl.edu

‡Department of Statistics, University of Florida, Gainesville, FL, http://www.stat.ufl.edu/~ghoshm/

§Department of Statistics, University of Florida, Gainesville, FL, http://www.stat.ufl.edu/~casella/

 DOI:10.1214/10-BA607

Y and p predictors:

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

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ is the $n \times p$ matrix of *standardized* regressors, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and $\boldsymbol{\epsilon} \sim N_n\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right)$. Penalized regression approaches have been used in cases where $p < n$, and in the ever-more-common case with $p \gg n$. In the former case, penalized regression, and its accompanying variable selection features, can lead to finding smaller groups of variables with good prediction accuracy. If $p \gg n$, ordinary least-squares regression (OLS), which minimizes the residual sum of squares $\text{RSS} = (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})' (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})$ where $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \bar{y}\mathbf{1}_n$, will yield an estimator that is not unique since $\boldsymbol{X}$ is not of full rank. Moreover, the variances will be artificially large. Here, again, penalized regression approaches can guide us to good subsets of predictors.

We are going to address both of these cases through a Bayesian approach to penalized regression. However, we are not specifically interested in variable selection, but rather in accurate prediction and determining which predictors are meaningful. That is, typically substantive subject-matter conclusions are based on the interpretation of coefficient estimates, and uncovering meaningful coefficients is a desired outcome of any model fitting approach.

1.1 Penalized Regression

To achieve better prediction in the face of multicollinearity, Hoerl and Kennard (1970) proposed ridge regression, which minimizes RSS subject to $\sum_{i=1}^{p} |\beta_j|^2 \leq t$ (L_2 norm). For the problem of multicollinearity, ridge regression improves the prediction performance, but it cannot produce a model with only the relevant predictors. Frank and Friedman (1993) introduced bridge regression which, subject to $\sum_{i=1}^{p} |\beta_j|^\gamma \leq t$ with $\gamma \geq 0$, minimizes RSS. The estimator from bridge regression is not explicit, but Frank and Friedman argued that the optimal choice of the parameter γ yields reasonable predictors. This is because it controls the degree of preference for the true coefficient $\boldsymbol{\beta}$ to align with the original variable axis directions in the predictor space. Fan and Li (2001) proposed the Smoothly Clipped Absolute Deviation (SCAD) penalty for penalized least squares to reduce bias and satisfy certain conditions to yield continuous solutions. Also, they derived the fixed tuning parameter asymptotic distribution of the estimator and showed that the estimator satisfies the oracle property (consistent model selection).

Among methods that do both continuous shrinkage and variable selection, a promising technique called the *Least Absolute Shrinkage and Selection Operator* (lasso) was proposed by Tibshirani (1996). The lasso is a penalized least squares procedure that minimizes RSS subject to the non-differentiable constraint expressed in terms of the L_1 norm of the coefficients. That is, the lasso estimator is given by

$$\hat{\boldsymbol{\beta}}_{\text{L}} = \arg\min_{\boldsymbol{\beta}} (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})' (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2}$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \bar{y}\mathbf{1}_n$, $\boldsymbol{X}$ is the matrix of standardized regressors and $\lambda \geq 0$ is a tuning parameter. Fu (1998) compared the lasso, ridge regression and bridge regression using

the criteria of prediction performance. He argued that because of the nonlinearity of the bridge operator, the bridge model does not always perform the best in estimation and prediction compared to the other shrinkage models - the lasso and ridge regression. Knight and Fu (2000) have shown consistency for lasso type estimators (generally bridge estimators) with fixed p under some regularity conditions on the design. They obtained the asymptotic normal distribution with a fixed true parameter $\boldsymbol{\beta}$ and local asymptotics, that is, when the true parameter is small but nonzero in finite samples. Also, they derived asymptotic properties of lasso type estimators under nearly singular design matrices.

For the computation of the lasso, Osborne *et al.* (2000a) proposed two algorithms. A compact descent algorithm was derived to solve the selection problem for a particular value of the tuning parameter, then a homotopy method for the tuning parameter was developed to completely describe the possible selection. Later, Efron *et al.* (2004) proposed Least Angle Regression Selection (LARS) for a model selection algorithm. They showed that with a simple modification, the LARS algorithm implements the lasso, and one of the advantages of LARS is the short computation time compared to other methods.

For efficiently selecting the optimal fit, the effective degrees of freedom of the lasso were studied by Efron *et al.* (2004). They discovered that the size of the active set (the indices corresponding to covariates to be chosen) can be used as a measure of the degrees of freedom, which changes, not necessarily monotonically, along the solution paths of LARS. Zou *et al.* (2007) improved this and showed that the number of nonzero coefficients is an unbiased estimate for degrees of freedom of the lasso. In addition, Zou *et al.* (2007) showed that the unbiased estimator is asymptotically consistent, thus various model selection criteria can be used with the LARS algorithm for the optimal lasso fit.

1.2 Generalizations of the Lasso

The lasso has shown excellent performance in many situations, however it has some limitations. As Tibshirani (1996) argued, if there exists multicollinearity among predictors, ridge regression dominates the lasso in prediction performance. Also, in the $p > n$ case, the lasso cannot select more than n variables because it is the solution to a convex optimization problem. If there is a meaningful ordering of the features (such as specification of consecutive predictors), the lasso ignores it. Furthermore, if there is a group of variables among which the pairwise correlations are very high and if we consider the problem of selecting grouped variables for accurate prediction, the lasso tends to select individual variables from the group or the grouped variables (for example, dummy variables).

To compensate the ordering limitations of the lasso, Tibshirani *et al.* (2005) introduced the fused lasso. The fused lasso penalizes the L_1-norm of both the coefficients

and their differences:

$$\hat{\boldsymbol{\beta}}_{\mathrm{F}} = \arg\min_{\beta} (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})' (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|,$$

where λ_1 and λ_2 are tuning parameters. They provided the theoretical asymptotic limiting distribution and a degrees of freedom estimator.

For grouped variables, Yuan and Lin (2006) proposed a generalized lasso that is called the group lasso. The group lasso estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\mathrm{G}} = \arg\min_{\beta} \left(\tilde{\boldsymbol{y}} - \sum_{k=1}^{K} \boldsymbol{X}_k \boldsymbol{\beta}_k \right)' \left(\tilde{\boldsymbol{y}} - \sum_{k=1}^{K} \boldsymbol{X}_k \boldsymbol{\beta}_k \right) + \lambda \sum_{k=1}^{K} ||\boldsymbol{\beta}_k||_{G_k},$$

where K is the number of groups, $\boldsymbol{\beta}_k$ is the vector of βs in group k, the G_k's are given positive definite matrices and $||\boldsymbol{\beta}||_G = (\boldsymbol{\beta}' G \boldsymbol{\beta})^{1/2}$. In general, $G_k = I_{m_k}$, where m_k is the size of the coefficient vector in group k. This penalty function is intermediate between the L_1 penalty and the L_2 penalty. Yuan and Lin (2006) argued that it does variable selection at the group level and is invariant under orthogonal transformations.

Zou and Hastie (2005) proposed the elastic net, a new regularization of the lasso, for an unknown group of variables and for multicollinear predictors. The elastic net estimator can be expressed as

$$\hat{\boldsymbol{\beta}}_{\mathrm{EN}} = \arg\min_{\beta} (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})' (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2,$$

where λ_1 and λ_2 are tuning parameters. The elastic net estimator can be interpreted as a stabilized version of the lasso. Thus, it enjoys a sparsity of representation and encourages a grouping effect. Also, it is useful when $p \gg n$. They provided the algorithm LARS-EN to solve the elastic net efficiently based on LARS of Efron *et al.* (2004).

Fan and Li (2001) showed that the lasso can perform automatic variable selection but it produces biased estimates for the larger coefficients. Thus, they argued that the oracle properties do not hold for the lasso. To obtain the oracle property, Zou (2006) introduced the adaptive lasso estimator as

$$\hat{\boldsymbol{\beta}}_{\mathrm{AL}} = \arg\min_{\beta} (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})' (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \hat{w}_j |\beta_j|,$$

with the weight vector $\hat{\boldsymbol{w}} = 1/|\hat{\boldsymbol{\beta}}|^{\gamma}$ where $\hat{\boldsymbol{\beta}}$ is a $\sqrt{n}$ consistent estimator such as $\hat{\boldsymbol{\beta}}(OLS)$ and $\gamma > 0$. The adaptive lasso enjoys the oracle property and it leads to a near-minimax-optimal estimator.

Many other lasso variations exist that we will not directly address here. For example, Kim *et al.* (2006) proposed an extension of the group lasso, called blockwise sparse regression (BSR), and studied it for logistic regression models, Poisson regression and

the proportional hazards model. Park and Hastie (2007) introduced a path following algorithm for L_1 regularized generalized linear models, and provided computational solutions along the entire regularization path by using the predictor-corrector method of convex optimization. Recently, Meier *et al.* (2008) presented algorithms which are suitable for very high dimensional problems for solving the convex optimization problems, and showed that the group lasso estimator for logistic regression is statistically consistent with a sparse true underlying structure even if $p \gg n$.

1.3 The Bayesian Lasso

Tibshirani (1996) noted that with the L_1 penalty term in (2), the lasso estimates could be interpreted as the Bayes posterior mode under independent Laplace (double-exponential) priors for the β_js. One of the advantages of the Laplace distribution is that it can be expressed as a scale mixture of normal distributions with independent exponentially distributed variances (Andrews and Mallows, 1974). This connection encouraged a few authors to use Laplace priors in a hierarchical Bayesian approach. Figueiredo (2003) used the Laplace prior to obtain sparsity in supervised learning using an EM algorithm. In the Bayesian setting, the Laplace prior suggests the hierarchical representation of the full model. Bae and Mallick (2004) adopted a Markov chain Monte Carlo (MCMC) based computation with the hierarchical representation of the Laplace prior in a gene selection problem.

Recently, Park and Casella (2008) suggested Gibbs sampling for the lasso with the Laplace prior in the hierarchical model. Specifically, they considered a fully Bayesian analysis using a conditional Laplace prior specification of the form

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

and the noninformative scale-invariant marginal prior $\pi(\sigma^2) = 1/\sigma^2$. They pointed out that conditioning on σ^2 is important because it guarantees a unimodal full posterior. Lack of unimodality slows convergence of the Gibbs sampler and makes point estimates less meaningful.

Other Bayesian methods with a Laplace prior have been developed for the lasso. Yuan and Lin (2005) proposed an empirical Bayes method for variable selection and estimation in linear regression models using approximations to posterior model probabilities that are based on orthogonal designs. Their method is based on a hierarchical Bayesian formulation with Laplace prior and showed that the empirical Bayes estimator is closely related to the lasso estimator. Genkin *et al.* (2007) presented a simple Bayesian logistic lasso with Laplace prior to avoid overfitting and produce sparse predictive models for text data.

1.4 Model Selection and Estimation

In traditional statistical data analysis, after the optimal model is found, the subsequent inference is considered. It is well known that AIC and BIC possess different asymptotic optimality, that is, BIC is consistent in selecting the true model and AIC is minimax-rate optimal for estimating the regression function.

In adaptive model selection such as the lasso or bridge regression, in contrast to AIC and BIC, the penalty term is data-dependent. Some theoretical and empirical results have been obtained in support of adaptive model selection. Although the results of Yang (2005) suggest that there still might be a dichotomy, this may be more of a pathology than a statistical concern (Casella and Consonni 2009). However, the results of Leeb and Pötscher (2005) also suggest that the distributional properties of post-model-selection estimators are quite intricate, and are not properly captured by the usual pointwise large-sample analysis using the maximal scaled mean squared error (MSE) and the maximal absolute bias.

All this leads us to the fact that we may not have reliable standard errors for the zero coefficients in penalized regression models. Both the approximate covariance matrix and the bootstrap methods might not improve the problem. If prediction is the ultimate goal, and the model-selector/predictor is being used as a predictor, we thus become more interested in the standard errors of our predictions (and less interested in selecting the optimal model).

Of course, in the context of a procedure like the lasso, the model selector is, in fact, nothing more than a point estimate. In a Bayesian analysis using MCMC, we have a sample from the posterior distribution which we can summarize in any way we please. For example, we could report the posterior density, or the mean and standard deviation. Alternatively, we could use the posterior mode as a point estimator and, in the hierarchies that we will present, this mode is exactly the lasso point estimate. Moreover, as we can validly summarize the spread of the posterior, we have a valid measure of variability.

1.5 Summary

In this paper, we use hierarchical models and Gibbs sampling to get estimators and valid Bayesian standard errors for generalized lasso estimators. In Section 2 we look at standard errors of the lasso, seeing that the bootstrap is not a straightforward option of putting errors on the coefficient estimates. The hierarchical models resulting in group, fused and elastic net are developed in Section 3, and in Section 4 we show that the Markov chains are geometrically ergodic. We compare the Bayesian lassos to their frequentist counterparts in Section 5, using simulations and data sets that previous lasso papers have used. There is a discussion in Section 6, and technical details are in the Appendices.

2 Standard Errors of the Lasso

For inference using the lasso estimator, various standard error estimators have been proposed. Fan and Li (2001) presented the sandwich formula in the likelihood setting as an estimator for the covariance of the estimates. Building on their work, Zou (2006) derived a sandwich formula for the adaptive lasso. However, all of the above approximate covariance matrices give an estimated variance 0 for predictors with $\hat{\beta}_j = 0$.

Osborne *et al.* (2000b) derived an estimate of the covariance matrix of lasso estimators that yields a positive standard error for all coefficient estimates. They pointed out that, nevertheless, since the distribution of individual lasso coefficient estimates will typically have a concentration of probability at zero, the estimates may be far from normally distributed. More recently, Pötscher and Leeb (2007) studied the distribution of penalized likelihood estimators (lasso, SCAD and thresholding) and showed that the finite sample distribution of soft thresholding (lasso) is a mixture of a singular normal distribution and of an absolutely continuous part, which is the sum of two normal densities, each with a truncated tail. The truncation is at the location of the point mass at 0. Thus, the suggested estimators might not yield reasonable estimates for the covariance matrix of $\boldsymbol{\beta}$.

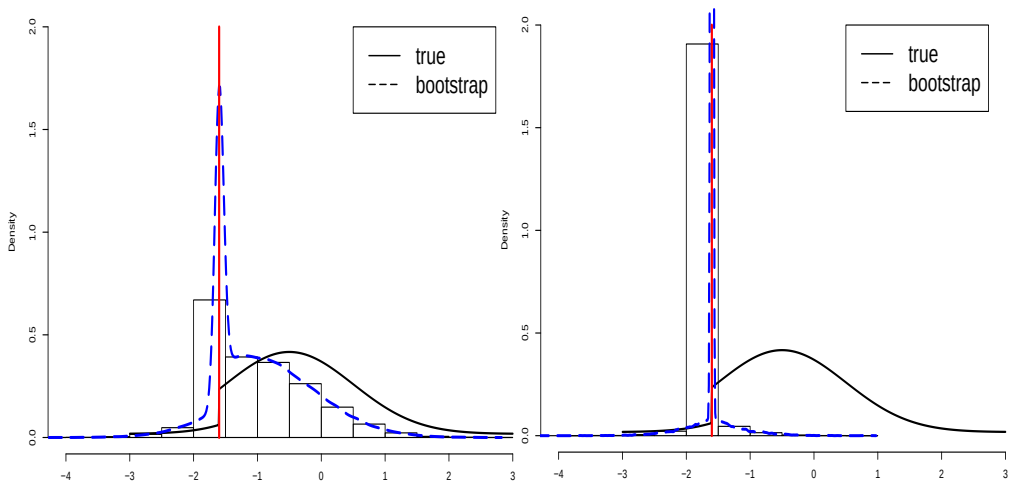
Instead of the approximate covariance matrix, as Tibshirani (1996) suggested, an alternative method of computing standard errors is the bootstrap. To study the behavior of the bootstrap estimates of the standard error of $\hat{\boldsymbol{\beta}}$, we conducted a simple simulation based on the finite sample lasso distribution of Pötscher and Leeb (2007). We set a univariate $\beta = 0.16$ and generated a sample of size $n = 100$ from a normal distribution with mean β and variance 1. We estimate the lasso in this setting and use bootstrapping for re-sampling of the lasso estimate. We also consider two settings for the parameter of penalty term, λ: λ is fixed at 0.05 and is estimated with the generalized cross-validation (GCV), $\hat{\lambda} = 0.19$. The histograms of the bootstrapped estimates are given in Figure 1.

In Figure 1, the true density is based on the finite sample distribution of the lasso from Pötscher and Leeb (2007). We observe that there are big peaks at the discontinuity points $-n^{1/2}\beta = -1.6$ and the distributions of the two histograms are highly non-normal. If we estimate λ by using GCV, most of bootstrap samples are in the area near around $-n^{1/2}\beta = -1.6$.

Samworth (2003) looked into the relation of pointwise asymptotics of consistency of bootstrap estimators and their finite sample behavior. He noted that the asymptotics can mask the finite-sample behavior, and inconsistent bootstrap estimators may in fact perform better than their consistent counterparts. He illustrated this point with reference to the parametric bootstrap and the Hodges-Lehmann super-efficient estimator (for details see Lehmann and Casella, 1998, Section 6.2) and Stein estimators. However, he argued that the inconsistent bootstrap of the super-efficient Hodges-Lehmann estimator, and of the Stein estimator, can only be improved in a very small neighborhood, and the improvements come at the expense of considerably worse performance outside this neighborhood.

Knight and Fu (2000) argued that the bootstrap may have some problems in esti-

Figure 1: Histogram of the bootstrapped estimates of β where a sample of size 100 is drawn with true $\beta = 0.16$, and $100,000$ bootstrap samples were used to estimate the distribution. The dashed lines are density estimators fitted to the bootstrapped values and the solid lines are the exact distribution of the lasso estimate, which has a spike at zero represented by the red line. For the exact distribution the left panel uses a fixed $\lambda = 0.05$, and the right panel fixes $\lambda = 0.19$, the GCV estimate.



Fixed Value of λ. λ Estimated by GCV.

mating the sampling distribution of bridge estimators for $\gamma \leq 1$ when the true parameter values are either exactly 0 or close to 0; in such cases, bootstrap sampling introduces a bias that does not vanish asymptotically. Also, they argued that the suggested standard error estimates for the bridge parameter estimates is nontrivial especially when $\gamma \leq 1$. (Bridge regression with $\gamma = 1$ is the lasso.). Thus, bootstrap estimates of the standard error may cause problems in practice. Problems were also uncovered by Leeb and Pötscher(2006) who detailed problems in estimating measures of precision of shrinkage-type estimators, such as James-Stein, lasso-type, and Hodges' super-efficient estimators. Lastly, Beran (1982) discussed the bootstrap estimate of the distribution function of statistics whose asymptotic distribution is normal. He showed that under certain assumptions the bootstrap estimate is asymptotically minimax. However, he proved that for a superefficient estimator, the bootstrap estimates are not consistent if the true parameter is fixed at the point of superefficiency.

From these findings, we see that the bootstrap estimates of the standard error of the lasso estimator might be unstable and not perform well. In fact, we can use the results of Knight and Fu (2000) together with those of Beran (1982) to formally establish that the bootstrap estimates based on the lasso are not consistent if the true $\beta = 0$.

Table 1: Penalty terms for the four types of lasso in model (4).

Model	λ_1	λ_2	$h_1(\boldsymbol{\beta})$	$h_2(\boldsymbol{\beta})$
lasso	λ	0	$\sum_{j=1}^{p} \lvert\beta_j\rvert$	0
Group lasso	λ	0	$\sum_{k=1}^{K} \lvert\lvert\boldsymbol{\beta}\rvert\rvert_G$	0
			positive definite G_k's and $\lvert\lvert\boldsymbol{\beta}\rvert\rvert_G = (\boldsymbol{\beta}'G\boldsymbol{\beta})^{1/2}$	
Fused lasso	λ_1	λ_2	$\sum_{j=1}^{p} \lvert\beta_j\rvert$	$\sum_{j=2}^{p} \lvert\beta_j - \beta_{j-1}\rvert$
Elastic net	λ_1	λ_2	$\sum_{j=1}^{p} \lvert\beta_j\rvert$	$\sum_{j=2}^{p} \lvert\beta_j\rvert^2$

Proposition 2.1. *Under model (1), the bootstrap standard errors of $\hat{\beta}_j$ are inconsistent if $\beta_j = 0$.*

Proof: Details in Appendix 6.

Thus, the bootstrap does not allow us to attach valid standard error estimates to the values of the lasso that are shrunk to zero, in the sense that these estimators are inconsistent. In this sense we still do not have a general, statistically valid method of obtaining standard errors of lasso estimates.

3 Hierarchical Models and Gibbs Samplers

The hierarchical representation of the full model, with the Laplace prior written as a scale mixture of normals with an exponential mixing density, was suggested by Park and Casella (2008). In this paper, we extend this model to a more general form that can represent the group lasso, the fused lasso, and the elastic net.

We first exhibit the hierarchical models that lead to the various types of lasso, and indicate a general strategy that may handle the new lassos that are sure to come. Technical calculations are deferred to Appendix 6. In Section 3.2 we address the problem of estimating the ubiquitous tuning parameters.

A general version of the lasso model can be expressed as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_1 h_1(\boldsymbol{\beta}) + \lambda_2 h_2(\boldsymbol{\beta}) \qquad \lambda_1,\ \lambda_2 > 0, \tag{4}$$

where the specific choices of $h_1(\boldsymbol{\beta})$ and $h_2(\boldsymbol{\beta})$ are given in Table 1. Other penalized regression models can be expressed similarly. However, in this paper, we consider the above four models.

The Bayesian formulation of the original lasso, as given in Park and Casella (2008),

is given by the following hierarchical model.

$$\begin{aligned}
\boldsymbol{y} \mid \mu, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 &\sim N_n\left(\mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I_n}\right) \\
\boldsymbol{\beta} \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2 &\sim N_p\left(\mathbf{0}_p, \sigma^2 \boldsymbol{D}_\tau\right), \qquad \boldsymbol{D}_\tau = \operatorname{diag}(\tau_1^2, \ldots, \tau_p^2) \\
\tau_1^2, \ldots, \tau_p^2 &\sim \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2}\, d\tau_j^2, \qquad \tau_1^2, \ldots, \tau_p^2 > 0 \\
\sigma^2 &\sim \pi(\sigma^2)\, d\sigma^2 \qquad \sigma^2 > 0,
\end{aligned} \tag{5}$$

where we standardize the $\boldsymbol{X}$ matrix. The parameter μ may be given an independent, flat prior, and posterior propriety can be maintained. After integrating out $\tau_1^2, \ldots, \tau_p^2$, the conditional prior on $\boldsymbol{\beta}$ has the desired form (3). Any inverted gamma prior for σ^2 would maintain conjugacy, and here we will use the limiting improper prior $\pi(\sigma^2) = 1/\sigma^2$, which also maintains propriety. Also note that we are using a conditional version of the Laplace distribution, that is, the resulting prior on $\boldsymbol{\beta}$ is a Laplace distribution with mean 0 and variance $\sigma^2\lambda^{-2}$. This variation assures unimodality of the posterior (Park and Casella 2008) while the unconditional version (without σ^2) does not.

3.1 Hierarchical Models

We now turn to the other types of lasso, and show how to represent them as a conjugate Bayesian hierarchy. For each model we describe the unconditional prior on $\boldsymbol{\beta}$, and how to represent it as a normal mixture with $\boldsymbol{\beta} \sim N(0, \Sigma_\beta)$, where Σ_β is parametrized with τ_is. We only need to specify the covariance matrix of $\boldsymbol{\beta}$, denote by Σ_β, and the distribution of the τ_i.

For the lasso, group lasso, and fused lasso, the covariance matrix Σ_β contains only τ_is and no λs. This not only results in $\boldsymbol{\beta}$ and λ being conditionally independent, it is important for the Gibbs sampler as it results in gamma conditionals for the λs. This is not the case for the elastic net; to accommodate the squared term we will need to put λ_2 in Σ_β. However, because of the normal form, this will also cause no problem for the Gibbs sampler. In this section we will only give the forms of the models; details on posterior distributions and Gibbs sampling are left to Appendix 6

Hierarchical Group Lasso In penalized linear regression with the group lasso, the conditional prior of $\boldsymbol{\beta}|\sigma^2$ can be expressed as

$$\pi(\boldsymbol{\beta}|\sigma^2) \propto \exp\left(-\frac{\lambda}{\sigma}\sum_{k=1}^{K} ||\boldsymbol{\beta}_{G_k}||\right).$$

This prior can be attained as a gamma mixture of normals, leading to the group lasso hierarchy

$$\begin{aligned} \boldsymbol{y} \mid \mu, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 &\sim N_n\left(\mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I_n}\right) \\ \boldsymbol{\beta}_{G_k} | \sigma^2, \tau_k^2 &\overset{ind}{\sim} N_{m_k}\left(\mathbf{0}, \sigma^2 \tau_k^2 \boldsymbol{I}_{m_k}\right) \\ \tau_k^2 &\overset{ind}{\sim} \text{gamma}\left(\frac{m_k+1}{2}, \frac{\lambda^2}{2}\right) \quad \text{for } k = 1, \ldots, K \end{aligned} \tag{6}$$

where m_k is the dimension of G_k, the grouping matrix. Note that for the group lasso we need to use a gamma prior on the τ_i, but the calculations are quite similar to those of the ordinary lasso. Details of the model are discussed in Appendix 6.

Hierarchical Fused Lasso In penalized linear regression with the fused lasso, the conditional prior of $\boldsymbol{\beta}|\sigma^2$ can be expressed as

$$\pi(\boldsymbol{\beta}|\sigma^2) \propto \exp\left(-\frac{\lambda_1}{\sigma}\sum_{j=1}^{p}|\beta_j| - \frac{\lambda_2}{\sigma}\sum_{j=1}^{p-1}|\beta_{j+1} - \beta_j|\right). \tag{7}$$

This prior can also be obtained as a gamma mixture of normals, leading to the hierarchical model

$$\begin{aligned} \boldsymbol{y} \mid \mu, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 &\sim N_n\left(\mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I_n}\right) \\ \boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2, \omega_1^2, \ldots, \omega_{p-1}^2 &\sim N_p\left(\mathbf{0}, \sigma^2 \Sigma_\beta\right) \\ \tau_1^2, \ldots, \tau_p^2 &\sim \prod_{j=1}^{p} \frac{\lambda_1^2}{2} e^{-\lambda_1 \tau_j^2/2}\, d\tau_j^2, \qquad \tau_1^2, \ldots, \tau_p^2 > 0 \\ \omega_1^2, \ldots, \omega_{p-1}^2 &\sim \prod_{j=1}^{p-1} \frac{\lambda_2^2}{2} e^{-\lambda_2 \omega_j^2/2}\, d\omega_j^2, \qquad \omega_1^2, \ldots, \omega_{p-1}^2 > 0 \end{aligned} \tag{8}$$

where $\tau_1^2, \ldots, \tau_p^2, \omega_1^2 \ldots, \omega_{p-1}^2$ are mutually independent, and Σ_β is a tridiagonal matrix with

$$\begin{aligned} \text{Main diagonal} &= \left\{\frac{1}{\tau_i^2} + \frac{1}{\omega_{i-1}^2} + \frac{1}{\omega_i^2}, i = 1, \ldots, p\right\}, \\ \text{Off diagonals} &= \left\{-\frac{1}{\omega_i^2}, i = 1, \ldots, p-1\right\}, \end{aligned}$$

where, for convenience, we define $(1/\omega_0^2) = (1/\omega_p^2) = 0$.

Here for the first time we have correlation in the prior for $\boldsymbol{\beta}$, adding some difficulty to the calculations. Details about the derivation of the fused lasso prior and it Gibbs sampler are discussed in Appendix 6.

Hierarchical Elastic Net In penalized linear regression with the elastic net, the conditional prior of $\boldsymbol{\beta}|\sigma^2$ can be expressed as

$$\pi\left(\boldsymbol{\beta}|\sigma^2\right) \propto \exp\left\{-\frac{\lambda_1}{\sigma}\sum_{j=1}^{p}|\beta_j| - \frac{\lambda_2}{2\sigma^2}\sum_{j=1}^{p}\beta_j^2\right\}. \tag{9}$$

This prior can also be written as a normal mixture of gammas, leading to the hierarchical model

$$\begin{aligned} \boldsymbol{y} \mid \mu, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 &\sim N_n\left(\mu\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2\mathbf{I_n}\right) \\ \boldsymbol{\beta} \mid \sigma^2, \boldsymbol{D}_\tau^* &\sim N_p(\mathbf{0}_p, \sigma^2\boldsymbol{D}_\tau^*), \\ \tau_1^2, \ldots, \tau_p^2 &\sim \prod_{j=1}^{p}\frac{\lambda_1^2}{2}e^{-\lambda_1^2\tau_j^2/2}\,d\tau_j^2, \qquad \tau_1^2, \ldots, \tau_p^2 > 0, \end{aligned} \tag{10}$$

where $\boldsymbol{D}_\tau^*$ is a diagonal matrix with diagonal elements $(\tau_i^{-2} + \lambda_2)^{-1}$, $i = 1\ldots, p$. Note that, in this case, $\boldsymbol{\beta}$ is not conditionally independent of λ_2, as it appears in the covariance matrix. However, as this term comes into the exponent of the posterior as a "normal" component, the Gibbs sampler is still straightforward. Details are in Appendix 6.

3.2 Tuning Parameters

The lassos of the previous section all have tuning parameters λ_1 or λ_2. Typical approaches for estimation of these parameters are cross-validation, generalized cross-validation and ideas based on Stein's unbiased risk estimate (Tibshirani, 1996). In the Bayesian framework, Park and Casella (2008) suggested some alternatives based on empirical Bayes using marginal maximum likelihood, putting λ_1 or λ_2 into the Gibbs sampler with an appropriate hyperprior. In this paper, we use the suggested gamma prior for a proper posterior from Park and Casella (2008), and also for comparison, estimate the tuning parameters with marginal maximum likelihood.

We use gamma priors on λ^2 given by

$$\pi(\lambda^2) \;=\; \frac{\delta^r}{\Gamma(r)}\left(\lambda^2\right)^{r-1}e^{-\delta\lambda^2}, \qquad (r > 0,\ \delta > 0). \tag{11}$$

The prior is put on λ^2 for convenience - because of the way λ enters into the posterior. The one exception is the λ_2 parameter in the elastic net. As that parameter refers to the normal piece of the prior on $\boldsymbol{\beta}$, it is given a typical gamma prior. (It enters as a rate parameter, so a gamma prior on the rate is equivalent to the usual conjugate inverted gamma prior on the variance.)

When the prior (11) is used in the hierarchy, the full conditional distributions of all the λs are gamma distributions, and are listed in Appendix 6.

The tuning parameters can also be estimated through marginal likelihood, which can be implemented with an EM/Gibbs algorithm (Casella, 2001). This approach was

also used by Park and Casella (2008), where more details about the empirical Bayes update of the hyperparameters are discussed. In our setting and notation, iteration t uses the output from the Gibbs sampler with hyperparameter $\lambda^{2(t-1)}$ (the estimate from iteration $t-1$) to update the estimate $\lambda^{2(t)}$. Details are given in Appendix 6.

4 The Lasso Gibbs Sampler

In this section we investigate the convergence properties of the Gibbs sampler from the Bayesian lasso (5) and, by simple extension, the convergence of the Gibbs samplers of the other lassos that we consider. We use the convergence relationship between joint and marginal models, first developed by Liu, Wong and Kong (1994), to adapt the results of Hobert and Geyer (1998) to the hierarchy considered here. Specifically, Hobert and Geyer (1998) proved geometric ergodicity of the block Gibbs sampler in a oneway random effects model with conjugate priors. We adapt their proof to a oneway random effects model that reflects the lasso priors. We then show that the hierarchy (5) can be obtained as a marginal model from this oneway random effects model. Lastly, we adapt and extend the results of Liu, Wong and Kong (1994) to bound the convergence rate of the lasso Gibbs sampler.

If we just consider a oneway model on Y and θ, a regression model such as in (5) can be obtained as a marginal model. Specifically, start with

$$\boldsymbol{y} \sim N\left(\boldsymbol{\theta}, \lambda_e^{-1} I\right), \quad \boldsymbol{\theta} \sim N\left(\mathbf{1}\mu, \Sigma_\theta\right) \tag{12}$$

and make the transformation $\theta \to \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{\beta} + \hat{\boldsymbol{X}}\eta$ where $\boldsymbol{X}'\mathbf{1} = \hat{\boldsymbol{X}}'\mathbf{1} = \hat{\boldsymbol{X}}'\boldsymbol{X} = 0$, and take Σ_θ to be of the form

$$\Sigma_\theta^{-1} = \lambda_0 J + \boldsymbol{X} A \boldsymbol{X}' + \hat{\boldsymbol{X}}\hat{\boldsymbol{X}}', \tag{13}$$

where J is the matrix of 1s, $\boldsymbol{D}_\tau$ is the covariance matrix of the βs in (5), and $A = (\lambda/\sigma^2)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{D}_\tau^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}$. These choices zero out all the difficult cross terms when $||\boldsymbol{y} - \mathbf{1}\mu - \boldsymbol{X}\boldsymbol{\beta} - \hat{\boldsymbol{X}}\eta||^2$ is expanded, and when η is integrated out; what remains is $\boldsymbol{y} \sim N_n\left(\mu\mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I_n\right)$, $\boldsymbol{\beta} \sim N_p\left(\mathbf{0}_p, \frac{\sigma^2}{\lambda}\boldsymbol{D}_\tau\right)$. Thus, the lasso is a marginal model when starting with a oneway random effects model.

Liu, Wong and Kong (1994) compare convergence rates of the following two scans

1. $x_1|y$, $y|x_1$
2. $(x_1, x_2)|y$, $y|(x_1, x_2)$,

where the first scan is based on the marginal distribution of (x_1, y), when x_2 has been integrated out. Note that, from the above discussion, this is exactly the situation that will occur when we apply a Gibbs sampler to the models in Section 3. Taking $(x_1, x_2) = \boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\eta})$, and y the vector of variances, scan (1) is in the form of the Gibbs sampler for the lasso, while scan (2) is the Gibbs sampler of the oneway model.

Liu, Wong and Kong (1994) (Theorem 5.1) state that the operator norm of scan (1) is less than or equal to the operator norm of scan (2), implying that scan (1) is better. However, we can say a bit more. For reversible chains, as are (1) and (2), the chain is geometrically ergodic if and only if the operator norm is less than one (Roberts and Rosenthal, 1998). Thus, if scan (2) is geometrically ergodic, then so is scan (1). We can further bound the total variation norm, where we recall that the total variation norm between two distributions P and Q is

$$||P(\cdot) - Q(\cdot)|| = \sup_A |P(A) - Q(A)| = \frac{1}{2}\int_{\mathcal{X}} |p(x) - q(x)|dx.$$

Denote the transition kernel of scan (2) by $K((x_1, x_2, y), \cdot)$ with stationary distribution $\pi(x_1, x_2, y)$. Then scan (1) has transition kernel $K((x_1, y), \cdot) = \int_{\mathcal{X}_2} K((x_1, x_2, y), \cdot)\, dx_2$ with stationary distribution $\pi(x_1, y) = \int \pi(x_1, x_2, y)dx_2$.

Proposition 4.1. *The n-step transition kernels of the scans* (1) *and* (2) *have the relationship*

$$||K^n((x_1, y), \cdot) - \pi(\cdot)|| \leq \int_{\mathcal{X}_2} ||K^n((x_1, x_2, y), \cdot) - \pi(\cdot)||f(x_2|x_1)dx_2,$$

where $f(x_2|x_1)$ *is the conditional distribution of* $x_2|x_1$.

Thus, the total variation norm of scan (1) is bounded by integrating the total variation norm of scan (2) over all possible initial values of the variable x_2, weighting by the conditional distribution of $x_2|x_1$. If scan (2) is geometrically ergodic with bound $\rho^n M(x_1, x_2, y)$, then $\rho^n \int M(x_1, x_2, y)f(x_2|x_1)dx_2$ bounds the total variation norm of scan (1). The proof of the proposition is given in Appendix 6

Finally, for the lasso model (5) we can put together the transformations in (12)-(13), together with Proposition 4.1 to obtain the geometric ergodicity of the lasso. The proof is in Appendix 6.

Proposition 4.2. *For the lasso model (5), the block Gibbs sampler with blocks* $(\boldsymbol{\beta}, \mu)$ *and* $(\lambda_e^{-1}, \boldsymbol{\tau})$ *is geometrically ergodic.*

Thus, the convergence of the Gibbs sampler is expected to be (and is) quite rapid. Moreover, if desired, we also have a Central Limit Theorem for the Monte Carlo estimates.

5 Applications

In this section, we carry out simulations to compare the performance of proposed models and apply models to real data sets for practical application in terms of prediction accuracy. We have based our simulations and data analyses to reflect those that have appeared in other papers that have developed lasso estimators. For the most part, we estimate the λ tuning parameters by using prior distributions and including them in the Gibbs sampler, as opposed to using marginal MLE. This is partly due to the faster speed of the Gibbs sampler, and the fact that the estimates were very close in all examples.

5.1 Simulation

We simulate data from the true model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}, \quad \epsilon_i \sim \text{iid } N(0,1) \text{ for } i = 1, \ldots, n.$$

Five models were considered in the simulations. The first three examples were used in the original lasso paper (Tibshirani, 1996) and in the elastic net paper (Zou and Hastie, 2005), to compare the prediction performance of the lasso and the elastic net systematically. The second and third examples are also appropriate for the fused lasso, while the the fourth and fifth examples create a grouped variable situation.

- **Example 1:** We simulated data sets with $n = 20$ to fit models and $n = 200$ to compare prediction errors of proposed models with eight predictors. We let $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ and $\sigma = 3$. The pairwise correlation between x_i and x_j was set to be $\text{corr}(i, j) = 0.5^{|i-j|}$.

- **Example 2:** Example 2 is the same as Example 1, except that $\beta_j = 0.85$ for all j.

- **Example 3:** We simulated data sets with $n = 100$ to fit models and $n = 400$ to compare prediction errors of proposed models with 40 predictors. We set $\boldsymbol{\beta} = (\mathbf{0}', \mathbf{2}', \mathbf{0}', \mathbf{2}')'$ where $\mathbf{0}$ is a 0 vector with length 10 and $\mathbf{2}$ is similarly a 2 vector with length 10, and $\sigma = 15$; $\text{corr}(i, j) = 0.5$.

- **Example 4:** 15 latent variables $Z_1, ..., Z_{15}$ were first simulated according to a centered multivariate normal distribution with $\text{corr}(i, j) = 0.5^{|i-j|}$. Then Z_i is trichotomized as 0, 1 or 2 if it is smaller than $\Phi^{-1}(1/3)$, larger than $\Phi^{-1}(2/3)$ or in between. We set
  $\boldsymbol{\beta} = (-1.2, 1.8, 0, 0, 0, 0, 0.5, 1, 0, 0, 0, 0, 1, 1, 0, \mathbf{0}')'$ where $\mathbf{0}$ is a 0 vector with length 30 and $\sigma = 3$. We simulated data sets with $n = 50$ to fit models and $n = 400$ to compare prediction errors of proposed models. We treat each Z_i as a group.

- **Example 5:** In this example, both main effects and second-order interactions between Z_1 and Z_2 were considered. Four categorical factors Z_1,Z_2,Z_3 and Z_4 were first generated as in example 4. We set
  $\boldsymbol{\beta} = (2, 3, 0, 2, 3, 0, 0, 0, 0, 0, 0, 0, 2.5, 2, 0, 1.5, 1, 0, 0, 0, 0)'$ and $\sigma = 3$. We simulated data sets with $n = 100$ to fit models and $n = 400$ to compare prediction errors of proposed models. We again treat each Z_i as a group, as well as grouping the interactions.

For each generated data set in the examples, we fit lasso, elastic net, fused lasso and grouped lasso models, as appropriate, using the Gibbs sampler. For the prior density of the parameter λ, we used a gamma distribution with shape parameter $a = 1$ and scale parameter $b = 0.1$, which is relatively flat and results in high posterior probability near the MLE. The Bayesian estimates are posterior means using 10,000 iterations of the Gibbs sampler (after 1,000 burn-in iterations). For the prediction errors, we calculate

Table 2: Median mean squared errors for Examples 1-3 and three methods, based on 50 replications: Bayesian lassos vs. LARS-EN from Zou and Hastie (2005) vs. Gibbs-Mode, standard errors in parentheses.

Method	Models			
		Example 1	Example 2	Example 3
Gibbs-Mean	Lasso	2.61(1.50)	2.17(1.68)	41.51(10.52)
	Elastic Net	6.80(2.42)	2.89(2.02)	22.70(6.20)
	Fused Lasso	———	1.17(1.61)	16.13(4.53)
LARS	Lasso	3.06(0.31)	3.87(0.38)	65.0(2.82)
	Elastic Net	2.51(0.29)	3.16(0.27)	56.6(1.75)
Gibbs-Mode	Lasso	3.13(1.92)	3.76(2.64)	56.69(16.27)

Table 3: Average mean squared errors for Examples 4-5 and three methods, based on 200 replications: Bayesian lassos vs. Group LARS from Yuan and Lin (2006) vs. Gibbs Group LARS, standard errors in parentheses.

Method	Models		
		Example 4	Example 5
Gibbs-Mean	Lasso	1.27(0.28)	0.50(0.06)
	Elastic Net	0.53(0.11)	0.34(0.03)
	Group Lasso	0.37(0.02)	0.11(0.03)
Group LARS	Lasso	1.17(0.47)	0.13(0.05)
	Group Lasso	0.83(0.4)	0.09(0.04)
Gibbs -Mode	Group Lasso	0.59(0.42)	0.14(0.06)

the average of mean squared error for the simulated examples and four methods based on 50 replications.

Table 2 summarizes median mean squared errors (MSEs) in Examples 1-3 based on 50 replications and for comparison, median mean-squared errors from 50 replications of the LARS-Lasso and LARS-EN from Zou and Hastie (2005) are added. The Gibbs-Mean uses the posterior mean as the point estimate. Comparing these estimates to the lasso counterparts, we see that the median MSE of the Bayesian lasso is smaller than LARS-lasso, but the LARS-elastic net does better than the Bayesian elastic net for Example 1. The Bayesian elastic net is based on the naïve elastic net model of Zou and Hastie (2005); this might a the reason for the larger MSE, but this is only a conjecture. However, in all, the MSE of the Bayesian lassos are quite competitive. The fused lasso model, which is reasonable for Examples 2 and 3, has excellent MSE there.

In Table 2 we also show the MSEs of Gibbs-Mode, which also can be used as a Bayesian lasso point estimate. We estimate λ with its posterior mean from Gibbs sampler, then estimate $\boldsymbol{\beta}$ using the LARS algorithm with this value of λ (instead of cross-validation). Gibbs-Mode can be used as an alternative model selector. Compared to LARS-Lasso, the median MSE of the Gibbs-Mode is smaller for Examples 2 and 3, and comparable for Example 1. However, the posterior mean is a substantially better point estimator.

Table 3 summarizes the average[1] model errors over 200 runs in Example 4-5. For comparison, the average mean squared errors of LARS, Group LARS from Yuan and Lin (2006) and Gibbs-Mode Group lasso are added. We observe that the models that were selected by lasso (or LARS) are larger than those selected by other methods in Example 4-5, which is in line with the findings of Yuan and Lin (2006). Gibbs-Mode Group Lasso estimates λ using the posterior mean from the Gibbs sampler, and then estimates coefficients using the `grplasso` package (Meier, 2009; see also Meier *et al.* 2008) . Gibbs-Mode Group lasso does better than Group lasso for Example 4, but is slightly less precise than Group lasso for Example 5.

Overall, the Bayesian Hierarchical lassos perform as well as, or better than the LARS fit in most of the examples. This is, in one sense, a comment on the method of choosing the tuning parameter λ, and shows that putting λ into the Gibbs sampler seems to be as effective as choosing it by cross-validation.

5.2 Data Analysis

In this section, we consider two different data sets, again choosing ones that have been analyzed in previous lasso papers. The intent is to fit this work into the existing literature and highlight differences.

Prostate Cancer Data

The prostate cancer data is from Stamey *et al.* (1989), who examined the correlation between the level of a prostate specific antigen and a number of clinical measures in men who were about to receive a radical prostatectomy. The factors were log(cancer volume), log(prostate weight), age, log(benign prostatic hyperplasia amount), seminal vesicle invasion, log(capsular penetration), Gleason score and percentage Gleason scores 4 or 5. We fit the above models to log(prostate specific antigen) using the Bayesian lasso. Details about fitting the lasso and elastic net to these data are in Tibshirani (1996) and in Zou and Hastie (2005), respectively.

The previous papers divided the data into a training set with 67 observations and a test set with 30 observations. Model fitting was carried out on the training data, and performance is evaluated with the prediction error (MSE) on the test data. For the group lasso, we made two arbitrary groups: log(cancer volume), log(prostate weight), age

[1] Table 2 uses median mean squared errors and Table 3 uses average MSE to enable comparison with the previously published results. This is also the reason for the different number of iterations.

Table 4: Mean squared prediction errors for the prostate cancer data based on 30 observations of the test set for both the Bayesian lassos and the elastic net of Zou and Hastie (2005). Predictions are based on the posterior mean of β for the Bayesian method, and on the model fitting and tuning parameter selection by ten-fold CV for LARS.

Gibbs			LARS		
Lasso	Naïve elastic net	Fused Lasso	Lasso	Naïve elastic net	elastic net
0.478	0.474	0.483	0.499	0.566	0.381

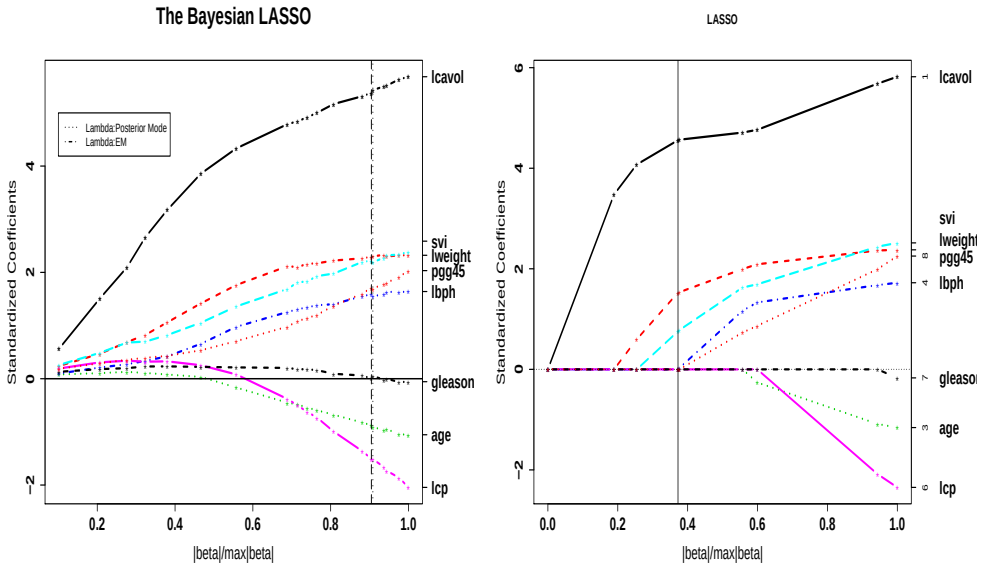
and log(benign prostatic hyperplasia amount) vs. seminal vesicle invasion, log(capsular penetration), Gleason score and percentage Gleason scores 4 or 5. Thus, we expect inaccurate prediction for the group lasso.

Table 4 shows similar prediction accuracy of the Bayesian lasso, elastic net and fused lasso, all out performing the lasso and the naïve elastic net. Among all models and methods, the elastic net has the smallest prediction error. Zou and Hastie (2006) argued that in this case, the elastic net is actually univariate soft thresholding (UST), because the selected λ_2 is very big (1000). However, UST totally ignores the dependence between predictors and treats them as independent variables. Thus, it might not work well if there exist collinearity among independent variables.

Figure 2 compares posterior median estimates for Bayesian lasso model (left) with the ordinary lasso (right) for the prostate cancer data. The figure shows the paths of these estimates as their respective shrinkage parameters (λ) are varied. The paths of Bayesian lasso appear to be smooth, but are similar in shape to the lasso paths. The vertical line in the lasso panel represents the estimate chosen by n-fold (leave-one-out) cross-validation, whereas the vertical lines in the Bayesian lasso panel represent the estimate chosen by marginal maximum likelihood (EM Gibbs) and the posterior mode (Gibbs). The estimated λs from Gibbs and EM Gibbs are $\lambda = 2.968$ and $\lambda = 3.107$ respectively. Note that, although the lines corresponding to the estimates of λ appear in different places in the corresponding plots, this is a result of the choice of horizontal scale: $|\beta|/\max(|\beta|)$, that is, the horizontal axis is each coefficient divided by its own maximum over the path, so the axis goes from $0-1$ for all coefficients (as done in Park and Casella 2008). The resulting point estimates are quite similar.

Figure 3 shows the 95% equal-tailed credible intervals for regression parameter βs based on the posterior mean Bayesian lasso estimates with, for comparison, overlaid point estimates of original lasso, posterior mode lasso, posterior means of elastic net, and fused and lasso. It is interesting to note that all estimates are all inside the credible intervals, showing that the substantive conclusion will be quite similar no matter which approach is used. However, we again remark that only the Bayesian lassos provide valid standard errors for the zero-estimated coefficients.

Figure 2: Bayesian Lasso (left panel) ($\lambda = 2.986$) and Lasso (right panel) trace plots for estimates of the prostate cancer data regression parameters, with vertical lines indicating the estimates chosen by n-fold cross-validation (lasso $\lambda = 2.940$) and posterior mode (Bayesian lasso $\lambda = 2.986$). The Bayesian Lasso estimates were posterior means computed over a grid of λ values, using 10,000 consecutive iterations of the Gibbs sampler after 1,000 burn-in iterations for each λ.

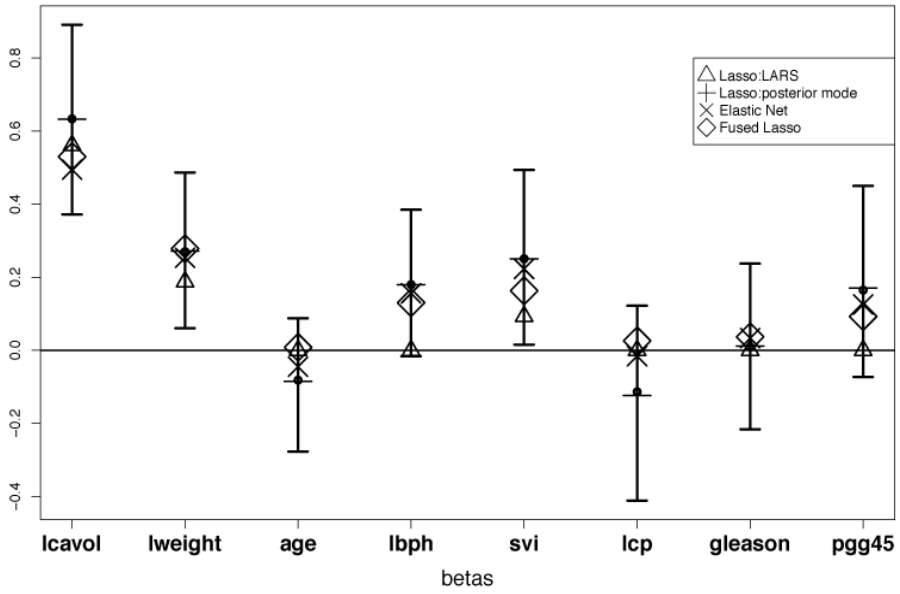


Birth Weight Data

The birth weight data set is from Hosmer and Lemeshow (2000), and records the birth weights of 189 babies and eight predictors concerning the mother. Among the eight predictors, two are continuous (mother's age in years and mother's weight in pounds at the last menstrual period) and six are categorical (mother's race, smoking status during pregnancy, number of precious premature labors, history of hypertension, presence of uterine irritability, number of physician visits). This data set was used to fit the group lasso (Yuan and Lin 2006). As in Yuan and Lin, we model both mother's age and weight by using third-order polynomials, and fit the Bayesian lasso, elastic net, and group lasso. The data were divided into a randomly chosen training set with 151 observations and a test set with 37 observations. Model fitting and shrinkage parameter estimation was carried out on the training data, and mean squared prediction errors were calculated on the test data.

The prediction errors are reported in Table 5, but note that our training set may not be the same as that of Yuan and Lin (2006). However, estimation and prediction can be compared. From Table 5, we observe that all of the Bayesian lassos give better

Figure 3: For the prostate cancer data, posterior mean Bayesian lasso estimates ($\circ$) and corresponding 95% equal-tailed credible intervals. Overlaid are the lasso estimates based on n-fold cross-validation ($\triangle$), and posterior mode estimates from the Bayesian lasso ($+$), elastic net ($\times$), and fused lasso ($\diamond$) based on $10,000$ Gibbs samples.



prediction than the LARS fit, with the Bayesian group lasso giving the most precise prediction. One reason for this performance difference may be the fact that we used a more extensive dummy variable structure than Yuan and Lin (2006). For example, for an explanatory variable like "physician's visit", coded $(0, 1, 2, 3$ or more), we fit the four means instead of one coefficient. This resulted in a much smaller mean-squared error in Table 5.

From Figure 4 we observe that lasso, elastic net and fused lasso seem to treat each category of regressors as independent variables. However, the group lasso estimates are very consistent within categories. Yuan and Lin (2006) note that "the number of physician visits should be excluded from the final model, ... the backward stepwise method excludes two more factors: mothers weight and history of hypertension". The Bayesian group lasso suggests excluding race and number of physician visits, but probably not the other estimates. Premature labor experience increases the birth weight of baby, history of hypertension and uterine irritability decrease the birth weight, and smoking during pregnancy mildly decreases the birth weight. The age of mothers has positive a

Table 5: Mean-squared prediction errors for the birth weight data based on 37 observations of test set: For comparison, mean-squared errors based on the randomly selected 37 observations of test set from Yuan and Lin (2005) are added. Predictions are based on the posterior mean of β for the Bayesian method, and on the model fitting and tuning parameter selection by C_p-criterion for group LARS.

Gibbs				Lassos			
Lasso	Naïve elastic net	Fused Lasso	Group Lasso	Group LARS	Group garrotte	Group Lasso	Stepwise
171944.8	102651.8	89787.5	36620.3	609092.8	579413.6	579413.6	646664.1

relationship but mothers weight at the last menstrual period has a negative relationship with the birth weight.

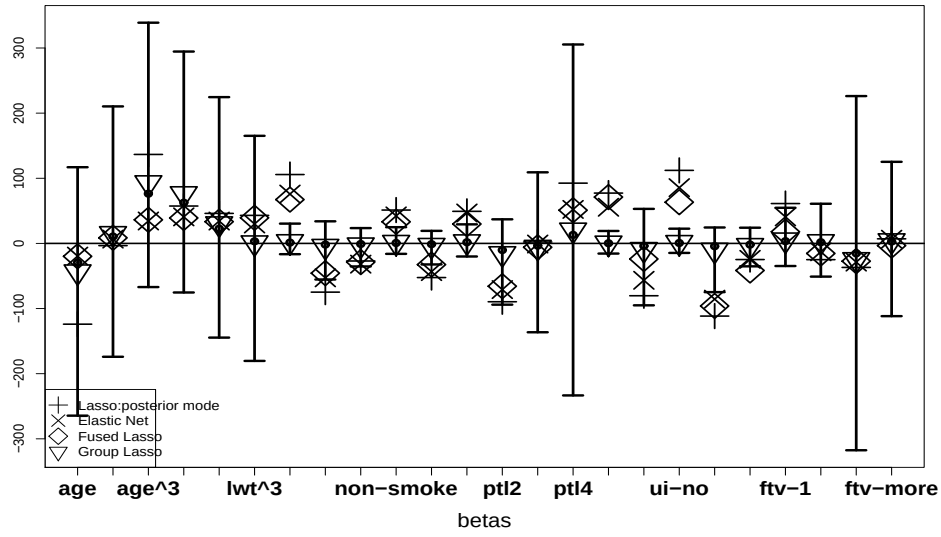
State Failures in Asia

These data are collected by the State Failure Task Force (SFTF, Esty *et al.* 1999), which is a U.S. government funded group of interdisciplinary researchers whose objective is to understand and forecast when governments cease to function effectively (usually collapsing in violence and disarray). Through a series of reports they have created a warning system of state failures based on the analysis of a huge collection of covariates (about 1,200) on all independent states around the world with a population of at least 500,000, from 1955 to 1998. Thus the greatest challenge is to consider a vast number of potential model specifications using prior theoretical knowledge and model-fitting comparisons. The final results of the SFTF team are controversial because they end up using only three explanatory variables, democracy, trade openness and infant mortality. Their findings are criticized on substantive grounds for being oversimplified (Millien and Krause 2003, Parris and Kate 2003, Sachs 2001), and on methodological ground for their treatment of missing data and forecasting procedures (King and Zeng 2001).

One consistent criticism of the SFTF approach is the use of all global regions in a single analysis. It is clear to area studies scholars that state failures occur with strong regional explanations that can differ significantly. Therefore we choose to concentrate on Asia, which constitutes both an interesting and important set of cases. In the original data for this subset, there are $p = 128$ explanatory variables with $n = 23$ observations. Here we omit four cases that had greater than 70% missing data, making standard missing data tools inoperable. In addition, eighteen variables provided no useful information and were dropped as well. Thus, for the final analysis, we have $p = 110$ explanatory variables with $n = 19$ observations to apply to the LARS algorithm, the Bayesian lasso, and the Bayesian Elastic Net, all with a probit link function for the dichotomous outcome of state failure.

Interestingly, the LARS lasso picks three variables, just like the stepwise procedure of the State Failure Task Force, but they are three *different* explanatory variables:

Figure 4: For the birth weight data, posterior mean Bayesian group Lasso estimates ($\circ$) and corresponding 95% equal-tailed credible intervals for mothers age (age, age^2,age^3),weight at the last menstrual period (lwt, lwt^2,lwt^3), race(white, black, other), smoking(no-yes), history of premature labor (ptl1, ptl2, ptl3), history of hypertension (ht-no, ht-yes), uterine irritability (ui-no, ui-yes),physician visit during the first trimester (ftv-no, ftv-1, ftv-2, ftv-3 or more). Overlaid are the lasso ($+$), elastic net ($\times$), fused lasso($\diamond$) and posterior mode group lasso ($\bigtriangledown$) posterior means based on $10,000$ Gibbs samples.



- ▷ `polxcons`: the level of constraints on the political executive, from low to high
- ▷ `sftpeind`: an indicator for ethnic war, 0 = none, 1 = at least one
- ▷ `sftpmmax`, the maximum yearly conflict magnitude scale

(see `http://globalpolicy.gmu.edu/pitf` for a more detailed explanation of these variables). Recall that the LARS lasso is a variable weighter not a variable selector, where the weights are either zero or one. Thus the LARS lasso zeros-out 107 variables here in favor of the 3 listed above. From a political science context, the LARS lasso asserts that the key determinants of state failure are: the degree to which the key leaders of government are constrained from making unilateral policy choices, the presence of an ethnic war on its borders (*not* an ethnic civil war, however), and the severity level of the greatest conflict that the nation is exposed to. As a group, these variable suggest a crises and reaction theory whereby governments that are faced with severe conflicts, perhaps of an ethnic nature, and have a limited ability for the political executive to quickly and non-consultatively respond, are more likely to fail.

Table 6 provides the top ten variables by absolute median effect from the Bayesian lasso, and also include for these variables the LARS lasso conclusion. The Bayesian Elastic Net produces results that are virtually indistinguishable from the Bayesian lasso for these data. We therefore omit the results summary for this approach from the table.

Table 6: LARS Lasso and Bayesian Lasso Results, Top Ten Effects By Bayesian Lasso Posterior Median (absolute value).

	Bayesian Lasso Quantiles					LARS Lasso
Variable	0.05	0.10	0.50	0.90	0.95	Weight
`sftpeind`	-0.2387	-0.0888	0.3823	1.6498	2.3219	0.1999
`sftpmmax`	-0.3282	-0.1686	0.2257	1.2086	1.6969	0.0307
`sftpmag`	-0.4095	-0.2380	0.1927	1.1228	1.6081	0.0000
`sftpcons`	-0.4197	-0.2315	0.1846	1.0907	1.5559	0.1750
`sftpnum`	-0.4430	-0.2407	0.1657	1.0253	1.4062	0.0000
`sftpem1`	-0.4421	-0.2636	0.1480	1.0140	1.4609	0.0000
`dispop1`	-1.3756	-0.9410	-0.1213	0.3031	0.5050	0.0000
`sftpeth`	-0.5091	-0.2951	0.1194	0.9251	1.3498	0.0000
`sftgreg2`	-0.4616	-0.2811	0.1137	0.9115	1.3047	0.0000
`polpacmp`	-0.5071	-0.3106	0.1098	0.8568	1.2240	0.0000

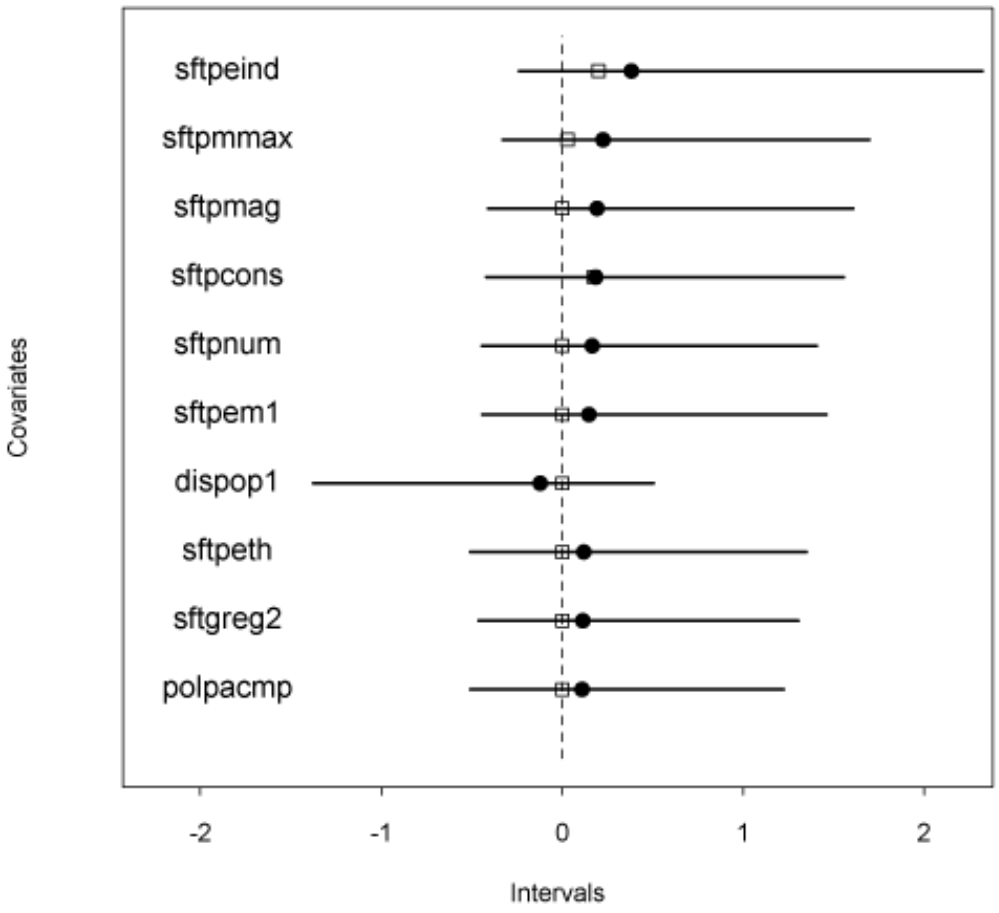
Every coefficient credible interval of the Bayesian lasso covers zero, including the ten given in Table 6. This indicates a broad lack of traditional statistical reliability despite the three choices of the LARS lasso. Recall that we cannot produce corresponding credible intervals for the LARS lasso, so the credible intervals from the Bayesian lasso are the only available measure of uncertainty. Thus, given the evidence from the Bayesian lasso, we are inclined to believe that the LARS lasso is overly-optimistic with these choices.

There is an interesting and confirmatory finding from Table 6. In terms of posterior effect size (absolute value of the posterior median), the Bayesian lasso's top four variables contain the three picked by the LARS lasso. The sole exception is

- ▷ `sftpmag`: the magnitude of conflict events of all types,

which is closely related to `sftpmmax`. This is reassuring since it implies that the two approaches are focusing on a small core of potentially important explainers. While picking the top ten effect sizes is arbitrary, there is a noticeable drop in magnitude after this era. The Bayesian lasso therefore brings some additional variables to our attention:

Figure 5: For the top ten coefficients of Table 6, the solid circle is the median from the Bayesian Lasso, and the square is the LARS Lasso estimate. The solid line is the 95% credible interval.



- ▷ `sftpnum`: the number of critical (negative political events).
- ▷ `sftpem1`: the ethnic war magnitude indicator number 1.
- ▷ `dispop1`: the population proportion of the largest politically significant communal group seeking autonomy and subject to discrimination.

- ▷ `sftpeth`: the ethnic wars score.
- ▷ `sftgreg2`: the subregion used by sftf. . . scores.
- ▷ `polpacmp`: a 0-10 point indicator with increasing levels of autocratic governmental control.

These variables reinforce the themes in the first first four: ethnic groups, ethnic conflict, magnitude of conflict, and the level of executive control of government, without broad consultation, appear to be important determinants of state failure. Moreover, looking at Figure 5 also suggests that, although the credible intervals cover zero, all of them a clearly skewed right or left, suggesting the existence of an effect in that direction.

It appears that these data contain many small, overlapping explanatory effects that contribute to the outcome of state failure in incremental ways that are difficult to sort-out with lasso approaches or conventional tools. This example is interesting because of the difficulty in picking from among the many $p \gg n$ possible right-hand-side variables and the suspect stepwise manner taken by the creators of the data producing only `poldemoc`, `pwtopen`, and `sfxinfm`. The LARS lasso picked none of these three but also picked a very parsimonious set of explanations. The Bayesian lasso, as indicated in the table, finds little support for the the stepwise-chosen result. Another problem with adopting this three-variable model is that political scientists have strong theory and evidence to support the causal power of: executive constraints, the level of autocracy, the extent of civil violence, population proportion of ethnic groups, multilateral interventions, ongoing conflict with neighboring states, the level of ethnic war/genocide, and so on. The LARS lasso picks-up some of these effects, such as the level constraints on the executive and the presence of ethnic war in the region. These contrasting results (stepwise regression, LARS lasso, Bayesian lasso) also demonstrate a difficulty faced by the SFTF team: many variables are seemingly important for theoretical reasons but remain unhelpful in terms of statistical prediction due to their small contributory power and highly overlapping explanations.

6 Discussion

Using the basic identity (14), it is relatively straightforward to form a (near) conjugate hierarchical model that will result in a posterior distribution reflecting a penalized least squares approach. Varieties of the lasso fall into this category.

A goal of the lasso is to both select and estimate, thus it will set some coefficient estimates equal to zero (selection) and estimate others to be non-zero. Although the lasso is not a consistent estimator (Zou 2006), some of its variations are, for example, the adaptive lasso. Bayesian estimation, based on the output of a Gibbs sampler, typically consists of estimating means and standard errors through Monte Carlo averages. How does a Bayesian approach achieve the goals of the lasso?

First, we must realize that a model selector such as the lasso is no more than a point estimator of the coefficient vector. In fact, it is exactly the posterior mode from the hierarchical models that we have used. Having the MCMC output allows us to

summarize the posterior in any manner that we choose - although it is typical to use the posterior mean, we could also use the posterior mode. Moreover, we have some assessment of how sure we are that such coefficients are actually zero. As the lasso cannot produce valid standard errors if the true coefficients are zero, it cannot give any confidence assessment of these estimates.

Of course, we are trading the Bayesian standard error for the frequentist standard error. As our hierarchies all use proper priors, the Bayesian credible intervals will not maintain a guaranteed coverage against all parameter values. However, coverage probability infima are typically attained at atypical values of the parameters - for example, as they approach infinity - and for reasonable parameter values the Bayesian intervals should have adequate frequentist coverage (if that is desired). However, perhaps the more important point is that, at this time, we have no default frequentist analysis that will provide valid confidence intervals.

With the introduction of the LARS algorithm, computation of the lasso path has become quite rapid. The Gibbs samplers presented in this paper are also quite fast. Of course we have shown that they are geometrically ergodic, but the other point is that, in application, the sampler runs very quickly. Moreover, after one run of the Gibbs sampler we not only have our point estimates, but also estimates of standard error.

To use lasso estimates, the tuning parameter λ must be estimated, typically by cross-validation. In the hierarchical model we consider two choices: marginal MLE using an EM/Gibbs sampler, and putting a prior on λ and entering it in the Gibbs iterations. We find that the latter is more attractive, mainly due to the facts that it is much faster and, in all the examples that we saw, the posterior mean from the Gibbs iterations is very close to the marginal MLE. We note, however, that when we put λ into the Gibbs iterations, our estimates of the regression coefficients are not based on a fixed value of λ, but rather are marginalized over all λ, leading to somewhat of a robustness property.

Throughout, we have used hierarchical models of the form

$$\boldsymbol{y} \mid \mu, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \ \sim \ N_n\left(\mu \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I_n}\right), \quad \boldsymbol{\beta} \sim N(0, \Sigma_\beta),$$

where Σ_β is parametrized with τ_is that are given gamma priors. We have shown how to parameterize Σ_β to obtain the lassos in this paper but, of course, there can be many more parameterizations. These would arise from other practical restrictions on the β_i, and in many cases will be easily implementable. This follows because for all the lassos, with the exception of the elastic net, λ and $\boldsymbol{\beta}$ are conditionally independent given the τ_is, leading to a straightforward Gibbs sampler.

Acknowledgments

Kyung, Gill, and Casella were supported by National Science Foundation Grants DMS-0631632 and SES-0631588, and Ghosh was supported by National Science Foundation Grant SES-0631426 and National Security Agency Grant MSPF-07G-097.

References

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

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

Beran, R. (1982). Estimated Sampling Distributions: The Bootstrap and Competitors. *The Annals of Statistics* **10**, 212-225.

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

Casella, G. and Consonni, G. (2009). Reconciling Model Selection and Prediction. Technical Report, Department of Statistics, University of Florida. Available at `http://www.stat.ufl.edu/ casella/Papers/AICBIC-3.pdf`

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **7**, 1-26.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics* **32**, 407-451.

Esty, Daniel C., Goldstone, Jack A., Gurr, Ted Robert, Harff, Barbara, Levy, Marc, Dabelko, Geoffrey D., Surko, Pamela T., and Unger, Alan N. (1999). *State Failure Task Force Report: Phase II Findings Environmental Change & Security Project Report 5*. MacClean, VA: Science Applications International Corporation.

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.

Genkin, A., Lewis, D. D. and Madigan, D. (2007). Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics* **49**, 291-304.

Hobert, J. P. and Casella, G. . (1998). Functional Compatibility, Markov Chains, and Gibbs Sampling With Improper Posteriors.' *Journal of Computational and Graphical Statistics* **7**, 42-60.

Hobert, J. P. and Geyer, C. J. (1998). Geometric Ergodicity of Gibbs and Block Gibbs Samplers for a Hierarchical Random Effects Model. *Journal of Multivariate Analysis* **67** 414-430.

Hoerl, A. E and Kennard, R. W (1970). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics* **12**, 55-68.

Hosmer, D. W. and Lemeshow, S. (2000). *Applied Logistic Regression, Second Edition*. NewYork: Wiley.

Kim, Y., Kim, J. and Kim, Y. (2006). Blockwise Sparse Regression. *Statistica Sinica* **16**, 375-390.

King, G. and Zeng, L. (2001). Improving Forecasts of State Failure. *World Politics* **53** 623-658.

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

Leeb, H. and Pötscher, B. M. (2005). Model Selection and Inference: Facts and Fiction. *Econometric Theory***21**, 21-59.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*, 2nd Edition, New York: Springer-Verlag.

Liu, J. S. , Wong, W. H. and Kong, A. (1994). Covariance Structure of the Gibbs Sampler with Applications to the Comparison of Estimators and Augmentation Schemes. *Biometrika* **81** 27-40.

Meier, L. (2009). *The grplasso Package.*
`http://cran.r-project.org/web/packages/grplasso/index.html`

Meier, L., van de Geer, S. and Buhlmann, P. (2008). The Group Lasso for Logistic Regression. *Journal of the Royal Statistical Society. Series B* **70**, 53-71.

Meyn, S.P. and Tweedie, R.L. (1993) *Markov Chains and Stochastic Stability*. New York: Springer-Verlag

Milliken, Jennifer and Krause, Keith. (2003). State Failure, State Collapse, and State Reconstruction: Concepts, Lessons and Strategies. In *State Failure, Collapse & Reconstruction.*, J. Milliken (ed.), pp.1-19. Oxford, UK: Blackwell Publishers.

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.

Osborne, M. R., Presnell, B. and Turlach, B. A. (2000b). On the Lasso and Its Dual. *Journal of Computational and Graphical Statistics* **9**, 319-337.

Parris, Thomas M. and Kate, Robert W. (2003). Characterizing and Measuring Sustainable Development. *Annual Review of Environment and Resources.* **28**, 559-586.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association* **103**, 681-686.

Park, M. and Hastie, T. (2007). L_1-Regularization Path Algorithm for Generalized Linear Models. *Journal of the Royal Statistical Society. Series B* **69**, 659-677.

Roberts, G.O. and Rosenthal, J.S. (1998) Markov chain Monte Carlo: Some practical implications of theoretical results (with discussion). *Can. J. Statist.* **26**, 5–32.

Sachs, Jeffrey D. (2001). The Strategic Significance of Global Inequality. In *The Washington Quarterly* **24**, 187-198.

Samworth, R. (2003). A Note on Methods of Restoring Consistency to The Bootstrap. *Biometrika* **90**, 985-990.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989). Prostate Specific Antigen in the Diagnosis and Treatment of Adenocarcinoma of the Prostate ii: Radical Prostatectomy Treated Patients. *Journal of Urology* **16**, 1076-1083.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society. Series B* **67**, 91-108.

Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *Journal of the Royal Statistical Society. Series B* **68**, 49-67.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B* **67**, 301-320.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association* **101**, 1418-1429.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the Degrees of Freedom of the Lasso. *The Annals of Statistics* **35**, 2173-2192.

Appendix

1. Inconsistency of Lasso Estimates

We show that the lasso is a superefficient estimator when $\beta = 0$ based on Theorem 2 of Knight and Fu (2000). Assume the following regularity conditions for the design,

1. $\boldsymbol{C}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i'\boldsymbol{x}_i \to \boldsymbol{C}$,
2. $\frac{1}{n}\max_{1\le i\le n} \boldsymbol{x}_i\boldsymbol{x}_i' \to \boldsymbol{0}$,

where $\boldsymbol{x}_i$ is $1 \times p$ design vector of i^{th} response and $\boldsymbol{C}$ is a nonnegative definite matrix. We note that $\boldsymbol{X}$ is a $n \times p$ matrix of standardized regressors so that the diagonal elements of $\boldsymbol{C}_n$ (and hence those of $\boldsymbol{C}$) are all identically 1. Recall that the bridge estimator (Frank and Friedman, 1993) is given by

$$\hat{\boldsymbol{\beta}}_{\text{B}} = \arg\min_{\boldsymbol{\beta}} (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})' (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|^{\gamma},$$

for a given λ where $\gamma > 0$.

Theorem 6.1. *(Knight and Fu, 2000) Suppose that $\gamma \geq 1$. If $\lambda_n/\sqrt{n} \to \lambda_0 \geq 0$ and $\boldsymbol{C}$ is nonsingular then*

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}\right) \xrightarrow{d} \arg\min(V),$$

where

$$V(\boldsymbol{u}) = -2\boldsymbol{u}'\boldsymbol{W} + \boldsymbol{u}'\boldsymbol{C}\boldsymbol{u} + \lambda_0 \sum_{j=1}^{p} u_j \text{sgn}(\beta_i)|\beta_j|^{\gamma-1},$$

if $\gamma > 1$,

$$V(\boldsymbol{u}) = -2\boldsymbol{u}'\boldsymbol{W} + \boldsymbol{u}'\boldsymbol{C}\boldsymbol{u} + \lambda_0 \sum j = 1^{p} \left[u_j \text{sgn}(\beta_i) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)\right],$$

if $\gamma = 1$ and $\boldsymbol{W} \sim N\left(\boldsymbol{0}, \sigma^2 \boldsymbol{C}\right)$, where sgn denotes the sign of the parenthetical quantity.

We use Theorem 6.1 with $p = 1$, $\sigma^2 = 1$ and $\beta = 0$, so the Cramér-Rao lower bound is 1. The limiting distribution of the lasso is that of the random variable V^*, where

$$V^* = \arg\min_u V(u) = \arg\min_u u^2 - 2uW + \lambda_0|u|,$$

where $W \sim N(0,1)$. It is straightforward to check that $V(u)$ is convex, so it has a unique minimum. A little calculus will show that

$$V^* = \begin{cases} \frac{2W+\lambda_0}{2} & \text{if } W < -\lambda_0/2 \\ 0 & \text{if } -\lambda_0/2 \leq W \leq \lambda_0/2 \\ \frac{2W-\lambda_0}{2} & \text{if } W > \lambda_0/2. \end{cases}$$

Of course $\text{E}V^* = 0$, so

$$\text{Var}(V^*) = 2\int_{\lambda_0/2}^{\infty} \left(\frac{2w - \lambda_0}{2}\right)^2 \phi(w)\, dw,$$

where $\phi(w)$ is the standard normal pdf. Now

$$\frac{d}{d\lambda_0}\text{Var}(V^*) = -\int_{\lambda_0/2}^{\infty} (2w - \lambda_0)\phi(w)\, dw < 0,$$

so $\text{Var}(V^*)$ is a strictly decreasing function of λ_0, with a maximum at $\lambda_0 = 0$, where it is equal to 1. Thus, for any $\lambda_0 > 0$, $\text{Var}(V^*) < 1$, showing that the lasso is superefficient.

Beran (1982) showed that the bootstrap is not consistent for superefficient estimators (such as the classic Hodges estimator). This is exactly the situation that we are in here, and we conclude that the bootstrap is not consistent for the lasso if the parameter is zero.

2. Lasso Hierarchies

2.1 Basic Identity

The Laplace (double-exponential) distribution is a scale mixture of a normal distribution with an exponential mixing density (Andrews and Mallows 1974), that is

$$\frac{a}{2}\exp(-a|z|) = \int_0^\infty \frac{1}{\sqrt{2\pi s}}\exp\left(-\frac{z^2}{2s}\right)\frac{a^2}{2}\exp\left(-\frac{a^2}{2}s\right)\mathrm{d}s. \tag{14}$$

To prove this, rewrite

$$\int_0^\infty \exp\left(-\frac{z^2}{2s}\right)\frac{a^2}{2}\exp\left(-\frac{a^2}{2}s\right)\mathrm{d}s = \frac{a^2}{2}\exp(-a|z|)\int_0^\infty \exp\left(-\frac{1}{2s}(|z|-as)^2\right)\mathrm{d}s.$$

Now recall the inverse Gaussian density

$$f(x|\mu,\lambda) \quad = \quad \frac{\lambda^{1/2}}{(2\pi x^3)^{1/2}}\exp\left(-\frac{\lambda}{2\mu^2 x}(x-\mu)^2\right).$$

Making the transformation $as = 1/r$, it can be shown that the integral above is an inverse Gaussian density, which leads directly to (14). We will use similar techniques with appropriate modifications for group lasso and fused lasso. The main idea is to introduce the appropriate latent parameters.

2.2 Original Lasso

As Park and Casella (2008) argued, because the columns of $\boldsymbol{X}$ are centered, it is easy to analytically integrate μ from the joint posterior, under its independent, flat prior. In this paper, we marginalize it out in the interest of simplicity and speed. Thus, we use $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \overline{y}$, where $\overline{y} = \sum_{i=1}^n y_i/n$.

Given λ, the full conditional posterior of the full hierarchical lasso is given by

$$\begin{aligned}
\boldsymbol{\beta} \mid \mu, \sigma^2, \tau_1^2, \ldots, \tau_p^2, \boldsymbol{X}, \boldsymbol{y} &\sim N_p\left(\left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1}\right)^{-1}\boldsymbol{X}'\tilde{\boldsymbol{y}}, \sigma^2\left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}_\tau^{-1}\right)^{-1}\right), \\
1/\tau_j^2 = \gamma_j \mid \mu, \boldsymbol{\beta}, \sigma^2, \boldsymbol{X}, \boldsymbol{y} &\sim \text{inverse Gaussian}\left(\frac{\lambda^2\sigma}{|\beta_j|}, \lambda^2\right)\boldsymbol{I}(\gamma_j > 0), \qquad (15) \\
&\qquad \text{for } j = 1, \ldots, p, \\
\sigma^2 \mid \mu, \boldsymbol{\beta}, \tau_1^2, \ldots, \tau_p^2, \boldsymbol{X}, \boldsymbol{y} &\sim \\
&\text{inverted gamma}\left(\frac{n-1+p}{2}, \frac{1}{2}\left(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}\right)'\left(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}\right) + \frac{\lambda}{2}\boldsymbol{\beta}'\boldsymbol{D}_\tau^{-1}\boldsymbol{\beta}\right).
\end{aligned}$$

2.3 Group Lasso

Partition the $\boldsymbol{\beta}$ vector into K groups $G_1, \ldots, G_K$ of sizes $m_1, \ldots, m_K$ where $\sum_{k=1}^K m_k = p$. Denote by $\boldsymbol{\beta}_{G_k}$ the vector of β_js in group k ($k = 1, \ldots, K$). Let the conditional prior of $\boldsymbol{\beta}$ be

$$\pi(\boldsymbol{\beta}|\sigma^2) \propto \exp\left(-\frac{\lambda}{\sigma}\sum_{k=1}^K ||\boldsymbol{\beta}_{G_k}||\right),$$

where $||\boldsymbol{\beta}_{G_k}|| = \left(\boldsymbol{\beta}_{G_k}'\boldsymbol{\beta}_{G_k}\right)^{1/2}$. We introduce the latent parameters $\tau_1^2|\sigma^2, \ldots, \tau_K^2|\sigma^2$ such that

$$\boldsymbol{\beta}_{G_k}|\tau_k^2, \sigma^2 \overset{ind}{\sim} N_{m_k}\left(\boldsymbol{0}, \sigma^2\tau_k^2\boldsymbol{I}_{m_k}\right) \quad \text{and} \quad \tau_k^2 \overset{ind}{\sim} \text{gamma}\left(\frac{m_k+1}{2}, \frac{\lambda^2}{2}\right),$$

for $k = 1, \ldots, K$. Now, similar to (14),

$$\begin{aligned}
&\int_0^\infty \left(\frac{1}{2\pi\sigma^2\tau_k^2}\right)^{m_k/2} \exp\left[-\frac{||\boldsymbol{\beta}_{G_k}||^2}{2\sigma^2\tau_k^2}\right] \frac{\left(\frac{\lambda^2}{2}\right)^{\frac{m_k+1}{2}}\left(\tau_k^2\right)^{\frac{m_k+1}{2}-1}}{\Gamma\left(\frac{m_k+1}{2}\right)} \exp\left[-\lambda^2\tau_k^2/2\right] \mathrm{d}\tau_k^2 \\
&= \exp\left[-\lambda||\boldsymbol{\beta}_{G_k}||/\sigma\right], \qquad (16)
\end{aligned}$$

where we needed the gamma prior on τ_k^2 to get the correct Jacobian.

Full Conditionals The full conditional posteriors of the hierarchical grouped lasso model are

$$\begin{aligned}
\boldsymbol{\beta}_{G_k} \mid \boldsymbol{\beta}_{-G_k}, \sigma^2, \tau_1^2, \ldots, \tau_K^2, \lambda, \boldsymbol{X}, \tilde{\boldsymbol{y}} &\sim N_p\left(A_k^{-1}\boldsymbol{X}_k'\left(\tilde{\boldsymbol{y}} - \frac{1}{2}\sum_{k'\neq k} X_{k'}\boldsymbol{\beta}_{G_{k'}}\right), \sigma^2 A_k^{-1}\right), \\
1/\tau_k^2 = \gamma_k \mid \boldsymbol{\beta}, \sigma^2, \lambda, \boldsymbol{X}, \tilde{\boldsymbol{y}} &\sim \text{inverse Gaussian}\left(\sqrt{\frac{\lambda^2\sigma^2}{\|\boldsymbol{\beta}_{G_k}\|^2}}, \lambda^2\right)\boldsymbol{I}(\gamma_k > 0), \\
&\qquad \text{for } k = 1, \ldots, K, \qquad (17) \\
\sigma^2 \mid \boldsymbol{\beta}, \tau_1^2, \ldots, \tau_K^2, \lambda, \boldsymbol{X}, \tilde{\boldsymbol{y}} &\sim \text{inverted gamma}\left(\frac{n-1+p}{2},\right. \\
&\qquad \left.\frac{1}{2}\|\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \frac{1}{2}\sum_{k=1}^{K}\frac{1}{\tau_k^2}\|\boldsymbol{\beta}_{G_k}\|^2\right).
\end{aligned}$$

where $\boldsymbol{\beta}_{-G_k} = (\boldsymbol{\beta}_{G_1}, \ldots, \boldsymbol{\beta}_{G_{k-1}}, \boldsymbol{\beta}_{G_{k+1}}, \ldots, \boldsymbol{\beta}_{G_K})$ and $A_k = \boldsymbol{X}_k'\boldsymbol{X}_k + (1/\tau_k^2)\boldsymbol{I}_{m_k}$.

Estimation of λ With a gamma(r, δ) prior, the full conditional distribution of λ^2 is

$$\pi(\lambda^2|\boldsymbol{\beta}, \sigma^2, \tau_1^2, \ldots, \tau_K^2, \boldsymbol{X}, \tilde{\boldsymbol{y}}) \sim \text{gamma}\left(\frac{p+K}{2} + r, \frac{1}{2}\sum_{k=1}^{K}\tau_k^2 + \delta\right),$$

which can be added into the Gibbs sampler.

To estimate λ from the marginal likelihood, the E-step in the EM algorithm involves taking the expected value of the log likelihood, conditional on $\tilde{\boldsymbol{y}}$ and under $\lambda^{(t-1)}$, to get

$$Q\left(\lambda^2|\lambda^{2(t-1)}\right) = \frac{p+K}{2}\ln\left(\lambda^2\right) - \frac{\lambda^2}{2}\sum_{k=1}^{K} E_{\lambda^{(t-1)}}\left[\tau_k^2|\tilde{\boldsymbol{y}}\right] + c,$$

where $c =$ terms not involving λ. At the M-step:

$$\lambda^{(t)} = \sqrt{\frac{p+K}{\sum_{k=1}^{K} E_{\lambda^{(t-1)}}\left[\tau_k^2|\tilde{\boldsymbol{y}}\right]}},$$

where the expectation is over the marginal distribution of the τ_k^2. The output from the Gibbs sampler can be used to calculate this expectation through a Monte Carlo average.

2.4 Fused Lasso

The conditional prior of $\boldsymbol{\beta}$ given σ^2 is given in (7). We introduce $2p-1$ latent parameters $\tau_1^2, \ldots, \tau_p^2, \sigma_1^2 \ldots, \sigma_{p-1}^2$ and, as in basic identity, use the fact that

$$\frac{\lambda_1}{2\sigma}e^{-\lambda_1|\beta_j|/\sigma} = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2\tau_j^2}}\exp\left(-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\right)\frac{\lambda_1^2}{2}e^{-\lambda_1^2\tau_j^2/2}\mathrm{d}\tau_j^2,$$

for $j = 1, \ldots, p$ and

$$\frac{\lambda_2}{2\sigma} e^{-\lambda_2 |\beta_{j+1} - \beta_j| / \sigma} = \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2\omega_j^2}} \exp\left[-\frac{(\beta_{j+1} - \beta_j)^2}{2\sigma^2\omega_j^2} \right] \frac{\lambda_2^2}{2} e^{-\lambda_2^2 \omega_j^2 / 2} \mathrm{d}\omega_j^2,$$

for $j = 1, \ldots, p-1$. From the identity

$$\begin{aligned} \sum_{j=1}^{p} \frac{\beta_j^2}{\tau_j^2} + \sum_{j=1}^{p-1} \frac{(\beta_{j+1} - \beta_j)^2}{\sigma_j^2} = \beta_1^2 \left(\frac{1}{\tau_1^2} + \frac{1}{\sigma_1^2} \right) + \beta_2^2 \left(\frac{1}{\tau_2^2} + \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right) + \ldots \\ \beta_{p-1}^2 \left(\frac{1}{\tau_{p-1}^2} + \frac{1}{\sigma_{p-2}^2} + \frac{1}{\sigma_{p-1}^2} \right) + \ldots \\ \beta_p^2 \left(\frac{1}{\tau_p^2} + \frac{1}{\sigma_{p-1}^2} \right) - 2 \sum_{j=1}^{p-1} \frac{\beta_j \beta_{j+1}}{\sigma_j^2}, \end{aligned}$$

we see that $\boldsymbol{\beta} | \tau_1^2, \ldots, \tau_p^2, \omega_1^2 \ldots, \omega_{p-1}^2, \sigma^2$ is multivariate normal with mean vector $\mathbf{0}$ and variance-covariance matrix $\sigma^2 \Sigma_\beta$ with

$$\Sigma_\beta^{-1} = \begin{bmatrix} \frac{1}{\tau_1^2} + \frac{1}{\omega_1^2} & -\frac{1}{\omega_1^2} & 0 & 0 & \cdots & 0 & 0 \\ -\frac{1}{\omega_1^2} & \frac{1}{\tau_2^2} + \frac{1}{\omega_1^2} + \frac{1}{\omega_2^2} & -\frac{1}{\omega_2^2} & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\frac{1}{\omega_{p-2}^2} & \frac{1}{\tau_{p-1}^2} + \frac{1}{\omega_{p-2}^2} + \frac{1}{\omega_{p-1}^2} & -\frac{1}{\omega_{p-1}^2} \\ 0 & 0 & 0 & \cdots & 0 & -\frac{1}{\omega_{p-1}^2} & \frac{1}{\tau_p^2} + \frac{1}{\omega_{p-1}^2} \end{bmatrix}.$$

Full Conditionals Given λ_1 and λ_2, the full conditional posteriors of the hierarchical fused lasso are

$$\begin{aligned} &\boldsymbol{\beta} \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2, \omega_1^2, \ldots, \omega_{p-1}^2, \boldsymbol{X}, \tilde{\boldsymbol{y}} \\ &\quad \sim N_p \left(\left(\boldsymbol{X}^\mathsf{T} \boldsymbol{X} + \Sigma_\beta^{-1} \right)^{-1} \boldsymbol{X}^\mathsf{T} \tilde{\boldsymbol{y}}, \sigma^2 \left(\boldsymbol{X}^\mathsf{T} \boldsymbol{X} + \Sigma_\beta^{-1} \right)^{-1} \right), \\ &1/\tau_j^2 = \gamma_j \mid \boldsymbol{\beta}, \sigma^2, \omega_1^2, \ldots, \omega_{p-1}^2, \boldsymbol{X}, \tilde{\boldsymbol{y}} \ \sim \ \text{inverse Gaussian} \left(\sqrt{\frac{\lambda_1^2 \sigma^2}{\beta_j^2}}, \lambda_1^2 \right), \\ &\qquad \text{with } \gamma_j > 0 \text{ for } j = 1, \ldots, p \qquad (18) \\ &1/\omega_j^2 = \eta_j \mid \boldsymbol{\beta}, \sigma^2, \tau_1^2, \ldots, \tau_p^2, \boldsymbol{X}, \tilde{\boldsymbol{y}} \ \sim \ \text{inverse Gaussian} \left(\sqrt{\frac{\lambda_2^2 \sigma^2}{(\beta_{j+1} - \beta_j)^2}}, \lambda_2^2 \right), \\ &\qquad \text{with } \eta_j > 0 \text{ for } j = 1, \ldots, p-1, \\ &\sigma^2 \mid \boldsymbol{\beta}, \tau_1^2, \ldots, \tau_p^2, \omega_1^2, \ldots, \omega_{p-1}^2, \boldsymbol{X}, \tilde{\boldsymbol{y}} \\ &\quad \sim \ \text{inverted gamma} \left(\frac{n - 1 + p}{2}, \frac{1}{2} (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})^\mathsf{T} (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{1}{2} \boldsymbol{\beta}^\mathsf{T} \Sigma^{-1} \boldsymbol{\beta} \right). \end{aligned}$$

Estimation of λ_1 and λ_2. With gamma(r, δ) priors, the full conditional distributions of λ_1^2 and λ_2^2 are

$$\pi(\lambda_1^2|\boldsymbol{\beta}, \sigma^2, \tau_1^2, \ldots, \tau_p^2, \omega_1, \ldots, \omega_{p-1}^2, \lambda_1, \boldsymbol{X}, \tilde{\boldsymbol{y}}) \sim \text{gamma}\left(p + r, \frac{1}{2}\sum_{j=1}^{p} \tau_j^2 + \delta\right),$$

$$\pi(\lambda_2^2|\boldsymbol{\beta}, \sigma^2, \tau_1^2, \ldots, \tau_p^2, \omega_1, \ldots, \omega_{p-1}^2, \lambda_2, \boldsymbol{X}, \tilde{\boldsymbol{y}}) \sim \text{gamma}\left(p - 1 + r, \frac{1}{2}\sum_{j=1}^{p-1} \sigma_j^2 + \delta\right),$$

which can be added to the Gibbs sampler. Although we use the same prior parameters for λ_1 and λ_2, different values can be used.

For marginal likelihood estimation, the E-step of the EM algorithm involves taking the expected value of the log likelihood, conditional on $\tilde{\boldsymbol{y}}$ and under $\lambda_h^{(t-1)}$ (h=1,2), to get

$$Q\left(\lambda_1^2|\lambda_1^{2(t-1)}\right) = p \ln\left(\lambda_1^2\right) - \frac{\lambda_1^2}{2}\sum_{j=1}^{p} E_{\lambda_1^{(t-1)}}\left[\tau_j^2|\tilde{\boldsymbol{y}}\right] + c,$$

$$Q\left(\lambda_2^2|\lambda_2^{2(t-1)}\right) = (p-1) \ln\left(\lambda_2^2\right) - \frac{\lambda_2^2}{2}\sum_{j=1}^{p-1} E_{\lambda_2^{(t-1)}}\left[\omega_j^2|\tilde{\boldsymbol{y}}\right] + c^*,$$

where c = terms not involving λ_1 and c^* = terms not involving λ_2. At the M-step:

$$\lambda_1^{(t)} = \sqrt{\frac{2p}{\sum_{j=1}^{p} E_{\lambda_1^{(t-1)}}\left[\tau_j^2|\tilde{\boldsymbol{y}}\right]}}, \quad \lambda_2^{(t)} = \sqrt{\frac{2(p-1)}{\sum_{j=1}^{p-1} E_{\lambda_2^{(t-1)}}\left[\omega_j^2|\tilde{\boldsymbol{y}}\right]}},$$

where the expectations are over the marginal distributions of the τ_k^2 and ω_k^2. The output from the Gibbs sampler can be used to calculate these expectations through Monte Carlo averages.

2.5 Elastic Net

The conditional prior of $\boldsymbol{\beta}$ given σ^2 is given in (9). Using the basic identity (14), the conditional prior can be expressed as

$$\pi\left(\boldsymbol{\beta}|\sigma^2, \sigma_1^2, \ldots, \sigma_p^2\right) \propto \prod_{j=1}^{p} \frac{\sqrt{\lambda_2}}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{\lambda_2}{2\sigma^2}\beta_j^2\right] \times \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2\tau_j^2}} \exp\left[-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\right] \frac{\lambda_1^2}{2} \exp\left[-\frac{\lambda_1^2\tau_j^2}{2}\right] \mathrm{d}\tau_j^2,$$

so we see that, conditional on τ_j^2, $\beta_j \sim N(0, \sigma^2(\tau_j^{-2} + \lambda_2)^{-1})$.

Full Conditionals Given λ_1 and λ_2, the full conditional posteriors of the hierarchical elastic net are

$$\begin{aligned}
\boldsymbol{\beta} \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2, \boldsymbol{X}, \tilde{\boldsymbol{y}} &\sim N_p\left(\left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \boldsymbol{D}_\tau^{*-1}\right)^{-1}\boldsymbol{X}^\mathsf{T}\tilde{\boldsymbol{y}}, \sigma^2\left(\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \boldsymbol{D}_\tau^{*-1}\right)^{-1}\right), \\
1/\tau_j^2 = \gamma_j \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{X}, \tilde{\boldsymbol{y}} &\sim \text{inverse Gaussian}\left(\sqrt{\frac{\lambda_1^2\sigma^2}{\beta_j^2}}, \lambda_1^2\right)\boldsymbol{I}(\gamma_j > 0). \qquad (19) \\
&\qquad \text{for } j = 1, \ldots, p, \\
\sigma^2 \mid \boldsymbol{\beta}, \tau_1^2, \ldots, \tau_p^2, \boldsymbol{X}, \tilde{\boldsymbol{y}} &\sim \text{inverted gamma}\left(\frac{n-1+p}{2}, \qquad (20)\right. \\
&\qquad \left.\frac{1}{2}\left(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}\right)'\left(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta}\right) + \frac{1}{2}\boldsymbol{\beta}'\boldsymbol{D}_\tau^{*-1}\boldsymbol{\beta}\right), \qquad (21)
\end{aligned}$$

where $\boldsymbol{D}_\tau^*$ is a diagonal matrix with diagonal elements $(\tau_i^{-2} + \lambda_2)^{-1}$, $i = 1. \ldots, p$.

Estimation of λ_1 and λ_2. With gamma(r_h, δ_h), $(h = 1, 2)$ priors, the full conditional distributions of λ_1^2 and λ_2 are

$$\begin{aligned}
\pi(\lambda_1^2|\boldsymbol{\beta}, \sigma^2, \tau_1^2, \ldots, \tau_p^2, \lambda_1, \boldsymbol{X}, \tilde{\boldsymbol{y}}) &\sim \text{gamma}\left(p + r_1, \frac{1}{2}\sum_{j=1}^p \tau_j^2 \; + \; \delta_1\right), \\
\pi(\lambda_2|\boldsymbol{\beta}, \sigma^2, \tau_1^2, \ldots, \tau_p^2, \lambda_2, \boldsymbol{X}, \tilde{\boldsymbol{y}}) &\sim \text{gamma}\left(\frac{p}{2} + r_2, \frac{1}{2\sigma^2}\sum_{j=1}^p \beta_j^2 \; + \; \delta_2\right).
\end{aligned}$$

Here we have used different prior parameters, as λ_1 and λ_2 come into the posterior in slightly different ways.

For marginal likelihood estimation using the EM algorithm, the E-step involves taking the expected value of the log likelihood, conditional on $\tilde{\boldsymbol{y}}$ and under $\lambda_h^{(t-1)}$ (h=1,2), to get

$$\begin{aligned}
Q\left(\lambda_1^2|\lambda_1^{2(t-1)}\right) &= p\,\ln\left(\lambda_1^2\right) \; - \; \frac{\lambda_1^2}{2}\sum_{j=1}^p E_{\lambda_1^{(t-1)}}\left[\tau_j^2|\tilde{\boldsymbol{y}}\right] \; + \; c, \\
Q\left(\lambda_2|\lambda_2^{(t-1)}\right) &= \frac{p}{2}\,\ln\left(\lambda_2\right) \; - \; \frac{\lambda_2}{2}\sum_{j=1}^p E_{\lambda_2^{(t-1)}}\left[\frac{\beta_j^2}{\sigma^2}|\tilde{\boldsymbol{y}}\right] \; + \; c^*,
\end{aligned}$$

where $c =$ terms not involving λ_1 and $c^* =$ terms not involving λ_2. At the M-step:

$$\lambda_1^{(t)} = \sqrt{\frac{2p}{\sum_{j=1}^p E_{\lambda_1^{(t-1)}}\left[\tau_j^2|\tilde{\boldsymbol{y}}\right]}}, \quad \lambda_2^{(t)} = \frac{p}{\sum_{j=1}^p E_{\lambda_2^{(t-1)}}\left[\frac{\beta_j^2}{\sigma^2}|\tilde{\boldsymbol{y}}\right]}.$$

3. Proof of Proposition 4.1

A typical transition on scan (2) is $y \to (x_1, x_2) \to y'$, where we note that given y, (x_1, x_2) is independent of the past xs. This is a key property of the two-stage Gibbs sampler (or Data Augmentation). The kernel of this transition is $f(y'|x_1, x_2)f(x_1, x_2|y)$, and integrating over x_2 gives

$$\int f(y'|x_1, x_2)f(x_1, x_2|y)dx_2 = \frac{f(x_1, y')}{f(y)} \int f(y|x_1, x_2)f(x_2|x_1, y')dx_2.$$

Exploiting the conditional independence property we have

$$\int f(y|x_1, x_2)f(x_2|x_1, y')dx_2 = f(y|x_1),$$

and hence

$$\int f(y'|x_1, x_2)f(x_1, x_2|y)dx_2 = \frac{f(x_1, y')}{f(y)} f(y|x_1) = f(x_1|y)f(y'|x_1),$$

which is the one-step transition kernel from scan (1). To complete the relationship between the n-step kernels, we need to look at the initial transition and the final transition. We have

$$\text{Initial: } f(y|x_1) = \textstyle\int f(y|x_1, x_2)f(x_2|x_1)dx_2,$$

$$\text{Final: } f(x_1|y) = \textstyle\int f(x_1, x_2|y)dx_2,$$

and

$$K^n((x_1, y), (x_1', y')) = \int_{\mathcal{X}_2'} \int_{\mathcal{X}_2} K^n((x_1, x_2, y), (x_1', x_2', y'))f(x_2|x_1)dx_2dx_2'.$$

Lastly, if we write $\pi(x_1, y) = \int_{\mathcal{X}_2'} \int_{\mathcal{X}_2} \pi(x_1, x_2, y)f(x_2'|x_1)dx_2dx_2'$, then (ignoring the constant 1/2)

$$\begin{aligned}
&||K^n((x_1, y), \cdot) - \pi(\cdot)|| \\
&= \int_{\mathcal{X}_1'} \int_{\mathcal{Y}'} |K^n((x_1, y), (x_1', y') - \pi(x_1', y')|dy'dx_1' \\
&= \int_{\mathcal{X}_1'} \int_{\mathcal{Y}'} \left| \int_{\mathcal{X}_2'} \int_{\mathcal{X}_2} |K^n((x_1, x_2, y), (x_1', x_2', y')) - \pi(x_1', x_2', y')|f(x_2|x_1)dx_2dx_2' \right| dy'dx_1' \\
&\leq \int_{\mathcal{X}_2} \left| \int_{\mathcal{X}_2'} \int_{\mathcal{X}_1'} \int_{\mathcal{Y}'} K^n((x_1, x_2, y), (x_1', x_2', y')) - \pi(x_1', x_2', y')dy'dx_1'dx_2' \right| f(x_2|x_1)dx_2 \\
&= \int_{\mathcal{X}_2} ||K^n((x_1, x_2, y), \cdot) - \pi(\cdot)||f(x_2|x_1)dx_2.
\end{aligned}$$

4. Proof of Proposition 4.2

Hobert and Geyer (1998) proved geometric ergodicity of the two-stage Gibbs sampler from the model

$$\begin{aligned} \boldsymbol{y} &\sim N\left(\boldsymbol{\theta}, \lambda_e^{-1} I\right), & \lambda_e &\sim \text{gamma}(a_2, b_2), \\ \boldsymbol{\theta} &\sim N\left(\mathbf{1}\mu, \lambda_\theta^{-1} I\right), & \lambda_\theta &\sim \text{gamma}(a_1, b_1) \\ \mu &\sim N(\mu_0, \lambda_0^{-1}), \end{aligned} \tag{22}$$

when the two blocks were taken as $(\mu, \boldsymbol{\theta})$ and $(\lambda_e, \lambda_\theta)$, and all other parameters are known and set so that all priors are proper. By taking $A = \lambda_\theta I$ in (13), the marginal model is

$$\boldsymbol{y} \sim N\left(\mathbf{1}\mu + \boldsymbol{X}\boldsymbol{\beta}, \lambda_e^{-1} I\right), \quad \boldsymbol{\beta} \sim N(0, \lambda_\theta^{-1} I). \tag{23}$$

From Section 4, and in particular Proposition 4.1, we have the following corollary.

Corollary 6.2. *If the block Gibbs sampler for model (22) is geometrically ergodic, then the block Gibbs sampler for model (23) is also geometrically ergodic.*

In each case, the blocks in the sampler are composed of the mean parameters $(\boldsymbol{\theta}, \mu)$ or $(\boldsymbol{\beta}, \mu)$, and the variance parameters $(\lambda_e^{-1}, \lambda_\theta^{-1})$.

Rather than prove this corollary, we will present a similar corollary with the lasso hierarchy. The proof of Corollary 6.2 is very similar. For the lasso model (5) we take $A = (\lambda/\sigma^2)(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{D}_\tau^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}$ in (13), which means that we start with the hierarchical model

$$\begin{aligned} \boldsymbol{y}_{p\times 1} &\sim N\left(\boldsymbol{\theta}, \lambda_e^{-1}\boldsymbol{I}\right), \\ \boldsymbol{\theta} &\sim N\left(\mathbf{1}\mu, \Sigma(\boldsymbol{\tau})\right), \\ \mu &\sim N\left(\mu_0, \lambda_0^{-1}\right), \\ \lambda_e &\sim \text{gamma}(a_2, b_2), \\ 1/\tau_j^2 &\sim \text{iid gamma}(a_1, b_2^*) \Leftrightarrow \lambda_{\theta_j} = \tau_j^2 \text{ iid inverted gamma}(a_1, b_2), \end{aligned} \tag{24}$$

where $\Sigma(\boldsymbol{\tau})$ is of the form

$$\Sigma(\boldsymbol{\tau}) = \text{diag}(\tau_1^2, \ldots, \tau_p^2) \sim \prod_{j=1}^{p} \frac{\lambda}{2}\, e^{-\lambda\tau_j^2/2}\, d\tau_j^2. \tag{25}$$

Note that we can assume $\Sigma(\boldsymbol{\tau})$ is diagonal by making the transformation $\boldsymbol{y} \to P\boldsymbol{y}$, $\boldsymbol{\theta} \to P\boldsymbol{\theta}$, and $\Sigma(\boldsymbol{\tau}) \to P\Sigma(\boldsymbol{\tau})P' = \boldsymbol{D}_{\boldsymbol{\tau}} = \text{diagonal}$. This results in

$$\begin{aligned} \boldsymbol{y}^*_{p\times 1} = \boldsymbol{P}\boldsymbol{y} &\sim N\left(\boldsymbol{\theta}^*, \lambda_e^{-1}\boldsymbol{I}\right), \\ \boldsymbol{\theta}^* = \boldsymbol{P}\boldsymbol{\theta} &\sim N\left(\boldsymbol{v}\mu, \boldsymbol{D}_{\boldsymbol{\tau}}\right), \end{aligned} \tag{26}$$

where $\boldsymbol{v} = P\mathbf{1}$ and the diagonal elements of $\boldsymbol{D_\tau}$ depend on all of the τ_j^2.

We will show that, for this model with variance priors as in (5), the block Gibbs sampler is geometrically ergodic. Geometric ergodicity can be demonstrated by using the following lemma, given in Hobert and Geyer (1998), who note that it is a special case of Lemma 15.2.8 of Meyn and Tweedie (1993).

Lemma 6.3. *Suppose the Markov chain $\{X_n : n = 0, 1, ...\}$ is Feller continuous. If for some positive function w that is unbounded off compact sets, $E[w(Xn+1)|Xn = x] \leq \rho w(x) + L$, for some $\rho < 1$ and $L < \infty$, then the Markov chain is geometrically ergodic.*

Feller continuity (Meyn and Tweedie, 1993, Chap. 6) is easily verified here (see Hobert and Casella, 1998). We now adapt the proof of Hobert and Geyer (1998) to establish the drift condition of Lemma 6.3, which will establish Proposition 4.2. Here, $a_1, b_1, a_2, b_2, \mu_0$ and λ_0 are known. Let $\boldsymbol{\theta^*} = \left(\theta_1^*, \ldots, \theta_p^*\right)'$ and let $\boldsymbol{\lambda_{\theta^*}} = \left(\lambda_{\theta_1^*}, \ldots, \lambda_{\theta_p^*}\right)'$. This transformed model is similar to the two stage hierarchy in Hobert and Geyer (1998) except for the relationship between μ and $\boldsymbol{\theta}$, and the inverted gamma prior on λ_{η_j}.

The univariate conditional densities that are required to use the Gibbs sampler are given by

$$
\begin{aligned}
\lambda_e|\boldsymbol{\lambda_{\theta^*}}, \boldsymbol{\theta^*}, \mu, \boldsymbol{y^*} &\sim \text{gamma}\left(\frac{p}{2} + a_2, b_2 + \frac{1}{2}\sum_{j=1}^{p}\left(y_j^* - \theta_j^*\right)^2\right), \\
\lambda_{\theta_j^*}|\lambda_e, \boldsymbol{\theta^*}, \mu, \boldsymbol{y^*} &\sim \text{inverted gamma}\left(a_1 + \frac{1}{2}, b_1 + \frac{1}{2}\left(\theta_j^* - v_j\mu\right)^2\right) \text{ for } j = 1, \ldots, p, \\
\mu|\lambda_e, \boldsymbol{\lambda_{\theta^*}}, \boldsymbol{\theta^*}, \boldsymbol{y^*} &\sim N\left(\frac{\boldsymbol{v'D_\tau^{-1}\theta^*} + \lambda_0\mu_0}{\boldsymbol{v'D_\tau^{-1}v} + \lambda_0}, \frac{1}{\boldsymbol{v'D_\tau^{-1}v} + \lambda_0}\right), \qquad (27) \\
\theta_j^*|\lambda_e, \boldsymbol{\lambda_{\theta^*}}, \mu, \boldsymbol{y^*} &\sim N\left(\frac{\lambda_e y_j^* + \frac{v_j}{\lambda_{\theta_j^*}}\mu}{\lambda_e + \frac{1}{\lambda_{\theta_j^*}}}, \frac{1}{\lambda_e + \frac{1}{\lambda_{\theta_j^*}}}\right) \text{ for } j = 1, \ldots, p.
\end{aligned}
$$

Here, $\boldsymbol{v'D_\tau^{-1}\theta^*} = \sum_{j=1}^{p} v_j\theta_j^*/\lambda_{\theta_j^*}$ and $\boldsymbol{v'D_\tau^{-1}v} = \sum_{j=1}^{p} v_j^2/\lambda_{\theta_j^*}$.

We consider a block Gibbs to update all of the normal components simultaneously. Let $\boldsymbol{\xi} = \left(\theta_1^*, \ldots, \theta_p^*, \mu\right)'$. From (27), we deduce that

$$\boldsymbol{\xi}|\lambda_e, \boldsymbol{\lambda_\eta} \sim N_{p+1}\left(\boldsymbol{\mu_\xi}, \Sigma_\xi\right),$$

where

$$\Sigma_\xi^{-1} = \begin{bmatrix} \boldsymbol{D}^2 & -\boldsymbol{v} \\ -\boldsymbol{v}' & \lambda_0 + \boldsymbol{v'D_\tau^{-1}v} \end{bmatrix},$$

and the mean $\boldsymbol{\mu_\xi}$ is the solution to

$$\Sigma_\xi^{-1}\boldsymbol{\mu_\xi} = \begin{pmatrix} \lambda_e y_1^* \\ \vdots \\ \lambda_e y_p^* \\ \lambda_0\mu_0 \end{pmatrix}, \tag{28}$$

where $\boldsymbol{D} = \text{diag}\left(\sqrt{\lambda_e + \frac{1}{\lambda_{\theta_j^*}}}\right)$ for $j = 1, \ldots, p$. Let

$$t = \sum_{j=1}^{p} \frac{\lambda_e \frac{v_j^2}{\lambda_{\theta_j^*}}}{\lambda_e + \frac{1}{\lambda_{\theta_j^*}}} = \sum_{j=1}^{p} \frac{v_j^2}{\lambda_{\theta_j^*}} - \sum_{j=1}^{p} \left(\frac{v_j}{\lambda_{\theta_j^*}}\right)^2 \left(\lambda_e + \frac{1}{\lambda_{\eta_j}}\right)^{-1},$$

then by using the Cholesky factorization, the precision matrix of $\boldsymbol{\xi}$ can be expressed as

$$\Sigma_\xi^{-1} = \boldsymbol{L}\boldsymbol{L}' = \begin{pmatrix} \boldsymbol{D} & 0 \\ \boldsymbol{c}' & \sqrt{\lambda_0 + t} \end{pmatrix} \begin{pmatrix} \boldsymbol{D} & \boldsymbol{c} \\ 0 & \sqrt{\lambda_0 + t} \end{pmatrix},$$

where $\boldsymbol{c} = \left(-\frac{v_1}{\lambda_{\theta_1^*}\sqrt{\lambda_e + \frac{1}{\lambda_{\theta_1^*}}}}, \ldots, -\frac{v_p}{\lambda_{\theta_p^*}\sqrt{\lambda_e + \frac{1}{\lambda_{\theta_p^*}}}}\right)'$. Thus, it can be shown that

$$\Sigma_\xi = \boldsymbol{L}^{-1\prime}\boldsymbol{L}^{-1} \text{ where } \boldsymbol{L}^{-1} = \begin{pmatrix} \boldsymbol{D}^{-1} & 0 \\ -\frac{\boldsymbol{c}'\boldsymbol{D}^{-1}}{\sqrt{\lambda_0+t}} & \frac{1}{\sqrt{\lambda_0+t}} \end{pmatrix}.$$

The upper bounds of the variance and covariances are

$$\begin{aligned}
\text{Var}\left(\mu|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right) &= \frac{1}{\lambda_0 + t} \leq \frac{1}{\lambda_0}, \\
\text{Var}\left(\theta_j^*|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right) &= \frac{1}{\lambda_e + \frac{1}{\lambda_{\theta_j^*}}} \left\{1 + \frac{v_j^2}{\lambda_{\theta_j^*}\left(\lambda_e + \frac{1}{\lambda_{\theta_j^*}}\right)(\lambda_0 + t)}\right\} \leq \frac{1}{\lambda_e + \frac{1}{\lambda_{\theta_j^*}}} + \frac{v_j^2}{\lambda_0}, \\
\text{Cov}\left(\theta_i^*, \theta_j^*|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right) &= \frac{\frac{v_i}{\lambda_{\theta_i^*}}\frac{v_j}{\lambda_{\theta_j^*}}}{\left(\lambda_e + \frac{1}{\lambda_{\eta_i}}\right)\left(\lambda_e + \frac{1}{\lambda_{\theta_j^*}}\right)(\lambda_0 + t)} \leq \frac{v_i v_j}{\lambda_0}, \\
\text{Cov}\left(\mu, \theta_j^*|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right) &= \frac{v_j}{\lambda_{\theta_j^*}\left(\lambda_e + \frac{1}{\lambda_{\theta_j^*}}\right)(\lambda_0 + t)} \leq \frac{v_j}{\lambda_0}.
\end{aligned}$$

We use (28) to calculate $E\left[\mu|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right]$ and $E\left[\theta_j^*|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right]$ for $j = 1, \ldots, p$.

$$\begin{aligned}
E\left[\mu|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right] &= \sum_{j=1}^{p} \lambda_e y_j^* \mathrm{Cov}\left(\mu, \theta_j^*|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right) + \lambda_0\mu_0 \mathrm{Var}\left(\mu|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right) \\
&= \frac{1}{\lambda_0 + t}\left(\sum_{j=1}^{p} \frac{\lambda_e y_j^* + \frac{v_j}{\lambda_{\theta_j^*}}}{\lambda_e + \frac{1}{\lambda_{\theta_j^*}}} + \lambda_0\mu_0\right) \leq C_1 < \infty \\
E\left[\theta_i^*|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right] &= \sum_{j=1}^{p} \lambda_e y_j^* \mathrm{Cov}\left(\theta_i^*, \theta_j^*|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right) + \lambda_0\mu_0 \mathrm{Cov}\left(\mu, \theta_i^*|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right) \\
&= \frac{\lambda_e y_i^*}{\lambda_e + \frac{1}{\lambda_{\theta_i^*}}} + \frac{v_i/\lambda_{\theta_i^*}}{\lambda_e + \frac{1}{\lambda_{\theta_i^*}}}\left\{\frac{1}{\lambda_0 + t}\left(\sum_{j=1}^{p} \frac{\lambda_e y_j^* + \frac{v_j}{\lambda_{\theta_j^*}}}{\lambda_e + \frac{1}{\lambda_{\theta_j^*}}} + \lambda_0\mu_0\right)\right\} \\
&\leq C_2 < \infty,
\end{aligned}$$

for $i = 1, \ldots, p$ and C_1 and C_2 are constants. $E\left[\mu|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right]$ is a convex combination of y_j^* and μ_0, and $E\left[\theta_i^*|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right]$ for $i = 1, \ldots, p$ is a convex combination of $E\left[\mu|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right]$ and y_i^*. Thus, these are uniformly bounded by a constant.

We now show that Lemma 6.3 works for our proposed model. Note that for our model, we have a block Gibbs sampler with a fixed scan that updates $\boldsymbol{\xi}$ then $(\lambda_e, \boldsymbol{\lambda_{\theta^*}})$. Given $\boldsymbol{\xi}$, λ_e and $\boldsymbol{\lambda_{\theta^*}}$ are independent, thus the order of update does not matter. To construct an energy function for a drift condition, we need to calculate some conditional expectations. Here, we compute the conditional expectation given the variables of the last iteration ("last") as

$$E\left[w\left(\lambda_e, \boldsymbol{\lambda_{\theta^*}}, \mu, \boldsymbol{\theta^*}\right)|\text{last}\right] = E\left\{\left[w\left(\lambda_e, \boldsymbol{\lambda_{\theta^*}}, \mu, \boldsymbol{\theta^*}\right)|\mu, \boldsymbol{\theta^*}, \text{last}\right]|\text{last}\right\}. \tag{29}$$

Define the functions

$$\begin{aligned}
\boldsymbol{w}_1 &= (w_{11}, \ldots, w_{1p})' = \left(\lambda_{\theta_1^*}, \ldots, \lambda_{\theta_p^*}\right)', \\
w_2 &= 1/\lambda_e, \\
\boldsymbol{w}_3 &= (w_{31}, \ldots, w_{3p})' = \left[(\theta_1^* - v_1\mu)^2, \ldots, (\theta_p^* - v_p\mu)^2\right]', \\
w_4 &= \sum_{j=1}^{p}\left(y_j^* - \theta_j^*\right)^2, \\
\boldsymbol{w}_5 &= (w_{51}, \ldots, w_{5p})' = \left(e^{c/\lambda_{\theta_1^*}}, \ldots, e^{c/\lambda_{\theta_p^*}}\right)', \\
w_6 &= e^{c\lambda_e},
\end{aligned}$$

where c is a positive constant. Consider the energy function $w = \boldsymbol{A}_1\boldsymbol{w}_1 + A_2 w_2 + \boldsymbol{A}_3\boldsymbol{w}_3 + A_4 w_4 + \boldsymbol{A}_5\boldsymbol{w}_5 + A_6 w_6$ where $\boldsymbol{A}_1$, $\boldsymbol{A}_3$, $\boldsymbol{A}_5$ are positive vectors and A_2, A_4 and A_6 are positive constants to be determined. With the same argument in Hobert and Geyer, it can be shown that the level set

$$\left\{(\mu, \boldsymbol{\theta^*}, \lambda_e, \boldsymbol{\lambda_{\theta^*}}) : w\left(\mu, \boldsymbol{\theta^*}, \lambda_e, \boldsymbol{\lambda_{\theta^*}}\right) \leq \gamma\right\}$$

is unbounded off compact sets.

Let $0 < c < \min\{b_1, b_2\}$, then

$$E\left(e^{c\lambda_e}|\mu, \boldsymbol{\theta^*}, \text{last}\right) = \left(\frac{b_2 + \frac{1}{2}\sum_{j=1}^{p}\left(y_j^* - \theta_j^*\right)^2}{b_2 + \frac{1}{2}\sum_{j=1}^{p}\left(y_j^* - \theta_j^*\right)^2 - c}\right)^{a_2+p/2} \leq \left(\frac{b_2}{b_2 - c}\right)^{a_2+p/2} = c_6, \tag{30}$$

where c_6 is a constant, and

$$E\left(e^{c/\lambda_{\theta_j^*}}|\mu, \boldsymbol{\theta^*}, \text{last}\right) = \left(\frac{b_1 + \frac{1}{2}\left(y_j^* - \theta_j^*\right)^2}{b_1 + \frac{1}{2}\left(y_j^* - \theta_j^*\right)^2 - c}\right)^{a_1+1/2} \leq \left(\frac{b_1}{b_1 - c}\right)^{a_1+1/2} = c_5, \tag{31}$$

where c_5 is a constant for $j = 1, \ldots, p$. To make this procedure works, a_1 needs to be $a_1 > 1/2$, then for $j = 1, \ldots, p$,

$$E\left(w_{1j}|\mu, \boldsymbol{\theta^*}, \text{last}\right) = \frac{b_1 + 1/2\left(\theta_j^* - v_j\mu\right)^2}{a_1 - 1/2} = \frac{2b_1 + w_{3j}}{2a_1 - 1}. \tag{32}$$

Since $a_2 + p/2 > 1$, we have

$$E\left(w_2|\mu, \boldsymbol{\theta^*}, \text{last}\right) = \frac{b_2 + \frac{1}{2}\sum_{j=1}^{p}\left(y_j^* - \theta_j^*\right)^2}{a_2 + p/2 - 1} = \frac{2b_2 + w_4}{2a_2 + p - 2}. \tag{33}$$

Now, for $j = 1, \ldots, p$

$$\begin{aligned} E\left(w_{3j}|\text{last}\right) &= \text{Var}\left(\theta_j^* - v_j\mu|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right) + \left\{E\left(\theta_j^* - v_j\mu|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right)\right\}^2 \\ &\leq \text{Var}\left(\theta_j^*|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right) \leq c_3^* + \left(\lambda_e + \frac{1}{\lambda_{\theta_j^*}}\right)^{-1} \leq c_3 + w_2, \end{aligned} \tag{34}$$

where c_3^* and c_3 are constants. Similarly,

$$\begin{aligned} E\left(w_4|\text{last}\right) &= \sum_{j=1}^{p}\text{Var}\left(y_j^* - \theta_j^*|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right) + \sum_{j=1}^{p}\left\{E\left(y_j^* - \theta_j^*|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right)\right\}^2 \\ &\leq \sum_{j=1}^{p}\text{Var}\left(\theta_j^*|\lambda_e, \boldsymbol{\lambda_{\theta^*}}\right) + c_4^* \leq c_4^{**} + \sum_{j=1}^{p}\left(\lambda_e + \frac{1}{\lambda_{\theta_j^*}}\right)^{-1} \\ &\leq c_4 + pw_2. \end{aligned} \tag{35}$$

Therefore, $E\left(w_{1j}|\text{last}\right) \leq C_1 + \delta_1 w_2$ and $E\left(w_2|\text{last}\right) \leq C_2 + \delta_2 w_2$ where

$$\delta_1 = \frac{1}{2a_1 - 1} < \infty \quad \text{and} \quad \delta_2 = \frac{p}{2a_2 + p - 2} < 1.$$

Now, there exists an $\epsilon > 0$ and a $0 < \rho < 1$ such that $\epsilon\left(p\delta_1 + 2p\right) + \delta_2 < \rho$. Therefore

$$\begin{aligned} & E\left[\epsilon\left(\sum_{j=1}^{p} w_{1j} + \sum_{j=1}^{p} w_{3j} + w_4\right) + w_2 + \sum_{j=1}^{p} w_{5j} + w_6 | \text{last}\right] \\ & \quad \leq C^* + \epsilon\left(p\delta_1 + p + p\right) w_2 + \delta_2 w_2 \\ & \quad \leq C^* + \rho w_2 \end{aligned}$$

which implies geometric ergodicity by Lemma 6.3.