Regularization Methods for Dimension Reduction

JEFF GILL
Distinguished Professor
Department of Government, Department of Mathematics & Statistics
Member, Center for Neuroscience and Behavior
Founding Director, Center for Data Science
American University

Introduction

- It is increasingly common to have more covariates (p) than cases (n).
- This is "big data" in the p sense.
- Regularization is the simultaneous process of selecting a subset of p and estimating the regression coefficients.
- This is done with penalties on the covariates such that the more important ones are featured.

Setup

- Consider a linear regression model with n observations on a dependent variable Y and p predictors:

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

  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 $\text{RSS} = (\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})'(\tilde{\boldsymbol{y}} - \boldsymbol{X}\boldsymbol{\beta})$ 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.

- Also, the variances will be artificially large.

The Basic LASSO

- 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

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

  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.

- Knight and Fu (2000, Annals) have shown consistency for lasso type 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.

The Basic LASSO

- For the computation of the lasso, Osborne *et al.* (2000a) proposed two algorithms:
  - A compact descent algorithm (solves a constrained optimization problem with constraint linearization) was derived to solve the selection problem for a particular value of the tuning parameter.
  - A homotopy method (an analytic approximation method for highly nonlinear problems using series expansion) for the tuning parameter was developed to completely describe the possible selection.

- Efron *et al.* (2004) proposed Least Angle Regression Selection (LARS) for a model selection algorithm.

- This algorithm is a piecewise linear solution path using a modification of forward stagewise and least angle regression paths.

- 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.

The Basic LASSO, Contrived Example

```
library(glmnet)
X <- matrix(rnorm(200),10,20)
y <- rnorm(10)
lasso.lm <- glmnet(X,y,family="gaussian",alpha=1)
lasso.lm
```

```
   Df  %Dev  Lambda
1   0  0.00 0.60190
2   1  4.03 0.57450
3   1  7.70 0.54840
4   1 11.04 0.52350
5   1 14.09 0.49970
6   1 16.86 0.47700
7   1 19.39 0.45530
:
93  9 99.83 0.00834
94  9 99.84 0.00796
95  9 99.86 0.00760
96  9 99.87 0.00725
97  9 99.88 0.00692
98  9 99.89 0.00661
99  9 99.90 0.00630
```

- (1) the number of nonzero coefficients (Df), (2) the percent (of null) deviance explained (%Dev), (3) and the value of Lambda.

The Basic LASSO, Contrived Example

```
coef(lasso.lm, s=0.5)
```

```
                  s1
(Intercept) 0.134222
V1          .
V2          .
V3          .
V4          .
V5          .
V6          .
V7          .
V8          .
V9          .
V10         .
V11         .
V12         .
V13         .
V14         .
V15         .
V16         .
V17         .
V18         .
V19         .
V20         .
```

The Basic LASSO, Contrived Example

```
coef(lasso.lm, s=0.005)
```

```
                     s1
(Intercept)  0.30546423
V1           0.31145715
V2           .
V3           .
V4          -0.78251180
V5           .
V6           .
V7           0.40352899
V8           .
V9           .
V10          0.59668660
V11         -0.20360788
V12          .
V13          .
V14          0.04669743
V15         -0.35153569
V16         -0.15161324
V17          .
V18          0.15144882
V19          .
V20          .
```

The Basic LASSO, State Failures Example

- 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, to produce a model with about 75% correct predictions of state failure (0/1) using the naïve criteria.

- 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).

The Basic LASSO, State Failures Data Setup

- These data are from the State Failure Task Force, 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.
- The final results of the SFTF team are controversial because they end up using only three explanatory variables, democracy, trade openness and infant mortality, to produce a model with about 75% correct predictions of state failure (0/1) using the na"ıve criteria.
- 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.

The Basic LASSO, State Failures Data Setup

- State Failures Data for 23 Asian countries:

```
sf.asia <- read.table("ARTICLES/Article.Lasso/Data.State.Failures/asia.dat",
           sep=" ", header=TRUE)
sf.asia <- sf.asia[,-1]
dim(sf.asia)
[1]  23 129
```

- Messy issues with missing data:

```
library(mice)
m <- 5; covars <- 50
mice.out <- mice(sf.asia[,1:covars],m)
mice.array <- array(NA,c(nrow(sf.asia),covars,m))
for (i in 1:m) mice.array[,,i] <- as.matrix(complete(mice.out,i))
for (i in 1:m) { for (j in 1:covars) mice.array[,j,i] <- random.imp.vec( mice.array
sum(is.na(mice.array))
[1] 0
```

The Basic LASSO, State Failures Data Setup

- Run the LASSO:

```
library{glmnet}
X <- mice.array[,-15,1]
y <- mice.array[,15,1]
y[y > 0] <- 1
lasso.lm <- glmnet(X,y,family=binomial(link=probit),alpha=1)
lasso.lm
```

```
   Df  %Dev  Lambda
1   0  0.00 0.60190
2   1  4.03 0.57450
3   1  7.70 0.54840
4   1 11.04 0.52350
5   1 14.09 0.49970
6   1 16.86 0.47700
7   1 19.39 0.45530
8   1 21.70 0.43460
:
92  9 99.81 0.00873
93  9 99.83 0.00834
94  9 99.84 0.00796
95  9 99.86 0.00760
96  9 99.87 0.00725
97  9 99.88 0.00692
98  9 99.89 0.00661
99  9 99.90 0.00630
```

The Basic LASSO, State Failures Data Setup

```
coef.glmnet(lasso.lm,s=0.05)
```

```
                     s1
(Intercept) -0.34062463
V1           0.35651339
V2           .
V3           .
V4           0.27364535
V5           .
V6          -0.04733199
V7          -0.04360697
V8           .
V9          -0.39910277
V10          0.40747002
V11          .
V12          .
V13          .
V14          .
V15         -0.57301993
V16         -0.13147005
V17          .
V18          0.30959403
V19          .
V20          .
```

The Basic LASSO, State Failures Data Setup

```
summary(sf.lm.out)$coefficients
              Estimate Std. Error    t value  Pr(>|t|)
(Intercept)  0.1449359  0.4164622  0.3480169 0.7610439
X[, 1]      -1.5169291  0.9988976 -1.5186032 0.2681881
X[, 4]      -1.1655855  1.7075531 -0.6826057 0.5653119
X[, 6]       0.9049659  1.1646306  0.7770412 0.5184506
X[, 7]      -0.1233321  0.9126464 -0.1351367 0.9048772
X[, 15]     -1.1244055  1.4816114 -0.7589072 0.5271529
X[, 16]     -0.6992929  0.6822413 -1.0249935 0.4131493
X[, 18]     -0.1458726  0.6178490 -0.2360975 0.8353328

summary(sf.lm.out)$r.squared
[1] 0.8151111

summary(sf.lm.out)$fstatistic
   value    numdf    dendf
1.259615 7.000000 2.000000
```

Fused LASSO

- 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.
- With meaningful ordering of the features (specification of consecutive predictors), the lasso ignores it.
- 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 these 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.

Group LASSO

- 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 argued that it does variable selection at the group level and is invariant under orthogonal transformations.

Elastic Net

- 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

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

  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).

Elastic Net with the State Failures Data, Equal Weighting of Penalties

```
elnet.lm <- glmnet(X,y,family=binomial(link=probit),alpha=0.5)
elnet.lm
```

```
      Df  %Dev Lambda
1     0  0.00 1.04600
2     1  2.52 0.99880
3     1  5.00 0.95340
4     1  7.39 0.91010
5     1  9.73 0.86870
6     1 12.00 0.82920
7     1 14.20 0.79150
8     1 16.34 0.75550
9     1 18.41 0.72120
10    2 20.58 0.68840
:
91   16 96.98 0.01590
92   16 97.13 0.01518
93   16 97.28 0.01449
94   16 97.41 0.01383
95   16 97.54 0.01320
96   16 97.66 0.01260
97   16 97.78 0.01203
98   17 97.88 0.01148
99   17 97.99 0.01096
100  17 98.09 0.01046
```

Elastic Net with the State Failures Data, Very Large Lambda

```
coef.glmnet(elnet.lm,s=200)

    Df  %Dev  Lambda
V1            .
V2            .
V3            .
V4            .
V5            .
V6            .
V7            .
V8            .
V9            .
V10           .
:
V40           .
V41           .
V42           .
V43           .
V44           .
V45           .
V46           .
V47           .
V48           .
V49           .
```

Adaptive LASSO

- 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 (an estimator that is consistent in variable selection is not necessarily consistent in parameter estimation terms of the asymptotic distribution) do not hold for the lasso.

- To obtain the oracle property, Zhou (2006) introduced the adaptive lasso estimator as

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

  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 ("performs as well as if the true underlying model were given in advance") and it leads to a near-minimax-optimal estimator (has near the least worst range of risk).

Other LASSOs

- 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.

- Balakrishnan and Madigan (2007) proposed a data-driven approach, the lasso with attribute partition search (LAPS) algorithm, by combining the fused lasso and the group lasso in particular types of structure classification problems.

- 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 $p \gg n$.

The Bayesian LASSO

- Tibshirani (1996) noted that with the L_1 penalty term in the basic form, 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.

The Bayesian LASSO of Park and Casella (2008)

- Estimation is actually easier with Gibbs sampling for the lasso with the Laplace prior in the hierarchical model.

- Park and Casella (2008) considered a fully Bayesian analysis using a conditional Laplace prior specification of the form:
$$\pi(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\lambda|\beta_j|/\sigma}$$
and the uninformative 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.

- Their point estimate recommendation is the posterior median.

Other Bayesian LASSOs

- Yuan and Lin (2005) give 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, and

- Raman *et al.* (2009) used a Bayesian group lasso for a contingency table analysis.

Standard Errors of the Lasso

- Tibshirani (1996) suggested the bootstrap (Efron, 1979) for the estimation of standard errors and also derived an approximate closed form estimate.
- Fan and Li (2001) presented the sandwich formula in the likelihood setting as an estimator for the covariance of the estimates.
- 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, but the distributions will be far from normally distributed.
- Pötscher and Leeb (2007) showed that the finite sample distribution of lasso parameters 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 at the location of the point mass at 0 so this does not give reasonable estimates for the covariance matrix of $\boldsymbol{\beta}$.
- Kyung, Gill, Casella, and Ghosh (2010) proved that the bootstrap standard errors of $\hat{\beta}_j$ are inconsistent if $\beta_j = 0$.

Basic Hierarchical Specification

- Consider 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.

- It is important to carefully parameterize Σ_β to obtain the lassos.

- 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.

Hierarchical Models and Gibbs Samplers

- A general version of the lasso model can be expressed as

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

where the specific choices of $h_1(\boldsymbol{\beta})$ and $h_2(\boldsymbol{\beta})$ are given by:

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	$\sum_{j=1}^{p} \lvert\beta_j\rvert$	$\sum_{j=2}^{p} \lvert\beta_j\rvert^2$

Bayesian Specification

- The Bayesian formulation of the original lasso, as given in Park and Casella (2008), is given by the following hierarchical model.

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

- 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.

- Any inverted gamma prior for σ^2 would maintain conjugacy, but here we will use the improper prior density $\pi(\sigma^2) = 1/\sigma^2$, with which we also can maintain propriety.

- The resulting prior on $\boldsymbol{\beta}$ is a Laplace distribution with mean 0 and variance $\sigma^2\lambda^{-2}$to assure uni-modality of the posterior while the unconditional version (without σ^2) does not.

Hierarchical Models

- Now consider how to represent lassos as a conjugate Bayesian hierarchy.
- For each model we need an 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 $\boldsymbol{\beta}$ denoted 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. This can easily be done within the Gibbs sampler. models; details on posterior distributions and Gibbs sampling are left to Appendix **??**

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

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

  where m_k is the dimension of G_k, the grouping matrix.

- 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.

Hierarchical Fused Lasso

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

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

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

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

Hierarchical Fused Lasso

- Here the $\tau_1^2, \ldots, \tau_p^2, \omega_1^2 \ldots, \omega_{p-1}^2$ 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.

Hierarchical Elastic Net

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

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

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

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

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

- In this case, $\boldsymbol{\beta}$ is not conditionally independent of λ_2, as it appears in the covariance matrix.

Tuning Parameters

- The lassos just discussed all have tuning parameters: λ_1 and λ_2.

- 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 can 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 implemented with an EM/Gibbs algorithm (Casella, 2001).

- The gamma priors on λ^2 are given by:

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

State Failures in Asia Again

- The final results of the SFTF team are controversial because they end up using only three explanatory variables: democracy, trade openness and infant mortality, to produce a model with about 75% correct predictions of state failure $(0/1)$ using the naïve criteria.

- 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, 18 variables provided no useful information and were dropped as well.

- So 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.

State Failures in Asia Again

- Interestingly, the LARS lasso picks three like the stepwise procedure of the State Failure Task Force, but they are three *different* explanatory variables:
  - `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
- For a more detailed explanation of these variables see `http://globalpolicy.gmu.edu/pitf`).
- 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.

State Failures in Asia Again

- This table provides the top ten variables by absolute posterior median effect from the Bayesian lasso, and also for these variables in the LARS lasso conclusion.

- The Bayesian Elastic Net produces results that are virtually indistinguishable from the Bayesian lasso for these data.

	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

State Failures in Asia Again

- Every coefficient credible interval of the Bayesian lasso covers zero, including the ten given in the table.
- 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.
- 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.

State Failures in Asia Again

- 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.

State Failures in Asia Again

- While picking the top ten effect sizes is arbitrary (in the table), there is a noticeable drop in magnitude after this era.

- The Bayesian lasso therefore brings some additional variables to our attention:

  - `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.

Final Notes on Bayesian Lassos

- Posterior means will never be exactly zero, so a method of selection based on that assumption cannot be achieved.

- If selection is desired, one strategy is to set equal to zero any coefficient estimate whose confidence (credible) interval includes zero.

- As the non-Bayesian lasso cannot produce valid standard errors if the true coefficients are zero, it cannot give any confidence assessment of these selections.

- 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 *and* a measure of uncertainty..

- Putting λ into the Gibbs iterations, the estimates of the regression coefficients are not based on a fixed value, but rather are marginalized over all λ, leading to somewhat of a robustness property.