The Bayesian Lasso

Rebecca C. Steorts
Predictive Modeling and Data Mining: STA 521

November 2015

- Recall the Lasso
- The Bayesian Lasso

The lasso

The lasso[1] estimate is defined as

$$
\begin{aligned}
\hat{\beta}^{\text{lasso}} &= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \; \|y - X\beta\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j| \\
&= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \; \underbrace{\|y - X\beta|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}
\end{aligned}
$$

The only difference between the lasso problem and ridge regression is that the latter uses a (squared) ℓ_2 penalty $\|\beta\|_2^2$, while the former uses an ℓ_1 penalty $\|\beta\|_1$. But even though these problems look similar, their solutions behave very differently

Note the name "lasso" is actually an acronym for: Least Absolute Selection and Shrinkage Operator

[1] Tibshirani (1996), "Regression Shrinkage and Selection via the Lasso"

$$\hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \; \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

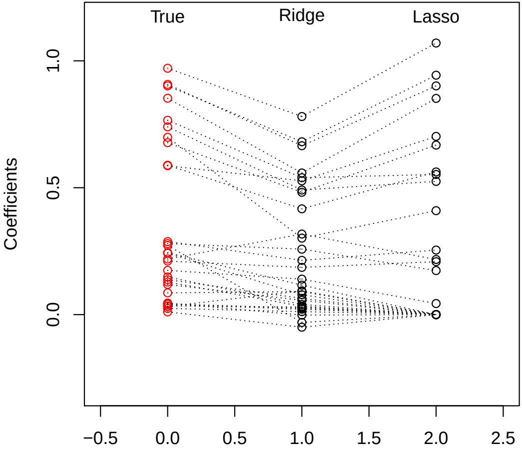
The tuning parameter λ controls the strength of the penalty, and (like ridge regression) we get $\hat{\beta}^{\text{lasso}} =$ the linear regression estimate when $\lambda = 0$, and $\hat{\beta}^{\text{lasso}} = 0$ when $\lambda = \infty$

For λ in between these two extremes, we are balancing two ideas: fitting a linear model of y on X, and shrinking the coefficients. But the nature of the ℓ_1 penalty causes some coefficients to be shrunken to zero exactly

This is what makes the lasso substantially different from ridge regression: it is able to perform variable selection in the linear model. As λ increases, more coefficients are set to zero (less variables are selected), and among the nonzero coefficients, more shrinkage is employed

Example: visual representation of lasso coefficients

Our running example from last time: $n = 50$, $p = 30$, $\sigma^2 = 1$, 10 large true coefficients, 20 small. Here is a visual representation of lasso vs. ridge coefficients (with the same degrees of freedom):



Important details

When including an intercept term in the model, we usually leave this coefficient unpenalized, just as we do with ridge regression. Hence the lasso problem with intercept is

$$\hat{\beta}_0, \hat{\beta}^{\text{lasso}} = \underset{\beta_0 \in \mathbb{R},\, \beta \in \mathbb{R}^p}{\operatorname{argmin}} \; \|y - \beta_0 \mathbb{1} - X\beta\|_2^2 + \lambda \|\beta\|_1$$

As we've seen before, if we center the columns of X, then the intercept estimate turns out to be $\hat{\beta}_0 = \bar{y}$. Therefore we typically center y, X and don't include an intercept them

As with ridge regression, the penalty term $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ is not fair is the predictor variables are not on the same scale. Hence, if we know that the variables are not on the same scale to begin with, we scale the columns of X (to have sample variance 1), and then we solve the lasso problem

Bias and variance of the lasso

Although we can't write down explicit formulas for the bias and variance of the lasso estimate (e.g., when the true model is linear), we know the general trend. Recall that

$$\hat{\beta}^{\text{lasso}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \; \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

Generally speaking:

- The bias increases as λ (amount of shrinkage) increases
- The variance decreases as λ (amount of shrinkage) increases

What is the bias at $\lambda = 0$? The variance at $\lambda = \infty$?

In terms of prediction error (or mean squared error), the lasso performs comparably to ridge regression

Bayesian Lasso

Tibshirani (1996) suggested that Lasso estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace (i.e., double-exponential) priors.

Tibshirani and the Bayesian Lasso

Specifically, the lasso estimate can be viewed as the mode of the posterior distribution of β

$$\hat{\beta}_L = \arg\max_{\beta} p(\beta \mid y, \sigma^2, \tau)$$

when

$$p(\beta \mid \tau) = (\tau/2)^p \exp(-\tau ||\beta||_1)$$

and the likelihood on

$$p(y \mid \beta, \sigma^2) = N(y \mid X\beta, \sigma^2 I_n).$$

For any fixed values $\sigma^2 > 0, \tau > 0$, the posterior mode of β is the lasso estimate with penalty $\lambda = 2\tau\sigma^2$.

(Details – homework).

The Bayesian lasso[2] was motivated by a a conditional Laplace prior where

$$\pi(\beta \mid \sigma^2) = \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}}$$

Note: conditioning on σ^2 is important as it ensures that the full posterior is unimodal.

Lack of unimodality slows convergence of the Gibbs sampler and makes point estimates less meaningful.

[2] Park and Casella (2008)

Diabetes data

- The diabetes data contains 442 patients that we measured on 10 baseline variables.
- Examples are age, sex, BMU, BP, etc.
- The response is a measure of disease progression one year after baseline.

Full model

$$\mathbf{y}|\mu, \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}_n(\mu \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n),$$

$$\boldsymbol{\beta}|\sigma^2, \tau_1^2, \ldots, \tau_p^2 \sim \mathrm{N}_p(\mathbf{0}_p, \sigma^2 \mathbf{D}_\tau),$$

$$\mathbf{D}_\tau = \mathrm{diag}(\tau_1^2, \ldots, \tau_p^2),$$

$$\sigma^2, \tau_1^2, \ldots, \tau_p^2 \sim \pi(\sigma^2)\, d\sigma^2 \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2}\, d\tau_j^2,$$

$$\sigma^2, \tau_1^2, \ldots, \tau_p^2 > 0.$$

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

Comparisons

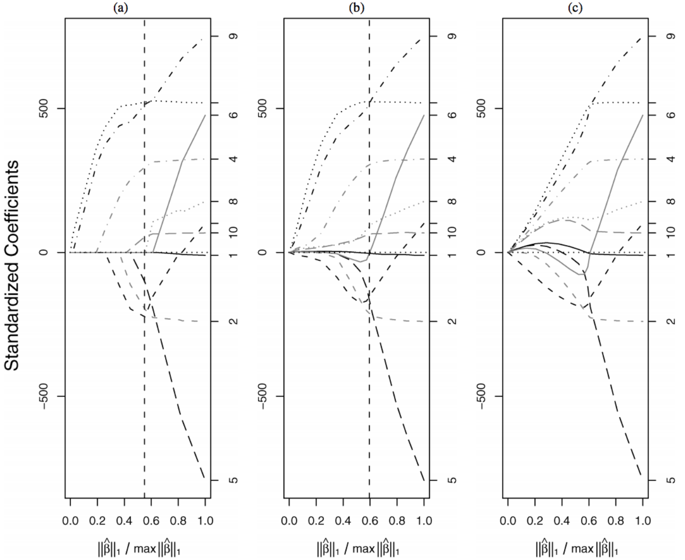


Figure: Lasso (a), Bayesian Lasso (b), and ridge regression (c) trace plots

Comparisons on the Diabetes data

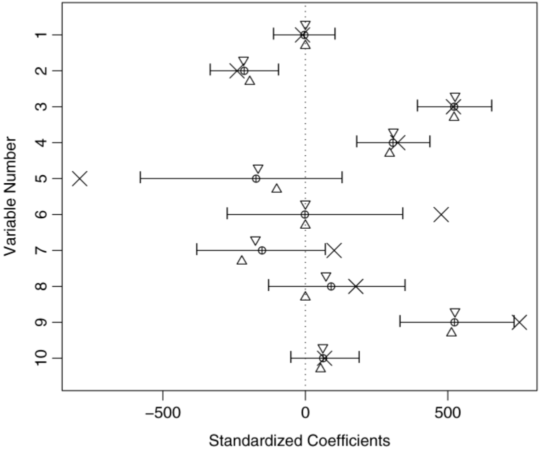


Figure: Posterior median Bayesian Lasso estimates, and corresponding 95% credible intervals (equal-tailed).

Running this in `R`

The lasso, Bayesian lasso, and extensions can be done using the `monomvn` package in `R`.

In lab we will do an example of comparing and contrasting the lasso with the Bayesian lasso.

- Results from the Bayesian Lasso are strikingly similar to those from the ordinary Lasso.
- Although more computationally intensive, the Bayesian Lasso is easy to implement and automatically provides interval estimates for all parameters, including the error variance.
- We did not cover this, but in the paper there are proposed methods for choosing λ (Bayesian lasso).
- These could aid in choosing λ for the lasso as well and results may be more stable.