The Bayesian Lasso

Trevor Park and George Casella

Annelise Wagner
University of Washington Statistics

April 4, 2016

A Familiar Framework

Have some response vector $\mathbf{y}$ $(n \times 1)$ which we describe as a linear combintation of standardized regressors $\mathbf{X}$ $(n \times p)$ and iid error terms $\epsilon \sim \mathrm{N}(0, \sigma^2)$ $(n \times 1)$

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

Interested in estimating regression parameters $\beta^T = (\beta_1, ..., \beta_p)$

Ordinary Least Squares Regression

Standard regression offers a simple solution

$$\min_{\beta}(\tilde{y} - \mathbf{X}\beta)^T(\tilde{y} - \mathbf{X}\beta)$$

This approach has a number of shortcomings:

- Intractable when $p > n$ (More predictors than data points)
- Superfluous predictors will have non-zero regression parameters
- Correlation structure in $\mathbf{X}$ hampers our ability to estimate β.

Audience Participation: How do I make those βs bold?

Alternative Approaches

- Subset Regression

$$\beta_k = \beta_k^{OLS} \text{ if } |\hat{\beta}_k^{OLS}| > t \text{ and } \hat{\beta}_k = 0 \text{ otherwise}$$

- Non-negative Garrote

$$\min \sum_{i=1}^{n} (y_i - \sum_{k}^{p} c_k \hat{\beta}_k^{OLS} x_{ik}) \text{ with } c_k \geq 0, \sum c_k \leq t < \sum |\hat{\beta}_k^{OLS}|$$

- Lasso

$$\min \sum_{i=1}^{n} (y_i - \sum_{k}^{p} \hat{\beta}_k x_{ik}) \text{ with } \sum |\hat{\beta}_k| \leq t$$

- Ridge Regression

$$\sum_{i=1}^{n} (y_i - \sum_{k}^{p} \hat{\beta}_k x_{ik}) \text{ with } \sum |\hat{\beta}_k|^2 \leq t$$

Example

Considering a toy example data set where $X^T X = I$ reveals the differing behavior of estimates

Ridge Regression (b):

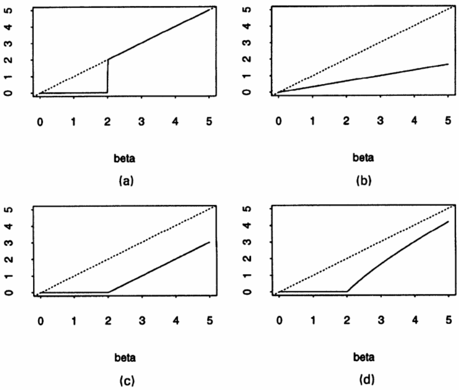
$$\hat{\beta}_k = \frac{1}{1+\gamma}\hat{\beta}_k^{OLS}$$

Lasso (c):

$$\hat{\beta}_k = \mathrm{sign}(\hat{\beta}_k^{OLS})(|\hat{\beta}_k^{OLS}| - \gamma)^+$$

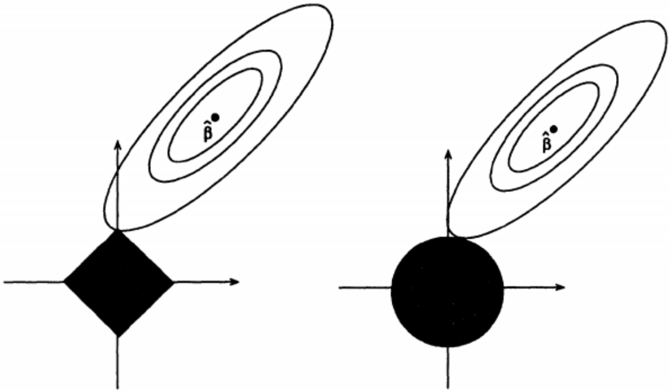
Garrote (d):

$$\hat{\beta}_k = \left(1 - \frac{\gamma}{\hat{\beta}_k^{2(OLS)}}\right)^+ \hat{\beta}_k^{OLS}$$



Driving Estimates to Zero

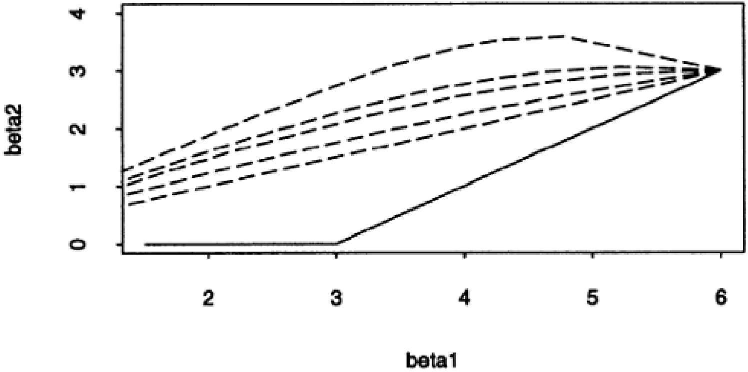
The behavior of the Lasso (left) and Garrote are preferable to Ridge Regression (right) as they drive estimates to 0 rather than near 0.



(a) (b)

Behavior with Correlation

The Lasso predictor (solid) also functions consistantly under correlated data, unlike Ridge Regression (dashed, displayed with varying levels of correlation)



Standard Form Representation

We can reexpress the Lasso as a single convex optimization problem

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{k}^{p} \hat{\beta}_k x_{ik}) \text{ with } \sum |\hat{\beta}_k| \leq t$$

is equivalent to

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \sum_{k}^{p} \hat{\beta}_k x_{ik}) + \lambda \sum_{k}^{p} |\hat{\beta}_k|$$

for some value of $\lambda \geq 0$ determined by t.

Bayesian Interpretation

We can interpret Lasso estimates as the modes of posterior distributions where the βs have double exponential prior distributions

The Bayesian Lasso expands on this idea by actually doing it, rather than just thinking about it

- Allows for Bayesian interpretation of parameter estimates
- Gibbs Sampler provides "reasonably fast" convergence (Fast enough for statistics)

Hyperparameter Selection

- Empirical Bayes approach meshes with Gibbs sampler to produce marginal maximum likelyhood estimates of hyperparameters
- Alternatively, hyperpriors provide credible intervals for estimation of ideal hyperparameter values
- These should produce consistant confidence intervals or credible intervals (But we shall see...)

Moving Forward

- Ridge Regression has a similar Bayesian interpretation, as mean estimates of posteriors when βs are given normal priors
- Can expand this framework and Gibbs sampling design beyond Lasso to compare against other methods

References

- George Casella, *Empirical bayes gibbs sampling*, Biostatistics **2** (2001), no. 4, 485–500.
- Wenjiang J Fu, *Penalized regressions: the bridge versus the lasso*, Journal of computational and graphical statistics **7** (1998), no. 3, 397–416.
- Trevor Park and George Casella, *The bayesian lasso*, Journal of the American Statistical Association **103** (2008), no. 482, 681–686.
- Robert Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) (1996), 267–288.