

The Bayesian Lasso

Trevor Park & George Casella

Dusty Turner

2023-04-17

Before we begin...

MODIFIED BAYES' THEOREM:

$$P(H|X) = P(H) \times \left(1 + P(C) \times \left(\frac{P(X|H)}{P(X)} - 1 \right) \right)$$

H: HYPOTHESIS

X: OBSERVATION

P(H): PRIOR PROBABILITY THAT H IS TRUE

P(X): PRIOR PROBABILITY OF OBSERVING X

P(C): PROBABILITY THAT YOU'RE USING
BAYESIAN STATISTICS CORRECTLY

The Bayesian Lasso¹

1. Formulation
 - Classical Regression
 - Classical Lasso
 - Bayesian Lasso
2. Selecting λ
 - Classical Regression
 - Classical Lasso
3. Comparison
4. Extensions
5. Synthetic Example

¹Park and Casella (2008)

Classical Regression

$$y = \mu 1_n + X\beta + \epsilon$$

- ▶ y is an $n \times 1$ vector of responses
- ▶ μ is the overall mean
- ▶ X is the $n \times p$ matrix of **standardized** regressors
- ▶ $\beta = (\beta_1, \dots, \beta_p)^T$
- ▶ ϵ is an $n \times 1$ vector of $\stackrel{iid}{\sim} N(0, \sigma^2)$

Satisfies

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

Classical Lasso

Formulation

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

$$\lambda \geq 0$$

Classical Lasso

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

Notes

1. Often called “penalized regression”
2. L1 penalty
3. “Shrinkage” β values are shrunk towards 0
4. Tune λ through cross validation

Motivation

1. Model selection - often as a precursor to other models
2. Reduce overfitting
3. Easily extendable to generalized linear models

Classical Lasso

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

Drawbacks

1. Biases β
2. Unreliable standard errors (issues with statistical tests)
3. Correlated features
4. Tuning issues / time

Bayesian Lasso²

Hierarchical Specification 1 (1 of 2)

$$y|\mu, X, \beta, \sigma^2 \sim N_n(\mu 1_n + X\beta, \sigma^2 I_n)$$

$$\beta|\sigma^2, \tau_1^2, \dots, \tau_p^2 \sim N_p(0_p, \sigma^2 D_t)$$

$$D_t = \text{diag}(\tau_1^2, \dots, \tau_p^2)$$

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

$$\sigma^2, \tau_1^2, \dots, \tau_p^2 > 0$$

²Andrews and Mallows (1974)

Bayesian Lasso

Hierarchical Specification 1 (2 of 2)

The authors integrate out $\tau_1^2, \dots, \tau_p^2$ which yields the conditional prior for β as a Laplace (double-exponential) distribution:

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

$$\pi(\sigma^2) = IG(\alpha, \beta)$$

$$\pi(\mu) = U(a, b)$$

Bayesian Lasso³

Hierarchical Specification 2 (1 of 2)

$$y|\mu, X, \beta, \sigma^2 \sim N_n(\mu 1_n + X\beta, \sigma^2 I_n)$$

Authors integrate out μ

$$p(\beta) = N(A^{-1}X^T\tilde{y}, \sigma^2 A^{-1})$$

where

$$A = X^T X + D_\tau^{-1}$$

³Bae and Mallick (2004)

Bayesian Lasso

Hierarchical Specification 2 (2 of 2)

$$p(\sigma^2) = IG(\frac{n-1}{2} + \frac{p}{2}, (\tilde{y} - X\beta)^T \frac{(\tilde{y} - X\beta)}{2} + \beta^T D_\tau^{-1} \frac{\beta}{2})$$

$$p(\tau_1^2, \dots, \tau_p^2) = \sqrt{\frac{\lambda'}{2\pi}} x^{-\frac{3}{2}} \exp\left\{-\frac{\lambda'(x - \mu')^2}{2(\mu')^2 x}\right\}$$

where

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

$$\lambda' = \lambda^2$$

Choosing the Lasso Parameter: Classical Lasso

Cross Validation

1. Cross validate over a grid of λ where $\lambda \geq 0$
2. For each λ value find the error metric of interest
3. Select the λ value that minimizes the metric of interest

Choosing the Lasso Parameter: Bayesian Lasso

Technique 1: Empirical Bayes

Choosing the Lasso Parameter: Bayesian Lasso

Technique 1: Empirical Bayes



Choosing the Lasso Parameter: Bayesian Lasso

Technique 1: Empirical Bayes⁴

- ▶ Solve for a marginal maximum likelihood for λ using estimates of the hyperparameters
- ▶ λ is updated for each iteration using estimates from the sample of the previous iteration

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

- ▶ Recommended initial value of:

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

- ▶ $\hat{\beta}_j^{LS}$ and $\hat{\sigma}_{LS}^2$ are estimated from least squares

⁴Casella (2001)

Choosing the Lasso Parameter: Bayesian Lasso

Technique 2: Hyperpriors

Authors recommend the diffuse Gamma hyperprior of λ^2 in the following form

$$\pi(\lambda^2) = \frac{\delta^r}{\Gamma(r)} (\lambda^2)^{r-1} e^{-\delta \lambda^2}$$

$$\lambda^2 > 0, r > 0, \delta > 0$$

- ▶ Select r and δ such that there is high probability near the maximum likelihood estimate to avoid mixing problems
- ▶ $r = 0$ and $\delta = 0$ are tempting but lead to an improper posterior
- ▶ This formulation allows easy integration into a Gibbs sampler

Comparison

Consider the following data:⁵ ⁶

obs	age	sex	bmi	bp	s1	s2	s3	s4	s5
-0.01	0.80	1.06	1.30	0.46	-0.93	-0.73	-0.91	-0.05	0.42
-1.00	-0.04	-0.94	-1.08	-0.55	-0.18	-0.40	1.56	-0.83	-1.43
-0.14	1.79	1.06	0.93	-0.12	-0.96	-0.72	-0.68	-0.05	0.06
0.70	-1.87	-0.94	-0.24	-0.77	0.26	0.52	-0.76	0.72	0.48
-0.22	0.11	-0.94	-0.76	0.46	0.08	0.33	0.17	-0.05	-0.67
-0.72	-1.95	-0.94	-0.85	-0.41	-1.45	-1.67	0.87	-1.60	-0.86

- ▶ Measurements of 440 diabetic patients
- ▶ 10 baseline variables (centered and scaled)
- ▶ Response variable is a measure of disease progression one year after baseline

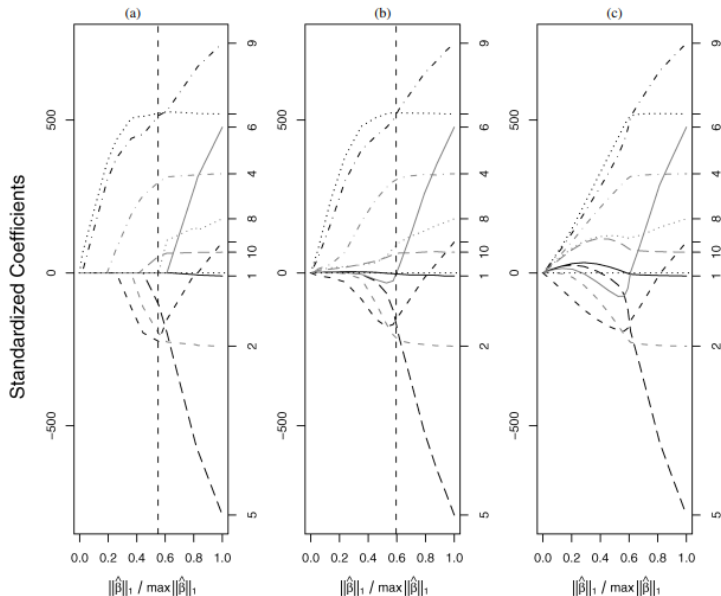
⁵{care} Efron et al. (2004)

⁶Zuber and Strimmer. (2021)

Trace Plot of Coefficients by Lasso

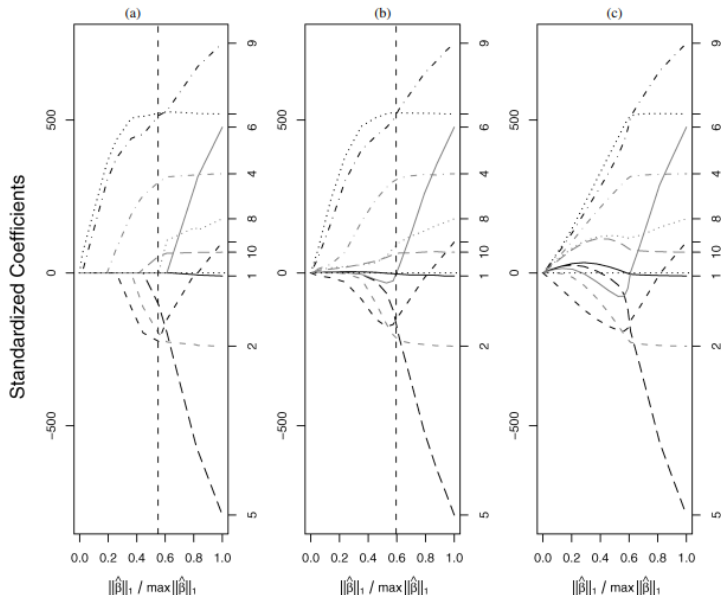
- a) Lasso
- b) Bayesian Lasso
- c) Ridge Regression

Vertical lines for the Lasso and Bayesian Lasso indicating the estimates chosen by n-fold cross-validation and marginal maximum likelihood



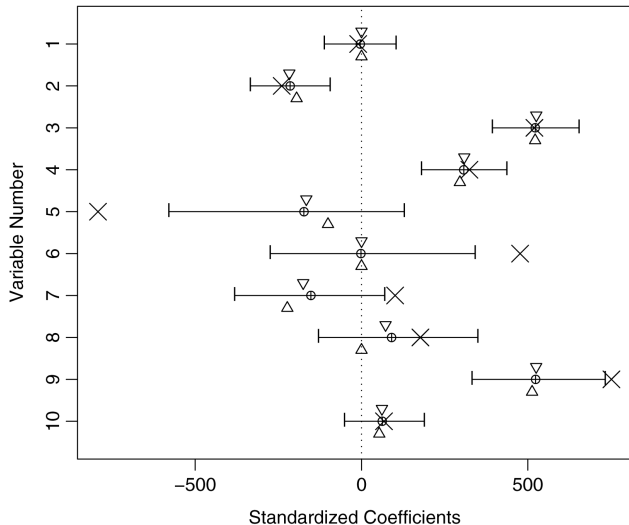
Comparison

- ▶ The Bayesian Lasso estimates appear to be a compromise between the Lasso and ridge regression estimates
- ▶ The Bayesian Lasso appears to pull the more weakly related parameters to 0 faster than ridge regression



Comparison

- ▶ Least squares estimates (\times),
- ▶ Lasso estimates based on n-fold cross-validation (\triangle),
- ▶ Lasso estimates chosen to match the L1 norm of the Bayes estimates ()
- ▶ Posterior median Bayesian Lasso estimates (\otimes) and corresponding 95% credible intervals (equal-tailed) with λ selected according to marginal maximum likelihood



Extensions

“Bridge” Regression⁷

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

$$\pi(\beta|\sigma^2) \propto \prod_{j=1}^P e^{-\lambda(\frac{|\beta_j|}{\sqrt{\sigma^2}})^2}$$

Huberized Lasso⁸

$$\min_{\beta} \sum_{i=1}^n L(\tilde{y}_i - x_i^T \beta) + \lambda \sum_{j=1}^p |\beta_j|$$

⁷Knight and Fu (2000)

⁸Rosset and Zhu (2007)

Example

To compare the results, we will generate synthetic data in the form:

$$\text{logit}(p) = \log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

```
num <- 100
```

```
x1 <- rnorm(num)
```

```
x2 <- rnorm(num)
```

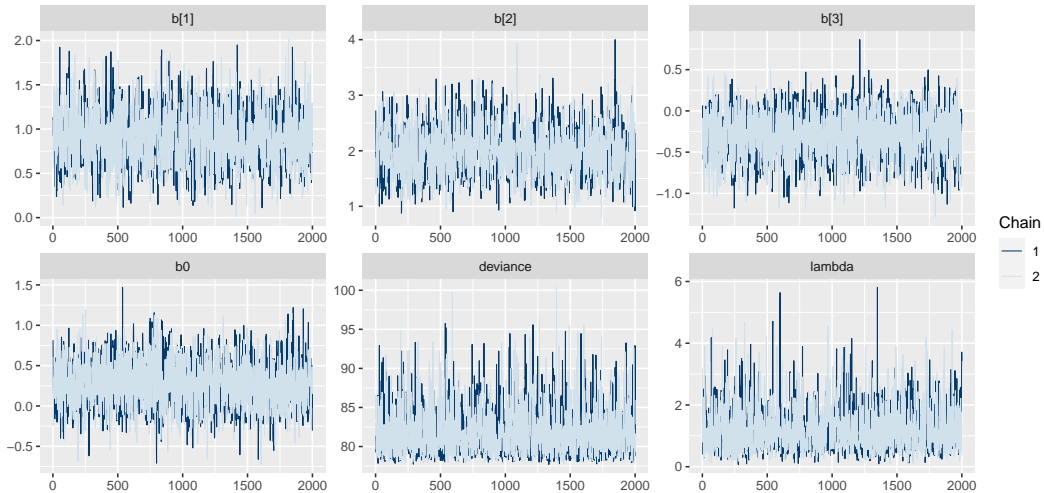
```
x3 <- rnorm(num)
```

```
prob <- exp(x1+2*x2) / (1+exp(x1+2*x2))
```

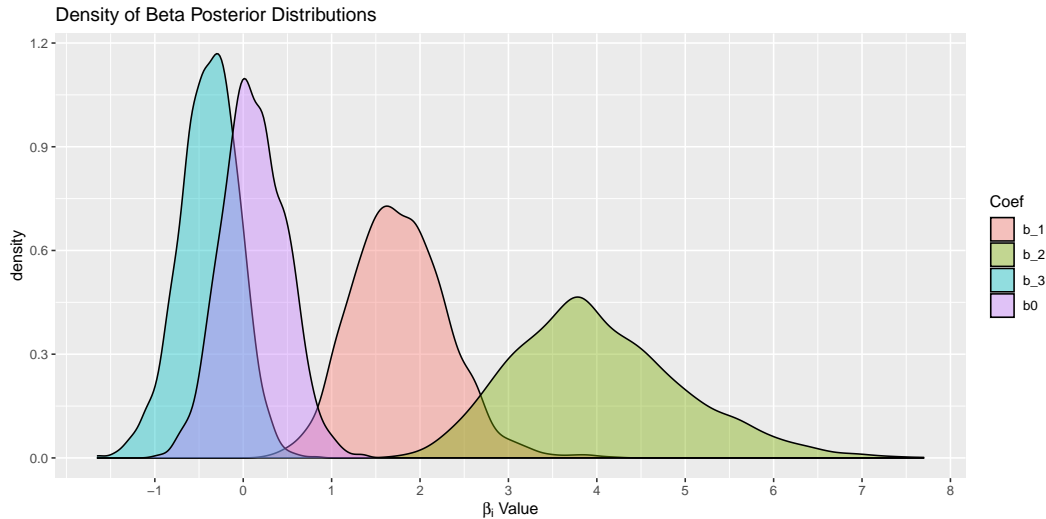
```
y <- rbinom(num, 1, prob)
```

Example

Trace Plots



Example



Example

model	b0	b_1	b_2	b_3
Logistic Regression	0.15	2.01	4.36	-0.49
Lasso	0.10	1.70	3.71	-0.36
Bayes Lasso	0.12	1.78	4.02	-0.39
Truth	0.00	2.00	4.00	0.00

Questions

IN MATH,
IT'S A ROTATED V;
IN SOCIETY,
IT'S A FEELING OF
SOME MARGINALIZED OR
UNDERREPRESENTED
PEOPLE

P IS FOR THIS IN
BAYES' THEOREM,
WHICH CAN BE USED
TO JUDGE HOW LIKELY
RAIN IS TODAY OR
YOUR CHANCES OF
GETTING MUMPS

Bibliography

- Andrews, D. F., and C. L. Mallows. 1974. "Scale Mixtures of Normal Distributions." *Journal of the Royal Statistical Society. Series B (Methodological)* 36 (1): 99–102. <http://www.jstor.org/stable/2984774>.
- Bae, Kyoung-hwa, and Bani K. Mallick. 2004. "Gene selection using a two-level hierarchical Bayesian model." *Bioinformatics* 20 (18): 3423–30. <https://doi.org/10.1093/bioinformatics/bth419>.
- Casella, George. 2001. "Empirical Bayes Gibbs sampling." *Biostatistics* 2 (4): 485–500. <https://doi.org/10.1093/biostatistics/2.4.485>.
- Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. "Least angle regression." *The Annals of Statistics* 32 (2): 407–99. <https://doi.org/10.1214/0090536040000000067>.
- Knight, Keith, and Wenjiang Fu. 2000. "Asymptotics for Lasso-Type Estimators." *The Annals of Statistics* 28 (5): 1356–78. <http://www.jstor.org/stable/2674097>.
- Park, Trevor, and George Casella. 2008. "The Bayesian Lasso." *Journal of the American Statistical Association* 103 (482): 681–86. <https://doi.org/10.1198/0162145080000000337>.
- Rosset, Saharon, and Ji Zhu. 2007. "Piecewise linear regularized solution paths." *The Annals of Statistics* 35 (2): 1012–30.