

BIOS 835: Bayesian Variable Selection Methods

Alexander McLain

September 14, 2023

Other penalties

- ▶ Last class, we went over Ridge, LASSO, and elastic net regression models.
- ▶ The optimization functions were all of the form

$$\operatorname{argmin}_{\beta} \{ (\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta) + p(\beta; \lambda) \},$$

where $p(\beta; \lambda)$ is a penalty function.

- ▶ There are other options for $p(\beta; \lambda)$.

Minimax Concave Penalty (MCP)

- ▶ Given a tuning parameter $\lambda > 0$ and a parameter $a > 1$, the MCP penalty for a coefficient β is defined as:

$$p_{MCP}(\beta; \lambda, a) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2a} & \text{if } |\beta| \leq a\lambda \\ \frac{a\lambda^2}{2} & \text{if } |\beta| > a\lambda \end{cases}$$

- ▶ The first segment is concave, and the penalty becomes constant after $|\beta| > a\lambda$.
- ▶ The parameter a controls the concavity and thus influences the amount of regularization applied to the coefficients.

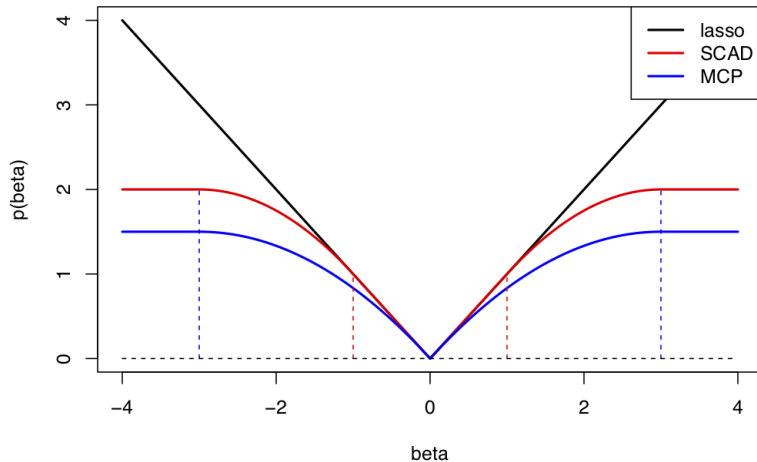
Smoothly Clipped Absolute Deviation (SCAD)

- ▶ Given a tuning parameter $\lambda > 0$ and a parameter $a > 2$, the SCAD penalty for a coefficient β is defined as:

$$p_{SCAD}(\beta; \lambda, a) = \begin{cases} \lambda|\beta| & \text{if } |\beta| \leq \lambda \\ -\left(\frac{\beta^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)}\right) & \text{if } \lambda < |\beta| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta| > a\lambda \end{cases}$$

- ▶ The SCAD penalty starts linearly (similar to the L1 penalty) but then bends down and becomes flat after $|\beta| > a\lambda$.
- ▶ The SCAD penalty is continuous everywhere, making it smoother than the MCP.

Lasso, SCAD and MCP penalties



Bayesian Variable Selection Methods



$PR(DATA|HYP.)$



$PR(HYP.|DATA)$

imgflip.com

Intro to Bayes

- ▶ The foundation of Bayesian analysis is **Bayes' theorem**, which relates the prior, likelihood, and posterior distributions:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

Where:

- ▶ $P(\theta|D)$ is the **posterior distribution** of the parameters θ given data D .
- ▶ $P(D|\theta)$ is the **likelihood** of the data D given the parameters θ .
- ▶ $P(\theta)$ is the **prior distribution** of the parameters θ , representing our beliefs or knowledge about θ before observing the data.
- ▶ $P(D)$ is the **marginal likelihood** or evidence, a normalization constant ensuring the posterior integrates (or sums) to 1.

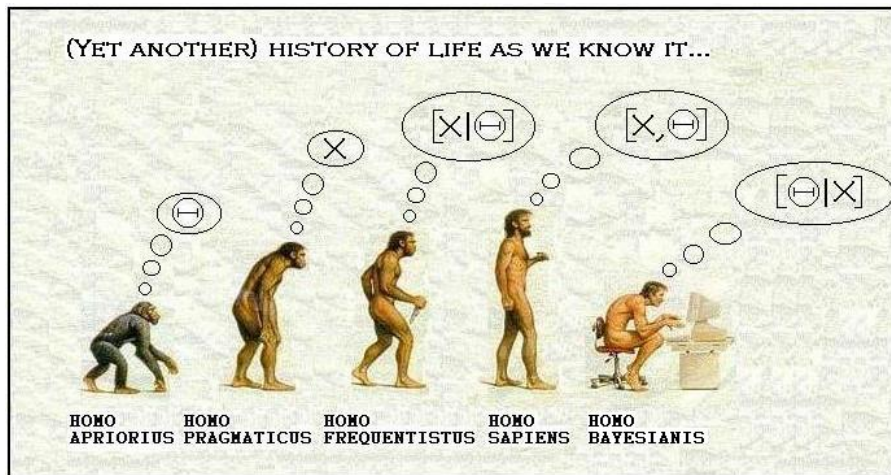


Figure: “A Bayesian is one who – vaguely expecting a horse – and catching a glimpse of a donkey, strongly believes he has seen a mule.”

Bayesian Linear Regression

- ▶ Bayesian Linear Regression (BLR) offers a probabilistic perspective on linear regression, focusing on distributions over parameters rather than point estimates.
- ▶ Assume a linear relationship between a response y and predictor variables \mathbf{x} :

$$y = \boldsymbol{\beta}^T \mathbf{x} + \epsilon$$

Where:

- ▶ $\boldsymbol{\beta}$ is the vector of coefficients.
- ▶ ϵ is the error term, typically assumed to be normally distributed with zero mean and variance σ^2 : $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Bayesian Linear Regression

- ▶ Given the error term's assumption, the likelihood of observing data $\mathbf{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ given parameters β is:

$$P(\mathbf{y}|\mathbf{X}, \beta, \sigma^2) = \prod_{i=1}^N \mathcal{N}(y_i | \beta^T \mathbf{x}_i, \sigma^2)$$

- ▶ We place a prior on the weights β . A common choice is a Gaussian prior:

$$\beta \sim \mathcal{N}(\mathbf{0}, \lambda^2 \mathbf{I})$$

Where λ^2 is a hyperparameter controlling the strength of the prior (smaller values result in stronger regularization).

Bayesian Linear Regression

- Using Bayes' theorem, the posterior distribution of β given the data is proportional to the product of the likelihood and the prior:

$$P(\beta|\mathbf{D}, \sigma^2) \propto P(\mathbf{y}|\mathbf{X}, \beta, \sigma^2)P(\beta)$$

where

$$P(y|X, \beta, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right)$$

- As a result,

$$P(\beta|\mathbf{D}, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) + \log P(\beta)\right)$$

Bayesian Linear Regression

- Notice that the β value that maximizes the posterior – called the maximum a posteriori probability (MAP) estimate – is

$$\beta_{map} = \min_{\beta} \left[(y - X\beta)^T (y - X\beta) - 2\sigma^2 \log P(\beta) \right]$$

- If $P(\beta) = MVN(0, \lambda^2 \mathbf{I})$, then the posterior is a *MVN* with mean and covariance:

$$\mu_{post} = \left(\mathbf{X}^T \mathbf{X} + \frac{\sigma^2}{\lambda^2} \mathbf{I} \right)^{-1} \mathbf{X}^T \mathbf{y}$$

$$\Sigma_{post} = \sigma^2 \left(\mathbf{X}^T \mathbf{X} + \frac{\sigma^2}{\lambda^2} \mathbf{I} \right)^{-1}$$

Connection to lasso and other penalization methods

- ▶ If we assume a Laplace (or double-exponential) prior on β , the MAP is analogous to Lasso regression.
- ▶ The Laplace distribution with scale parameter b is defined as:

$$P(\beta_j) \propto \exp\left(-\frac{|\beta_j|}{b}\right).$$

- ▶ The “penalty” for β_j is then

$$-2\sigma^2 \log P(\beta_j) \equiv \frac{2\sigma^2}{b} |\beta_j|$$

Again, the regularization strength λ can be linked with the scale parameter b and the error variance σ^2 .

Spike-and-slab priors

- ▶ Simply replicating penalization methods with Bayesian estimation techniques has benefits.
- ▶ That being said, Bayesian regression can go further.
- ▶ A common prior to use is the spike-and-slab prior for β , where

$$p(\beta_j) = \gamma_j f_1(\beta_j) + (1 - \gamma_j) f_0(\beta_j)$$

Some examples are: $f_j = N(0, \sigma_{bj}^2)$ where $\sigma_{b1}^2 \gg \sigma_{b0}^2$, or $f_0 = \delta_0$ a point mass at zero.

- ▶ Here, the problem changes to estimating γ and β , however, given γ estimating β is straightforward.
- ▶ What is $E(\gamma_j = 1) = \Pr(\gamma_j = 1)$?

Horseshoe priors

- ▶ The horseshoe prior for a coefficient β_j is defined as follows:

$$\beta_j | \lambda_j, \tau \sim \mathcal{N}(0, \lambda_j^2 \tau^2)$$

where

- ▶ λ_j^2 are local shrinkage parameters, one for each coefficient.
- ▶ τ is a global shrinkage parameter common to all coefficients.
- ▶ The horseshoe behavior is then induced by assigning specific priors to λ_j and τ . The prior on λ_j^2 is a half-Cauchy distribution:

$$\lambda_j \sim \text{Cauchy}^+(0, 1)$$

$$\tau \sim \text{Cauchy}^+(0, \tau_0)$$

where Cauchy^+ denotes the positive half of the Cauchy distribution.

Other priors

- ▶ We could spend weeks going over all the priors that have been proposed for Bayesian linear regression.
- ▶ Some other options include:
 - ▶ **Group Lasso Prior:** encourages sparsity at the group level, meaning that it's likely for all coefficients in a group to be zero or non-zero together.
 - ▶ **Matrix Factorization Priors:** These are useful when there's a belief that the design matrix can be decomposed into lower-dimensional latent structures (think PCA).
 - ▶ **Structured Priors:** When there's some known structure or relationship among the predictors, structured priors can be employed. For instance, in time-series or spatial data, the predictors might have temporal/spatial relationships that can be captured using autoregressive priors.

Bayesian Linear Regression: Advantages

1. **Probabilistic Interpretation:** Bayesian methods provide a full posterior distribution over the model parameters. This allows for natural quantification of uncertainty about variable importance, coefficient estimates, and predictions.
2. **Flexible Priors:** Bayesian frameworks allow for the inclusion of various prior distributions, which can be informed by domain knowledge.
3. **Natural Incorporation of Hierarchical Structures:** Hierarchical Bayesian models can be easily formulated to share information across groups of related variables, aiding in variable selection when there's a known structure in predictors.
4. **Variable Selection with probabilities.**
5. **Model Averaging:** Bayesian methods allow for model averaging over a range of models, weighing them by their posterior probabilities.

Bayesian Linear Regression: Challenges


- ▶ Markov Chain Monte Carlo (MCMC) is a powerful and widely used method for Bayesian inference in a variety of statistical models.
- ▶ MCMC has been used with high-dimensional Bayesian linear models (Liang et al., 2008; Bondell and Reich, 2012; Chae et al., 2019); however, MCMC doesn't scale well.

Bayesian Linear Regression: Challenges

- ▶ Markov Chain Monte Carlo (MCMC) is a powerful and widely used method for Bayesian inference in a variety of statistical models.
- ▶ MCMC has been used with high-dimensional Bayesian linear models (Liang et al., 2008; Bondell and Reich, 2012; Chae et al., 2019); however, MCMC doesn't scale well.
- ▶ Other methods to estimate the components of posterior distributions include:
 - ▶ MAP estimation via the expectation maximization (EM) algorithm (Rockova and George, 2014; Rockova, 2018; Rockova and George, 2018, McLain, et al. 2022),
 - ▶ Variational Bayes (VB, Carbonetto and Stephens, 2012; Blei et al., 2017, Ray and Szabo, 2022).
- ▶ VB is the only one that gives posterior variances but comes with assumptions.

Bayesian Linear Regression in R

- ▶ Bayesian linear regression techniques can be fitted in R (among others):
 - ▶ **VB packages:** sparsevb, and varbvs,
 - ▶ **MAP estimation via EM:** probe, EMVS, SSLASSO, and lmmprobe,
 - ▶ **Full MCMC:** ebreg, horseshoe,



**Yesterday's
posterior is
today's prior**