STAT 153 & 248 - Time Series
Lecture Twelve

Spring 2025, UC Berkeley

Aditya Guntuboyina

February 27, 2025

In this lecture, we discuss a Bayesian treatment for regularization. Before that, let us recap the high-dimensional regression model from the last two lectures, and ridge regression.

1 Recap: Ridge Regression

Our model from the last two lectures is given by:

$$y_t = \beta_0 + \beta_1(t-1) + \beta_2 \text{ReLU}(t-2) + \cdots + \beta_{n-1}\text{ReLU}(t-(n-1)) + \epsilon_t \tag{1}$$

where, as always, $\epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. Here $\text{ReLU}(t-c) = (t-c)_+$ equals 0 if $t \leq c$ and equals $(t-c)$ if $t > c$.

The unknown parameters in this model are $\beta_0, \beta_1, \ldots, \beta_{n-1}$ as well as σ.

Alternatively, (1) can be written as:

$$y = X\beta + \epsilon$$

where

$$X = \begin{pmatrix} 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 2 & 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & n-1 & n-2 & \cdot & \cdot & \cdot & 1 \end{pmatrix} \text{ and } \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_{n-1} \end{pmatrix}. \tag{2}$$

The ridge regression estimator $\hat{\beta}^{\text{ridge}}(\lambda)$ for β is given by the minimizer of:

$$\begin{aligned} &\sum_{t=1}^{n} \left(y_t - \beta_0 - \beta_1(t-1) - \beta_2\text{ReLU}(t-2) - \cdots - \beta_{n-1}\text{ReLU}(t-(n-1))\right)^2 \\ &+ \lambda\left(\beta_2^2 + \beta_3^2 + \cdots + \beta_{n-1}^2\right). \end{aligned} \tag{3}$$

This is equivalent to the Hodrick-Prescott filter as we discussed in the last lecture. The objective function can also be written as

$$\|y - X\beta\|^2 + \lambda \sum_{t=2}^{n-1} \beta_j^2.$$

It turns out that $\hat{\beta}^{\text{ridge}}(\lambda)$ can be written in closed form using matrix notation. To see this, note first that the gradient of the above objective function with respect to β is given by

$$\nabla\left(\|y - X\beta\|^2 + \lambda\sum_{t=2}^{n-1}\beta_j^2\right) = -2X^Ty + 2X^TX\beta + 2\lambda\begin{pmatrix}0\\0\\\beta_2\\\cdot\\\cdot\\\cdot\\\beta_{n-1}\end{pmatrix}$$

Let J denote the $n \times n$ diagonal matrix whose diagonal entries are $0, 0, 1, \ldots, 1$. In other words, the first two diagonal entries of J are 0 and the rest of the diagonal entries equal 1:

$$J = \begin{pmatrix}0&0&0&0&\cdot&\cdot&\cdot&0\\0&0&0&0&\cdot&\cdot&\cdot&0\\0&0&1&0&\cdot&\cdot&\cdot&0\\0&0&0&1&\cdot&\cdot&\cdot&0\\\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot\\\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot\\\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot\\0&0&0&0&\cdot&\cdot&\cdot&1\end{pmatrix}$$

With this matrix, we can write

$$\nabla\left(\|y - X\beta\|^2 + \lambda\sum_{t=2}^{n-1}\beta_j^2\right) = -2X^Ty + 2X^TX\beta + 2\lambda\begin{pmatrix}0\\0\\\beta_2\\\cdot\\\cdot\\\cdot\\\beta_{n-1}\end{pmatrix} = -2X^Ty + 2X^TX\beta + 2\lambda J\beta.$$

Setting this gradient equal to zero, we get

$$-2X^Ty + 2X^TX\beta + 2\lambda J\beta = 0 \implies \left(X^TX + \lambda J\right)\beta = X^Ty.$$

which gives

$$\hat{\beta}^{\text{ridge}}(\lambda) = (X^TX + \lambda J)^{-1}X^Ty. \tag{4}$$

This looks very similar to the usual linear regression least squares formula $(X^TX)^{-1}X^Ty$ with the only difference being the presence of the λJ term.

2 Bayesian Regularization

We now treat regularization in high-dimensional regression from the Bayesian point of view. Before discussing regularization, let us first recap the basics of Bayesian regression in the model:

$$y = X\beta + \epsilon \qquad \text{with } \epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2).$$

The basic prior that we used previously is

$$\beta_j \overset{\text{i.i.d}}{\sim} \text{Unif}(-C, C).$$

for a large positive constant C. For this prior, we showed (see e.g., problem 4 in Homework 1) that

$$\beta \mid \text{data}, \sigma \sim N\left((X^TX)^{-1}X^Ty, \sigma^2(X^TX)^{-1}\right) \tag{5}$$

when $C \to \infty$. This fact is not quite true if C is not very large.

A slightly different prior which allows exact formulae even for finite C is the Gaussian prior:

$$\beta_j \overset{\text{i.i.d}}{\sim} N(0, C). \tag{6}$$

Under this prior, it turns out that

$$\beta \mid \text{data}, \sigma \sim N\left(\left(\frac{X^TX}{\sigma^2} + \frac{I}{C}\right)^{-1}\frac{X^Ty}{\sigma^2}, \left(\frac{X^TX}{\sigma^2} + \frac{I}{C}\right)^{-1}\right) \tag{7}$$

where I is the identity matrix. It is instructive to compare (5) and (7). Unlike (5) which is only true for large C, the fact (7) is true for every $C > 0$. It is also clear that when $C \to \infty$, then (7) is the same as (5). Observe that when C is large, there is not much difference qualitatively between unif$(-C, C)$ and $N(0, C)$ (they are both uninformative priors).

We shall prove a more general form of (7) later in this lecture.

Now let us specialize to the case of the high dimensional regression (2). If we use the prior (6) with $C \to \infty$, then the posterior mean becomes the unregularized least squares (or unregularized MLE) estimator $(X^TX)^{-1}X^Ty$. The fitted values will then perfectly interpolate the data leading to overfitting. From the Bayesian perspective, this is happening because the prior (6) with very large C is not useful for this dataset. The prior needs to be changed for a more meaningful analysis. In the frequentist analysis, the main motivation for the ridge regularization (3) is the need to obtain smaller estimates for $\beta_2, \ldots, \beta_{n-1}$ which will lead to a smoother fit to the data. This same effect can be obtained by the following modification of the prior (6):

$$\beta_0, \beta_1 \overset{\text{i.i.d}}{\sim} N(0, C) \quad \text{and} \quad \beta_2 \ldots, \beta_{n-1} \overset{\text{i.i.d}}{\sim} N(0, \tau^2) \tag{8}$$

for a small parameter τ (in the above, we also assume that $\beta_0, \ldots, \beta_{n-1}$ are all independent). The prior (8) can be written as

$$\beta \sim N(0, Q) \tag{9}$$

where Q is the diagonal matrix with diagonal entries $C, C, \tau^2, \ldots, \tau^2$. Under the prior (9), the posterior of β is given by

$$\beta \mid \text{data}, \sigma \sim N\left(\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}\frac{X^Ty}{\sigma^2}, \left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}\right) \tag{10}$$

We will prove this result later in this lecture. The posterior mean therefore is given by

$$\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}\frac{X^Ty}{\sigma^2} = \left(X^TX + \sigma^2Q^{-1}\right)^{-1}X^Ty. \tag{11}$$

This expression is closely related to the ridge estimator (4). Note that Q^{-1} is diagonal with diagonal entries $1/C, 1/C, 1/\tau^2, \ldots, 1/\tau^2$. When C is very large, the first two diagonal entries of Q^{-1} are very close to zero so that

$$Q^{-1} \approx \frac{1}{\tau^2}J.$$

Thus the posterior mean (11) is therefore

$$\left(X^TX + \frac{\sigma^2}{\tau^2}J\right)^{-1} X^Ty$$

which matches (4) if

$$\lambda = \frac{\sigma^2}{\tau^2} \text{ or, equivalently } \tau = \frac{\sigma}{\sqrt{\lambda}}.$$

Ridge regularization therefore can be understood as Bayesian regression with the prior (8). The precise equivalence is obtained if λ is related to τ^2 via $\lambda = \sigma^2/\tau^2$.

3 Bayesian approach for dealing with unknown τ and σ

One gets smooth fits to the data by working with the prior (8) for small τ. This is not very surprising because the prior injects a strong amount of bias in favor of smooth fits. The real power of the Bayesian approach lies in the ability to automatically infer τ from the data. This is done by simply placing a prior on τ (along with the priors on β and σ). We shall use the following prior:

$$\log\tau, \log\sigma \overset{\text{i.i.d}}{\sim} \text{unif}(-C, C)$$

and

$$\beta \mid \tau, \sigma \sim N(0, Q)$$

where Q is the same as in (9). Note that this prior implies that we are allowing essentially (because C is large) all possible values of τ and σ. In particular, we are **not** *a priori* ruling out large τ just because we don't like wiggly fits.

The prior joint density for β, τ, σ is

$$\begin{aligned} f_{\beta,\tau,\sigma}(\beta,\tau,\sigma) &= f_\tau(\tau) f_\sigma(\sigma) f_{\beta|\tau}(\beta) \\ &= \frac{I\{e^{-C} < \tau < e^C\}}{2C\tau} \frac{I\{e^{-C} < \sigma < e^C\}}{2C\sigma} \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{\sqrt{\det Q}} \exp\left(-\frac{1}{2}\beta^TQ^{-1}\beta\right) \\ &\propto \frac{I\{e^{-C} < \tau, \sigma < e^C\}}{\tau\sigma} \frac{1}{\sqrt{\det Q}} \exp\left(-\frac{1}{2}\beta^TQ^{-1}\beta\right). \end{aligned}$$

We will also ignore the indicator because C will be very large. It is important to note that Q is not a constant matrix as it depends on τ. The likelihood is (as usual in linear regression)

$$\left(\frac{1}{\sqrt{2\pi}}\right)^n \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\|y - X\beta\|^2\right).$$

The posterior for β, τ, σ is therefore

$$f_{\beta,\tau,\sigma|\text{data}}(\beta,\tau,\sigma) \propto \frac{\sigma^{-n-1}\tau^{-1}}{\sqrt{\det Q}} \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}\|y - X\beta\|^2 + \beta^TQ^{-1}\beta\right)\right).$$

The term inside the exponent is a quadratic in β and it is natural to complete the square which is done as follows:

$$\begin{aligned} \frac{1}{\sigma^2}\|y - X\beta\|^2 + \beta^TQ^{-1}\beta &= \frac{y^Ty}{\sigma^2} - \frac{2\beta^TX^Ty}{\sigma^2} + \beta^T\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)\beta \\ &= (\beta - \mu)^T\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)(\beta - \mu) + \frac{y^Ty}{\sigma^2} - \mu^T\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)\mu \end{aligned}$$

where

$$\mu := \left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1} \frac{X^Ty}{\sigma^2}$$

We thus have

$$\begin{aligned}
&\frac{1}{\sigma^2}\|y - X\beta\|^2 + \beta^T Q^{-1}\beta \\
&= (\beta-\mu)^T \left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)(\beta-\mu) + \frac{y^Ty}{\sigma^2} - \frac{y^TX}{\sigma^2}\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}\frac{X^Ty}{\sigma^2}.
\end{aligned}$$

Plugging this in the posterior formula, we deduce

$$\begin{aligned}
&f_{\beta,\tau,\sigma|\text{data}}(\beta,\tau,\sigma) \\
&\propto \frac{\sigma^{-n-1}\tau^{-1}}{\sqrt{\det Q}} \exp\left(-\frac{1}{2}\left((\beta-\mu)^T\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)(\beta-\mu) + \frac{y^Ty}{\sigma^2} - \frac{y^TX}{\sigma^2}\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}\frac{X^Ty}{\sigma^2}\right)\right) \\
&= \frac{\sigma^{-n-1}\tau^{-1}}{\sqrt{\det Q}} \exp\left(-\frac{1}{2}(\beta-\mu)^T\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)(\beta-\mu)\right)\exp\left(-\frac{y^Ty}{2\sigma^2}\right) \\
&\times \exp\left(\frac{y^TX}{2\sigma^2}\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}\frac{X^Ty}{\sigma^2}\right).
\end{aligned}$$

This expression may look complicated but the dependence on β is simple through the quadratic which implies that

$$\beta \mid \text{data}, \sigma, \tau \sim N\left(\mu, \left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}\right) = N\left(\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}\frac{X^Ty}{\sigma^2}, \left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}\right)$$

This proves (7) and (10). It is also straightforward to integrate β from the joint posterior to obtain the posterior of τ, σ:

$$\begin{aligned}
&f_{\tau,\sigma|\text{data}}(\tau,\sigma) \\
&\propto \frac{\sigma^{-n-1}\tau^{-1}}{\sqrt{\det Q}}\sqrt{\det\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}} \exp\left(-\frac{y^Ty}{2\sigma^2}\right)\exp\left(\frac{y^TX}{2\sigma^2}\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}\frac{X^Ty}{\sigma^2}\right).
\end{aligned}$$

In practice, inference can be carried out by first taking a grid of σ and τ values and computing the above posterior (on the logarithmic scale) at the grid points. We can obtain point estimates of σ and τ by taking the posterior maximizers. Alternatively, we can obtain posterior samples of σ and τ by sampling from the grid points with posterior weights. For each (σ, τ) sample, one can sample β using the multivariate normal distribution (10).

This grid approach can be avoided by using MCMC methods such as the Gibbs sampler. We shall not be discussing these.

4 Comments on Bayesian Regularization

In practice, the posterior $f_{\tau,\sigma|\text{data}}(\tau,\sigma)$ tends to prefer τ values which are neither too small nor too large. Because

$$f_{\tau,\sigma|\text{data}}(\tau,\sigma) \propto f_{\text{data}|\tau,\sigma}(\tau,\sigma) f_{\tau,\sigma}(\tau,\sigma),$$

and the prior $f_{\tau,\sigma}(\tau,\sigma)$ is quite flat, the likelihood $f_{\text{data}|\tau,\sigma}(\tau,\sigma)$ must prefer values of τ which are neither too small nor too large. Note that there is a big difference between the two likelihoods:

$$f_{\text{data}|\beta,\sigma}(\text{data}) \quad \text{and} \quad f_{\text{data}|\tau,\sigma}(\text{data}).$$

Maximizing $f_{\text{data}|\beta,\sigma}(\text{data})$ leads to the unregularized least squares estimate which leads to overfitting. On the other hand, maximizing $f_{\text{data}|\tau,\sigma}(\text{data})$ often leads to a fairly small estimate of $\hat{\tau}$ leading to a smooth trend function. The reason for this discrepancy can be understood by noting that

$$f_{\text{data}|\tau,\sigma}(\text{data}) = \int f_{\text{data}|\beta,\sigma}(\text{data}) f_{\beta|\tau}(\beta) d\beta.$$

When τ is large, the term $f_{\beta|\tau}(\beta)$ will be small simply because the normal density with variance τ^2 will be flat for large τ. On the other hand, when τ is too small, the weight $f_{\beta|\tau}(\beta)$ will be significant only for very smooth βs but these βs will have poor values for $f_{\text{data}|\beta,\sigma}(\text{data})$.