

STAT 238 - Bayesian Statistics

Lecture Twenty

Spring 2026, UC Berkeley

Aditya Guntuboyina

11 March 2026

1 Bayesian Inference in a High-Dimensional Linear Regression Model

We studied the following model in the last lecture.

We have a response variable y and a single covariate x . In our example, y denotes weekly earnings and x denotes years of experience. The covariate x takes the values $0, 1, \dots, m$ for some fixed integer m .

Our data is $(x_i, y_i), i = 1, \dots, n$. The model is:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \text{ReLU}(x_i - 1) + \dots + \beta_m \text{ReLU}(x_i - (m - 1)) + \epsilon_i \quad (1)$$

where $\text{ReLU}(u) := \max(u, 0)$, and $\epsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. Note we did not include $\text{ReLU}(x_i - m)$ because it always equals 0.

The model (1) can be rewritten in the usual regression way as:

$$y = X\beta + \epsilon \quad \text{with } \epsilon \sim N(0, \sigma^2 I_d).$$

Here X is the $n \times (m + 1)$ matrix with columns $1, x, \text{ReLU}(x - j)$ for $j = 1, \dots, m - 1$, where x denotes observed values of the experience variable.

The usual least squares analysis coincides with Bayesian inference using the prior:

$$\beta_0, \dots, \beta_m \stackrel{\text{i.i.d.}}{\sim} N(0, C) \text{ and } \log \sigma \sim \text{uniform}(-C, C). \quad (2)$$

for $C \rightarrow \infty$. In Problem 7 of Homework 3, it is proved that the posterior for the above prior is:

$$\beta \mid \text{data}, \sigma \sim N_{m+1}(\hat{\beta}, \sigma^2 (X^T X)^{-1}) \text{ \& } f_{\sigma \mid \text{data}}(\sigma) \propto \sigma^{-n+m} \exp\left(-\frac{y^T y - y^T X (X^T X)^{-1} X^T y}{2\sigma^2}\right) I\{\sigma > 0\}$$

where $\hat{\beta} = (X^T X)^{-1} X^T y$ is the least squares estimator. If we marginalize σ and write the posterior distribution of β , we will get a t -distribution. The posterior distribution for σ above can be written in terms of the Gamma (or chi-squared) distribution (this is problem 7(d) in Homework 3) as:

$$\frac{1}{\sigma^2} \mid \text{data} \sim \text{Gamma}\left(\frac{n - m - 1}{2}, \frac{y^T y - y^T X (X^T X)^{-1} X^T y}{2}\right).$$

Because least squares does not give sensible results in this example, we change the prior in (2) to:

$$\beta_0 \sim N(0, C), \beta_1 \sim N(0, C), \beta_2, \dots, \beta_m \stackrel{\text{i.i.d.}}{\sim} N(0, \tau^2) \quad \text{and} \quad \log \sigma \sim \text{uniform}(-C, C).$$

Therefore, we are bringing in a new parameter τ which controls the scale of β_2, \dots, β_m . For now, treat τ as fixed. So the prior on β and σ is now:

$$f_{\beta, \sigma}(\beta, \sigma) \propto \frac{1}{\sigma \sqrt{\det Q}} \exp\left(-\frac{1}{2} \beta^T Q^{-1} \beta\right)$$

where Q is the $(m+1) \times (m+1)$ diagonal matrix with diagonal entries $C, C, \tau^2, \dots, \tau^2$.

The likelihood is unchanged:

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

So the posterior for β, σ is:

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

Using the completing the square formula:

$$\begin{aligned} \frac{1}{\sigma^2} \|y - X\beta\|^2 + \beta^T Q^{-1} \beta &= (\beta - \hat{\beta})^T \left(\frac{X^T X}{\sigma^2} + Q^{-1}\right) (\beta - \hat{\beta}) \\ &\quad + \frac{y^T y}{\sigma^2} - \left(\frac{y^T X}{\sigma^2}\right) \left(\frac{X^T X}{\sigma^2} + Q^{-1}\right)^{-1} \left(\frac{X^T y}{\sigma^2}\right), \end{aligned}$$

where

$$\hat{\beta} = \left(\frac{X^T X}{\sigma^2} + Q^{-1}\right)^{-1} \frac{X^T y}{\sigma^2}$$

one can then show that

$$\beta \mid \text{data}, \sigma, \tau \sim N\left(\left(\frac{X^T X}{\sigma^2} + Q^{-1}\right)^{-1} \frac{X^T y}{\sigma^2}, \left(\frac{X^T X}{\sigma^2} + Q^{-1}\right)^{-1}\right) \quad (3)$$

Note that when $Q \rightarrow \infty$ (i.e., when $C \rightarrow \infty$ and $\tau \rightarrow \infty$), this reverts to $N((X^T X)^{-1} X^T y, \sigma^2 (X^T X)^{-1})$. To get the posterior of σ given τ , we simply integrate β from the joint posterior to obtain:

Integrating β from the joint posterior gives the posterior of γ, σ :

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

Simplifying by using $\det Q \propto (\tau^2)^{m-1}$ and $Q^{-1} \approx J/\tau^2$ where J is the $(m+1) \times (m+1)$ diagonal matrix with diagonals $0, 0, 1, \dots, 1$, we get

$$\begin{aligned} f_{\sigma | \text{data}, \tau}(\sigma) \\ \propto \frac{\sigma^{-n+m}}{\tau^{m-1}} \left|X^T X + \frac{\sigma^2}{\tau^2} J\right|^{-1/2} \exp\left(-\frac{y^T y - y^T X (X^T X + \sigma^2 \tau^{-2} J)^{-1} X^T y}{2\sigma^2}\right) \end{aligned}$$

This distribution is not easy to handle because of the presence of the term σ^2/τ^2 inside the determinant as well as inside the inverse (in the exponent). One trick to simplify this is to reparametrize and assume $\tau = \sigma\gamma$. With this, the above conditional density becomes

$$f_{\sigma|\text{data},\gamma}(\sigma) \propto \frac{1}{\gamma^{m-1}\sigma^{n-1}} |X^T X + \gamma^{-2}J|^{-1/2} \exp\left(-\frac{y^T y - y^T X (X^T X + \gamma^{-2}J)^{-1} X^T y}{2\sigma^2}\right)$$

Ignoring terms above which do not depend on γ , we get

$$f_{\sigma|\text{data},\gamma}(\sigma) \propto \sigma^{-n+1} \exp\left(-\frac{y^T y - y^T X (X^T X + \gamma^{-2}J)^{-1} X^T y}{2\sigma^2}\right).$$

The right hand side above is actually the pdf of an inverse gamma density (see https://en.wikipedia.org/wiki/Inverse-gamma_distribution). This can be seen by converting it into the density of $1/\sigma^2$:

$$\begin{aligned} f_{1/\sigma^2|\text{data},\gamma}(x) &\propto f_{\sigma|\text{data},\gamma}(x^{-1/2})x^{-3/2} \\ &\propto (x^{-1/2})^{-n+1} \exp\left(-x \frac{y^T y - y^T X (X^T X + \gamma^{-2}J)^{-1} X^T y}{2}\right) x^{-3/2} \\ &= x^{(n-4)/2} \exp\left(-x \frac{y^T y - y^T X (X^T X + \gamma^{-2}J)^{-1} X^T y}{2}\right). \end{aligned}$$

Thus

$$\frac{1}{\sigma^2} | \text{data}, \gamma \sim \text{Gamma}\left(\frac{n}{2} - 1, \frac{y^T y - y^T X (X^T X + \gamma^{-2}J)^{-1} X^T y}{2}\right) \quad (4)$$

Thus when γ is fixed, we can perform inference on β, σ using the above closed form formulae. Since γ is also unknown, we can place a prior $p(\gamma)$ on it, and then calculate its posterior.

Finally, we can marginalize σ to obtain the posterior of γ alone as follows:

$$\begin{aligned} f_{\gamma|\text{data}}(\gamma) &\propto \int_0^\infty p(\gamma) \frac{1}{\gamma^{m-1}\sigma^{n-1}} |X^T X + \gamma^{-2}J|^{-1/2} \exp\left(-\frac{y^T y - y^T X (X^T X + \gamma^{-2}J)^{-1} X^T y}{2\sigma^2}\right) d\sigma \\ &= p(\gamma)\gamma^{-m+1} |X^T X + \gamma^{-2}J|^{-1/2} \int_0^\infty \sigma^{-n+1} \exp\left(-\frac{y^T y - y^T X (X^T X + \gamma^{-2}J)^{-1} X^T y}{2\sigma^2}\right) d\sigma. \end{aligned}$$

Letting

$$A := y^T y - y^T X (X^T X + \gamma^{-2}J)^{-1} X^T y,$$

we get

$$f_{\gamma|\text{data}}(\gamma) \propto p(\gamma)\gamma^{-m+1} |X^T X + \gamma^{-2}J|^{-1/2} \int_0^\infty \sigma^{-n+1} \exp\left(-\frac{A}{2\sigma^2}\right) d\sigma$$

By the change of variable $\sigma = s\sqrt{A}$, we obtain

$$\begin{aligned} f_{\gamma|\text{data}}(\gamma) &\propto p(\gamma)\gamma^{-m+1} |X^T X + \gamma^{-2}J|^{-1/2} A^{-(n/2)+1} \int_0^\infty s^{-n+1} \exp\left(-\frac{1}{2s^2}\right) ds \\ &\propto p(\gamma)\gamma^{-m+1} |X^T X + \gamma^{-2}J|^{-1/2} A^{-(n/2)+1} \\ &= \frac{p(\gamma)\gamma^{-m+1} |X^T X + \gamma^{-2}J|^{-1/2}}{\left(y^T y - y^T X (X^T X + \gamma^{-2}J)^{-1} X^T y\right)^{(n/2)-1}}. \end{aligned}$$

We usually choose $p(\gamma)$ as:

$$p(\gamma) \propto \frac{1}{\gamma} I\{\text{low} < \gamma < \text{high}\}$$

for two fixed values low and high. These could be the values of γ which lead to the posterior mean (which coincides with ridge regression) leading to underfitting and overfitting respectively.

Inference can be carried out by first taking a grid of γ values and computing the above posterior (on the logarithmic scale) at the grid points. This posterior can be used to obtain posterior samples of γ . For each sample of γ , we can then sample σ using the distribution (4). Given samples from both γ and σ , we can then sample β using (3).

2 Induced Prior for the Regression Function

Our regression function is being modeled as:

$$f(x) = \beta_0 + \beta_1 x + \beta_2(x-1)_+ + \cdots + \beta_m(x-(m-1))_+ \quad (5)$$

where $\beta_0, \beta_1, \dots, \beta_m$ are all independent with

$$\beta_0, \beta_1 \stackrel{\text{i.i.d.}}{\sim} N(0, C) \quad \text{and} \quad \beta_2, \dots, \beta_m \stackrel{\text{i.i.d.}}{\sim} N(0, \tau^2).$$

This can also be seen as a prior on the regression function f more directly. For every $0 \leq u_1 < \cdots < u_k < \infty$, the joint distribution of $(f(u_1), \dots, f(u_k))$ induced by (5) will be multivariate normal. This implies that $(f(u), u \geq 0)$ is a Gaussian Process. The description of the GP can be done in terms of the mean function $\mathbb{E}f(u)$ and the covariance kernel $\text{Cov}(f(u), f(v))$.

The mean function is clearly zero (because each $\mathbb{E}\beta_j = 0$). The covariance kernel is given by:

$$\begin{aligned} & \text{cov}(f(u), f(v)) \\ &= \text{cov}(\beta_0 + \beta_1 u + \beta_2(u-1)_+ + \cdots + \beta_m(u-(m-1))_+, \beta_0 + \beta_1 v + \beta_2(v-1)_+ + \cdots + \beta_m(v-(m-1))_+) \\ &= C(1+uv) + \tau^2 \sum_{j=1}^{m-1} (u-j)_+(v-j)_+ \\ &= C(1+uv) + \tau^2 \sum_{j=1}^k (u-j)(v-j) \\ &= C(1+uv) + \tau^2 \left(uvk + \frac{1}{6}k(k+1)(2k+1) - (u+v)\frac{k(k+1)}{2} \right) \end{aligned}$$

where $k = \lfloor u \wedge v \rfloor$ (and $u \wedge v := \min(u, v)$).

Suppose now we use the approximation

$$k \approx u \wedge v \quad \text{and} \quad k(k+1)(2k+1) \approx 2k^3 \approx 2(u \wedge v)^3 \quad \text{and} \quad k(k+1) \approx k^2 \approx (u \wedge v)^2.$$

Then the covariance kernel becomes:

$$\text{cov}(f(u), f(v)) \approx C(1+uv) + \tau^2 \left(uv(u \wedge v) + \frac{(u \wedge v)^3}{3} - \frac{(u \wedge v)^2}{2}(u+v) \right).$$

It turns out that the right hand side above is precisely the covariance kernel of:

$$G_x = \beta_0 + \beta_1 x + \tau I_x$$

where $\beta_0, \beta_1 \stackrel{\text{i.i.d}}{\sim} N(0, C)$ and I_x is **Integrated Brownian Motion** on $[0, \infty)$. We will revisit this in the next lecture.