STAT 153 & 248 - Time Series Lecture Thirteen

Fall 2025, UC Berkeley

Aditya Guntuboyina

October 9, 2025

1 More on Bayesian Regularization

1.1 Recap

In the last lecture, we discussed Bayesian regularization in the context of the following high-dimensional linear regression model:

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

The unknown parameters in this model are $\beta_0, \dots, \beta_{n-1}$ as well as σ. We can write this model in regression form as $y = X\beta + \epsilon$ for the very special matrix X:

$$X = \begin{pmatrix} 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 2 & 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & n-1 & n-2 & \cdot & \cdot & \cdot & 1 \end{pmatrix}$$

We discussed Bayesian inference with the prior (throughout C denotes a very large constant):

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

This prior depends on the unknown parameter (sometimes called 'hyperparameter') τ. We treat τ also as an unknown parameter (along with σ) and place the following prior on τ, σ:

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

The prior assumption (2) can be written in matrix notation as:

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

where Q is the diagonal matrix with diagonal entries $C, C, \tau^2, \dots, \tau^2$.

We calculated the posterior distribution of all the parameters β, σ, τ in the last lecture. This posterior can be described as follows. First the conditional posterior distribution of β given the other two parameters τ, σ is given by:

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

Because Q is diagonal with diagonal entries $C, C, 1/\tau^2, \dots, 1/\tau^2$, it is easy to see that Q^{-1} is also diagonal with diagonal entries $1/C, 1/C, \tau^{-2}, \dots, \tau^{-2}$. Because C is large, we can approximate Q^{-1} by the matrix with diagonal entries $0, 0, \tau^{-2}, \dots, \tau^{-2}$. In other words:

$$Q^{-1} \approx \frac{1}{\tau^2} J \quad \text{where } J = \begin{pmatrix} 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 0 & 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 1 \end{pmatrix}$$

With this approximation, the conditional posterior of β (given τ, σ) in (3) becomes:

$$\begin{aligned} \beta \mid \text{data}, \sigma, \tau &\sim N\left(\left(\frac{X^TX}{\sigma^2} + \frac{J}{\tau^2}\right)^{-1} \frac{X^Ty}{\sigma^2}, \left(\frac{X^TX}{\sigma^2} + \frac{J}{\tau^2}\right)^{-1}\right) \\ &\sim N\left(\left(X^TX + \frac{\sigma^2}{\tau^2}J\right)^{-1} X^Ty, \sigma^2\left(X^TX + \frac{\sigma^2}{\tau^2}J\right)^{-1}\right) \end{aligned} \tag{4}$$

We saw in the last lecture that the mean of this conditional posterior distribution:

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

coincides with the Ridge Regularized estimate with tuning parameter λ provided $\lambda = \sigma^2/\tau^2$.

We also derived that the posterior of τ and σ is given by the formula:

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

Using again the approximation $Q^{-1} \approx (1/\tau^2)J$ as well as $\det Q = C^2\tau^{2(n-2)} \propto \tau^{2(n-2)}$, we get

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

This posterior lets us figure out the value of τ (as well as σ) from the data. Specifically, we can take a grid of σ and τ values and compute the above posterior (on the logarithmic scale) at the grid points. We then obtain point estimates of σ and τ by taking the posterior mode or mean. 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 (6).

In most cases, the posterior $f_{\tau,\sigma|\text{data}}(\tau, \sigma)$ will prefer values of τ that are not too large. Note that large τ would mean that the resulting β estimates (see (4)) will lead to fitted values that overfit the data. In other words, this Bayesian approach for selecting τ automatically avoids overfitting.

Further, unless the data is very simple (in the sense of being well-explained by a single line), the posterior $f_{\tau,\sigma|\text{data}}(\tau, \sigma)$ will also prefer values of τ that are not too small. Note that small τ will mean that the resulting fitted values will be very close to the least squares line.

Thus, this Bayesian approach offers protection from both overfitting and underfitting. To get some insight as to why $f_{\tau,\sigma|\text{data}}(\tau, \sigma)$ protects from overfitting and underfitting, consider the following alternative expression for it:

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

In the above, $f_{\tau,\sigma}(\tau, \sigma)$ is a very flat function (because we are using uninformative priors for τ and σ) so the behavior of $f_{\tau,\sigma|\text{data}}(\tau, \sigma)$ will be mainly driven by the first term: $f_{\text{data}|\tau,\sigma}(\text{data})$. This term can be seen as the likelihood of the data only in terms of τ and σ. It is related to the original likelihood $f_{\text{data}|\beta,\sigma}(\text{data})$ via

$$f_{\text{data}|\tau,\sigma}(\text{data}) = \int f_{\text{data}|\beta,\sigma}(\text{data}) f_{\beta|\tau}(\beta) d\beta. \tag{5}$$

This integrated likelihood $f_{\text{data}|\tau,\sigma}(\text{data})$ tends to prefer values of τ which are neither too small nor too large. For example, suppose τ is very small. The the prior $f_{\beta|\tau}(\beta)$ will be concentrated on values of β for which the corresponding fitted values are close to a single line. But for such β, the data likelihood $f_{\text{data}|\beta,\sigma}(\text{data})$ will be small (unless the dataset is simple enough to be explained by a single line).

On the other hand, suppose τ is very large. Then the prior $f_{\beta|\tau}(\beta)$ will be quite flat over a large region so it will assign a small value (because of the factor $1/\tau$) for most values of β. This will bring down the overall integral value for (5).

It is important to note that the Bayesian approach for selecting hyperparameters such as τ is very different from the frequentist approach for selecting the tuning parameter λ using Cross-Validation (CV). CV relies on splitting the data into two parts: training and test, while the Bayesian calculation uses the entire data (no splits are necessary). For more on the relation between Bayesian approaches for hyperparameter tuning and model selection, and frequentist methods, see `http://www.inference.org.uk/mackay/Bayes_FAQ.html#gcv` and `https://statmodeling.stat.columbia.edu/2011/12/04/david-mackay-and-occams-razor/`.

1.2 Calculations with a slightly different prior

The Bayesian method described above involves placing a grid on both τ and σ. This grid approach can be avoided by using MCMC methods such as the Gibbs sampler. We shall not be discussing these (see the 248 problem in Homework 3 if you are interested in Gibbs sampling).

There is a slightly different prior which lets us integrate over σ in closed form. So the grid approach would then be needed only for τ (and not σ) which reduces the computational burden to some extent. This method is described here.

The idea is to reparametrize τ as $\tau = \sigma\gamma$ for a new parameter γ, and change the prior to

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

and

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

where Q is the diagonal matrix with diagonal entries $C, C, \gamma^2\sigma^2, \dots, \gamma\sigma^2$ (this is the same Q as before; the only difference is that now we are writing $\gamma\sigma$ for τ). The prior joint density for β, γ, σ now becomes:

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

This new parameter γ is almost directly related to the tuning parameter λ in ridge regression. Specifically, $\lambda = 1/\gamma^2$ or $\gamma = 1/\sqrt{\lambda}$.

The likelihood function is exactly the same as before:

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

So the posterior for β, γ, σ is:

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

From here, we proceed exactly as in last lecture to derive:

$$\beta \mid \text{data}, \sigma, \gamma \sim N\left(\left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1} \frac{X^Ty}{\sigma^2}, \left(\frac{X^TX}{\sigma^2} + Q^{-1}\right)^{-1}\right) \tag{6}$$

Again, as in last lecture, we integrate β from the joint posterior to obtain the posterior of γ, σ:

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

Simplying by use of $\det Q \propto (\gamma^2\sigma^2)^{n-2}$ and $Q^{-1} \approx J/(\gamma^2\sigma^2)$. we can write

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

The advantage of this parametrization (i.e., in terms of γ, σ as opposed to τ, σ) is that the conditional posterior distribution of σ given γ (and the data) can be written in closed form. This is because from the above expression, we can write

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

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 $1/\sigma^2$:

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

Thus

$$\frac{1}{\sigma^2} \mid \text{data}, \gamma \sim \text{Gamma}\left(\frac{n}{2} - 1, \frac{y^Ty - y^TX\left(X^TX + \gamma^{-2}J\right)^{-1}X^Ty}{2}.\right) \tag{7}$$

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

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

Letting

$$A := y^Ty - y^TX\left(X^TX + \gamma^{-2}J\right)^{-1}X^Ty,$$

we get

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

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

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

With this, 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 (7). Given samples from both γ and σ, we can then sample β using (6).

In the connection with usual ridge regression, we noted previously that $\lambda = \sigma^2/\tau^2$. With the new parametrization $\tau^2 = \gamma^2\sigma^2$, we have $\lambda = 1/\gamma^2$ so that $\gamma = 1/\sqrt{\lambda}$.

2 Variance Models

Our next topic involves Variance Models (especially in the context of spectral analysis). To motivate variance models, note first that all the models that we studied in the class so far can be seen as "Mean Models".

For example, consider the high-dimensional linear regression model (1) that we have been studied for the past 3 lectures. This model and the resulting estimation procedures can be written as:

$$y_t \overset{\text{ind}}{\sim} N(\mu_t, \sigma^2)$$

where ind stands for "independently distributed as". Note that the right hand side depends on t so the distribution of y_t changes with t and we cannot therefore use "i.i.d".

The parameters in this model are $\mu_1, \dots, \mu_n$ and σ^2. Clearly this is a high-dimensional because the number of parameters is large.

If we attempt to estimate the parameters by maximizing the likelihood without any regularization, we get $\mu_t = y_t$ and $\sigma^2 = 0$, leading to full interpolation (overfitting) to the data. Regularization is therefore necessary to obtain something useful. If we want to obtain "smooth" trend estimates, we can employ regularization terms which force neighboring values or neighboring slopes of μ_t to be close. If we focus on slopes (which leads to more smoothness compared to just imposing closeness of values), we obtain the estimators $\hat{\mu}_t^{\text{ridge}}(\lambda)$ and $\hat{\mu}_t^{\text{lasso}}(\lambda)$ which minimize:

$$\sum_{t=1}^{n}(y_t - \mu_t)^2 + \lambda \sum_{t=2}^{n-1} \left((\mu_{t+1} - \mu_t) - (\mu_t - \mu_{t-1})\right)^2$$

and

$$\sum_{t=1}^{n}(y_t - \mu_t)^2 + \lambda \sum_{t=2}^{n-1} |(\mu_{t+1} - \mu_t) - (\mu_t - \mu_{t-1})|$$

respectively. We have already studied these estimators last week where we observed, among other things, that they can be alternatively represented as $\hat{\mu}_t^{\text{ridge}}(\lambda) = X\hat{\beta}^{\text{ridge}}(\lambda)$ and $\hat{\mu}_t^{\text{lasso}}(\lambda) = X\hat{\beta}^{\text{lasso}}(\lambda)$ where

$$X = \begin{pmatrix} 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 1 & 2 & 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & n-1 & n-2 & \cdot & \cdot & \cdot & 1 \end{pmatrix} \text{ and } \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_{n-1} \end{pmatrix} \tag{8}$$

and $\hat{\beta}^{\text{ridge}}(\lambda)$ and $\hat{\beta}^{\text{lasso}}(\lambda)$ minimize

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

and

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

respectively.

This model is an example of a "Mean Model" where the focus is on estimating the mean parameters μ_t. In contrast, we shall now study "Variance Models" which model the variances of the data. The simplest variance model is given by:

$$y_t \overset{\text{ind}}{\sim} N(0, \tau_t^2) \tag{9}$$

The parameters are $\tau_1^2, \dots, \tau_n^2$. Since these represent variances, we refer to this as a "variance model". We shall study parameter estimation in this model (and other related models) next week.