

STAT 238 - Bayesian Statistics

Lecture Twenty Three

Spring 2026, UC Berkeley

Aditya Guntuboyina

18 March 2026

1 Last lecture: interpolation with Gaussian processes

The interpolation problem is: Suppose we are given values of f at points x_1, \dots, x_n in the domain Ω of f . Let $x \in \Omega$ be a new point (i.e., distinct from x_1, \dots, x_n). What can we say about $f(x)$?

In the last lecture, we saw how to use Gaussian processes to solve this problem. We model $\{f(x), x \in \Omega\}$ as a Gaussian process with mean zero and covariance function or *kernel* given by $K(x, x')$ i.e.,

$$\text{Cov}(f(x), f(x')) = K(x, x') \quad \text{for all } x, x' \in \Omega.$$

Under this modeling assumption, the answer to the interpolation question is given by the conditional distribution of $f(x)$ given $f(x_1), \dots, f(x_n)$ which is calculated as follows. Note that:

$$(f(x_1), \dots, f(x_n), f(x)) \sim N \left(0, \begin{pmatrix} (K(x_i, x_j))_{n \times n} & (K(x_i, x))_{n \times 1} \\ (K(x, x_i))_{1 \times n} & K(x, x) \end{pmatrix} \right).$$

Using the notation $K = (K(x_i, x_j))_{n \times n}$ and $\mathbf{k} = (K(x_i, x))_{n \times 1}$, we can write the conditional distribution of $f(x)$ given $f(x_1), \dots, f(x_n)$ as

$$f(x) \mid f(x_1), \dots, f(x_n) \sim N(\mathbf{k}^T K^{-1} (f(x_1), \dots, f(x_n))^T, K(x, x) - \mathbf{k}^T K^{-1} \mathbf{k}).$$

Thus the posterior mean (or mode) estimate of $f(x)$ is given by

$$\widehat{f(x)} = \mathbf{k}^T K^{-1} (f(x_1), \dots, f(x_n))^T. \tag{1}$$

2 Regression with Gaussian Processes

Today, we will see how to perform regression using Gaussian processes. The key difference between **regression** and **interpolation** is that, in regression, the values $f(x_1), \dots, f(x_n)$ are not observed exactly; instead, they are observed with noise.

More precisely, we are given observations y_1, \dots, y_n modeled as

$$y_i = f(x_i) + \epsilon_i, \quad \text{where } \epsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

The parameter σ controls the level of noise, i.e., how much each observation y_i deviates from the true value $f(x_i)$.

As in the interpolation setting, our goal is to estimate $f(x)$ at a test point x . This test point may be different from the observed inputs x_1, \dots, x_n , or it may coincide with one of them.

Importantly, because the observations are noisy, it is meaningful to estimate $f(x_i)$ even at observed points. In fact, by combining information from all observations, we can often obtain a better estimate of $f(x_i)$ than the raw observation y_i .

Another key difference from interpolation is that the input points x_1, \dots, x_n need not necessarily be distinct. Since each observation contains noise, having repeated measurements at the same input can help improve the overall estimate.

The solution to the regression problem is very similar to that of the interpolation problem with only one difference. The goal is to estimate $f(x)$ at the test point x given the data $(x_1, y_1), \dots, (x_n, y_n)$. We will use the conditional distribution of $f(x)$ given y_1, \dots, y_n (we will assume that x_1, \dots, x_n, x are deterministic). To calculate this conditional distribution, first note that the marginal distribution of $(y_1, \dots, y_n, f(x_*))$ is given by

$$(y_1, \dots, y_n, f(x)) \sim N \left(0, \begin{pmatrix} (K(x_i, x_j))_{n \times n} + \sigma^2 I_n & (K(x_i, x))_{n \times 1} \\ (K(x, x_i))_{1 \times n} & K(x, x) \end{pmatrix} \right).$$

Using the notation $K = (K(x_i, x_j))_{n \times n}$ and $\mathbf{k} = (K(x_i, x))_{n \times 1}$, we can write the conditional distribution of $f(x)$ given y_1, \dots, y_n as

$$f(x) \mid \text{data} \sim N \left(\mathbf{k}^T (K + \sigma^2 I_n)^{-1} Y, K(x, x) - \mathbf{k}^T (K + \sigma^2 I_n)^{-1} \mathbf{k} \right).$$

Thus the posterior mean (or mode) estimate of $f(x)$ is given by

$$\widehat{f(x)} = \mathbf{k}^T (K + \sigma^2 I_n)^{-1} y, \tag{2}$$

where y is the $n \times 1$ vector with entries y_1, \dots, y_n .

So the only difference between (2) and (1) is the presence of the additional $\sigma^2 I_n$ term in case of regression.

Often, we would need to estimate σ from the observed data (and also additional hyperparameters present in the kernel K). For these, the marginal likelihood of y_1, \dots, y_n is important. This is simply the multivariate normal distribution with mean vector 0 and covariance $K + \sigma^2 I_n$.

To illustrate the calculations, let us take the special case of the Integrated Brownian Motion prior.

3 Calculations for the IBM prior

We take the prior

$$f(x) = \beta_0 + \beta_1 x + \tau I(x)$$

where β_0, β_1 are i.i.d $N(0, C)$ and $I(x)$ is integrated Brownian Motion.

This is a Gaussian process prior with mean zero and covariance kernel:

$$K(u, v) = C(1 + uv) + \tau^2 K_I(u, v)$$

where $K_I(u, v)$ is the kernel corresponding to IBM, which is:

$$\begin{aligned} K_I(u, v) &= \frac{1}{2} (\min(u, v))^2 \max(u, v) - \frac{1}{6} (\min(u, v))^3 \\ &= \frac{1}{6} (\min(u, v))^2 (3 \max(u, v) - \min(u, v)) \\ &= uv(\min(u, v)) - \frac{u+v}{2} (\min(u, v))^2 + \frac{1}{3} (\min(u, v))^3. \end{aligned}$$

Note that the kernel has the unknown parameter τ (which we write as $\tau = \gamma\sigma$). The constant C is assumed to be large and it is not to be estimated (ideally we want to take $C = +\infty$).

Given data $(x_1, y_1), \dots, (x_n, y_n)$, what is the posterior of $f(x)$? First let us assume that τ, σ are given. The estimate is simply (2). We simplify this expression below. Let X denote the $n \times 2$ matrix with columns 1 and x_i (this is the usual X matrix in the simple linear regression of y on x based on the data $(x_1, y_1), \dots, (x_n, y_n)$).

Note that

$$\begin{aligned} \mathbf{k}^T &= (K(x, x_1), \dots, K(x, x_n)) \\ &= (C(1 + xx_i) + \tau^2 K_I(x, x_i), i = 1, \dots, n) \\ &= C(1, x)X^T + \tau^2 (K_I(x, x_1), \dots, K_I(x, x_n)) \\ &= C(1, x)X^T + \tau^2 \mathbf{k}_I^T. \end{aligned}$$

where

$$\mathbf{k}_I^T := (K_I(x, x_1), \dots, K_I(x, x_n)),$$

and $(1, x)$ denotes the row vector with entries 1 and x .

Further the $n \times n$ matrix K has the (i, j) -th entry:

$$C(1 + x_i x_j) + \gamma^2 \sigma^2 K_I(x_i, x_j)$$

so that

$$K = CXX^T + \tau^2 K_I \tag{3}$$

where K_I is the $n \times n$ matrix with (i, j) -th entry $K_I(x_i, x_j)$.

As a result,

$$\widehat{f(x)} = \mathbf{k}^T (K + \sigma^2 I_n)^{-1} y = (C(1, x)X^T + \tau^2 \mathbf{k}_I^T) (CXX^T + \tau^2 K_I + \sigma^2 I_n)^{-1} y.$$

This expression depends on the large constant C . Direct computation with a large C might make it unstable. It is therefore natural to compute the limit as $C \rightarrow \infty$. Using the Sherman-Morrison-Woodbury identity,

$$\begin{aligned} (K + \sigma^2 I_n)^{-1} &= (CXX^T + \tau^2 K_I + \sigma^2 I_n)^{-1} \\ &= (CXX^T + \Sigma)^{-1} \\ &= \Sigma^{-1} - \Sigma^{-1} X (C^{-1} I_2 + X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \end{aligned} \tag{4}$$

where

$$\Sigma = \tau^2 K_I + \sigma^2 I_n. \quad (5)$$

Further

$$\mathbf{k} = C(1, x)X^T + \tau^2 \mathbf{k}_I^T.$$

We thus get

$$\widehat{f(x)} = (C(1, x)X^T + \tau^2 \mathbf{k}_I^T) \left(\Sigma^{-1} - \Sigma^{-1}X (C^{-1}I_2 + X^T \Sigma^{-1}X)^{-1} X^T \Sigma^{-1} \right) y.$$

Write $\tau = \gamma\sigma$. In Fact 4.1, it is proved that, as $C \rightarrow \infty$ the above converges to:

$$\widehat{f(x)} := (1, x) (X^T A_\gamma^{-1} X)^{-1} X^T A_\gamma^{-1} y + \gamma^2 \mathbf{k}_I^T (A_\gamma^{-1} - A_\gamma^{-1} X (X^T A_\gamma^{-1} X)^{-1} X^T A_\gamma^{-1}) y$$

where

$$A_\gamma = I_{n \times n} + \gamma^2 K_I.$$

This expression for $\widehat{f(x)}$, which only depends on $\gamma = \tau/\sigma$ and not on τ and σ individually, can be used for computation.

3.1 Estimation of γ and σ

The hyperparameters γ and σ also need to be estimated from the observed data $(x_1, y_1), \dots, (x_n, y_n)$ (recall that $\tau = \gamma\sigma$). For this, the marginal likelihood of the data given σ, γ is important. This is calculated using:

$$y \mid \sigma, \gamma \sim N(0, K + \sigma^2 I_n)$$

where K is given by (3). In other words,

$$f_{y \mid \sigma, \gamma}(y) \propto \frac{1}{\sqrt{\det(K + \sigma^2 I_n)}} \exp\left(-\frac{1}{2} y^T (K + \sigma^2 I_n)^{-1} y\right).$$

For $(K + \sigma^2 I_n)^{-1}$, we use (4) and let $C \rightarrow \infty$ to get

$$(K + \sigma^2 I_n)^{-1} \rightarrow \Sigma^{-1} - \Sigma^{-1}X (X^T \Sigma^{-1}X)^{-1} X^T \Sigma^{-1}.$$

Further, using the matrix determinant lemma, we get

$$\begin{aligned} |K + \sigma^2 I_n| &= |\Sigma + CXX^T| \\ &= |\Sigma| |I + CX^T \Sigma^{-1}X| \\ &\approx |\Sigma| |CX^T \Sigma^{-1}X| \text{ when } C \text{ is large} \\ &= |\Sigma| C^2 |X^T \Sigma^{-1}X| \propto |\Sigma| |X^T \Sigma^{-1}X|. \end{aligned}$$

So the marginal likelihood of y given τ, σ becomes:

$$f_{y \mid \tau, \sigma}(y) \propto |\Sigma|^{-1/2} |X^T \Sigma^{-1}X|^{-1/2} \exp\left(-\frac{1}{2} y^T \left[\Sigma^{-1} - \Sigma^{-1}X (X^T \Sigma^{-1}X)^{-1} X^T \Sigma^{-1} \right] y\right)$$

with Σ defined in (5).

We now take $\tau = \gamma\sigma$ so that

$$\Sigma = \sigma^2 (I_n + \gamma^2 K_I) = \sigma^2 A_\gamma \quad \text{where } A_\gamma := I_n + \gamma^2 K_I.$$

This gives

$$f_{y|\gamma,\sigma}(y) \propto \sigma^{-(n-2)} |A_\gamma|^{-1/2} |X^T A_\gamma^{-1} X|^{-1/2} \exp\left(-\frac{y^T [A_\gamma^{-1} - A_\gamma^{-1} X (X^T A_\gamma^{-1} X)^{-1} X^T A_\gamma^{-1}] y}{2\sigma^2}\right).$$

Combining this with the prior $\log \sigma \mid \gamma \sim \text{uniform}(-\infty, \infty)$ gives

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

Finally if we take a prior $p(\gamma)$ for γ , then the posterior of γ becomes:

$$p(\gamma \mid y) \propto \frac{p(\gamma) |A_\gamma|^{-1/2} |X^T A_\gamma^{-1} X|^{-1/2}}{(y^T [A_\gamma^{-1} - A_\gamma^{-1} X (X^T A_\gamma^{-1} X)^{-1} X^T A_\gamma^{-1}] y)^{(n/2)-1}}.$$

4 Proof of the $C \rightarrow \infty$ fact

Fact 4.1. *The quantity*

$$\hat{f}(x) = (C(1, x)X^T + \tau^2 \mathbf{k}_I^T) \left(\Sigma^{-1} - \Sigma^{-1} X (C^{-1}I_2 + X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \right) y$$

converges, as $C \rightarrow \infty$, to:

$$\hat{f}(x) := (1, x) (X^T A_\gamma^{-1} X)^{-1} X^T A_\gamma^{-1} y + \gamma^2 \mathbf{k}_I^T (A_\gamma^{-1} - A_\gamma^{-1} X (X^T A_\gamma^{-1} X)^{-1} X^T A_\gamma^{-1}) y.$$

Here $\tau = \gamma\sigma$ and $\Sigma = \tau^2 K_I + \sigma^2 I_n$.

$$\Sigma = \tau^2 K_I + \sigma^2 I_n = \sigma^2 \gamma^2 K_I + \sigma^2 I_n = \sigma^2 A_\gamma$$

where $A_\gamma = I + \gamma^2 K_I$.

Proof of Fact 4.1. First note that

$$\begin{aligned} \hat{f}(x) &= (C(1, x)X^T + \tau^2 \mathbf{k}_I^T) \left(\Sigma^{-1} - \Sigma^{-1} X (C^{-1}I_2 + X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \right) y \\ &= T_1 + T_2 \end{aligned}$$

where

$$T_1 := C(1, x)X^T \left(\Sigma^{-1} - \Sigma^{-1} X (C^{-1}I_2 + X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \right) y$$

and

$$\begin{aligned} T_2 &:= \tau^2 \mathbf{k}_I^T \left(\Sigma^{-1} - \Sigma^{-1} X (C^{-1}I_2 + X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \right) y \\ &\rightarrow \tau^2 \mathbf{k}_I^T \left(\Sigma^{-1} - \Sigma^{-1} X (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} \right) y \quad \text{as } C \rightarrow \infty \\ &= \tau^2 \mathbf{k}_I^T \left(\sigma^{-2} A_\gamma^{-1} - \sigma^{-2} A_\gamma^{-1} X (X^T A_\gamma^{-1} X)^{-1} X^T A_\gamma^{-1} \right) y \\ &= \gamma^2 \mathbf{k}_I^T \left(A_\gamma^{-1} - A_\gamma^{-1} X (X^T A_\gamma^{-1} X)^{-1} X^T A_\gamma^{-1} \right) y. \end{aligned}$$

The limit of T_1 as $C \rightarrow \infty$ is calculated as follows. First note that

$$\begin{aligned}
T_1 &= C(1, x)X^T \left(\Sigma^{-1} - \Sigma^{-1}X (C^{-1}I_2 + X^T\Sigma^{-1}X)^{-1} X^T\Sigma^{-1} \right) y \\
&= C(1, x)X^T\Sigma^{-1}y - C(1, x)X^T\Sigma^{-1}X (C^{-1}I_2 + X^T\Sigma^{-1}X)^{-1} X^T\Sigma^{-1}y \\
&= C(1, x)X^T\Sigma^{-1}y - C(1, x) \left[(C^{-1}I_2 + X^T\Sigma^{-1}X) (X^T\Sigma^{-1}X)^{-1} \right]^{-1} X^T\Sigma^{-1}y \\
&= C(1, x)X^T\Sigma^{-1}y - C(1, x) \left[I_2 + C^{-1}(X^T\Sigma^{-1}X)^{-1} \right]^{-1} X^T\Sigma^{-1}y
\end{aligned}$$

Using the fact:

$$(I + A^{-1}/C)^{-1} = I - A^{-1}/C + O(1/C^2)$$

for $A = X^T\Sigma^{-1}X$, we obtain

$$\begin{aligned}
T_1 &= C(1, x)X^T\Sigma^{-1}y - C(1, x) \left(I - (X^T\Sigma^{-1}X)^{-1}/C + O(1/C^2) \right) X^T\Sigma^{-1}y \\
&= (1, x)(X^T\Sigma^{-1}X)^{-1}(X^T\Sigma^{-1}y) + (1, x)O(1/C)X^T\Sigma^{-1}y \\
&\rightarrow (1, x)(X^T\Sigma^{-1}X)^{-1}(X^T\Sigma^{-1}y) \text{ as } C \rightarrow \infty \\
&= (1, x)(X^T A_\gamma^{-1}X)^{-1}(X^T A_\gamma^{-1}y) \text{ because } \Sigma = \sigma^2 A_\gamma.
\end{aligned}$$

This completes the proof of Fact 4.1. □