

1 Standard Gaussian Process Fitting

Suppose we have a vector of output variables $y \in \mathbb{R}^N$, with associated matrix of input variables $X \in \mathbb{R}^{N \times k}$. Note that N represents the number of observations and k represents the number of input variables. Suppose there is a true relationship between y and X given by $y = f(X)$, where the function f is unknown to the researcher. Gaussian processes aim to approximate f via the following model:

$$\theta \sim g_\theta(\phi_\theta) \quad (1a)$$

$$\sigma \sim g_\sigma(\phi_\sigma) \quad (1b)$$

$$f(X) \sim N(0, K_\theta(X)) \quad (1c)$$

$$y \sim N(f(X), \sigma^2 I_N). \quad (1d)$$

Equation (1a) and eq. (1b) represent priors over parameters θ and σ . Equation (1c) is the prior over values of $f(X)$, formed using kernel function K_θ . A common choice of kernel function is the normal kernel, also referred to as the exponentiated quadratic kernel. This kernel function takes the form

$$K_\theta(X)_{ij} = \alpha^2 \exp\left(-\frac{1}{2}(X_i - X_j)'P^{-1}(X_i - X_j)\right) \quad \forall i = 1, \dots, N; j = 1, \dots, N, \quad (2)$$

where α is a positive scalar, P is a positive definite $k \times k$ matrix, and $X_i \in \mathbb{R}^k$ is the i th row of X . This implies that $K := K_\theta(X)$ is an $N \times N$ positive definite matrix. Finally, eq. (1d) relates observations of output, y , to their predicted values, $f(X)$.

The above form presents difficulties in sampling for a couple of reasons. First, f has a strong prior dependence on kernel parameters θ in eq. (1c) that can lead to inefficient sampling (Betancourt and Girolami 2013). To reduce this prior dependence, first denote the Cholesky decomposition of K as $K = LL'$. Then, we can rewrite the model as

$$\theta \sim g_\theta(\phi_\theta) \quad (3a)$$

$$\sigma \sim g_\sigma(\phi_\sigma) \quad (3b)$$

$$\eta \sim N(0, I_N) \quad (3c)$$

$$K_\theta(X) = LL' \quad (3d)$$

$$f(X) = L\eta \quad (3e)$$

$$y \sim N(f(X), \sigma^2 I_N). \quad (3f)$$

This model is equivalent to the one presented above by properties of multiplying a normally-distributed random variable by a constant matrix, which imply that

$$f(X) = L\eta \sim N(L \times E[\eta], L \times \text{var}(\eta) \times L') \quad (4a)$$

$$\equiv N(L \times 0, L \times I_N \times L') \quad (4b)$$

$$\equiv N(0, K). \quad (4c)$$

This form still presents difficulties in sampling because its posterior distribution is of high dimensionality; using a normal kernel, there are $N + 3$ parameters to sample. We can note, however, that

$$f(X) \sim N(0, K_\theta(X)) \quad (5)$$

and

$$y \sim N(f(X), \sigma^2 I_N), \quad (6)$$

imply that

$$y \sim N(0, K_\theta(X) + \sigma^2 I_N). \quad (7)$$

Altogether, the model becomes

$$\theta \sim g_\theta(\phi_\theta) \quad (8a)$$

$$\sigma \sim g_\sigma(\phi_\sigma) \quad (8b)$$

$$y \sim N(0, K_\theta(X) + \sigma^2 I_N). \quad (8c)$$

This model only requires sampling of θ and σ , greatly reducing the dimensionality of the model. Further, the mean and variance of $y|f(X)$ can still be derived, described in more detail in the following section.

2 Standard Gaussian Process Inference

Suppose that we have drawn S samples of θ and σ from their posterior distributions. Denote the s th sample of θ as $\theta^{[s]}$ and the s th sample of σ as $\sigma^{[s]}$. Suppose that we have a matrix of input variables $X^* \in \mathbb{R}^{N^* \times k}$ for which we want to predict output $y^* \in \mathbb{R}^{N^*}$. For given θ and σ , it is known that

$$y^*|x^*, y, x \sim N(A, B), \quad (9)$$

where

$$A = K_\theta(X^*, X)\Sigma^{-1}y \quad (10a)$$

$$B = K_\theta(X^*) - K_\theta(X^*, X)\Sigma^{-1}K_\theta(X^*, X)', \quad (10b)$$

where

$$\Sigma = K_\theta(X) + \sigma^2 I_N. \quad (11)$$

Using a normal kernel, the kernel functions above are defined as

$$K_\theta(X)_{ij} = \alpha^2 \exp\left(-\frac{1}{2}(X_i - X_j)'P^{-1}(X_i - X_j)\right) \quad \forall i = 1, \dots, N; j = 1, \dots, N \quad (12a)$$

$$K_\theta(X^*)_{ij} = \alpha^2 \exp\left(-\frac{1}{2}(X_i^* - X_j^*)'P^{-1}(X_i^* - X_j^*)\right) \quad \forall i = 1, \dots, N^*; j = 1, \dots, N^* \quad (12b)$$

$$K_\theta(X^*, X)_{ij} = \alpha^2 \exp\left(-\frac{1}{2}(X_i^* - X_j)'P^{-1}(X_i^* - X_j)\right) \quad \forall i = 1, \dots, N^*; j = 1, \dots, N. \quad (12c)$$

In practice it can be more computationally efficient and numerically stable to use the Cholesky decomposition of Σ in calculating A and B .

3 Examples

3.1 Homoskedastic Gaussian Process Over \mathbb{R}

Suppose that $y_i = \sin(x_i)/x_i + \varepsilon_i$ for $x_i \in (0, 10)$ and $\varepsilon \sim N(0, 0.05)$. The data used in this analysis is shown in fig. 1.

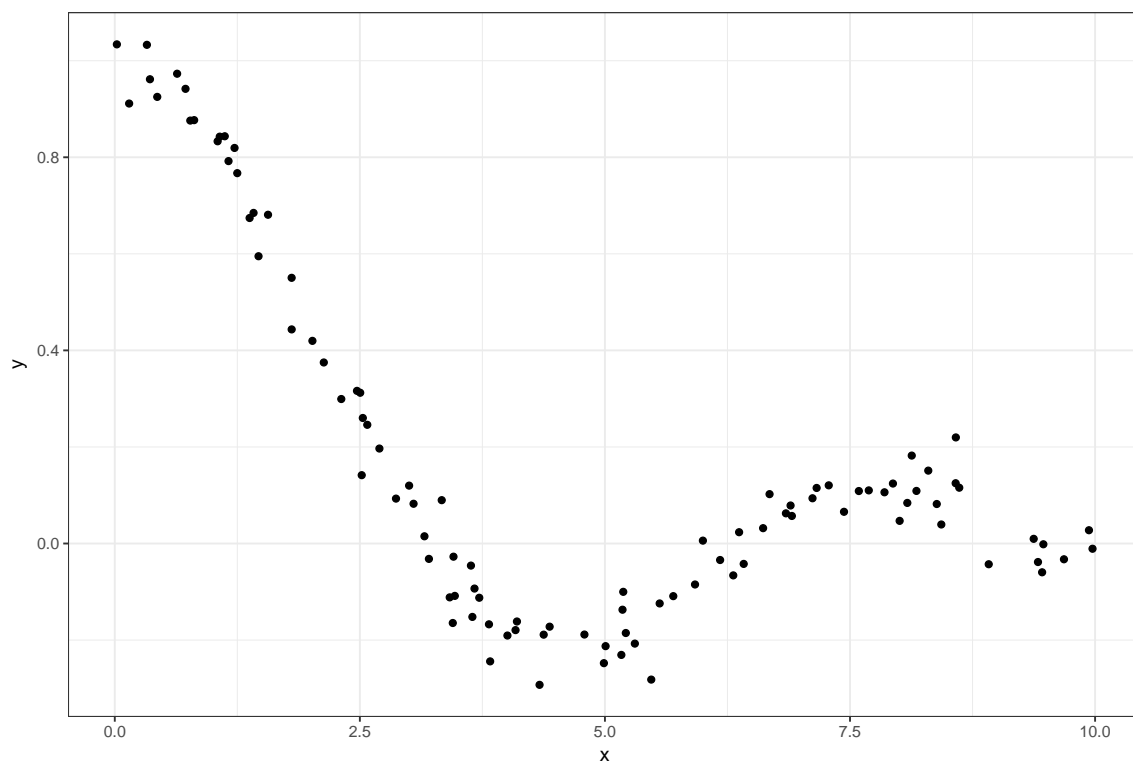


Figure 1: $y_i = \sin(x_i)/x_i + \varepsilon_i$; $\varepsilon_i \sim N(0, 0.05)$