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:

$$\begin{aligned} \theta &\sim g_\theta(\phi_\theta) && \text{(1a)} \\ \sigma &\sim g_\sigma(\phi_\sigma) && \text{(1b)} \\ f(X) &\sim N(0, K_\theta(X)) && \text{(1c)} \\ y &\sim N(f(X), \sigma^2 I_N). && \text{(1d)} \end{aligned}$$

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

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

where α is a positive scalar, P is a positive definite $k \times k$ matrix, and $X_i \in \mathbb{R}^k$ is the ith 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

$$\begin{aligned} \theta &\sim g_\theta(\phi_\theta) && \text{(3a)} \\ \sigma &\sim g_\sigma(\phi_\sigma) && \text{(3b)} \\ \eta &\sim N(0, I_N) && \text{(3c)} \\ K_\theta(X) &= LL' && \text{(3d)} \\ f(X) &= L\eta && \text{(3e)} \\ y &\sim N(f(X), \sigma^2 I_N). && \text{(3f)} \end{aligned}$$

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

$$\begin{aligned} f(X) = L\eta &\sim N(L \times E[\eta], L \times \text{var}(\eta) \times L') && \text{(4a)} \\ &\equiv N(L \times 0, L \times I_N \times L') && \text{(4b)} \\ &\equiv N(0, K). && \text{(4c)} \end{aligned}$$

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

$$f(X) \sim N(0, K_\theta(X)) \tag{5}$$

and

$$y \sim N(f(X), \sigma^2 I_N), \tag{6}$$

imply that

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

Altogether, the model becomes

$$\theta \sim g_\theta(\phi_\theta) \tag{8a}$$
$$\sigma \sim g_\sigma(\phi_\sigma) \tag{8b}$$
$$y \sim N(0, K_\theta(X) + \sigma^2 I_N). \tag{8c}$$

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 sth sample of θ as $\theta^{[s]}$ and the sth 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

$$y^*|x^*, y, x, \theta, \sigma \sim N(A, B), \tag{9}$$

where

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

where

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

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

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

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

Samples from the posterior predictive distribution of $f(y^*)|x^*, y, x$ can be formed by drawing samples of $y^{*[s]}|x^*, y, x, \theta^{[s]}, \sigma^{[s]}$ and applying f to each of those samples. For example, a distribution of 95% quantiles of predicted output is given as

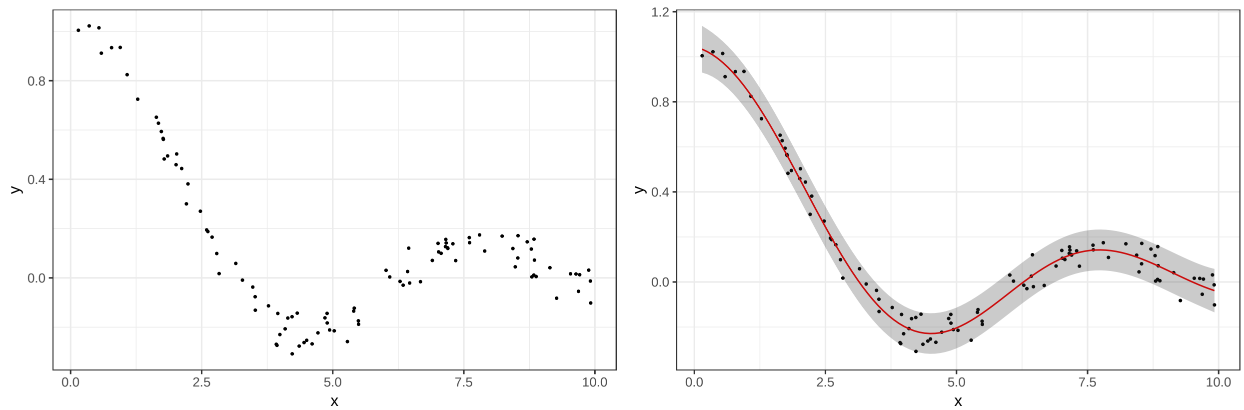
$$q_{0.95}^N(y^{*[s]})|x^*, y, x, \theta^{[s]}, \sigma^{[s]}. \tag{13}$$

Distributional statistics (mean, quantiles, etc.) can then be calculated from those posterior predictive distributions as usual.

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. 1a.



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

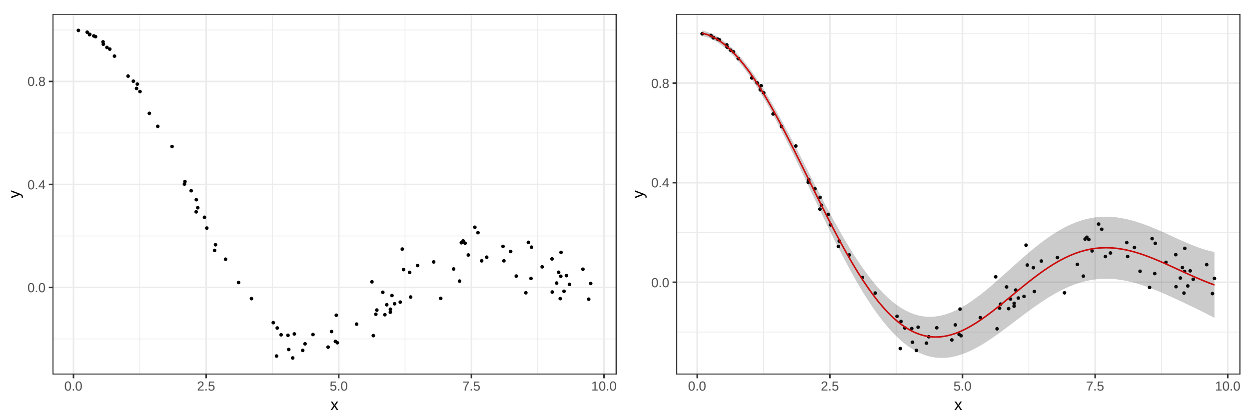
(b) Gaussian Process Fit and 90% Confidence Interval

Figure 1: Homoskedastic Gaussian Process Over $\mathbb{R}$

The model presented in section 1 was used to estimate a Gaussian process and methods described section 2 were used to draw inference on the mean and variance of predicted values. Plots of the prediction mean and associated 90% confidence interval are shown in fig. 1b.

3.2 Heteroskedastic Gaussian Process Over $\mathbb{R}$

Suppose that $y_i = \sin(x_i)/x_i + \varepsilon_i$ for $x_i \in (0, 10)$ and $\varepsilon_i \sim N(0, 0.01x_i)$. The data used in this analysis is shown in fig. 2a.



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

(b) Gaussian Process Fit and 90% Confidence Interval

Figure 2: Heteroskedastic Gaussian Process Over $\mathbb{R}$

This analysis uses a Gaussian process to model both the conditional mean and variance of y about its conditional mean. Specifically, the model is

$$\theta \sim g_\theta(\phi_\theta) \tag{14a}$$
$$\theta_\sigma \sim g_\sigma(\phi_\sigma) \tag{14b}$$
$$\xi \sim g_\xi(\phi_\xi) \tag{14c}$$
$$f_\sigma(X) \sim N(0, K_{\theta_\sigma}(X)) \tag{14d}$$
$$f(X) \sim N(0, K_\theta(X)) \tag{14e}$$
$$\ln \sigma \sim N(f_\sigma(X), \xi^2 I_N) \tag{14f}$$
$$y \sim N(f(X), \mathrm{diag}(\sigma^2)). \tag{14g}$$

In eq. (14g), the diag function creates a diagonal matrix from the vector σ^2, where the ith diagonal element of the resulting matrix is the ith element of σ^2. Plots of the prediction mean and associated 90% confidence interval are shown in fig. 2b. Further, estimated and true standard deviation are shown in fig. 3. It can be seen that heteroskedasticity in the data is accurately captured by the estimates.

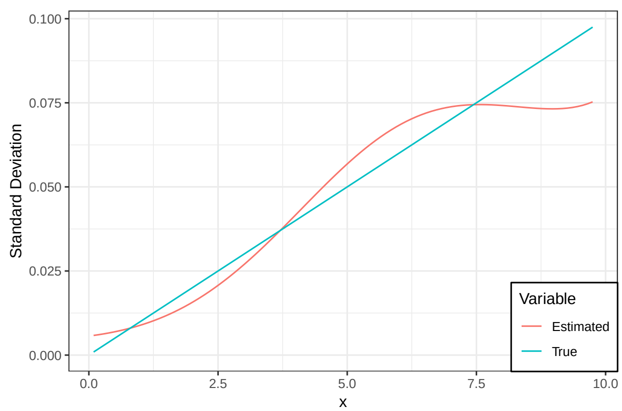


Figure 3: Estimated and True Variance