

# HOMework 3

REGRESSION, GAUSSIAN PROCESSES, AND BOOSTING

DANA VAN AKEN

## Problem 1: Gaussian Processes

- (a) A comparison of covariance functions: see figures 1, 2, and 3.
- (b)  $\sigma^2$  affects the “noisyness” of the output points,  $y_i$ . Figure 4 shows that as we increase  $\sigma^2$ , the amount of noise (or ”spikyness” in the graph) increases.
- (c) Show  $p(x_1|x_2) \propto p(x_1, x_2)$

We want to find  $\mu_{x_1|x_2}$  and  $\Sigma_{x_1|x_2}$  in:

$$\begin{aligned} p(x_1|x_2) &= Z \exp \left( -\frac{1}{2}(x - \mu_{x_1|x_2})^T \Sigma_{x_1|x_2}^{-1} (x - \mu_{x_1|x_2}) \right) \\ &= Z \exp \left( -\frac{1}{2}(x^T \Sigma_{x_1|x_2}^{-1} x - x^T \Sigma_{x_1|x_2}^{-1} \mu_{x_1|x_2} - \mu_{x_1|x_2}^T \Sigma_{x_1|x_2}^{-1} x + \mu_{x_1|x_2}^T \Sigma_{x_1|x_2}^{-1} \mu_{x_1|x_2}) \right) \\ &= Z \exp \left( -\frac{1}{2}x^T \Sigma_{x_1|x_2}^{-1} x + x^T \Sigma_{x_1|x_2}^{-1} \mu_{x_1|x_2} - \frac{1}{2}\mu_{x_1|x_2}^T \Sigma_{x_1|x_2}^{-1} \mu_{x_1|x_2} \right) \end{aligned} \quad (1)$$

where  $Z = \frac{1}{\sqrt{(2\pi)^k |\Sigma_{x_1|x_2}|}}$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

Let  $\Sigma^{-1} = \Lambda^{-1}$  such that

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} = \Lambda$$

We can focus on the exponent since we want to find  $\mu_{x_1|x_2}$  and  $\Sigma_{x_1|x_2}$ .

$$\begin{aligned}
exp &= -\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \\
&= -\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \\
&= -\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \\
&= -\frac{1}{2}(x_1 - \mu_1)^T \Lambda_{11}(x_1 - \mu_1) - \frac{1}{2}(x_1 - \mu_1)^T \Lambda_{12}(x_2 - \mu_2) \\
&\quad - \frac{1}{2}(x_2 - \mu_2)^T \Lambda_{21}(x_1 - \mu_1) - \frac{1}{2}(x_2 - \mu_2)^T \Lambda_{22}(x_2 - \mu_2)
\end{aligned}$$

We can call the last term,  $-\frac{1}{2}(x_2 - \mu_2)^T \Lambda_{22}(x_2 - \mu_2)$ ,  $C$  since it does not depend on  $x_1$  (constant).

$$\begin{aligned}
&= -\frac{1}{2}(x_1 - \mu_1)^T \Lambda_{11}(x_1 - \mu_1) - \frac{1}{2}(x_1 - \mu_1)^T \Lambda_{12}(x_2 - \mu_2) - \frac{1}{2}(x_2 - \mu_2)^T \Lambda_{21}(x_1 - \mu_1) + C \\
&= -\frac{1}{2}x_1^T \Lambda_{11}x_1 + \frac{1}{2}x_1^T \Lambda_{11}\mu_1 + \frac{1}{2}\mu_1^T \Lambda_{11}x_1 - \frac{1}{2}\mu_1^T \Lambda_{11}\mu_1 \\
&\quad - \frac{1}{2}x_1^T \Lambda_{12}x_2 + \frac{1}{2}x_1^T \Lambda_{12}\mu_2 + \frac{1}{2}\mu_1^T \Lambda_{12}x_2 - \frac{1}{2}\mu_1^T \Lambda_{12}\mu_2 \\
&\quad - \frac{1}{2}x_2^T \Lambda_{21}x_1 + \frac{1}{2}x_2^T \Lambda_{21}\mu_1 + \frac{1}{2}\mu_2^T \Lambda_{21}x_1 - \frac{1}{2}\mu_2^T \Lambda_{21}\mu_1 + C
\end{aligned}$$

Again, include any constants that do not depend on  $x_1$  in  $C$ .

$$= -\frac{1}{2}x_1^T \Lambda_{11}x_1 + \frac{1}{2}x_1^T \Lambda_{11}\mu_1 + \frac{1}{2}\mu_1^T \Lambda_{11}x_1 - \frac{1}{2}x_1^T \Lambda_{12}x_2 + \frac{1}{2}x_1^T \Lambda_{12}\mu_2 - \frac{1}{2}x_2^T \Lambda_{21}x_1 + \frac{1}{2}\mu_2^T \Lambda_{21}x_1 + C$$

We can use the fact that  $\Lambda_{21} = \Lambda_{12}^T$  to reduce the equation.

$$= -\frac{1}{2}x_1^T \Lambda_{11}x_1 + x_1^T \Lambda_{11}\mu_1 - x_1^T \Lambda_{12}x_2 + x_1^T \Lambda_{12}\mu_2 + C \tag{2}$$

By comparing the only second-order  $x_1$  term in equations 1 and 2, we can see that:

$$\Sigma_{x_1|x_2} = \Lambda_{11}^{-1} \tag{3}$$

Look at the first-order  $x_1$  terms and factor:

$$x_1^T \Lambda_{11} \mu_1 - x_1^T \Lambda_{12} x_2 + x_1^T \Lambda_{12} \mu_2 = x_1^T (\Lambda_{11} \mu_1 - \Lambda_{12} (x_2 - \mu_2))$$

Comparing this term to equation 1, we can see that:

$$\begin{aligned} \Sigma_{x_1|x_2}^{-1} \mu_{x_1|x_2} &= (\Lambda_{11} \mu_1 - \Lambda_{12} (x_2 - \mu_2)) \\ \mu_{x_1|x_2} &= \Sigma_{x_1|x_2} (\Lambda_{11} \mu_1 - \Lambda_{12} (x_2 - \mu_2)) \\ \mu_{x_1|x_2} &= \Lambda_{11}^{-1} (\Lambda_{11} \mu_1 - \Lambda_{12} (x_2 - \mu_2)) && \text{from equation 3} \\ \mu_{x_1|x_2} &= \mu_1 - \Lambda_{11}^{-1} \Lambda_{12} (x_2 - \mu_2) \end{aligned}$$

Use the **Schur Complement** to put  $\Sigma_{x_1|x_2}$  and  $\mu_{x_1|x_2}$  in terms of  $\Sigma$

$$\begin{aligned} \Sigma_{x_1|x_2} &= \Lambda_{11}^{-1} \\ &= [(\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})^{-1}]^{-1} \\ &= \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \\ \mu_{x_1|x_2} &= \mu_1 - \Lambda_{11}^{-1} \Lambda_{12} (x_2 - \mu_2) \\ &= \mu_1 - (\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}) (-(\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21})^{-1} \Sigma_{12} \Sigma_{22}^{-1}) (x_2 - \mu_2) \\ &= \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2) \end{aligned}$$

(d)

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim N \left( \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right) \iff \begin{bmatrix} f(X) \\ Y_* \end{bmatrix} \sim N \left( \mathbf{0}, \begin{bmatrix} k(X, X) & k(X, X_*) \\ k(X_*, X) & k(X_*, X_*) + \sigma^2 I \end{bmatrix} \right)$$

$$\begin{aligned} \mu_{f(X)|Y_*} &= 0 + k(X, X_*) (k(X_*, X_*) + \sigma^2 I)^{-1} (Y_* - 0) \\ &= k(X, X_*) (k(X_*, X_*) + \sigma^2 I)^{-1} Y_* \\ \Sigma_{f(X)|Y_*} &= k(X, X) - k(X, X_*) (k(X_*, X_*) + \sigma^2 I)^{-1} k(X_*, X) \end{aligned}$$

(e) A comparison of covariance functions sampled from  $p(f(X)|Y_*)$ : see figures 5, 6, and 7.

(f) Figures 8, 9, and 10 show  $f(X)|Y_*$  plotted for increasing values of  $\lambda^2$ . Smaller values of  $\lambda^2$  result in higher variance and lower bias, whereas larger values of  $\lambda^2$  result in lower variance and higher bias. The function is more unstable for smaller values of  $\lambda^2$ , (i.e., changing the training points would significantly

change the function), than for large values of  $\lambda^2$ .

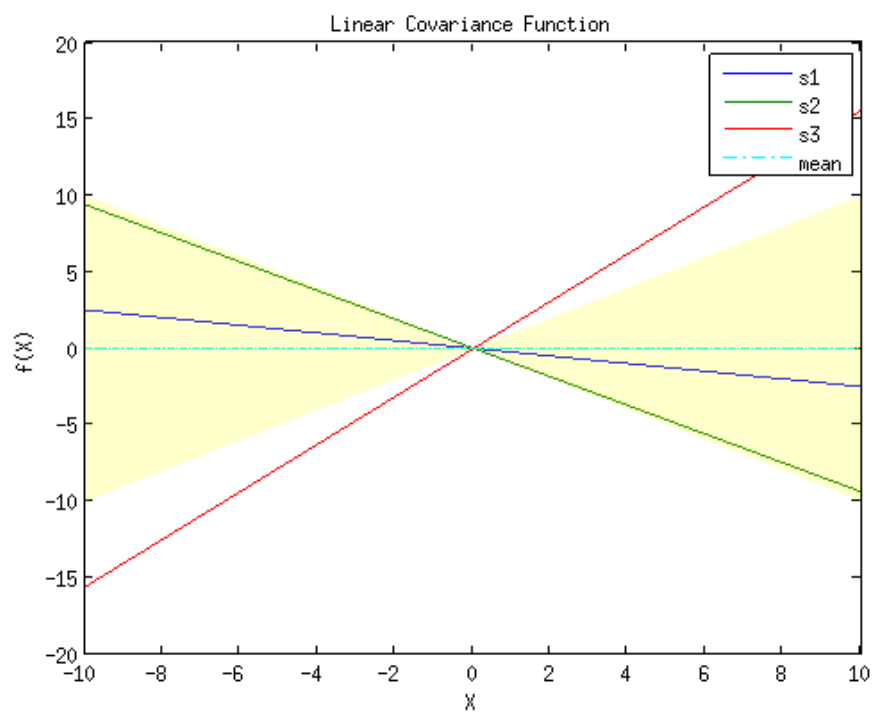


Figure 1: Linear Covariance Function

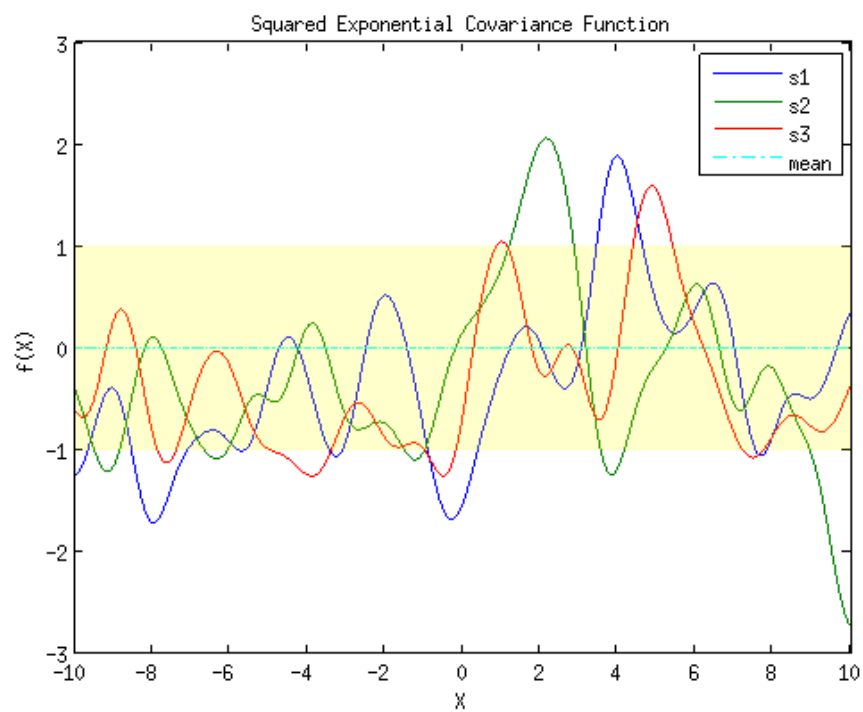


Figure 2: Square Exponential Covariance Function

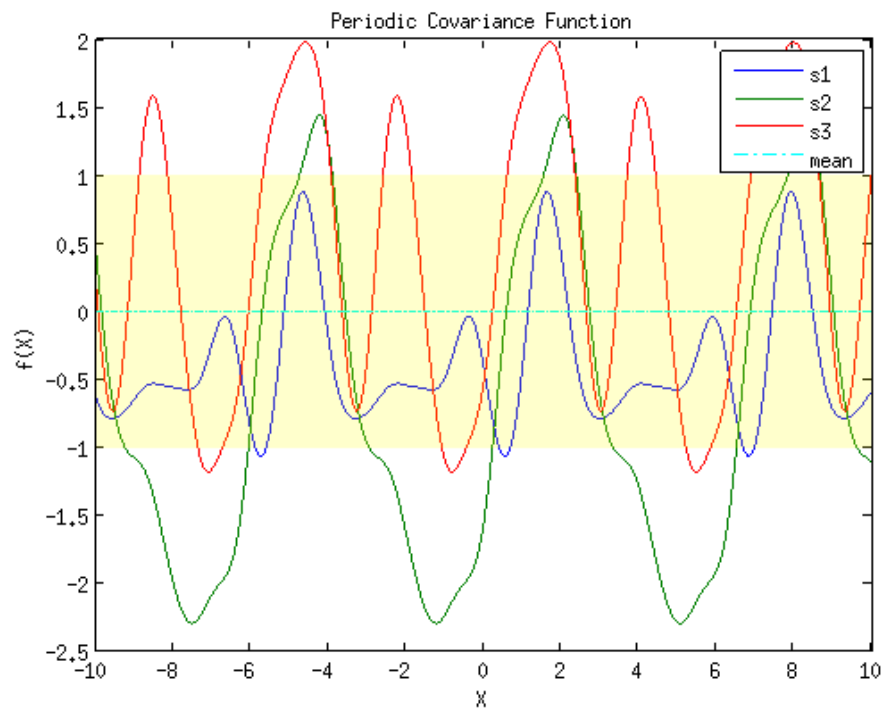


Figure 3: Periodic Covariance Function

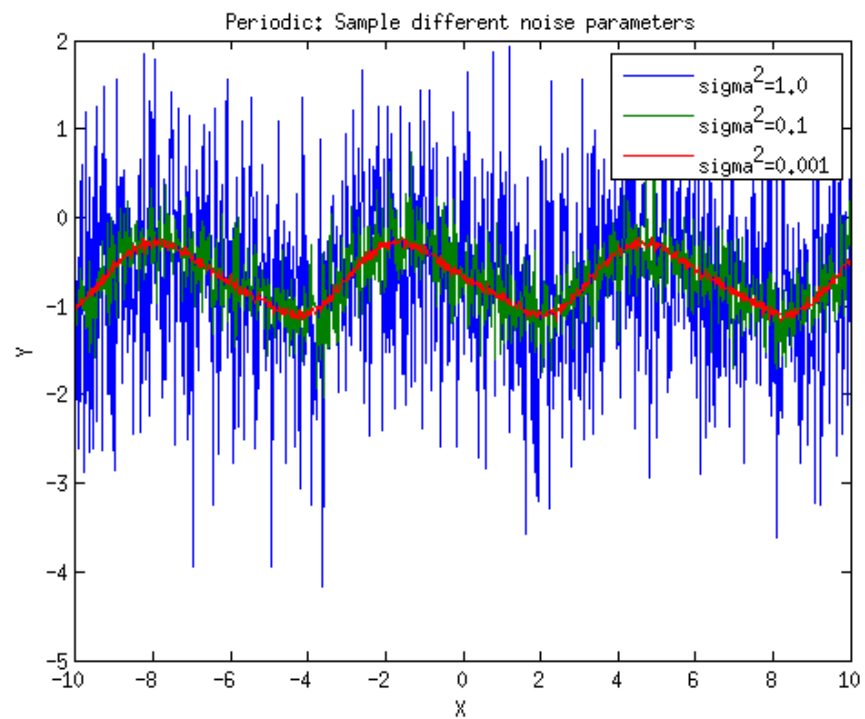


Figure 4: Sampling Different Gaussian Noise Parameters Using a Periodic Covariance Function

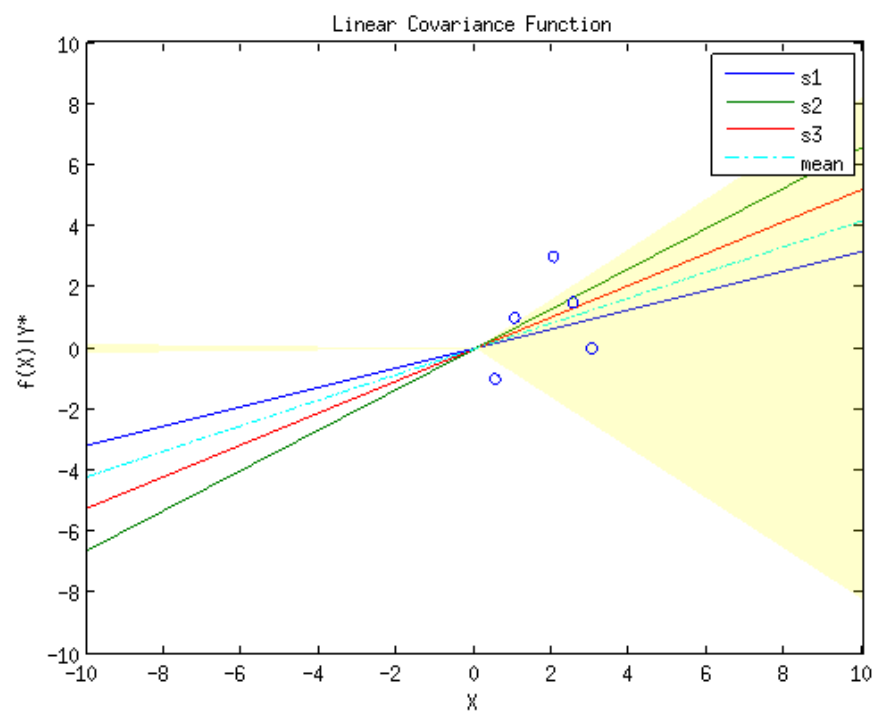


Figure 5: Linear Covariance Function

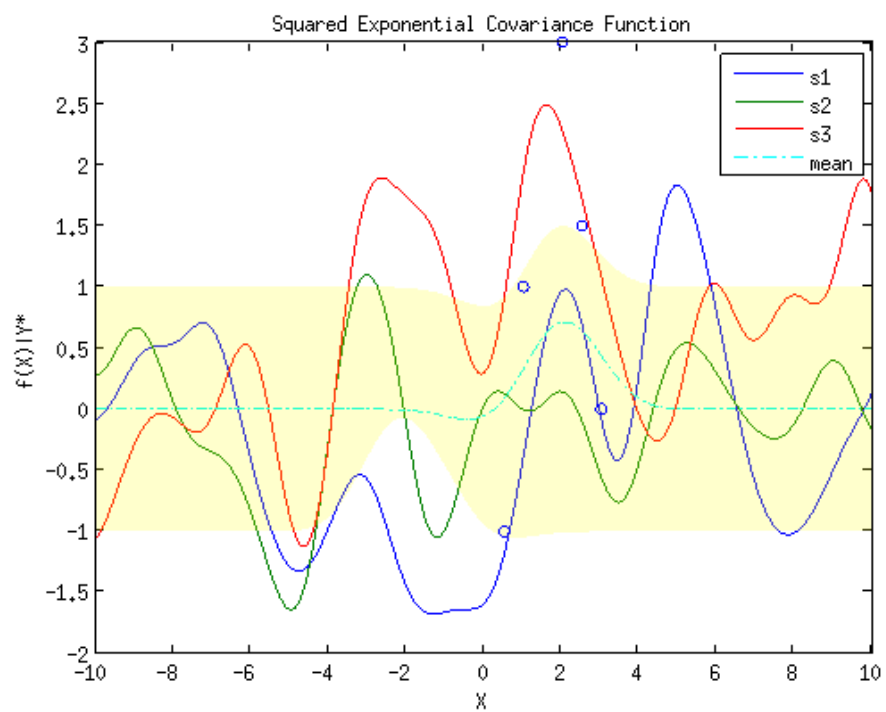


Figure 6: Square Exponential Covariance Function

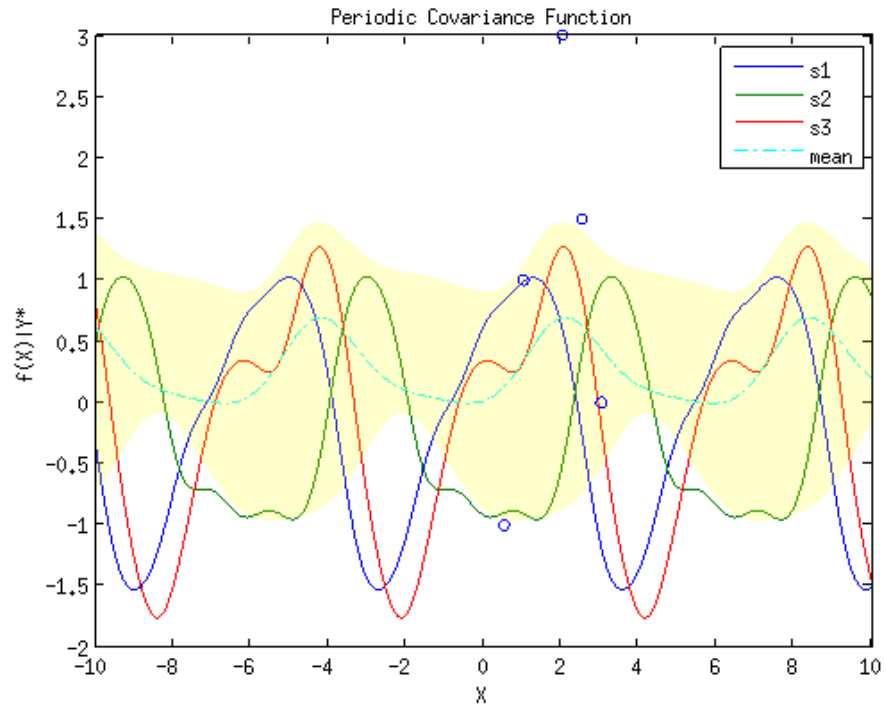


Figure 7: Periodic Covariance Function

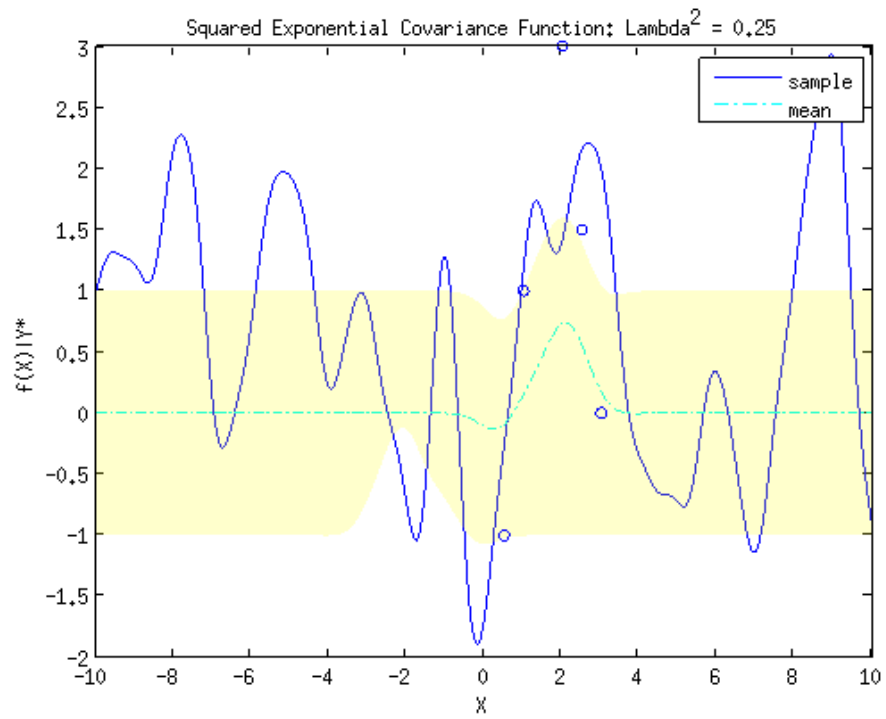


Figure 8: Sampling Different  $\lambda^2$  Parameters Using the Squared Exponential Function



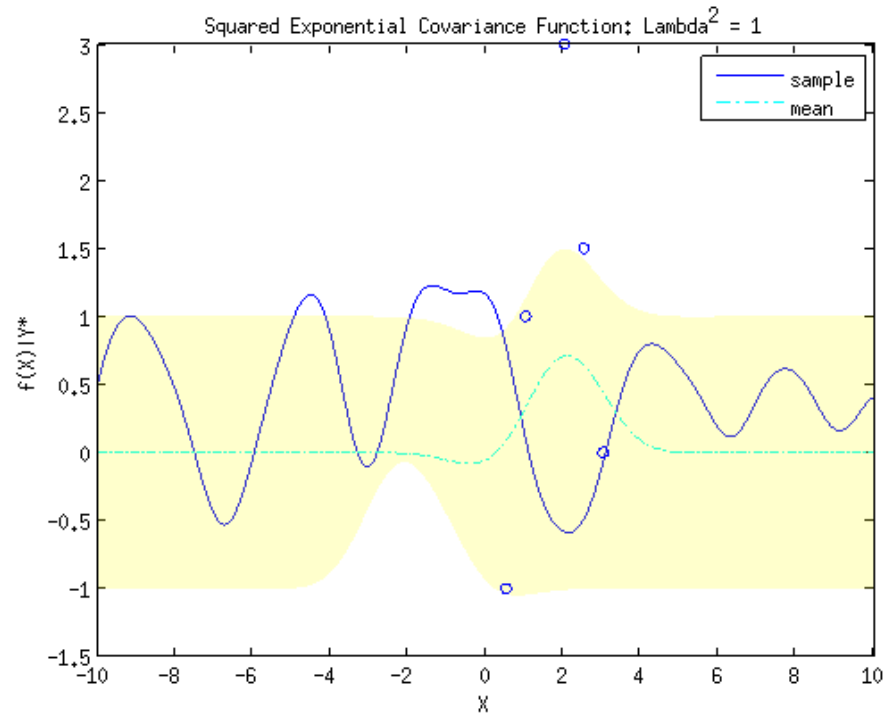


Figure 9: Sampling Different  $\lambda^2$  Parameters Using the Squared Exponential Function

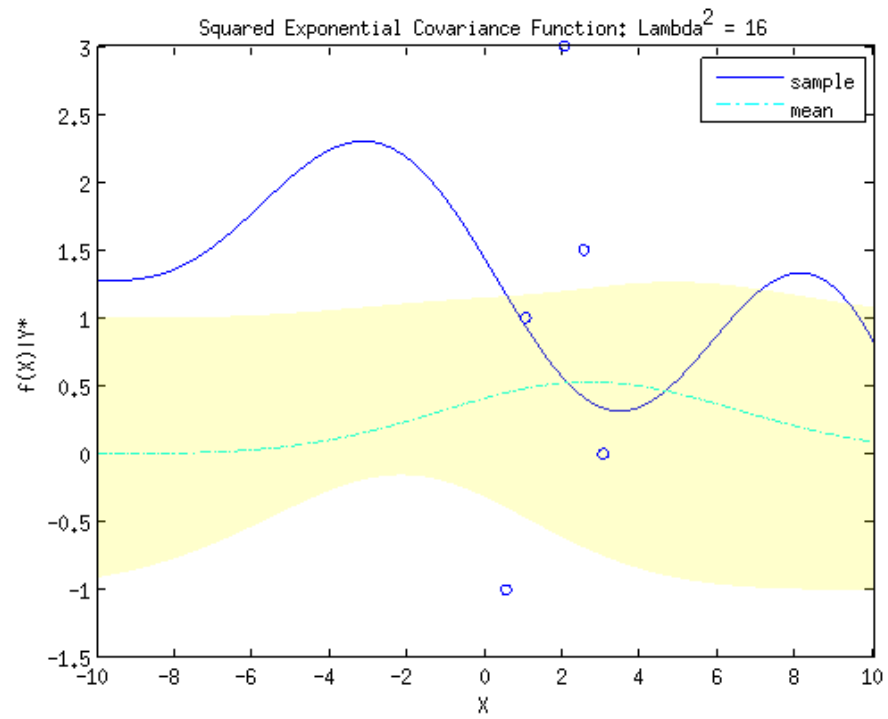


Figure 10: Sampling Different  $\lambda^2$  Parameters Using the Squared Exponential Function