1. Let $Y(\mathbf{x})$ be a Gaussian process (GP) with the following properties:

   - $Y(\mathbf{x})$ has a Gaussian distribution;
   - $\mathrm{E}(Y(\mathbf{x}))$ is given by a regression model,

   $$\sum_{j=1}^{k} \beta_j f_j(\mathbf{x}) \equiv \mathbf{f}^T(\mathbf{x})\boldsymbol{\beta},$$

     where $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_k(\mathbf{x}))^T$ is a vector of k given (known) functions, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ is a vector of unknown regression parameters;
   - $\mathrm{Var}(Y(\mathbf{x})) = \sigma^2$, i.e., constant;
   - The correlation $\mathrm{Cor}(Y(\mathbf{x}), Y(\mathbf{x}'))$ is given by $R(\mathbf{x}, \mathbf{x}')$, a known correlation function.

   We will be taking the variance and correlation structure as known in this question, but $\boldsymbol{\beta}$ will be estimated.

   The GP is observed at n distinct locations, $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$, in the $\mathbf{x}$ space, giving the random data vector $\mathbf{Y} = (Y(\mathbf{x}^{(1)}), \ldots, Y(\mathbf{x}^{(n)}))^T$.

   We define the $n \times n$ correlation matrix $\mathbf{R}$ with element i, j given by $R(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ and the $n \times 1$ vector $\mathbf{r}(\mathbf{x})$ with element i given by $R(\mathbf{x}, \mathbf{x}^{(i)})$. Also, let $\mathbf{F}$ be the $n \times k$ matrix with row i containing $\mathbf{f}^T(\mathbf{x}^{(i)})$.

   Consider predicting $Y(\mathbf{x}^*)$, where $\mathbf{x}^*$ is a specific value of the input vector, by a predictor that is a linear combination of the data: $\hat{Y}(\mathbf{x}^*) = \mathbf{w}^T(\mathbf{x}^*)\mathbf{Y}$, where $\mathbf{w}^T(\mathbf{x}^*) = (w_1(\mathbf{x}^*), \ldots, w_n(\mathbf{x}^*))$. We are taking a frequentist viewpoint here: The $Y(\mathbf{x}^{(i)})$ and hence $\hat{Y}(\mathbf{x}^*)$ are random variables that will vary from one sample realization to another according to the above probability model.

   (a) Define the bias of prediction as

   $$\mathrm{E}(\hat{Y}(\mathbf{x}^*)) - \mathrm{E}(Y(\mathbf{x}^*)).$$

   Show that the bias is

   $$(\mathbf{F}^T\mathbf{w}(\mathbf{x}^*) - \mathbf{f}(\mathbf{x}^*))^T\boldsymbol{\beta}.$$

   (b) Define the mean squared error (MSE) of prediction as

   $$\mathrm{MSE}(\hat{Y}(\mathbf{x}^*)) = \mathrm{E}(\hat{Y}(\mathbf{x}^*) - Y(\mathbf{x}^*))^2.$$

   Show that $\mathrm{MSE}(\hat{Y}(\mathbf{x}^*))$ is

   $$((\mathbf{F}^T\mathbf{w}(\mathbf{x}^*) - \mathbf{f}(\mathbf{x}^*))^T\boldsymbol{\beta})^2 + \sigma^2\left(1 + \mathbf{w}^T(\mathbf{x}^*)\mathbf{R}\mathbf{w}(\mathbf{x}^*) - 2\mathbf{w}^T(\mathbf{x}^*)\mathbf{r}(\mathbf{x}^*)\right).$$

(c) Minimizing the mean squared error subject to unbiasedness implies minimizing

$$1 + \mathbf{w}^T(\mathbf{x}^*)\mathbf{R}\mathbf{w}(\mathbf{x}^*) - 2\mathbf{w}^T(\mathbf{x}^*)\mathbf{r}(\mathbf{x}^*)$$

subject to

$$\mathbf{F}^T\mathbf{w}(\mathbf{x}^*) - \mathbf{f}(\mathbf{x}^*) = \mathbf{0}.$$

By introducing a $k \times 1$ vector of Lagrange multipliers, $\boldsymbol{\lambda}$, show that $\mathbf{w}(\mathbf{x}^*)$ and $\boldsymbol{\lambda}$ satisfy

$$\begin{pmatrix} \mathbf{R} & \mathbf{F} \\ \mathbf{F}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w}(\mathbf{x}^*) \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{r}(\mathbf{x}^*) \\ \mathbf{f}(\mathbf{x}^*) \end{pmatrix}.$$

Here, $\boldsymbol{\lambda}$ is really a function of $\mathbf{x}^*$, too. The notation can be made friendlier by dropping the dependence on $\mathbf{x}^*$ everywhere.

(d) Using standard results on the inverse of a partitioned matrix, we have

$$\begin{pmatrix} \mathbf{R} & \mathbf{F} \\ \mathbf{F}^T & \mathbf{0} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{F}\mathbf{K}^{-1}\mathbf{F}^T\mathbf{R}^{-1} & \mathbf{R}^{-1}\mathbf{F}\mathbf{K}^{-1} \\ \mathbf{K}^{-1}\mathbf{F}^T\mathbf{R}^{-1} & -\mathbf{K}^{-1} \end{pmatrix},$$

where $\mathbf{K} = \mathbf{F}^T\mathbf{R}^{-1}\mathbf{F}$. Hence, show that the coefficients $\mathbf{w}(\mathbf{x}^*)$ are given by

$$\mathbf{w}(\mathbf{x}^*) = \mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*) + \mathbf{R}^{-1}\mathbf{F}\mathbf{K}^{-1}(\mathbf{f}(\mathbf{x}^*) - \mathbf{F}^T\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*)).$$

(e) Hence, show that

$$\hat{Y}(\mathbf{x}^*) = \mathbf{f}^T(\mathbf{x}^*)\hat{\boldsymbol{\beta}} + \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{F}\hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{F}^T\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{R}^{-1}\mathbf{Y}$ is the generalized least squares estimator of $\boldsymbol{\beta}$.

(f) Substitute the optimal linear-combination coefficients, $\mathbf{w}(\mathbf{x}^*)$, from part 1d into the MSE of part 1b to show that the optimal MSE is

$$\sigma^2\left(1 - \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*) + \left(\mathbf{f}(\mathbf{x}^*) - \mathbf{F}^T\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*)\right)^T\mathbf{K}^{-1}\left(\mathbf{f}(\mathbf{x}^*) - \mathbf{F}^T\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*)\right)\right).$$

This is a generalization of the formula on Slide 12 of Module 3.

2. Let $Y(\mathbf{x})$ be a GP with the mean, variance, and correlation properties of Question 1. The same notation will also be used. Unlike Question 1, however, we will *not* be estimating the vector of regression parameters, $\boldsymbol{\beta}$, when developing a prediction formula; all parameters are taken as known.

Again, the GP is observed at n distinct locations, $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$, in the $\mathbf{x}$ space, giving the data vector $\mathbf{Y} = (Y(\mathbf{x}^{(1)}), \ldots, Y(\mathbf{x}^{(n)}))^T$. Hence, the joint density of $\mathbf{Y}$ is multivariate normal:

$$f_{\mathbf{Y}}(\mathbf{y} \,|\, \mu, \sigma^2, \mathbf{R}) = \frac{1}{(2\pi\sigma^2)^{n/2}}\frac{1}{\det^{1/2}(\mathbf{R})} \times \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})\right),$$

where $\det^{1/2}(\mathbf{R})$ is the square root of the determinant of $\mathbf{R}$.

Consider predicting $Y(\mathbf{x}^*)$, where $\mathbf{x}^*$ is a fixed location. The joint density of $(\mathbf{Y}, Y(\mathbf{x}^*))$ is multivariate normal for the $n+1$ random variables. We will use the conditional distribution of $Y(\mathbf{x}^*)$ given $\mathbf{Y}$, i.e.,

$$f_{Y(\mathbf{x}^*)\,|\,\mathbf{Y}}(y(\mathbf{x}^*)\,|\,\mathbf{y}, \mu, \sigma^2, \mathbf{R}),$$

as a predictive distribution.

(a) Using results on conditional distributions, show that the predictive distribution is (univariate) normal with mean

$$\mathbf{f}^T(\mathbf{x}^*)\boldsymbol{\beta} + \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta}),$$

and variance

$$\sigma^2\left(1 - \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*)\right).$$

Hints: The joint distribution of $(\mathbf{Y}, Y(\mathbf{x}^*))^T$ has correlation matrix

$$\begin{pmatrix} \mathbf{R} & \mathbf{r}(\mathbf{x}^*) \\ \mathbf{r}^T(\mathbf{x}^*) & 1 \end{pmatrix}.$$

The inverse of this partitioned matrix is

$$\begin{pmatrix} \mathbf{R}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \frac{1}{q(\mathbf{x}^*)}\begin{pmatrix} -\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*) \\ 1 \end{pmatrix}\begin{pmatrix} -\mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1} & 1 \end{pmatrix},$$

where $q(\mathbf{x}^*) = 1 - \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*)$, and its determinant is

$$\det(\mathbf{R})q(\mathbf{x}^*).$$

(b) In what sense is this a "posterior" distribution?

(c) Compare the predictive variance with that in part 1f of Question 1? Is it smaller or larger? Why?

3. Let $Y(x)$ be a GP indexed by one-dimensional input, x. Its properties are $\mathrm{E}(Y(x)) = 0$, $\mathrm{Var}(Y(x)) = \sigma^2$, and

$$\mathrm{Cor}(Y(x), Y(x+h)) = R(x, x+h) = \exp\left(-\theta h^2\right).$$

Thus, the correlation function is from the squared-exponential family.

(a) Find $\mathrm{E}(Y(x+h) - Y(x))^2$.

(b) Hence, show that

$$\lim_{h\to 0} \mathrm{E}\left(\frac{Y(x+h) - Y(x)}{h}\right)^2 = 2\sigma^2\theta.$$

(c) Hence contrast the behaviour of realizations from two GP models with the same value of σ^2 but different values of θ.

(d) Now suppose $Y(\mathbf{x})$ is indexed by d-dimensional input, $\mathbf{x}$. The correlation function is a product of one-dimensional squared-exponential correlation functions:

$$\mathrm{Cor}(Y(\mathbf{x}), Y(\mathbf{x}')) = R(\mathbf{x}, \mathbf{x}') = \prod_{j=1}^{d} \exp(-\theta_j (x_j - x_j')^2).$$

Let $\mathbf{x}' = \mathbf{x} + \boldsymbol{\delta}_j$, where $\boldsymbol{\delta}_j$ is a $d \times 1$ vector with h in element j and 0 elsewhere. Find

$$\lim_{h \to 0} \mathrm{E} \left(\frac{Y(\mathbf{x} + \boldsymbol{\delta}_j) - Y(\mathbf{x})}{h} \right)^2 .$$

(e) What does this result say about the interpretation of the parameters $\theta_1, \ldots, \theta_d$ in the squared-exponential correlation function?

4. Suppose $Y(x)$ for $x \in [0, 1]$ follows a GP with mean zero, variance σ^2, and correlation function $R(Y(x), Y(x')) = R(x, x') = \exp(-\theta(x - x')^2)$. In this question we will consider realizations of $Y(x)$ at the 11 equally spaced points $x^{(i)} = (i - 1)/10$ for $i = 1, \ldots, 11$, i.e., with $\mathbf{x}^{(1)} = 0$ and $\mathbf{x}^{(11)} = 1$. You will write your own `R` code to carry out all computations, which should be handed in.

(a) Generate a realization from the above GP with $\sigma^2 = 1$ and $\theta = 1$ at the 11 locations $x^{(i)}$. Plot the observations (keep your $\left(x^{(i)}, y(x^{(i)})\right)$ realization as we will use it below).

(b) Generate a realization from the above GP with $\sigma^2 = 5$ and $\theta = 1$ at the 11 locations $x^{(i)}$. Plot the observations and comment on the impact of the increase in variance from part 4a.

(c) Generate a realization from the above GP with $\sigma^2 = 1$ and $\theta = 5$ at the 11 locations $x^{(i)}$. Plot the observations and comment on the impact of the increase in θ from part 4a (keep your $\left(x^{(i)}, y(x^{(i)})\right)$ realization as we will use it below).

(d) Let $\mathbf{x}^{(\mathrm{O})} = (x^{(1)}, x^{(3)}, \ldots, x^{(11)})$ be the locations with odd indices, and similarly let $\mathbf{x}^{(\mathrm{e})} = (x^{(2)}, x^{(4)}, \ldots, x^{(10)})$. Consider the GP realization from part 4a with $\sigma^2 = 1$ and $\theta = 1$ (i.e., the parameters of the GP that generated the data are known to be $\sigma^2 = 1$ and $\theta = 1$). Use only the observations at $\mathbf{x}^{(\mathrm{O})}$ to predict the observations at $\mathbf{x}^{(\mathrm{e})}$. Compute the root mean square prediction error (RMSE).

(e) Consider the GP realization from part 4c with $\sigma^2 = 1$ and $\theta = 5$ (i.e., again the values of the parameters of the GP that generated the data are known). Use only the observations at $\mathbf{x}^{(\mathrm{O})}$ to predict the observations at $\mathbf{x}^{(\mathrm{e})}$. Compute the RMSE and compare it with that in part 4d.

(f) Consider the GP realization from part 4a. Predict the observations at $\mathbf{x}^{(\mathrm{e})}$ using the data at $\mathbf{x}^{(\mathrm{O})}$ only, but assume $\sigma^2 = 1$ and $\theta = 100$ for the GP used for prediction (i.e., θ is misspecified). Compute the RMSE and compare it with that in part 4d.

(g) Consider the GP realization from part 4a. Predict the observations at $\mathbf{x}^{(\mathrm{e})}$ using the data at $\mathbf{x}^{(\mathrm{O})}$ only, but assume $\sigma^2 = 1$ and $\theta = 0.1$ for the GP used for prediction (i.e., θ is misspecified again). Compute the RMSE and compare it with that in part 4d.