SDS 383D: Exercises 3 – Linear smoothing and Gaussian processes

March 1, 2017

Professor Scott

Spencer Woody

Problem 1

Basic Concepts

(A)

Problem 2

Curve fitting by linear smoothing

In this problem, consider a general nonlinear regression with one predictor and one response, $y_i = f(x_i) + \epsilon_i$, where ϵ_i are mean-zero random variables.

(A) For now, consider a linear regression on a response y_i with one predictor x_i, and both y_i and x_i have had their means subtracted, so the $y_i = \beta x_i + \epsilon_i$. Define $S_x := \sum_{i=1}^n x_i^2$. The least squares estimate for the coefficient, from Exercises 1, is

$$\begin{aligned}
\hat{\beta} &= (X^T X)^{-1} X^T y \\
&= (x^T x)^{-1} x^T y \\
&= \frac{\sum_{i=1}^n x_i \cdot y_i}{\sum_{i=1}^n x_i^2} \\
&= \frac{\sum_{i=1}^n x_i \cdot y_i}{S_x} \\
&= \sum_{i=1}^n \frac{x_i}{S_x} \cdot y_i.
\end{aligned}$$

So now our prediction $y^\star | x^\star$ is,

$$\begin{aligned}
\hat{y}^\star &= \hat{f}(x^\star) \\
&= \hat{\beta} x^\star \\
&= \left(\sum_{i=1}^n \frac{x_i}{S_x} \cdot y_i \right) \cdot x^\star \\
&= \sum_{i=1}^n \left(\frac{x_i}{S_x} \cdot x^\star \right) \cdot y_i,
\end{aligned}$$

which we recognize as being in the form of the general *linear smoother*

$$\hat{f}\,(x^\star) = \sum_{i=1}^n w(x_i, x^\star) \cdot y_i$$

for some weight function $w(x_i, x^\star)$. In particular, the weight function for linear regression gives weight to each y_i proportional to the value of x_i. Contrast this with the k-nearest neighbors smoothing weight function,

$$w_K(x_i, x^\star) = \begin{cases} 1/K & \text{if } x_i \text{ is one of the K closest sample points to } x^\star \\ 0 & \text{otherwise} \end{cases},$$

which gives *equal* weight to y_is but *only* to the k-nearest neighbors of $x^\star$.

(B) Now we have the very general weight function

$$w(x_i, x^\star) = \frac{1}{h} \cdot K\left(\frac{x_i - x^\star}{h}, \right)$$

where $K(\bullet)$ is some kernel function. The script `myfuns03.R` in the appendix shows an R function for linear smoothing, as well functions for the uniform and Gaussian kernels. Figure 1 shows an example of smoothing with a bandwidth of 0.75 for a cubic function $f(x)$ with iid residuals from the $\mathcal{N}(0, 15^2)$ distribution.

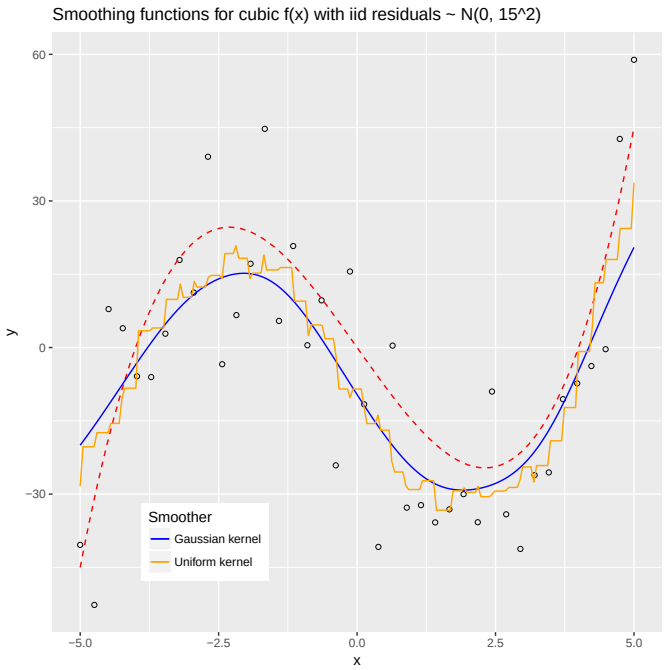

Figure 1: Uniform and Gaussian kernel smoothing for $y = f(x) + e$, $f(x) = x(x-4)(x+4)$, $h = 0.75$

Problem 3

Cross validation

(A) See attached R code for a script to return prediction error estimates for smoothing given a specified choice of bandwidth, h.

(B) For this exercise, I produced 500 data points on the x-space $[0, 1]$ from a sine function $f(x)$ with a given period and set the amplitude, and added Gaussian noise with a given standard deviation. Then I used 5-fold cross validation to select the optimal bandwidth for that given period and standard deviation of noise term. Figure 2 shows the optimal bandwidths for period ranging from 0.1 to 1, and standard deviation ranging from 0.001 to 0.5, and Figure 3 shows four example The highest bandwidths are chosen for functions with high "wigglyness" and high noise, and the smallest bandwidths are chosen for functions with low "wigglyness" and low noise. This makes sense. As the frequency increases (i.e., period decreases) then we need a tighter bandwidth because the value of the function is fluctuating at a greater rate. As noise increases, we need a greater bandwidth to smooth out the noise. Furthermore, we can see that in all cases we recover the underlying function pretty well.

(C) I'll get around to this eventually....

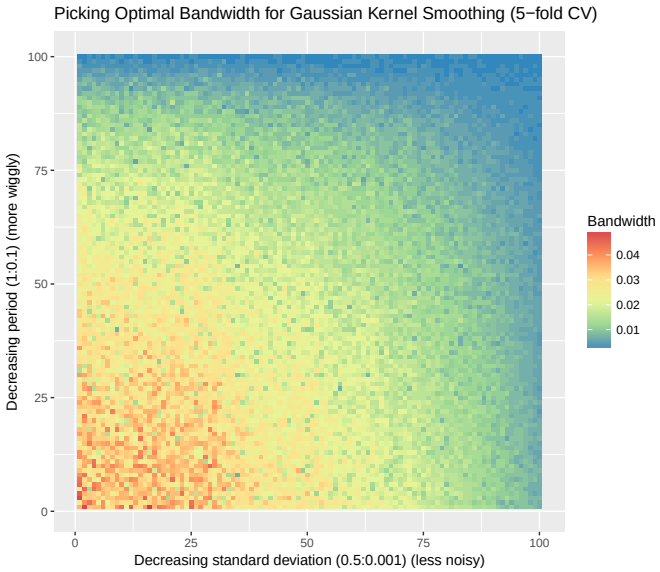

Figure 2: Optimal bandwidths for varying periods and standard deviations

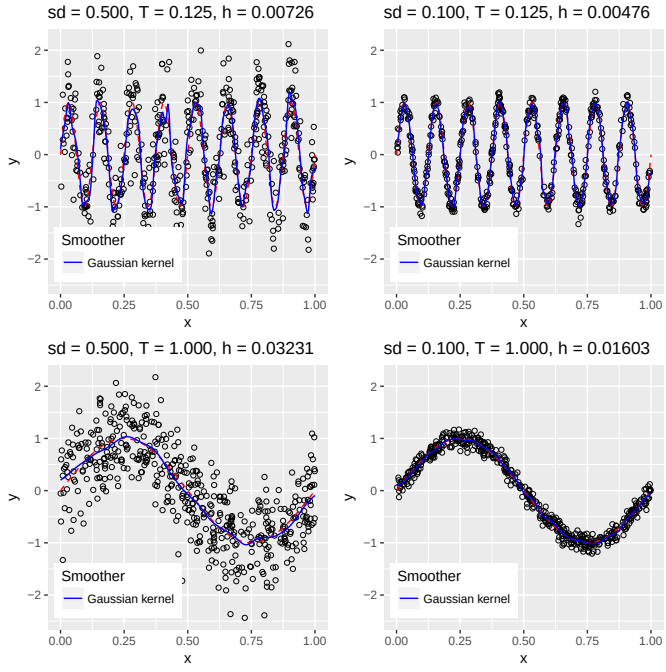

Figure 3: 2×2 example with fitted curves

Problem 4

Local polynomial regression

(A) Define

$$\begin{aligned} g_x(u;a) &= a_0 + \sum_{k=1}^{D} a_k (u-x)^k \\ &= \begin{cases} \sum_{j=0}^{D+1} a_j (u-x)^j & \text{if } u \neq x \\ a_0 & \text{if } u = x \end{cases}. \end{aligned}$$

The coefficients of a for the local polynomial regression with dimension D will come from the weighted least squares problem

$$\hat{a} = \arg\min_{a \in \mathcal{R}^{D+1}} \sum_{i=1}^{n} w_i \left[y_i - g_x(x_i, a) \right]^2 \tag{1}$$

Furthermore, define the matrix R_x whose (i,j) element is $(x_i - x)^2$. Then the estimate $\hat{f}(x)$ will be $R_x \hat{a}$. The solution of $\hat{a}$ may be found with

$$\begin{aligned} \hat{a} &= \arg\min_{a \in \mathcal{R}^{D+1}} (y - R_x a)^T W (y - R_x a) \\ &= (R_x^T W R_x)^{-1} R_x^T W y \end{aligned}$$

where $W = \text{diag}(w_1, \ldots, w_n)$ is a diagonal matrix of weights from some kernel,

$$w_i = \frac{1}{h} K\left(\frac{x_i - x}{h} \right)$$

and this solution is found following the same argument to find the WLS estimate of linear regression from Exercises 1.

(B) Define the matrix $B_x = (R_x^T W R_x)^{-1} R_x^T W$. The estimate of f at a point $x^\star$ is $\hat{f}(x) = \hat{a}_0$, the first element of the vector $\hat{a}$ whose form is derived above. Since our estimate is in the form of a linear smoother, this can also be written as $\hat{f}(x) = \frac{b_x^T y}{\sum_{i=1}^n b_{x,i}}$ if we take b^T to be the first row of B. In the special case of the local linear smoother ($D = 1$), the matrix R_x has dimension $n \times 2$ and can be written as

$$R_x = \begin{bmatrix} 1 & x_1 - x \\ \vdots & \vdots \\ 1 & x_n - x \end{bmatrix}.$$

$$R_x^T = \begin{bmatrix} 1 & \ldots & 1 \\ x_1 - x & \ldots & x_n - x \end{bmatrix}$$

$$W = \begin{bmatrix} w_1 & & \mathcal{O} \\ & \ddots & \\ \mathcal{O} & & w_n \end{bmatrix}$$

$$R_x^T W R_x = \begin{bmatrix} \sum_{i=1}^n w_i & \sum_{i=1}^n w_i (x_i - x) \\ \sum_{i=1}^n w_i (x_i - x) & \sum_{i=1}^n w_i (x_i - x)^2 \end{bmatrix}^{-1}$$

Problem 4 continued on next page...

$$R_x^T W = \begin{bmatrix} w_1 & \dots & w_n \\ w_1(x_1 - x) & \dots & w_n(x_n - x) \end{bmatrix}$$

Furthermore, define the term

$$s_j(x) = \sum_{i=1}^{n} w_i (x_i - x)^j.$$

Now we can find the exact form of B up to a proportionality constant,

$$\begin{aligned}
B &= (R_x^T W R_x)^{-1} R_x^T W \\
&= \left(\begin{bmatrix} 1 & \dots & 1 \\ x_1 - x & \dots & x_n - x \end{bmatrix} \begin{bmatrix} w_1 & & \mathcal{O} \\ & \ddots & \\ \mathcal{O} & & w_n \end{bmatrix} \begin{bmatrix} 1 & x_1 - x \\ \vdots & \vdots \\ 1 & x_n - x \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & \dots & 1 \\ x_1 - x & \dots & x_n - x \end{bmatrix} \begin{bmatrix} w_1 & & \mathcal{O} \\ & \ddots & \\ \mathcal{O} & & w_n \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^n w_i & \sum_{i=1}^n w_i (x_i - x) \\ \sum_{i=1}^n w_i (x_i - x) & \sum_{i=1}^n w_i (x_i - x)^2 \end{bmatrix}^{-1} \begin{bmatrix} w_1 & \dots & w_n \\ w_1(x_1 - x) & \dots & w_n(x_n - x) \end{bmatrix} \\
&= \begin{bmatrix} \sum_{i=1}^n w_i & s_1(x) \\ s_1(x) & s_2(x) \end{bmatrix}^{-1} \begin{bmatrix} w_1 & \dots & w_n \\ w_1(x_1 - x) & \dots & w_n(x_n - x) \end{bmatrix} \\
&\propto \begin{bmatrix} s_2(x) & -s_1(x) \\ -s_1(x) & \sum_{i=1}^n w_i \end{bmatrix} \begin{bmatrix} w_1 & \dots & w_n \\ w_1(x_1 - x) & \dots & w_n(x_n - x) \end{bmatrix} \\
&= \begin{bmatrix} w_1(s_2(x) - (x_1 - x)s_1(x)) & \dots & w_n(s_2(x) - (x_n - x)s_1(x)) \\ w_1((x_1 - x)\sum_{i=1}^n w_i - s_1(x)) & \dots & w_n((x_n - x)\sum_{i=1}^n w_i - s_1(x)) \end{bmatrix}.
\end{aligned}$$

From this we conclude that $\hat{f}(x)$ is a linear smoother with a weight on each y_i proportional to $b_{x,i} = w_i(s_2(x) - (x_i - x)s_1(x))$.

(C)

(D) With H a smoothing matrix (or "hat matrix"), let $r = y - Hy$ be the vector of residuals. If the random vector x with mean vector μ and covariance matrix Σ, then $E(x^T Q x) = \text{tr}(Q\Sigma) + \mu^T Q \mu$. By assumption, $E(\mathbf{y}) = f(\mathbf{x})$ and $\text{cov}(\mathbf{y}) = \sigma^2 I$ Then,

$$\begin{aligned}
E\left(\|r^2\|_2^2\right) &= E\left((\mathbf{y} - H\mathbf{y})^T(\mathbf{y} - H\mathbf{y})\right) \\
&= E\left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T H\mathbf{y} + \mathbf{y}^T H^T H\mathbf{y}\right) \\
&= E\left(\mathbf{y}^T\mathbf{y}\right) - 2E\left(\mathbf{y}^T H\mathbf{y}\right) + E\left(\mathbf{y}^T H^T H\mathbf{y}\right) \\
&= \left(\text{tr}\left[I\sigma^2 I\right] + f(\mathbf{x})^T f(\mathbf{x})\right) - 2\left(\text{tr}\left[H^T\sigma^2 I\right] + f(\mathbf{x})^T H^T f(\mathbf{x})\right) + \left(\text{tr}\left[H^T H\sigma^2 I\right] + f(\mathbf{x})^T H^T H f(\mathbf{x})\right) \\
&= \left(n\sigma^2 + f(\mathbf{x})^T f(\mathbf{x})\right) - 2\left(\sigma^2 \text{tr}[H] + f(\mathbf{x})^T H^T f(\mathbf{x})\right) + \left(\sigma^2 \text{tr}[H^T H] + f(\mathbf{x})^T H^T H f(\mathbf{x})\right) \\
&= \left(n - \text{tr}[H] + \text{tr}[H^T H]\right)\sigma^2 + \left(f(\mathbf{x})^T f(\mathbf{x}) - 2f(\mathbf{x})^T H^T f(\mathbf{x}) + f(\mathbf{x})^T H^T H f(\mathbf{x})\right) \\
&= \left(n - \text{tr}[H] + \text{tr}[H^T H]\right)\sigma^2 + \left(f(\mathbf{x}) - Hf(\mathbf{x})\right)^T \left(f(\mathbf{x}) - Hf(\mathbf{x})\right),
\end{aligned}$$

so the estimator

$$\hat{\sigma}^2 = \frac{\|r^2\|_2^2}{n - \text{tr}[H] + \text{tr}[H^T H]}$$

will be nearly unbiased in σ^2 when $f(\mathbf{x}) \approx Hf(\mathbf{x})$.

(E) See attached R code for implementation of the local polynomial regression, along with leave-one-out cross validation.

Problem 4 continued on next page...

(F) Fitting this model with $D = 1$, the assumption of homoscedasticity (constance variance of residuals) is not met. The residuals fan out towards the lower end of the range for temperature. Taking a logarithmic transform of the response variable, daily gas bill, makes the residuals more uniform, although there are still several outliers in the residuals, now towards the higher ejnd of the range for temperature. See figures below.

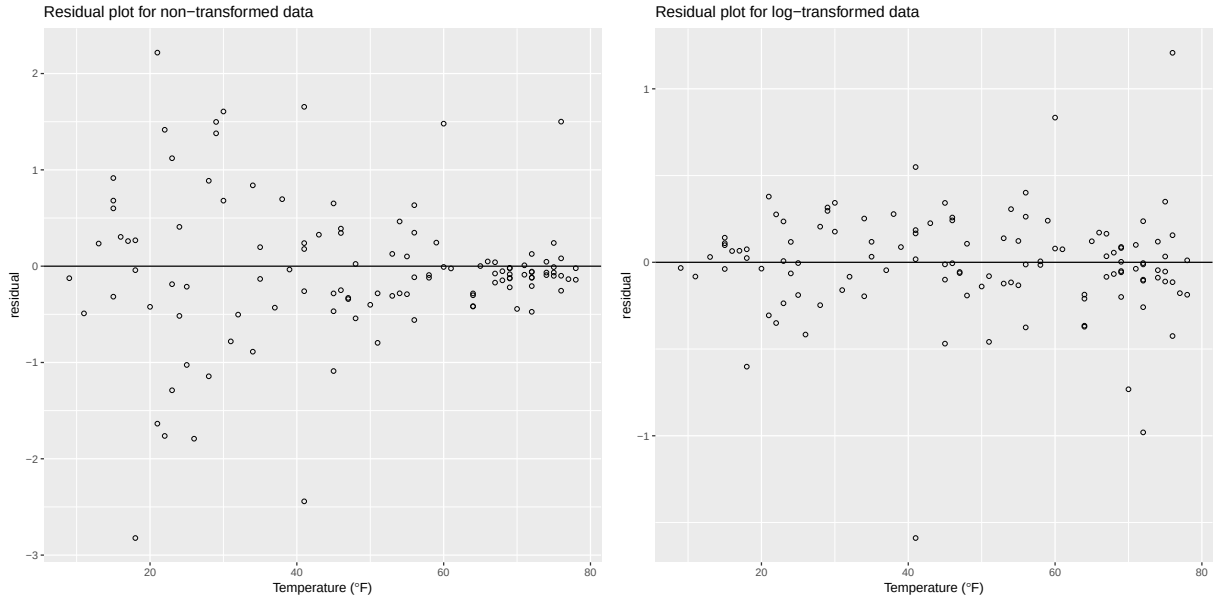

Figure 4: Residual plots for models fitted with non-transformed (left) and log-transformed (right) response variables

(G) The figure below shows the fitted model with 95% confidence bands (found with $\hat{y} \pm 1.96\hat{\sigma}$) and overlaid scatterplot of the data. The optimal bandwidth $h = 5.4168$ was chosen with leave-one-out cross validation. The fit is pretty good, with $R^2 = 0.88$, but there are four observations which fall outside the confidence band.

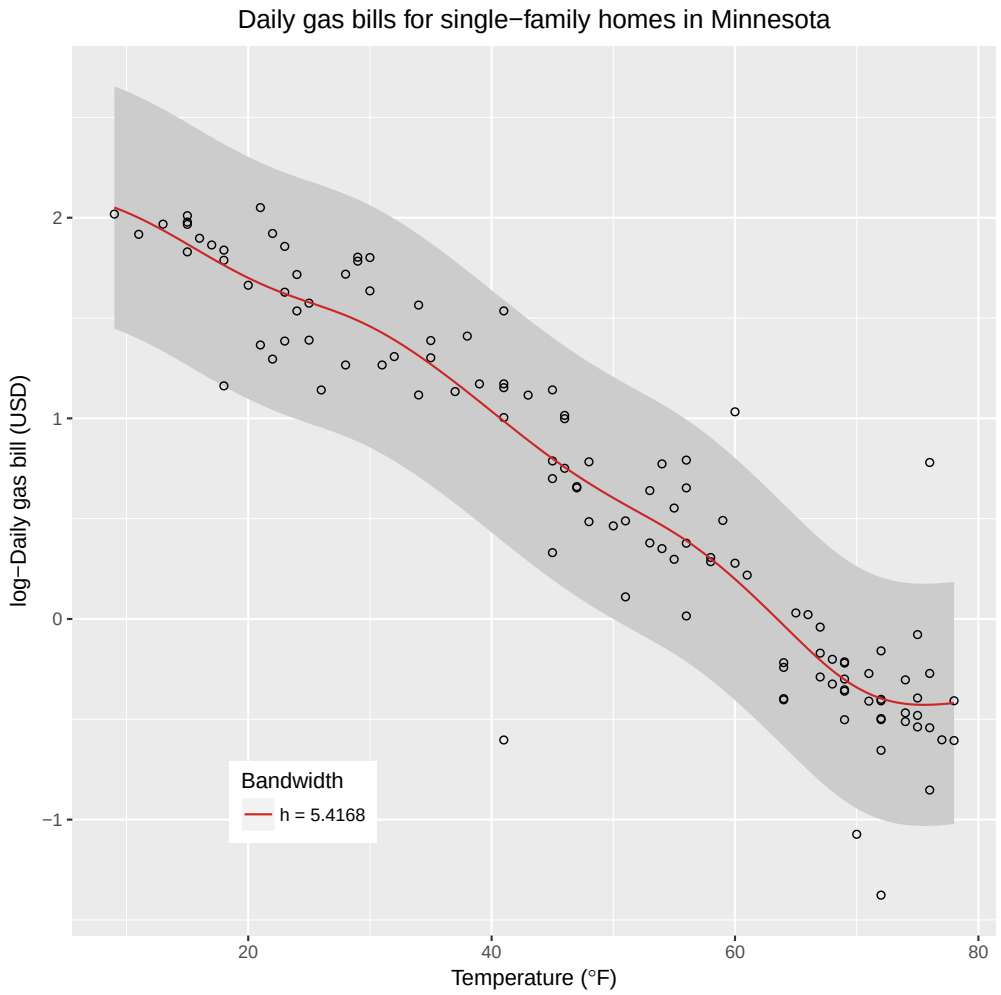

Figure 5: Local linear regression for Minnesota gas bill data

Problem 5

Gaussian processes

(A) Examples...

(B) Suppose that we have a sequence coming from a Gaussian process, $[x_1, x_2, \ldots, x_N, x^\star]^T \sim f \sim \text{GP}(m, C)$ for some mean function m and covariance function C. We want to find the distribution of $x^\star$ conditional on the observations $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$. First, define the three matrices,

$$\begin{aligned} C &= \begin{bmatrix} C(x_1, x_1) & C(x_1, x_2) & \ldots & C(x_1, x_N) \\ C(x_2, x_1) & C(x_2, x_2) & \ldots & C(x_2, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ C(x_N, x_1) & C(x_N, x_2) & \ldots & C(x_N, x_N) \end{bmatrix} \\ C_\star &= \begin{bmatrix} C(x^\star, x_1) & C(x^\star, x_2) & \ldots & C(x^\star, x_N) \end{bmatrix} \\ C_{\star\star} &= \begin{bmatrix} C(x^\star, x^\star) \end{bmatrix}. \end{aligned}$$

The joint distribution may be written as

$$\begin{bmatrix} \mathbf{x} \\ x^\star \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(x^\star) \\ m(\mathbf{x}) \end{bmatrix}, \begin{bmatrix} C & C_\star^T \\ C_\star & C_{\star\star} \end{bmatrix} \right).$$

In Exercises 1, we showed how to derive the conditional distribution from a multivariate normal distribution. Here, we can see that

$$x^\star | \mathbf{x} \sim \mathcal{N}\left(m(x^\star) + C_\star C^{-1}(\mathbf{x} - m(\mathbf{x})), C_{\star\star} - C_\star C^{-1} C_\star^T \right).$$

(C)

Problem 6

In nonparametric regression and spacial smoothing

We have the likelihood $y|\theta \sim \mathcal{N}(R\theta, \Sigma)$ and the prior distribution $\theta \sim \mathcal{N}(m, V)$. Let $\Omega = \Sigma^{-1}$ and $W = V^{-1}$. We can expand out the respective PDFs of these distributions up to a proportionality constant,

$$\begin{aligned} p(y|\theta) &\propto \exp\left[-\frac{1}{2}(y - R\theta)^T\Omega(y - R\theta)\right] \\ &= \exp\left[-\frac{1}{2}(R\theta - y)^T\Omega(R\theta - y)\right] \\ &= \exp\left[-\frac{1}{2}(\theta^T R^T\Omega R\theta - 2y^T\Omega R\theta + y^T\Omega y)\right] \\ p(\theta) &\propto \exp\left[-\frac{1}{2}(\theta - m)^T W(\theta - m)\right] \\ &= \exp\left[-\frac{1}{2}(\theta^T W\theta - 2m^T W\theta + m^T Wm)\right] \end{aligned}$$

Then, by dropping proportionality terms not containing θ and completing the square, the posterior of θ is

$$\begin{aligned} p(\theta|y) &\propto p(y|\theta)p(\theta) \\ &\propto \exp\left[-\frac{1}{2}(\theta^T[W + R^T\Omega R]\theta - 2[m^T W + y^T\Omega R]\theta)\right] \\ &\propto \exp\left[-\frac{1}{2}(\theta - [W + R^T\Omega R]^{-1}[Wm + R^T\Omega y])^T(W + R^T\Omega R)(\theta - [W + R^T\Omega R]^{-1}[Wm + R^T\Omega y])\right]. \end{aligned}$$

Thus we can see that the posterior distribution is $\theta|y \sim \mathcal{N}(m', (W')^{-1})$, where

$$\begin{aligned} m' &= (W')^{-1}(Wm + R^T\Omega y) \\ W' &= W + R^T\Omega R \end{aligned}$$

(A) We observe data $y_i = f(x_i) + \epsilon_i$, where $\epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$ and there is some unknown function f. In this problem, we assume σ is known. Now we put a mean-zero Gaussian process prior distribution on $f \sim \text{GP}(0, C(\bullet))$ for some covariance function $C(\bullet)$. Let $\mathbf{x} = [x_1, x_2, \ldots, x_N]^T$ denote the previous observed x points. Further, define the random vector $\mathbf{f} = [f(x_1), f(x_2), \ldots, f(x_N)]^T$. We want to find the posterior distribution of $\mathbf{f}$ given the observed responses $\mathbf{y} = [y_1, y_2, \ldots, y_N]^T$. The likelihood is

$$(\mathbf{y}|\mathbf{f}) \sim \mathcal{N}(\mathbf{f}, \sigma^2 I)$$

and, if we define the covariance matrix C such that its (i, j) element is $C(x_i, x_j)$ the prior for $\mathbf{f}$ is

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, C).$$

And with this, we can find the posterior for $\mathbf{f}$,

$$p(\mathbf{f}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{f})p(\mathbf{f}).$$

From the general result we have already shown above, we can see that the posterior of $\mathbf{f}$ will take the form

$$p(\mathbf{f}|y) \sim \mathcal{N}(m', (D')^{-1}), \tag{2}$$

where

$$\begin{aligned} m' &= \frac{1}{\sigma^2}(D')^{-1}y \\ D' &= C^{-1} + \frac{1}{\sigma^2}I \end{aligned}$$

Problem 6 continued on next page...

(B) In this section we are trying to predict the value of the function $f(x_j^\star)$ for $j = 1, 2, \ldots, M$ at new points $\mathbf{x}^\star = [x_1^\star, x_2^\star, \ldots, x_M^\star]^T$ given the observed data $\mathbf{y}$. First, since $\mathbf{y} = \mathbf{f} + \boldsymbol{\epsilon}$, it is distributed with

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, C + \sigma^2 I),$$

and furthermore, we can express the joint distribution of y and $\mathbf{f}^\star = [f(x_1^\star), f(x_2^\star), \ldots, f(x_M^\star)]$ as

$$\begin{bmatrix} y \\ \mathbf{f}^\star \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} C + \sigma^2 I & C_\star^T \\ C_\star & C_{\star\star} \end{bmatrix}\right),$$

where C is defined as before,

$$C_\star = \begin{bmatrix} C(x_1^\star, x_1) & C(x_1^\star, x_2) & \ldots & C(x_1^\star, x_N) \\ C(x_2^\star, x_1) & C(x_2^\star, x_2) & \ldots & C(x_2^\star, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ C(x_M^\star, x_1) & C(x_M^\star, x_2) & \ldots & C(x_M^\star, x_N) \end{bmatrix},$$

and,

$$C_{\star\star} = \begin{bmatrix} C(x_1^\star, x_1^\star) & C(x_1^\star, x_2^\star) & \ldots & C(x_1^\star, x_M^\star) \\ C(x_2^\star, x_1^\star) & C(x_2^\star, x_2^\star) & \ldots & C(x_2^\star, x_M^\star) \\ \vdots & \vdots & \ddots & \vdots \\ C(x_M^\star, x_1^\star) & C(x_M^\star, x_2^\star) & \ldots & C(x_M^\star, x_M^\star) \end{bmatrix}.$$

Note that the off-diagonal sub-matrices of the larger covariance matrix in the joint distribution does not include σ^2 because any residual ϵ_i is independent of any $f(x_i)$. From the properties of the conditional multivariate normal distribution, the conditional distribution of $\mathbf{f}^\star$ on $\mathbf{y}$ is

$$\mathbf{f}^\star | \mathbf{y} \sim \mathcal{N}\left(C_\star (C + \sigma^2 I)^{-1} \mathbf{y}, C_{\star\star} - C_\star (C + \sigma^2 I)^{-1} C_\star^T\right).$$

From this, we see that

$$E(\mathbf{f}^\star) = C_\star (C + \sigma^2 I)^{-1} \mathbf{y} \tag{3}$$

$$\text{cov}(\mathbf{f}^\star) = C_{\star\star} - C_\star (C + \sigma^2 I)^{-1} C_\star^T \tag{4}$$

(C) Stuff.

(D) In this problem, we find the general form of the marginal likelihood of $\mathbf{y}$, integrating out the unknown

f. Remember that $\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma^2 I)$ and $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, C)$.

$$
\begin{aligned}
p(\mathbf{y}) &= \int_{\mathbb{R}^N} p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}) d\mathbf{f} \\
&\propto \int_{\mathbb{R}^N} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{f})^T(\mathbf{y}-\mathbf{f})\right] \exp\left[-\frac{1}{2}\mathbf{f}^T C^{-1}\mathbf{f}\right] d\mathbf{f} \\
&= \int_{\mathbb{R}^N} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{f} + \mathbf{f}^T\mathbf{f})\right] \exp\left[-\frac{1}{2}\mathbf{f}^T C^{-1}\mathbf{f}\right] d\mathbf{f} \\
&= \exp\left[-\frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{y}\right] \cdot \int_{\mathbb{R}^N} \exp\left[-\frac{1}{2}\left(\mathbf{f}^T\left[\frac{1}{\sigma^2}I + C^{-1}\right]\mathbf{f} - 2\cdot\frac{\mathbf{y}^T}{\sigma^2}\mathbf{f}\right)\right] d\mathbf{f} \\
&= \exp\left[-\frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{y}\right] \times \dots \\
&\int_{\mathbb{R}^N} \exp\left[-\frac{1}{2}\left(\left[\mathbf{f} - \left(\frac{1}{\sigma^2}I + C^{-1}\right)^{-1}\frac{\mathbf{y}}{\sigma^2}\right]^T\left[\frac{1}{\sigma^2}I + C^{-1}\right]\left[\mathbf{f} - \left(\frac{1}{\sigma^2}I + C^{-1}\right)^{-1}\frac{\mathbf{y}}{\sigma^2}\right] - \frac{\mathbf{y}^T}{\sigma^2}\left[\frac{1}{\sigma^2}I + C^{-1}\right]^{-1}\frac{\mathbf{y}}{\sigma^2}\right)\right] d\mathbf{f} \\
&= \exp\left[-\frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{y} + \frac{1}{2}\cdot\frac{\mathbf{y}^T}{\sigma^2}\left[\frac{1}{\sigma^2}I + C^{-1}\right]^{-1}\frac{\mathbf{y}}{\sigma^2}\right] \times \dots \\
&\int_{\mathbb{R}^N} \underbrace{\exp\left[-\frac{1}{2}\left(\left[\mathbf{f} - \left(\frac{1}{\sigma^2}I + C^{-1}\right)^{-1}\frac{\mathbf{y}}{\sigma^2}\right]^T\left[\frac{1}{\sigma^2}I + C^{-1}\right]\left[\mathbf{f} - \left(\frac{1}{\sigma^2}I + C^{-1}\right)^{-1}\frac{\mathbf{y}}{\sigma^2}\right]\right)\right]}_{\text{multivariate normal kernel}} d\mathbf{f} \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\mathbf{y}^T\mathbf{y} + \frac{1}{2}\cdot\frac{\mathbf{y}^T}{\sigma^2}\left[\frac{1}{\sigma^2}I + C^{-1}\right]^{-1}\frac{\mathbf{y}}{\sigma^2}\right] \\
&= \exp\left[-\frac{1}{2}\mathbf{y}^T\left(\frac{1}{\sigma^2}I - \left[\frac{1}{\sigma^2}I\right]\left[\frac{1}{\sigma^2}I + C^{-1}\right]\left[\frac{1}{\sigma^2}I\right]\right)\mathbf{y}\right] \\
&= \exp\left[\mathbf{y}^T\left(\sigma^2 I + C\right)^{-1}\mathbf{y}\right],
\end{aligned}
$$

using the matrix inverse lemma,

$$\left(A^{-1} + B^{-1}\right)^{-1} = A - A(A+B)^{-1}A.$$

We recognize this as a mean-zero multivariate normal kernel. We can see that the marginal distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 I + C\right).$$