SDS 383D Ex 03:
Linear Smoothing and Gaussian Processes

February 18, 2016

Jennifer Starling

Basic Concepts

Bias-Variance Decomposition

Let $\hat{f}(x)$ be a noisy estimate of some function $f(x)$, evaluated at some point x. Define the mean-squared error of the estimate as

$$MSE(\hat{f}, f) = E\{[f(x) - \hat{f}]^2\}$$

Prove that $MSE\left(f, \hat{f}\right) = B^2 + v$, where

$$B = E\{\hat{f}(x)\} - f(x) \text{ and } v = var\{f(x)\} = E\left[\left(\hat{f} - E\left(\hat{f}\right)\right)^2\right]$$

Begin with the definition of MSE.

$$\begin{aligned}
MSE\left[f, \hat{f}\right] &= E\left[\left(f - \hat{f}\right)^2\right] \\
&= E\left[\left(\hat{f} - f\right)^2\right] \\
&= E\left[\left(\hat{f} - E\left(\hat{f}\right) + E\left(\hat{f}\right) - f\right)^2\right], \text{adding/subtracting } E\left(\hat{f}\right) \\
&= E\left[\left(\underbrace{\hat{f} - E\left(\hat{f}\right)}_{} + \underbrace{E\left(\hat{f}\right) - f}_{}\right)\left(\underbrace{\hat{f} - E\left(\hat{f}\right)}_{} + \underbrace{E\left(\hat{f}\right) - f}_{}\right)\right] \\
&= E\left[\left(\hat{f} - E\left(\hat{f}\right)\right)^2\right] - E\left[\left(E\left(\hat{f}\right) - f\right)^2\right] + 2E\left[\left(\hat{f} - E\left(\hat{f}\right)\right)\left(E\left(\hat{f}\right) - f\right)\right] \\
&= E\left[\left(\hat{f} - E\left(\hat{f}\right)\right)^2\right] - \left(E\left(\hat{f}\right) - f\right)^2 + 2E\left[\left(\hat{f} - E\left(\hat{f}\right)\right)\left(E\left(\hat{f}\right) - f\right)\right] \\
&\qquad \text{since } E\left(E\left(X\right)\right) = E\left(X\right) \\
&= Var\left(\hat{f}\right) - \left(Bias\left(\hat{f}, f\right)\right)^2 + 0
\end{aligned}$$

The last term reduces to zero as shown below.

$$\begin{aligned}
&2E\left[\left(\hat{f} - E\left(\hat{f}\right)\right)\left(E\left(\hat{f}\right) - f\right)\right] \\
&= E\left[\hat{f}E\left(\hat{f}\right) - E\left(\hat{f}\right)E\left(\hat{f}\right) - \hat{f}f + E\left(\hat{f}\right)f\right]
\end{aligned}$$

Then $\hat{f} = E\left(\hat{f}\right)$, giving us

$$\begin{aligned}
&= E\left[E\left(\hat{f}\right)E\left(\hat{f}\right) - E\left(\hat{f}\right)E\left(\hat{f}\right) - E\left(\hat{f}\right)f + E\left(\hat{f}\right)f\right] \\
&= 0
\end{aligned}$$

Part A

Suppose we observe $x_1, \ldots, x_n$ from some distribution F, and want to estimate $f(0)$, the value of the probability density function at 0. Let h be a small positive number, called the bandwidth, and define the quantity

$$\pi_h = P\left(-\frac{h}{2} < X < \frac{h}{2}\right) = \int_{-h/2}^{h/2} f(x)dx$$

Clearly $\pi_h \approx hf(0)$. Let Y be the number of observations in a sample of size n that fall within the interval $(-h/2, h/2)$. What is the distribution of Y? What are its mean and variance in terms of n and π_h? Propose a simple estimator $\hat{f}(0)$ involving Y.

Let Y be the number of x_i in $(-\frac{h}{2}, \frac{h}{2})$. Then

$$Y \sim Binom(n, \pi_h)$$

To estimate $\hat{\pi}_h$,

$$\hat{\pi}_h = \frac{y}{n}, \text{ the Binomial MLE}$$

Therefore $y = n\pi_h$.
Then our simple estimator for $\hat{f}(0)$ is

$$\hat{f}(0) = \frac{\hat{\pi}_h}{h} = \frac{y}{nh}$$

Then, since $Y \sim Binom(n, \pi_h)$, expectation and variance are

$$E(Y) = n\pi_h$$
$$Var(Y) = n\pi_h(1 - \pi_h)$$

Part B

Suppose we expand $f(x)$ in a second-order Taylor series about 0:

$$f(x) \approx f(0) + xf'(0) + \frac{x^2}{2}f''(0)\,.$$

Use this in the above expression for π_h, together with the bias–variance decomposition, to show that

$$MSE\{\hat{f}(0), f(0)\} \approx Ah^4 + \frac{B}{nh}$$

for constants A and B that you should (approximately) specify. What happens to the bias and variance when you make h small? When you make h big?

Plug in Taylor series approximation to definition of π_h.

$$\begin{aligned}\hat{\pi}_h &\approx \int_{-\frac{h}{2}}^{\frac{h}{2}} \left[f(0) + xf'(0) + \frac{x^2}{2}f''(0)\right] dx \\ &= hf(0) + \frac{h^3 f''(0)}{24}\end{aligned}$$

Plug in $\hat{\pi}_h$ to $E(Y)$ and $Var(Y)$ to obtain components of $MSE = Var + Bias^2$.
Mean:

$$E\left[\hat{f}(0)\right] = E\left[\frac{Y}{nh}\right] = \frac{1}{nh}E\left[Y\right] = \frac{1}{nh}n\pi_h \approx \frac{1}{h}\left(\frac{h^2 f''(0)}{24}\right) = f(0) + \frac{h^2 f''(0)}{24}$$

Variance:

$$Var\left[\hat{f}(0)\right] = Var\left[\frac{Y}{nh}\right] = \frac{Var\left[Y\right]}{n^2h^2} = \frac{n\pi_h(1-\pi_h)}{n^2h^2} = \frac{\pi_h(1-\pi_h)}{nh^2} \approx \frac{\pi_h}{nh^2}^{(**)}$$
$$= \frac{hf(0) + \frac{h^3f''(0)}{24}}{nh^2} = \frac{f(0)}{nh} + \frac{hf''(0)}{24n} \approx \frac{f(0)}{nh}^{(***)}$$

$(**)$ Because h small positive number, so π_h small, so $(1-\pi_h) \approx 1$.
$(***)$ Because h small and $h < 1$, first term bigger than second by 24x, so can simplify.
Bias:

$$bias = E\left[\hat{f}(0)\right] - f(0) = f(0) + \frac{h^2f''(0)}{24} - f(0) = \frac{h^2f''(0)}{24}$$

Then MSE is as follows.

$$MSE = var + bias^2 \approx \frac{f(0)}{nh} + \left(\frac{h^2f''(0)}{24}\right)^2 = \frac{f(0)}{nh} + + \frac{h^4\left(f''(0)\right)^4}{576}$$

- Smaller h $\rightarrow$ smaller bias, but larger variance.
- Large h $\rightarrow$ larger bias, but smaller variance.

Note: Could include all of the algebraic terms, but these two dominate in order of h.

Part C

Use this result to derive an expression for the bandwidth that minimizes mean-squared error, as a function of n. You can approximate any constants that appear, but make sure you get the right functional dependence on the sample size.

Could solve previous MSE expression for the optimal h by taking the first derivative, setting equal to zero, and getting an expression in terms of h. This expression would include n, so the optimal bandwidth depends on sample size.

Curve Fitting by Linear Smoothing

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

Part A

Suppose we want to estimate the value of the regression function y^* at some new point x^*, denoted $\hat{f}(x^*)$. Assume for the moment that $f(x)$ is linear, and that y and x have already had their means subtracted, in which case $y_i = \beta x_i + \epsilon_i$. Return to your least-squares estimator for multiple regression. Show that for the one-predictor case, your prediction $y^* = f(x^*) = \hat{\beta} x^*$ may be expressed as a linear smoother of the following form:

$$\hat{f}(x^*) = \sum_{i=1}^{n} w\left(x_i, x^*\right) y_i$$

for any x^*. Inspect the weighting function you derived. Briefly describe your understanding of how the resulting smoother behaves, compared with the smoother that arises from an alternate form of the weight function $w(x_i, w^*)$:

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

This is referred to as K-nearest-neighbor smoothing.

The weighting function for smoothing is derived using the following.

$$\begin{aligned} \hat{f}(x^*) &= \hat{\beta} x^* \\ &= X^*(X'X)^{-1}X'y \\ &= x^* \cdot \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\sum_{i=1}^{n} x_i x^* y_i}{\sum_{i=1}^{n} x_i^2} \end{aligned}$$

since we are working in the single-predictor, single-response, mean-zero case.

Therefore,

$$\hat{f}(x^*) = \sum_{i=1}^{n} w\left(x_i, x^*\right) y_i, \text{ with}$$
$$w\left(x_i, x^*\right) = \frac{\sum_{i=1}^{n} x_i x^*}{\sum_{i=1}^{n} x_i^2}$$

This is a linear smoother in the sense that all x^* points have their corresponding $y^* = \hat{f}(x^*)$ estimates 'smoothed' to the regression line represented by intercept 0, slope $\hat{\beta}$.

This is a different behavior than the *K-nearest-neighbor* smoothing. KNN smoothing is not fitting a line which is constructed by using all of the points. KNN smoothing is calculating each new y^* as the straight average of the closest K points y_i.

Part B

A kernel function $K(x)$ is a smooth function satisyfing

$$\int_{\mathbb{R}} K(x)dx = 1\,, \quad \int_{\mathbb{R}} xK(x)dx = 0\,, \quad \int_{\mathbb{R}} x^2K(x)dx > 0\,.$$

A very simple example is the uniform kernel,

$$K(x) = \frac{1}{2}I(x) \quad \text{where} \quad I(x) = \begin{cases} 1, & |x| \leq 1 \\ 0, & \text{otherwise}\,. \end{cases}$$

Another common example is the Gaussian kernel:

$$K(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,.$$

Kernels are used as weighting functions for taking local averages. Specifically, define the weighting function

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

where h is the bandwidth. Using this weighting function in a linear smoother is called kernel regression. (The weighting function gives the unnormalized weights; you should normalize the weights so that they sum to 1.)
Write your own R function that will fit a kernel smoother for an arbitrary set of x-y pairs, and arbitrary choice of (positive real) bandwidth h. Set up an R script that will simulate noisy data from some nonlinear function, $y = f(x) + \epsilon$; subtract the sample means from the simulated x and y; and use your function to fit the kernel smoother for some choice of h. Plot the estimated functions for a range of bandwidths large enough to yield noticeable changes in the qualitative behavior of the prediction functions.

See **R Appendix, R Functions**. Function is called **linear_smoother**. There are also functions to specify which kernel function the linear smoother should use. These functions are called **K_gaussian** and **K_uniform**.

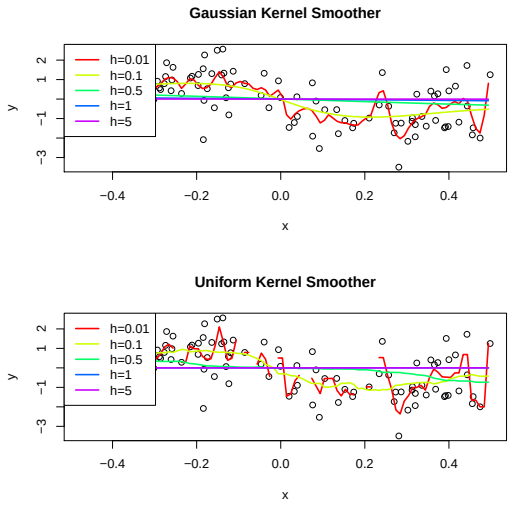

Figure 1: Kernel smoother for varying bandwidths h

Cross Validation

Left unanswered so far in our previous study of kernel regression is the question: how does one choose the bandwidth h used for the kernel? Assume for now that the goal is to predict well, not necessarily to recover the truth. (These are related but distinct goals.)

Part A

Presumably a good choice of h would be one that led to smaller predictive errors on fresh data. Write a function or script that will: (1) accept an old ("training") data set and a new ("testing") data set as inputs; (2) fit the kernel-regression estimator to the training data for specified choices of h; and (3) return the estimated functions and the realized prediction error on the testing data for each value of h. This should involve a fairly straightforward "wrapper" of the function you've already written.

See **R Appendix, R Functions**. Function is called **tune_h**.

Part B

Imagine a conceptual two-by-two table for the unknown, true state of affairs. The rows of the table are "wiggly function" and "smooth function," and the columns are "highly noisy observations" and "not so noisy observations." Simulate one data set (say, 500 points) for each of the four cells of this table, where the x's take values in the unit interval. Then split each data set into training and testing subsets. You choose the functions. Apply your method to each case, using the testing data to select a bandwidth parameter. Choose the estimate that minimizes the average squared error in prediction, which estimates the mean-squared error:

$$L_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n^\star}(y_i^\star - \hat{y}_i^\star)^2\,,$$

where $(y_i^\star, x_i^\star)$ are the points in the test set, and $\hat{y}_i^\star$ is your predicted value arising from the model you fit using only the training data. Does your out-of-sample predictive validation method lead to reasonable choices of h for each case?

My function found optimal bandwidths of h as below, using a 70/30 train-test split. My functions were $sin(2\pi x)$ for smooth, and $sin(2\pi x)$ for wiggly.

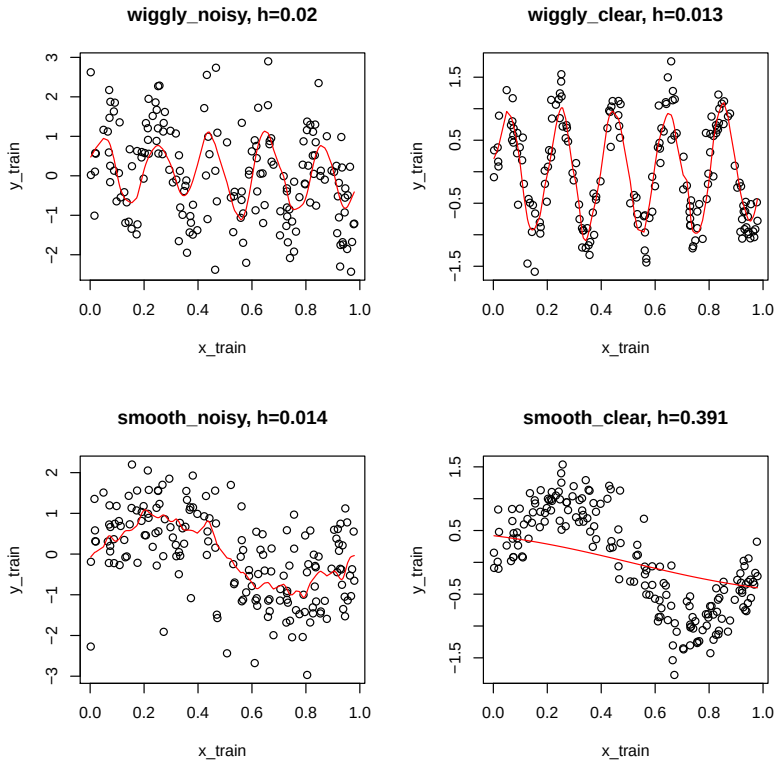

Figure 2: Bandwidth selection for various function types

These bandwidths look generally reasonable in terms of recovering the underlying functions, though the smooth/noisy function's bandwidth was smaller than I anticipated, and looks rather overfitted. My results also varied noticeably in quality as I reran the simulations for various random test/train splits, enough that I would recommend cross-validation each bandwidth selection via a few different test/train splits.

Part C

Splitting a data set into two chunks to choose h by out-of-sample validation has some drawbacks. List at least two. Then consider an alternative: leave-one-out cross validation. Define

$$\text{LOOCV} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i^{(-i)}\right)^2,$$

where $\hat{y}_i^{(-i)}$ is the predicted value of y_i obtained by omitting the ith pair and fitting the model to the reduced data set.

The intuition here is straightforward: for each possible choice of h, you have to predict each data point using all the others. The bandwidth that with the lowest prediction error is the "best" choice by the LOOCV criterion. This is contingent upon a particular bandwidth, and is obviously a function of x_i, but these dependencies are suppressed for notational ease. This looks expensive to compute: for each value of h, and for each data point to be held out, fit a whole nonlinear regression model. But you will derive a shortcut!

Observe that for a linear smoother, we can write the whole vector of fitted values as $\hat{y} = Hy$, where H is called the smoothing matrix (or "hat matrix") and y is the vector of observed outcomes.

Remember that in multiple linear regression this is also true:

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy\,.$$

Write $\hat{y}_i$ in terms of H and y, and show that $\hat{y}_i^{(-i)} = \hat{y}_i - H_{ii}y_i + H_{ii}\hat{y}_i^{(-i)}$. Deduce that, for a linear smoother,

$$\text{LOOCV} = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - H_{ii}} \right)^2 .$$

Test/Train Split Issues:

A few problems regarding splitting the data into test/train chunks as in previous problem:

1. Lose potentially valuable information about outliers or patterns in the data by only using a portion of your data to train the model.
2. As mentioned previously, optimal bandwidth selection and the resulting model fit depended on the test/train split.

Plotting of Optimal Bandwidth h:

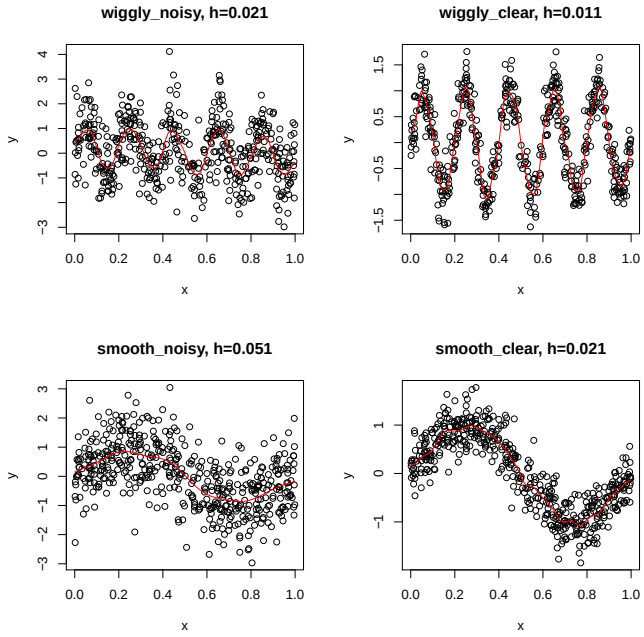

Figure 3: LOOCV Bandwidth selection for various function types

Derivation of H Matrix

Begin with definition of $\hat{y}$ for a single x^* value.

$$\hat{y} = \sum_{i=1}^{n} w(x_i, x^*) y_i \text{ for any value of } x^*$$

Extend to a vector of x^* values, and expression $\hat{y} = Hy$.

$$\begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix}_{(nx1)} = \begin{bmatrix} w(x_1,x_1^*) & \dots & w(x_n,x_1^*) \\ \vdots & & \vdots \\ w(x_1,x_p^*) & \dots & w(x_n,x_p^*) \end{bmatrix}_{(nxp)} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}_{(nx1)}$$

Note that p is the length of the x^* vector, which will be n in the case of LOOCV, since we are using the x values in the existing data set to compute LOOCV prediction error. So H is (nxn).

Must also normalize the weights, so that $H = H/rowsums(H)$. Can formalize this properly as follows.

$$H = \{H_{ij}\} = \frac{w(x_j, x_i^*)}{\sum_{j=1}^{n} w(x_j, x_i^*)}$$

Accompanying Proof

Goal: Show

$$\hat{y}_i^{(-i)} = \hat{y}_i - H_{ii}y_i + H_{ii}\hat{y}_i^{(-i)}$$

Begin with the following two definitions.

$$\hat{y}_i^{(-i)} = \frac{\sum_{j \neq i} h_{ij} y_j}{1 - h_{ii}} \tag{1}$$

This follows from the fact that $H1 = 1$, ie H smooths constants to constants.

$$H_i = (h_{i1}, ..., h_{ii}, ...h_{in})$$

When we remove h_{ii}, doesn't change entires for any other values in row; just renormalizes. This assumption holds for a large class of linear smoothers.
Then rearrange the above equation to obtain

$$\begin{aligned}
\sum_{j \neq i} h_{ij} y_j &= \hat{y}_i^{(-i)} - h_{ii}\hat{y}_i^{(-i)} + h_{ii}y_i \\
\hat{y}_i &= \hat{y}_i^{(-i)} - h_{ii}y_i + h_i i\hat{y}_i^{(-i)} \\
\hat{y}_i^{(-i)} &= \hat{y}_i - H_{ii}y_i + H_{ii}\hat{y}_i^{(-i)}
\end{aligned}$$

Local Polynomial Regression

Kernel regression has a nice interpretation as a "locally constant" estimator, obtained from locally weighted least squares. To see this, suppose we observe pairs (x_i, y_i) for $i = 1, \ldots, n$ from our new favorite model, $y_i = f(x_i) + \epsilon_i$ and wish to estimate the value of the underlying function $f(x)$ at some point x by weighted least squares. Our estimate is the scalar[1] quantity

$$\hat{f}(x) = a = \arg\min_{\mathbb{R}} \sum_{i=1}^{n} w_i (y_i - a)^2 ,$$

where the w_i are the normalized weights (i.e. they have been rescaled to sum to 1 for fixed x). Clearly if $w_i = 1/n$, the estimate is simply $\bar{y}$, the sample mean, which is the "best" globally constant estimator. Using elementary calculus, it is easy to see that if the unnormalized weights are

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

then the solution is exactly the kernel-regression estimator.

Part A

A natural generalization of locally constant regression is local polynomial regression. For points u in a neighborhood of the target point x, define the polynomial

$$g_x(u; a) = a_0 + \sum_{k=1}^{D} a_j (u - x)^k$$

for some vector of coefficients $a = (a_0, \ldots, a_D)$. As above, we will estimate the coefficients a in $g_x(u; a)$ at some target point x using weighted least squares:

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

where $w_i \equiv w(x_i, x)$ are the kernel weights defined just above, normalized to sum to one.[2] Derive a concise (matrix) form of the weight vector $\hat{a}$, and by extension, the local function estimate $\hat{f}(x)$ at the target value x.[3] Life will be easier if you define the matrix R_x whose (i, j) entry is $(x_i - x)^{j-1}$, and remember that (weighted) polynomial regression is the same thing as (weighted) linear regression with a polynomial basis.

Matrix form of $\hat{a}$

$$\hat{a} = \underset{a \in \mathbb{R}^{D+1}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left[y_i - g_x\left(x_i; a\right) \right]^2$$

Sub in expression for $g_x\left(x_i; a\right)$.

$$\hat{a} = \underset{a \in \mathbb{R}^{D+1}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left[y_i - a_0 - \sum_{k=1}^{D} a_j \left(x_i - x\right)^k \right]^2$$

[1] Because we are only talking about the value of the function at a specific point x, not the whole function.

[2] We are fitting a different polynomial function for every possible choice of x. Thus $\hat{a}$ depends on the target point x, but we have suppressed this dependence for notational ease.

[3] Observe that at the target point x, $g_x(u = x; a) = a_0$. That is, only the constant term appears. But this is not the same thing as fitting only a constant term!

Switch to j indices, instead of k, for ease of notation. For clarify, we can expand the summation expression.

$$\begin{aligned} a_0 - \sum_{k=1}^{D} a_j (x_i - x)^k &= a_0 + a_1(x_i - x) + a_2(x_i - x)^2 + ... + a_D(x_i - x)^{D+1} \\ &= a_0(x_i - x)^0 + a_1(x_i - x)^1 + a_2(x_i - x)^2 + ... + a_D(x_i - x)^{D+1} \end{aligned}$$

We will therefore define two matrices:

$W = diag(w_1, ..., w_n)$, a diagonal (nxn) matrix containing weights

$$R = \begin{bmatrix} (x_1-x)^0 & (x_1-x)^1 & (x_1-x)^2 & ... & (x_1-x)^D \\ \vdots & & & & \vdots \\ (x_n-x^0) & (x_n-x)^1 & (x_n-x)^2 & ... & (x_n-x)^D \end{bmatrix} \rightarrow \{R_{ij}\} = (x_i - x)^{j=1}, \text{ for } j = \{1, 2, ..., D+1\}$$

Then we can rewrite $\hat{a}$ as

$$\hat{a} = \underset{a \in \mathbb{R}^{D+1}}{\operatorname{argmin}} (y - Ra)^T W (y - Ra)$$

Minimize by taking derivative wrt a and solving for zero.

$$\begin{aligned} &\frac{d}{da} (y - Ra)^T W (y - Ra) \\ &\qquad = \frac{d}{da} \left[y^T W y - 2y^T W R a + a^T R^T W R a \right] \\ &\qquad - 2y^T W R + 2R^T W R a = 0 \\ &\qquad R^T W R a = R^T W y \\ &\qquad \hat{a} = (R^T W R)^{-1} R^T W y \end{aligned}$$

in a result with the same form as our usual weighted regression.

NOTE: Do not solve by inverting. Solve the linear system, for better speed/stability.

$$Ax = b \equiv A\hat{a} = R^T W y$$

Form of $\hat{f(x)}$

$$\hat{f(x)} = \hat{a}_0, \text{ the first element of } \hat{a}$$

This is because at target point x, only the constant term appears. This is not same as fitting only a constant term. A Taylor Approximation is the underlying intuition here. We can approximate polynomial $f(x)$ around some target point x_0 using a D-degree Taylor polynomial.

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + ...$$

When we approximate $f(x)$ centered at x_0 for the value x_0, only the first term $f(x_0)$ is left, and all other terms include $(x_0 - x_0)$ and so drop out.

Therefore, $\hat{f}(x^*) = \hat{a}_0$.

We can write this more compactly as:

$$\hat{f}(x) = e_1^T \hat{a} = e_1^T (R^T W R)^{-1} R^T W y, \text{ with } e_1 = (1, 0, 0, ...)_{(D+1x1)}$$

We can also write this as:

$$\hat{f}(x_i) = \sum_{j=1}^{n} H_{ij} y_i, \text{ since } \hat{y} = Hy$$

where

$$H = (R^T W R)^{-1} R^T W$$

Part B

From this, conclude that for the special case of the local linear estimator ($D = 1$), we can write $\hat{f}(x)$ as a linear smoother of the form

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} w_i(x) y_i}{\sum_{i=1}^{n} w_i(x)},$$

where the unnormalized weights are

$$\begin{aligned} w_i(x) &= K\left(\frac{x - x_i}{h}\right) \{s_2(x) - (x_i - x) s_1(x)\} \\ s_j(x) &= \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) (x_i - x)^j. \end{aligned}$$

Begin with the previous result for $\hat{f}(x)$.

$$\hat{f}(x) = e_1^T \hat{a} = e_1^T (R^T W R)^{-1} R^T W y \qquad \text{where } R = \begin{bmatrix} 1 & (x_1 - x) \\ \vdots & \vdots \\ 1 & (x_1 n - x) \end{bmatrix} \text{ and } w_i = K\left(\frac{x - x_i}{h}\right)$$

Then $\hat{f}(x)$ can be expanded as follows.

$$\hat{f}(x) = \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{e_1^T} \underbrace{\left[\begin{bmatrix} 1 & \dots & 1 \\ (x_1 - x) & \dots & (x_n - x) \end{bmatrix} \begin{bmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{bmatrix} \begin{bmatrix} 1 & (x_1 - x) \\ \vdots & \vdots \\ 1 & (x_1 n - x) \end{bmatrix} \right]^{-1}}_{(R^T W R)^{-1}} \underbrace{\begin{bmatrix} 1 & \dots & 1 \\ (x_1 - x) & \dots & (x_n - x) \end{bmatrix} \begin{bmatrix} w_1 & & 0 \\ & \ddots & \\ 0 & & w_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{R^T W y}$$

$$= \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{(2x1)} \left[\underbrace{\begin{bmatrix} w_1 & \dots & w_n \\ w_1(x_1 - x) & \dots & w_n(x_n - x) \end{bmatrix}}_{(2xn)} \underbrace{\begin{bmatrix} 1 & (x_1 - x) \\ \vdots & \vdots \\ 1 & (x_n - x) \end{bmatrix}}_{(nx2)} \right]^{-1} \underbrace{\begin{bmatrix} w_1 & \dots & w_n \\ w_1(x_1 - x) & \dots & w_n(x_n - x) \end{bmatrix}}_{(2xn)} \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{(nx1)}$$

$$= \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{(2x1)} \underbrace{\begin{bmatrix} \sum_{i=1}^n w_i & \sum_{i=1}^n w_i(x_i - x) \\ \sum_{i=1}^n w_i(x_i - x) & \sum_{i=1}^n w_i(x_i - x)^2 \end{bmatrix}^{-1}}_{(2x2)} \underbrace{\begin{bmatrix} \sum_{i=1}^n w_i y_i \\ \sum_{i=1}^n w_i(x_i - x) y_i \end{bmatrix}}_{(2x1)}$$

$$= \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{(2x1)} \underbrace{\begin{bmatrix} \sum_{i=1}^n w_i(x_i - x)^2 & -\sum_{i=1}^n w_i(x_i - x) \\ -\sum_{i=1}^n w_i(x_i - x) & \sum_{i=1}^n w_i \end{bmatrix}}_{(2x2)} \left(\frac{1}{\sum_{i=1}^n w_i \sum_{i=1}^n w_i(x_i - x)^2 - \left(\sum_{i=1}^n w_i(x_i - x)\right)^2} \right) \underbrace{\begin{bmatrix} \sum_{i=1}^n w_i y_i \\ \sum_{i=1}^n w_i(x_i - x) y_i \end{bmatrix}}_{(2x1)}$$

$$= \begin{bmatrix} \frac{\sum_{i=1}^n w_i(x_i - x)^2}{\sum_{i=1}^n w_i \sum_{i=1}^n w_i(x_i - x)^2 - \left(\sum_{i=1}^n w_i(x_i - x)\right)^2} & \frac{-\sum_{i=1}^n w_i(x_i - x)}{\sum_{i=1}^n w_i \sum_{i=1}^n w_i(x_i - x)^2 - \left(\sum_{i=1}^n w_i(x_i - x)\right)^2} \end{bmatrix} \begin{bmatrix} \sum_{i=1}^n w_i y_i \\ \sum_{i=1}^n w_i(x_i - x) y_i \end{bmatrix}$$

$$= \frac{\sum_{i=1}^n w_i(x_i - x)^2 \sum_{i=1}^n w_i y_i - \sum_{i=1}^n w_i(x_i - x) \sum_{i=1}^n w_i(x_i - x) y_i}{\sum_{i=1}^n w_i \sum_{i=1}^n w_i(x_i - x)^2 - \sum_{i=1}^n w_i \sum_{i=1}^n w_i(x_i - x)^2}$$

Let $s_1 = \sum_{i=1}^n w_i(x_i - x)$ and $s_2 = \sum_{i=1}^n w_i(x_i - x)^2$

$$= \frac{\sum_{i=1}^n w_i y_i s_2 - \sum_{i=1}^n w_i(x_i - x) y_i s_1}{\sum_{i=1}^n w_i \left(s_2 - (x_i - x)s_1\right)}$$

$$= \frac{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)\left[s_2 - (x_i - x)s_1\right] y_i}{\sum_{i=1}^n K\left(\frac{x-x_i}{h}\right)\left[s_2 - (x_i - x)s_1\right]}$$

We can further simplify by defining an updated 'polynomial weight', call it w_i^* so that we can write $\hat{f}(x)$ in linear smoother form.

$$w_i^* = K\left(\frac{x - x_i}{h}\right)\left[s_2 - (x_i - x)s_1\right]$$

Then rewrite the previous expression in the desired form.

$$\hat{f}(x) = \frac{\sum_{i=1}^n w_i^* y_i}{\sum_{i=1}^n w_i^*}$$

Part C

Suppose that the residuals have constant variance σ^2 (that is, the spread of the residuals does not depend on x). Derive the mean and variance of the sampling distribution for the local polynomial estimate $\hat{f}(x)$ at some arbitrary point x. Note: the random variable $\hat{f}(x)$ is just a scalar quantity at x, not the whole function.

Note that $E(y) = E(f(x) + e) = f(x)$ and $Var(y) = \sigma^2 I$. Cannot assume R is invertible.

Expectation.

$$E\left[\hat{f}(x)\right] = E\left[\sum_{j=1}^n H_{ij} y_i\right]$$
$$= \sum_{j=1}^n H_{ij} f(x_i)$$

because

$$y_i = f(x_i) + e_i$$
$$E(y_i) = f(x_i)$$
$$Var(y_i) = \sigma^2$$

This only ends up being unbiased if the first row of $(R^T W R)^{-1} R^T W$ is equal to $[1, 0, ..., 0]$. But unbiased is ok, remember the bias-variance tradeoff!

Variance:

$$Var\left[\hat{f}(x)\right] = Var\left[E\left[\sum_{j=1}^n H_{ij} y_i\right]\right]$$
$$= \sigma^2 \sum_{j=1}^n H_{ij}^{2\ (**)}$$

$(**)$ = row sums of the projection matrix.

Part D

We don't know the residual variance, but we can estimate it. A basic fact is that if x is a random vector with mean μ and covariance matrix Σ, then for any symmetric matrix Q of appropriate dimension, the quadratic form x^TQx has expectation

$$E(x^TQx) = \text{tr}(Q\Sigma) + \mu^TQ\mu\,.$$

Write the vector of residuals as $r = y - \hat{y} = y - Hy$, where H is the smoothing matrix. Compute the expected value of the estimator

$$\hat{\sigma}^2 = \frac{\|r\|_2^2}{n - 2\text{tr}(H) + \text{tr}(H^TH)}\,,$$

and simplify things as much as possible. Roughly under what circumstances will this estimator be nearly unbiased for large n? Note: the quantity $2tr(H) - tr(H^TH)$ is often referred to as the "effective degrees of freedom" in such problems.

$$\begin{aligned} E\left[\hat{\sigma}^2\right] &= E\left[\frac{||r||_2^2}{n - 2tr(H) + tr(H^TH)}\right] \\ &= E\left[\frac{(y - Hy)^T(y - Hy)}{n - 2tr(H) + tr(H^TH)}\right] \\ &= \frac{E\left[y^Ty\right] - 2E\left[y^THy\right] + E\left[y^TH^THy\right]}{n - 2tr(H) + tr(H^TH)} \end{aligned}$$

Then $E\left[x^TQx\right] = tr(Q\Sigma) + \mu^TQ\mu$, where $E(X) = \mu = f(x)$ and $Var(X) = \Sigma = \sigma^2 I$.

$$\begin{aligned} E\left[\hat{\sigma}^2\right] &= \frac{tr(\Sigma) + \mu^T\mu - 2tr(H\Sigma) - 2\mu^TH\mu + tr(H^TH\Sigma) + \mu H^TH\mu}{n - 2tr(H) + tr(H^TH)} \\ &= \frac{tr(\sigma^2 I) + f(x)^Tf(x) - 2tr(H\sigma^2 I) - 2f(x)^THf(x) + tr(H^TH\sigma^2 I) + f(x)H^THf(x)}{n - 2tr(H) + tr(H^TH)} \\ &= \frac{\sigma^2 n + f(x)^Tf(x) - 2\sigma^2 tr(H) - 2f(x)^THf(x) + \sigma^2 tr(H^TH) + f(x)^TH^THf(x)}{n - 2tr(H) + tr(H^TH)} \\ &= \frac{\sigma^2\left(n - 2tr(H) + tr(H^TH)\right) + f(x)^T\left[I - 2H + H^TH\right]f(x)}{n - 2tr(H) + tr(H^TH)} \\ &= \sigma^2 + \frac{f(x)^T\left[I - 2H + H^TH\right]f(x)}{n - 2tr(H) + tr(H^TH)} \\ &= \sigma^2 + \frac{f(x)^T\left(I - H\right)^T\left(I - H\right)f(x)}{n - 2tr(H) + tr(H^TH)} \\ &= \sigma^2 + \frac{\left(f(x) - Hf(x)\right)^T\left(f(x) - Hf(x)\right)}{n - 2tr(H) + tr(H^TH)} \end{aligned}$$

Intuition about this bias:

Denom is effective degrees of freedom. It reduces to $n - p$ in case of linear model. As n grows, the effective df of the model is not growing nearly as fast as n. The more aggressively you smooth (higher smoothing parameter), end up with fewer df, and the smaller p is.

Numerator: μ is true function value, and $H\mu$ is smoothed true value of the function. In general we would expect this to be small, bc function is smooth generally. (Unless function is crazy.) So if you apply smoothing matrix to a noiseless function, you aren't doing much smoothing.

So overall, we are not concerned about this bias term.

Part E

Write a new R function that fits the local linear estimator using a Gaussian kernel for a specified choice of bandwidth h. Then load the data in "utilities.csv" into R. This data set shows the monthly gas bill (in dollars) for a single-family home in Minnesota, along with the average temperature in that month (in degrees F), and the number of billing days in that month. Let y be the average daily gas bill in a given month (i.e. dollars divided by billing days), and let x be the average temperature. Fit y versus x using local linear regression and some choice of kernel. Choose a bandwidth by leave-one-out cross-validation.

See R code appendix for function details.

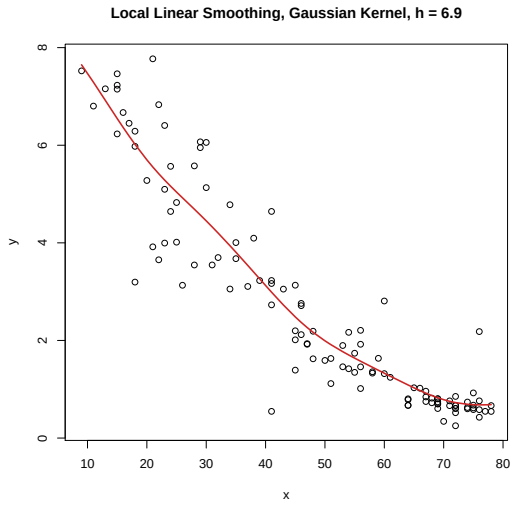

Figure 4: Local Linear Regression with LOOCV Bandwidth

For fun, an illustration of how fit changes with increased degrees.

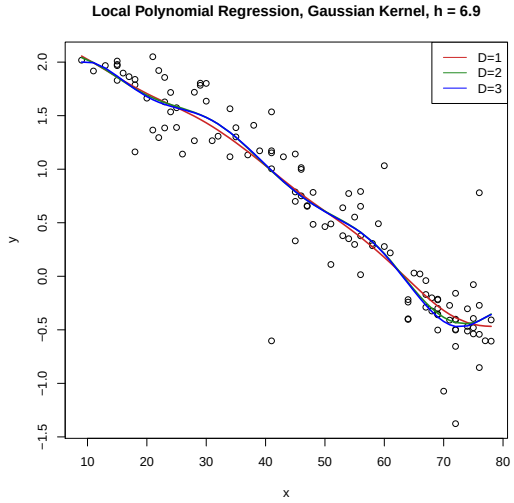

Figure 5: Local Polynomial Regression for Varying Degrees

Part F

Inspect the residuals from the model you just fit. Does the assumption of constant variance (homoscedasticity) look reasonable? If not, do you have any suggestion for fixing it?

The residuals did not look homoscedastic. We log-transform y and obtain homoscedastic residuals, as shown below. (All previous and subsequent analysis is performed using the log-transformed y. There are still one or two outliers, but this residuals plot looks much more homoscedastic.)

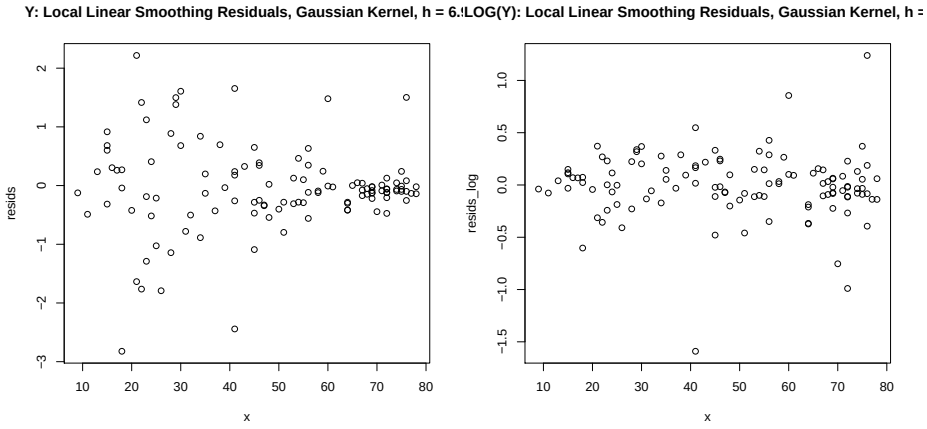

Figure 6: Local Linear Residuals

Part G

Put everything together to construct an approximate point-wise 95% confidence interval for the local linear model (using your chosen bandwidth) for the value of the function at each of the observed points x_i for the utilities data. Plot these confidence bands, along with the estimated function, on top of a scatter plot of the data. (It's fine to use Gaussian critical values for your confidence set.)

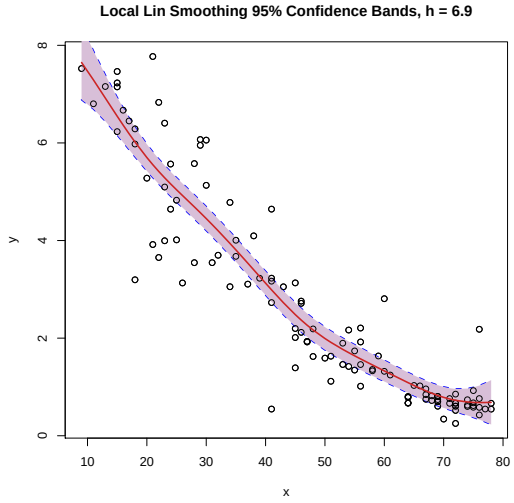

Figure 7: Confidence Bands

Gaussian Processes

A Gaussian Process is a collection of random variables $\{f(x) : x \in \mathcal{X}\}$ such that, for any finite collection of indices $x_1, \ldots, x_N \in \mathcal{X}$, the random vector $[f(x_1), \ldots, f(x_N)]^T$ has a multivariate normal distribution. It is a generalization of the multivariate normal distribution to infinite-dimensional spaces. The set $\mathcal{X}$ is called the index set or the state space of the process, and need not be countable.
A Gaussian process can be thought of as a random function defined over $\mathcal{X}$, often the real line or $\mathbb{R}^p$. We write $f \sim GP(m, C)$ for some mean function $m : \mathcal{X} \to \mathbb{R}$ and a covariance function $C : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. These functions define the moments[4] of all finite-dimensional marginals of the process, in the sense that

$$E\{f(x_1)\} = m(x_1) \quad \text{and} \quad \text{cov}\{f(x_1), f(x_2)\} = C(x_1, x_2)$$

for all $x_1, x_2 \in \mathcal{X}$. More generally, the random vector $[f(x_1), \ldots, f(x_N)]^T$ has covariance matrix whose (i, j) element is $C(x_i, x_j)$. Typical covariance functions are those that decay as a function of increasing distance between points x_1 and x_2. The notion is that $f(x_1)$ and $f(x_2)$ will have high covariance when x_1 and x_2 are close to each other.

Part A

Read up on the Matern Class of covariance functions. The Matern class has the squared exponential covariance function as a special case:

$$C_{SE}(x_1, x_2) = \tau_1^2 \exp\left\{-\frac{1}{2}\left(\frac{d(x_1, x_2)}{b}\right)^2\right\} + \tau_2^2 \delta(x_1, x_2)\,,$$

where $d(x_1, x_2) = ||x_1 - x_2||_2$ is Euclidean distance (or just $|x - y|$ for scalars). The constants (b, τ_1^2, τ_2^2) are often called hyperparameters, and $\delta(a, b)$ is the Kronecker delta function that takes the value 1 if $a = b$, and 0 otherwise. But usually this covariance function generates functions that are "too smooth," and so we use other covariance functions in the Matern class as a default. (See the speed comparison in kernel-benchmark.R on the class GitHub site if you want to see how Rcpp can be used to speed things up here. My code is for the squared-exponential covariance function.)

Let's start with the simple case where $\mathcal{X} = [0, 1]$, the unit interval. Write a function that simulates a mean-zero Gaussian process on $[0, 1]$ under the Matern(5/2) covariance function. The function will accept as arguments: (1) finite set of points $x_1, \ldots, x_N$ on the unit interval; and (2) a triplet (b, τ_1^2, τ_2^2). It will return the value of the random process at each point: $f(x_1), \ldots, f(x_N)$.

Use your function to simulate (and plot) Gaussian processes across a range of values for b, τ_1^2, and τ_2^2. Try starting with a very small value of τ_2^2 (say, 10^{-6}) and playing around with the other two first. On the basis of your experiments, describe the role of these three hyperparameters in controlling the overall behavior of the random functions that result. What happens when you try $\tau_2^2 = 0$? Why? If you can fix this, do—remember our earlier discussion on different ways to simulate the MVN.

Now simulating a few functions with a different covariance function, the Matérn with parameter $5/2$:

$$C_{M52}(x_1, x_2) = \tau_1^2 \left\{1 + \frac{\sqrt{5}d}{b} + \frac{5d^2}{3b^2}\right\} \exp\left(\frac{-\sqrt{5}d}{b}\right) + \tau_2^2 \delta(x_1, x_2)\,,$$

where $d =' ||x_1 - x_2||_2$ is the distance between the two points x_1 and x_2. Comment on the differences between the functions generated from the two covariance kernels. (The Matern covariance is actually a whole family of functions. See Wikipedia article.)

[4] And therefore the entire distribution, because it is normal

See R code appendix for functions.
Below are the results of varying each hyperparameter for both covariance functons.

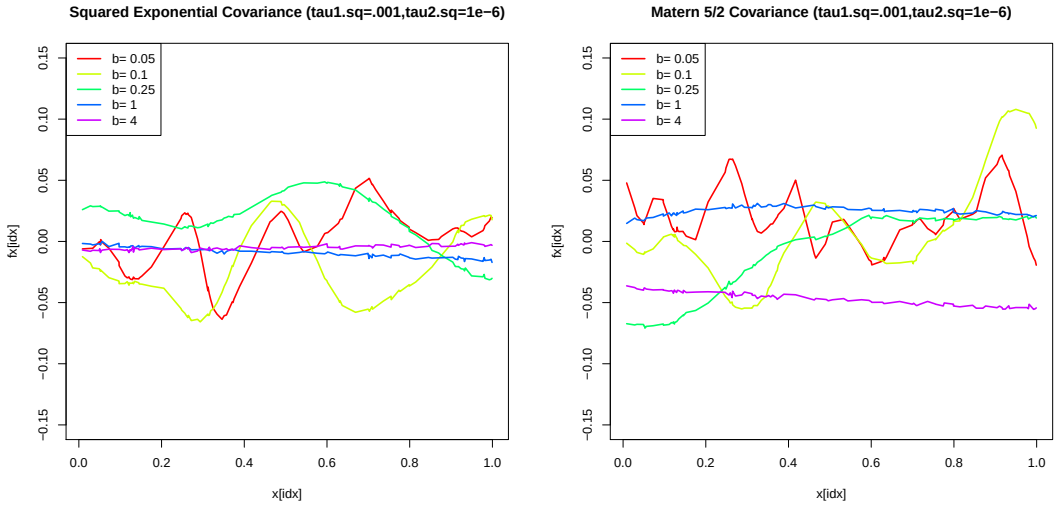

Figure 8: Varying b for Squared Exponential and Matern 5/2

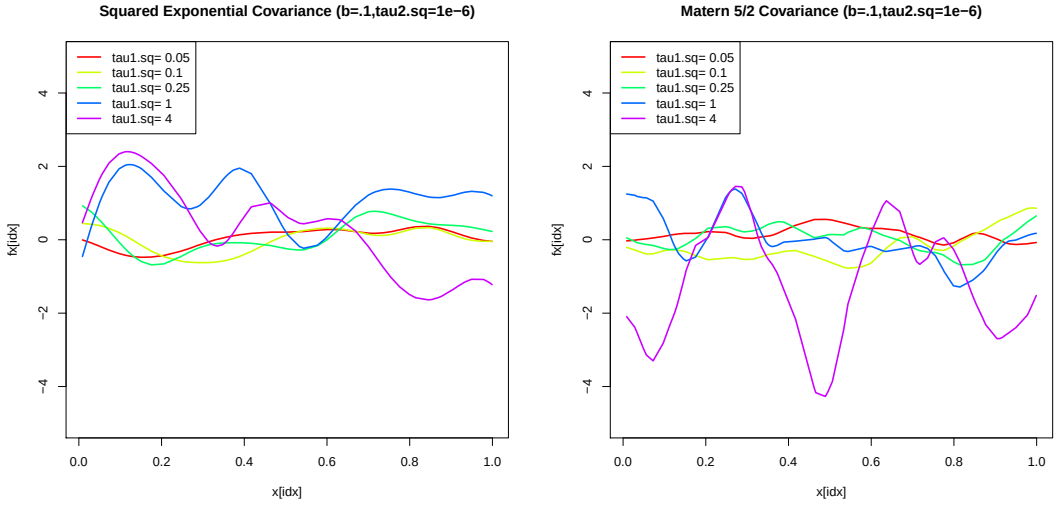

Figure 9: Varying τ_1^2 for Squared Exponential and Matern 5/2

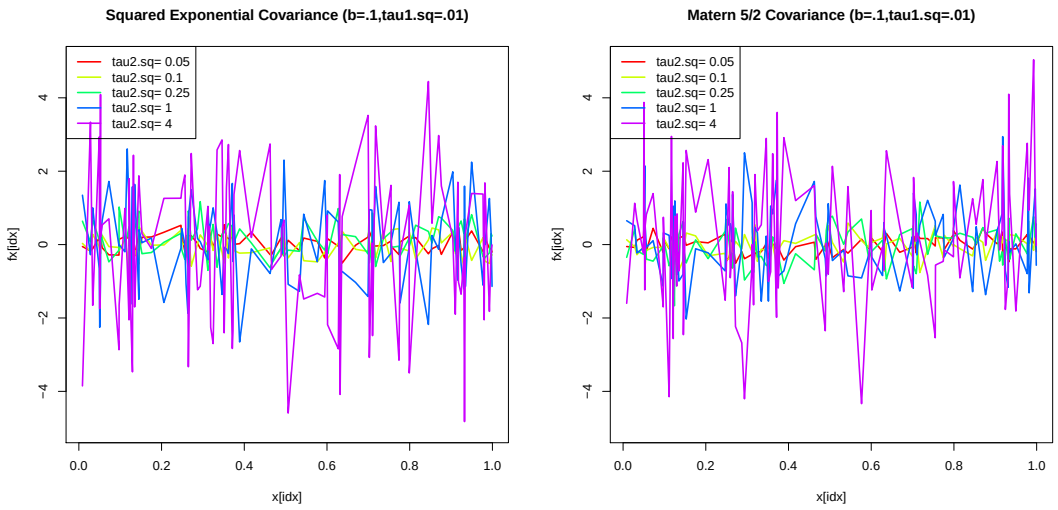

Figure 10: Varying τ_2^2 for Squared Exponential and Matern 5/2

The parameters appear to have the following effects on the function, ie the generation from the gaussian process.

1. b controls the how wiggly the function is, ie how long or short is its period.
2. τ_1^2 is the variance of the function; it determines how far on average the function output is from its mean.
3. τ_2^2 is a noise-level parameter; it is controlling how much noise appears in the function.

The two covariance kernel functions look similar, but squared exponential appears smoother. This could be a benefit due to nicer properties, but could also be a drawback in modeling real-life processes that are not as nicely smoothed.

Part B

Suppose you observe the value of a Gaussian process $f \sim GP(m, C)$ at points $x_1, \ldots, x_N$. What is the conditional distribution of the value of the process at some new point $x^\star$? For the sake of notational ease simply write the value of the (i, j) element of the covariance matrix as $C_{i,j}$, rather than expanding it in terms of a specific covariance function.

Begin with a Gaussian Process $f \sim Gp(m, C)$. We can write our Gaussian Process results in terms of points we have observed, x and $f(x)$, and points we have not yet observed, x^* and $f(x^*)$.

We know that the first n observations are generated from a Gaussian Process, $f \sim Gp(m, C)$, so

$$\left[f(x_1) \ \ldots \ f(x_n) \right]^T \sim N(m, C) = N\left(\begin{bmatrix} m_1 \\ \vdots \\ m_n \end{bmatrix}, \begin{bmatrix} C_{11} & \ldots & C_{1n} \\ \vdots & & \vdots \\ C_{n1} & \ldots & C_{nn} \end{bmatrix} \right) \tag{2}$$

We also know that the obs beginning at $(n+1)$ are generated from the same Gaussian Process, but with the new point added in, $f \sim Gp(m', C')$, so

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}^* \end{bmatrix} \sim N(m' = \begin{bmatrix} m \\ m^* \end{bmatrix}, C' = \begin{bmatrix} C(\mathbf{x},\mathbf{x}) & C(\mathbf{x}^*,\mathbf{x}) \\ C(\mathbf{x}^*,\mathbf{x})^T & C(\mathbf{x}^*,\mathbf{x}^*) \end{bmatrix})$$

where $C(\mathbf{y}, \mathbf{z})$ is the chosen covariance function evaluated for two vectors.

Conditional of Partitioned Multivarate Normal:
From Exercise 1, we recall how to obtain a conditional from a multivariate normal.

Let $X \sim N(\mu, \Sigma)$. Let $x = (x_1, x_2)^T$ be an arbitrary partition of x into two components, of lengths k and $q = p - k$ respectively. Partition $\mu = (\mu_1, \mu_2)^T$ and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$ where $\Sigma_{12} = \Sigma_{21}^T$.
Therefore, the conditional distribution of $f(x_1|x_2)$ is multivariate normal, with parameters

$$\mu_{x_1|x_2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$
$$\Sigma_{x_1|x_2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Using these results, we obtain the distribution for $f^*|f$:

$$f^*|f \sim N\left(m^* + C(x^*, x)C(\mathbf{x}, \mathbf{x})^{-1}(f - m), C(x^*, x^*) - C(x^*, x)^T C(\mathbf{x}, \mathbf{x})^{-1} C(x^*, x)\right)$$

Part C

Prove the following lemma.

Lemma: *Suppose that the joint distribution of two vectors y and θ has the following properties: (1) the conditional distribution for y given θ is multivariate normal, $(y \mid \theta) \sim N(R\theta, \Sigma)$; and (2) the marginal distribution of θ is multivariate normal, $\theta \sim N(m, V)$. Assume that R, Σ, m, and V are all constants. Then the joint distribution of y and θ is multivariate normal.*

We will work in terms of precision, for notational simplicity.

$$y|\theta \sim N(R\theta, \Lambda)$$
$$\theta \sim N(m, W)$$

The given pdfs for $y|\theta$ and θ are as follows.

$$p(y|\theta) \propto \exp\left[-\frac{1}{2}(y - R\theta)^T \Lambda (y - R\theta)\right]$$
$$p(\theta) \propto \exp\left[-\frac{1}{2}(\theta - m)^T W (\theta - m)\right]$$

The joint density is

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

Working with just the inner term:

$$\begin{aligned} &(y - R\theta)^T \Lambda (y - R\theta) + (\theta - m)^T W (\theta - m) \\ &\propto y^T \Lambda y - 2y^T \Lambda R\theta + \theta^T R^T \Lambda R\theta + \theta^T W\theta - 2\theta^T W m \end{aligned}$$

We can rewrite this in a 2-d matrix form, since we want our result to be a 2-d joint multivariate normal.

$$= \begin{bmatrix} y & \theta \end{bmatrix} \begin{bmatrix} \Lambda & 0 \\ 0 & R^T \Lambda R + W \end{bmatrix} \begin{bmatrix} y \\ \theta \end{bmatrix} - 2 \begin{bmatrix} \theta & y \end{bmatrix} \begin{bmatrix} \Lambda R\theta \\ Wm \end{bmatrix}$$

We can complete the square in the same multivate way we have done previously; the above result has form $x^T C x - 2x^T b + a$.

We can rewrite as usual, as $(x - m^*)^T C (x - m^*)$ where $C = C$ and $m^* = C^{-1} b$. In this case,

$$C = \begin{bmatrix} \Lambda & 0 \\ 0 & R^T \Lambda R + W \end{bmatrix}$$
$$m^* = C^{-1} \begin{bmatrix} \Lambda R \theta \\ W m \end{bmatrix}$$

Rewrite the joint pdf, including the exponential term.

$$p(y, \theta) \propto \exp\left[-\frac{1}{2} \left(\begin{bmatrix} y \\ \theta \end{bmatrix} - \begin{bmatrix} \Lambda R \theta \\ W m \end{bmatrix} \right)^T \begin{bmatrix} \Lambda & 0 \\ 0 & R^T \Lambda R + W \end{bmatrix} \left(\begin{bmatrix} y \\ \theta \end{bmatrix} - \begin{bmatrix} \Lambda R \theta \\ W m \end{bmatrix} \right) \right]$$

This is the form of a 2-d multivariate normal pdf for random variable $\begin{bmatrix} y \\ \theta \end{bmatrix}$, with

$$\text{mean} = \left(\begin{bmatrix} \Lambda R \theta \\ W m \end{bmatrix} \right)$$
$$\text{precision} = \begin{bmatrix} \Lambda & 0 \\ 0 & R^T \Lambda R + W \end{bmatrix}$$

Therefore, the joint distribution of (y, θ) is multivariate normal.

GPs in Nonparametric Regression and Spatial Smoothing

Part A

Suppose we observe data $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, for some unknown function f. Suppose that the prior distribution for the unknown function is a mean-zero Gaussian process: $f \sim GP(0, C)$ for some covariance function C. Let $x_1, \ldots, x_N$ denote the previously observed x points. Derive the posterior distribution for the random vector $[f(x_1), \ldots, f(x_N)]^T$, given the corresponding outcomes $y_1, \ldots, y_N$, assuming that you know σ^2.

Prior:

$$f \sim N(0, C) \text{ since prior for } f \text{ is the mean-zero Gaussian Process, } f \sim GP(0, C)$$

Likelihood:

$$y \sim N\left(\begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}, \sigma^2 I \right), \text{ since } y_i = f(x_i) + e_i$$

or more compactly,

$$y \sim N\left(f, \sigma^2 I\right)$$

Posterior:

$$\begin{aligned} p(f|y) &\propto p(y|f) p(f) \\ &= \exp\left[-\frac{1}{2} (y - f)^T \sigma^2 I (y - f) \right] \cdot \exp\left[-\frac{1}{2} (f - 0)^T C^{-1} (f - 0) \right] \end{aligned}$$

As we have seen previously many times, the precisions add, and the posterior mean is the precision-weighted average of the prior mean and data.

$$p(f|y) \sim N\left(\mu_*, D^{-1}\right), \text{with}$$
$$D = \left[\frac{1}{\sigma^2} I + C^{-1}\right]$$
$$\mu_* = \frac{1}{\sigma^2} D^{-1} y$$

Part B

As before, suppose we observe data $y_i = f(x_i) + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, for $i = 1, \ldots, N$. Now we wish to predict the value of the function $f(x^\star)$ at some new point $x^\star$ where we haven't seen previous data. Suppose that f has a mean-zero Gaussian process prior, $f \sim GP(0, C)$. Show that the posterior mean $E\{f(x^\star) \mid y_1, \ldots, y_N\}$ is a linear smoother, and derive expressions both for the smoothing weights and the posterior variance of $f(x^\star)$.

This is similar to Part B of the previous Gaussian Processes section, except now we are observing noisy $y_1 ... y_n$ observations isntead of denoised function values $f(x_1) ... f(x_n)$. We can take a similar approach to derive the posterior $f(x^*)|y$ for some new point x^* we wish to predict.

We know that $y_i = f(x_i) + e_i$, and $f \sim GP(0, C)$, and $e \overset{iid}{\sim} N(0, \sigma^2 I)$.

The sum of multivariate gaussians is multivariate gaussian, so

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

We can use a similar technique as Gaussian Procceses Part B to construct the joint (partitioned) distribution of $(f^*, \mathbf{y})$.

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

The only difference from the noise-free (Gaussian Processes, Part B) case is that the covariance matrix of the y's now has an extram σ^2 term added to diagonal. We can again use the multivariate conditional theory from exercise 1 to obtain the conditional $f^*|y$.

$$f^*|y \sim N\left(E\left[f^* y\right], Var\left[f^* y\right]\right)$$
$$E\left[f^*|y\right] = C(x^*, \mathbf{x})\left(C(\mathbf{x}, \mathbf{x}) + \sigma^2 I\right)^{-1} y$$
$$Var\left[f^*|y\right] = C(x^*, x^*) - C(x^*, x)^T \left(C(\mathbf{x}, \mathbf{x}) + \sigma^2 I\right)^{-1} C(x^*, x)$$

We can write the mean as a linear smoother:

$$E\left[f^*|y, x, x^*, \sigma^2\right] = \sum_{i=1}^{n} w_i y_i \qquad \text{with}$$
$$\mathbf{w} = C(x^*, \mathbf{x})\left(C(\mathbf{x}, \mathbf{x}) + \sigma^2 I\right)^{-1}$$

Part C

Go back to the utilities data, and plot the pointwise posterior mean and 95% posterior confidence interval for the value of the function at each of the observed points x_i (again, superimposed on top of the scatter plot of the data itself).

Choose τ_2^2 to be very small, say 10^{-6}, and choose (b, τ_1^2) that give a sensible-looking answer.

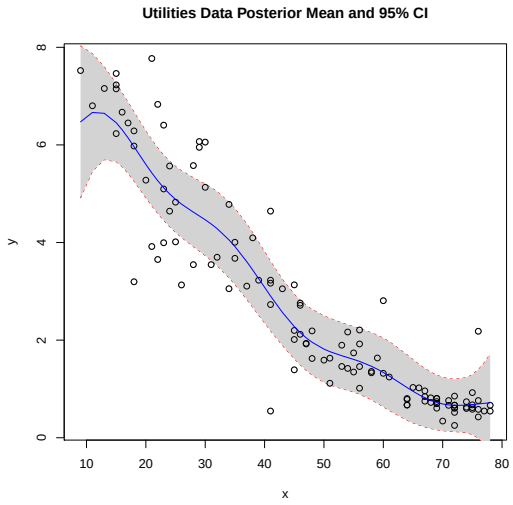

Figure 11: Fitting utilities data with Gaussian Process and non-optimized parameters

Part D

Let $y_i = f(x_i) + \epsilon_i$, and suppose that f has a Gaussian-process prior under the Matern(5/2) covariance function C with scale τ_2^1, range b, and nugget τ_2^2. Derive an expression for the marginal distribution of $y = (y_1 \ldots, y_N)$ in terms of (τ_1^2, b, τ_2^2), integrating out the random function f. This is called a marginal likelihood.

Recall that

$$\begin{aligned} y &\sim N(f, \sigma^2 I) \\ f &\sim N(0, C) \end{aligned}$$

Then

$$\begin{aligned} p(y) &= \int p(y|f)p(f)df \\ &\propto \int \exp\left[-\frac{1}{2}(y-f)^T\left(\frac{1}{\sigma^2}I\right)(y-f)\right]\exp\left[-\frac{1}{2}f^T C^{-1} f\right]df \\ &= \int \exp\left[-\frac{1}{2}\left\{y^T\left(\frac{1}{\sigma^2}I\right)y - 2f^T\underbrace{\left(\frac{1}{\sigma^2}I\right)y}_{\text{"b"}} + f^T\underbrace{\left(\left(\frac{1}{\sigma^2}I\right)+C^{-1}\right)}_{\text{"K"}}f\right\}\right]df \end{aligned}$$

We can complete the square as usual, using the form:

$$\begin{aligned} (f-m^*)^T K(f-m^*) - b^T K^{-1} b, \text{ where} \\ K &= \left(\left(\frac{1}{\sigma^2}I\right)+C^{-1}\right) \\ m^* &= K^{-1}b \\ b^T K^{-1} b &= \left(\left(\frac{1}{\sigma^2}I\right)+C^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}I\right)y \end{aligned}$$

We can then rewrite the density, separating out the non-f terms.

$$= \int \underbrace{\exp\left[-\frac{1}{2}(f-m^*)^T K(f-m^*)\right]}_{\text{multivariate normal kernel}} \exp\left[-\frac{1}{2}\left\{y^T\left(\frac{1}{\sigma^2}I\right)y - y^T\left[\left(\frac{1}{\sigma^2}I\right)\left(\frac{1}{\sigma^2}I + C^{-1}\right)\left(\frac{1}{\sigma^2}I\right)\right]y\right\}\right] df$$

The multivariate normal kernel integrates to $1/c$ where c is the constant of proportionality, which does not depend on y. We are left with

$$\begin{aligned} p(y) &\propto \exp\left[-\frac{1}{2}\left\{y^T\left(\frac{1}{\sigma^2}I\right)y - y^T\left[\left(\frac{1}{\sigma^2}I\right)\left(\frac{1}{\sigma^2}I + C^{-1}\right)\left(\frac{1}{\sigma^2}I\right)\right]y\right\}\right] \\ &= y^T\left[\frac{1}{\sigma^2}I - \left(\frac{1}{\sigma^2}I\right)\left(\frac{1}{\sigma^2}I + C^{-1}\right)\left(\frac{1}{\sigma^2}I\right)\right]y \end{aligned}$$

We can simplify this using the matrix inverse lemma:

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

Let $A = \left(\frac{1}{\sigma^2}I\right)$ and $b = C^{-1}$. Then our expression simplifies to

$$y^T\left(\sigma^2 I + C\right)^{-1}y$$

We recognize this as a mean-zero multivariate normal distribution. Therefore the marginal of y is

$$y \sim N\left(0, \sigma^2 I + C\right)$$

where C is the covariance matrix generated based on the selected function, which takes (τ_1^2, b, τ_2^2) as parameters.