SDS 383D: Homework 3

Giorgio Paulon

February 26, 2017

Problem 1. Basics 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

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

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

$$B = E\{\hat{f}(x)\} - f(x) \quad \textit{and} \quad V = Var\{\hat{f}(x)\}\,.$$

We can write

$$\begin{aligned}
\text{MSE}(\hat{f}, f) &= \text{E}\left[f(x)^2 + \hat{f}(x)^2 - 2f(x)\hat{f}(x)\right] \\
&= f(x)^2 + \text{E}\left[\hat{f}(x)^2\right] - 2f(x)\text{E}\left[\hat{f}(x)\right] \\
&= f(x)^2 - 2f(x)\text{E}\left[\hat{f}(x)\right] + \text{E}\left[\hat{f}(x)\right]^2 + \text{E}\left[\hat{f}(x)^2\right] - \text{E}\left[\hat{f}(x)\right]^2 \\
&= \left(f(x) - \text{E}\left[\hat{f}(x)\right]\right)^2 + \text{Var}\left[\hat{f}(x)\right] \\
&= B^2 + V.
\end{aligned}$$

A simple example

Some people refer to the above decomposition as the bias–variance tradeoff. Why a tradeoff? Here's a simple example to convey the intuition.

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 $\pi_h \approx hf(0)$.

(A) *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.*

The number of observation that fall in the interval $(-h/2, h/2)$ follows a Binomial distribution, i.e.

$$Y \sim \text{Bin}\,(n, \pi_h)\,.$$

Therefore, we get $\mathrm{E}[Y] = n\pi_h$ and $\mathrm{Var}(Y) = n\pi_h(1-\pi_h)$. A simple estimator for $f(0)$, say $\hat{f}(0)$, is

$$\hat{f}(0) = \frac{Y}{nh},$$

so that

$$\mathrm{E}[\hat{f}(0)] = \frac{1}{nh}\mathrm{E}[Y] = \frac{\pi_h}{h} \approx f(0)$$
$$\mathrm{Var}(\hat{f}(0)) = \frac{1}{n^2h^2}\mathrm{Var}(Y) = \frac{\pi_h(1-\pi_h)}{nh^2}$$

(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?

We can re-express the probabilities π_h using a second order Taylor approximation for $f(x)$, that is

$$\begin{aligned}
\pi_h &= \int_{-h/2}^{h/2} f(x)dx \\
&\approx \int_{-h/2}^{h/2} f(0)dx + \int_{-h/2}^{h/2} xf'(0)dx + \int_{-h/2}^{h/2} \frac{x^2}{2}f''(0)dx \\
&= f(0)\cdot h + \frac{1}{2}f'(0)\left[\left(\frac{h}{2}\right)^2 - \left(-\frac{h}{2}\right)^2\right] + \frac{1}{2}f''(0)\left[\frac{x^3}{3}\right]_{-h/2}^{h/2} \\
&= h\left(f(0) + \frac{f''(0)}{24}\cdot h^2\right)
\end{aligned}$$

Recall the formula for the MSE and write

$$\begin{aligned}
\mathrm{MSE}\left[\hat{f}(0), f(0)\right] &= \left[f(0) - \mathrm{E}\left(\hat{f}(0)\right)\right]^2 + \mathrm{Var}\left(\hat{f}(0)\right) \\
&= \left(f(0) - f(0) - \frac{f''(0)}{24}\cdot h^2\right)^2 + \frac{1}{nh}\left(f(0) + \frac{f''(0)}{24}\cdot h^2\right)\left(1 - f(0) - \frac{f''(0)}{24}\cdot h^2\right) \\
&= \left(\frac{f''(0)}{24}\right)^2 h^4 + \frac{1}{nh}\left(f(0) + \frac{f''(0)}{24}\cdot h^2 - f(0)^2 h - \frac{f(0)f''(0)}{12}\cdot h^3 - \frac{f''(0)^2}{24^2}\cdot h^5\right) \\
&\approx \underbrace{\left(\frac{f''(0)}{24}\right)^2\left(1 - \frac{1}{n}\right)}_{\mathrm{A}}\cdot h^4 + \frac{1}{nh}\cdot\underbrace{f(0)}_{\mathrm{B}}.
\end{aligned}$$

When h is small, the bias term gets small and the variance term gets large. The opposite happens when h is large.

(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.*

One can find the minimum of the MSE with a standard calculus procedure.

$$\begin{aligned}
&\underset{h}{\operatorname{argmin}}\, \text{MSE}(n) = \underset{h}{\operatorname{argmin}} \left(Ah^4 + \frac{B}{nh} \right) \\
&\Rightarrow \frac{d}{dh}\left(Ah^4 + \frac{B}{nh} \right) = 4Ah^3 - \frac{B}{nh^2} = 0 \\
&\Rightarrow h = \left(\frac{B}{4An} \right)^{1/5}
\end{aligned}$$

Problem 2. 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.

(A) *Suppose we want to estimate the value of the regression function $y^\star$ at some new point $x^\star$, denoted $\hat{f}(x^\star)$. 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 $\hat{y}^\star = f(x^\star) = \hat{\beta} x^\star$ may be expressed as a linear smoother of the following form:

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

for any $x^\star$. 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 $w(x_i, x^\star)$:

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

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

Let us suppose that $y_i = f(x_i) + \epsilon_i$, where $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. If the function f is linear, then $\text{E}[y_i] = a\text{E}[x] + b = 0$, which implies $b = 0$. Thus, $f(x) = \beta x$ and $y_i = \beta x + \epsilon_i$.

In the case of Least squares, the estimator is

$$\hat{y}^* = \hat{\beta} x^*,$$

where

$$\hat{\beta} = (\boldsymbol{x}^T \boldsymbol{x})^{-1} \boldsymbol{x}^T \boldsymbol{y} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}.$$

Therefore, we get

$$\hat{y}^* = \frac{\sum_{i=1}^n x_i x^*}{\sum_{i=1}^n x_i^2} y_i = \sum_{i=1}^n w(x_i, x^*) y_i$$

where the weights are

$$w(x_i, x^*) = \frac{x_i x^*}{\sum_{i=1}^n x_i^2}.$$

This estimator uses all of the points x_i's. In the k-nearest neighbours case, it is easy to see how the distance between $x^\star$ and x_i affects the weights $w(x_i, x^\star)$. In the one variable OLS case, the corresponding intuition is that it only matters how $x_i - \bar{x}$ is large, not the distance between $x^\star$ and x_i. The weight $w(x_i, x^\star)$ is negative if x_i and $x^\star$ lie to the opposite sides of $\bar{x}$.

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

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

A very simple example is the uniform kernel,

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

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.

In order to do kernel regression, we evaluate the smoothed function on a fine grid of points. For each point, we compute a set of weights that are given to the observations $x_1, \ldots, x_n$ given the distance to that particular point. The value of the bandwidth is crucial, as one can

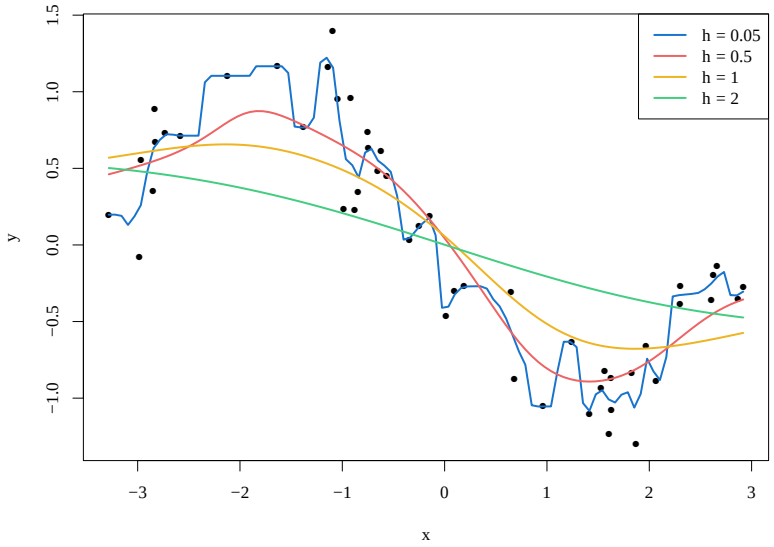


Figure 1: Gaussian kernel regression for the sampled black points. Different values of the bandwidth λ result in different smoothed curves.

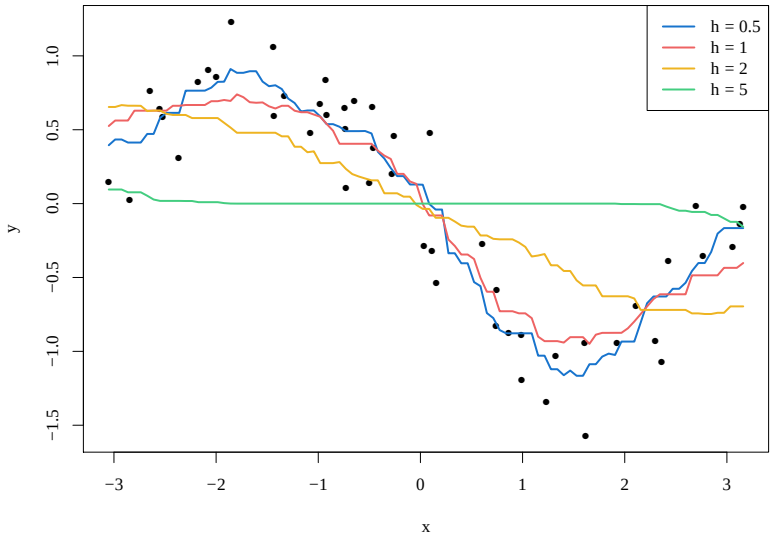


Figure 2: Uniform kernel regression for the sampled black points. Different values of the bandwidth λ result in different smoothed curves.

see in Figure 1 and in Figure 2. In particular, higher values for the bandwidth result in a smoother function, whereas smaller values of the bandwidth will tend to yield an interpolating and wiggly curve.

The linear smoothing problem can be reinterpreted, as we will see in the next section, as a locally constant regression. This is evident in Figure 1 and in Figure 2, where small values of the bandwidth produce flat lines.

Why would we use a uniform kernel instead of a gaussian one? The uniform kernel enforces sparsity in the smoothing matrix H, and so it could be preferable for high dimensional data.

Let us also remark that on the left edge, since the function is increasing, for small bandwidths the function will be biased upwards (**edge bias**). The same happens on the right edge.

Problem 3. 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.)

(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.*

The requested functions are displayed in Listing **??**.

(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:*

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

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?

In Table 1 the values of the bandwidth are displayed in the four different cases. We see that, as expected, the highest bandwidth will arise in the smooth and noisy case, since the corresponding function will be flat. On the other hand, a non noisy wiggly function will produce a small value of the bandwidth and therefore a not-so-smooth estimate.

In Figure 3 the results are displayed visually. As one can see, the cross-validation procedure yields to reasonable choices for the optimal bandwidth. The true mechanism generating the data seems to be reconstructed pretty well!

	Wiggly	Smooth
Noisy	0.01556	0.04780
Not noisy	0.00793	0.01947

Table 1: Comparison of the optimal bandwidths for different choices of the underlying function.

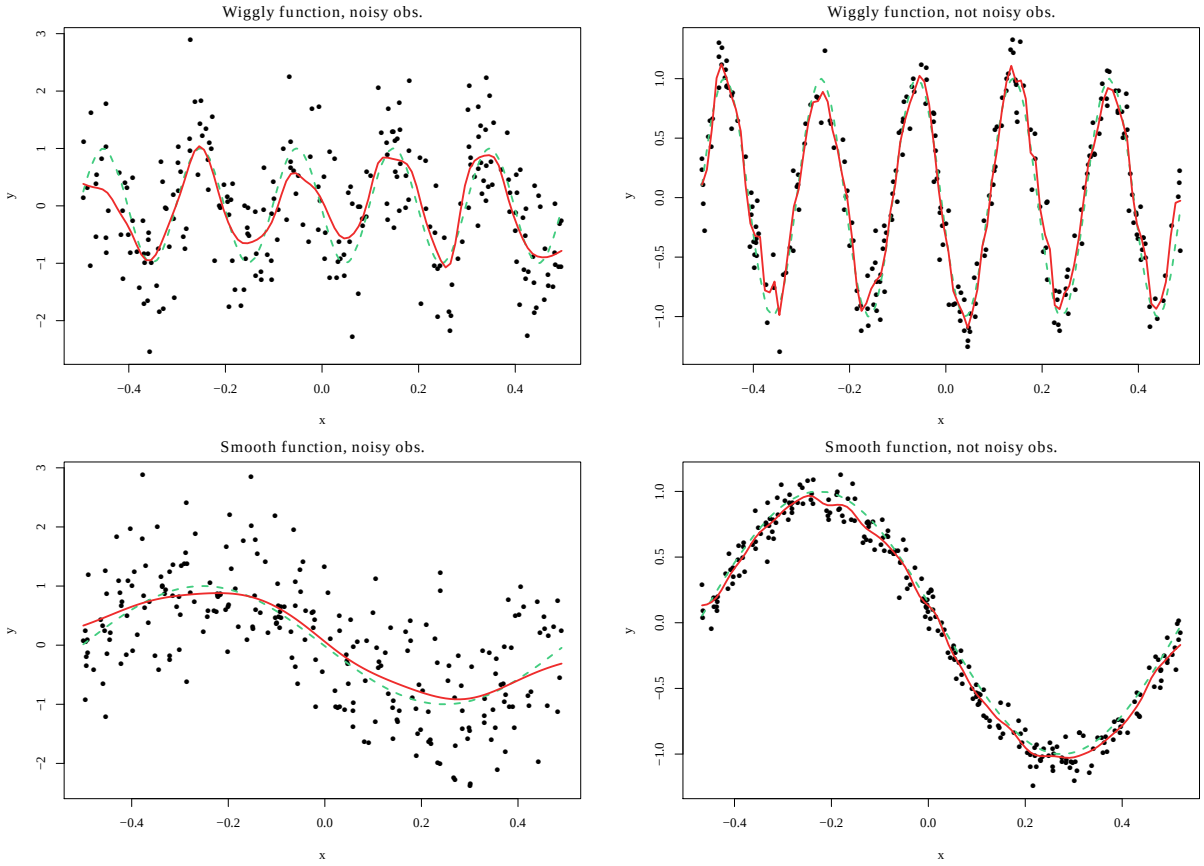


Figure 3: Four different smoothing problems. The smoothed version of the data is represented in red solid line, the "true" density generating the data is overlapped in green dashed line. The wiggly function is $f(x) = \sin(10\pi x)$, the smooth one is $g(x) = \sin(2\pi x)$. The noisy observations have a standard deviation $\sigma_a = 0.75$, the non noisy ones $\sigma_b = 0.25$.

(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*

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

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. 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. 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,

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

Two drawbacks of splitting the dataset in two chunks is surely the fact of not using all of the data to fit the model. In situations where we can't afford to have many observations, we might not want to do that. Moreover, the h chosen via this method leads to low bias for that particular partition of training/test set but high variance for other splits!

Let us now prove the LOOCV lemma:

Problem 4. 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 $i = 1, \dots, n$ 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 quantity

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

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.

We can prove this by simply solving the minimization problem, that is

$$\hat{a} = \operatorname*{argmin}_{\mathbb{R}} \sum_{i=1}^{n} \frac{1}{h} K \left(\frac{x_i - x}{h} \right) (y_i - a)^2,$$

which is equivalent to, taking the derivative with respect to a and setting it equal to 0, to

$$\sum_{i=1}^{n} \frac{1}{h} K \left(\frac{x_i - x}{h} \right) 2(y_i - \hat{a}) = 0$$

$$\Rightarrow \hat{a} = \hat{f}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) y_i}{K\left(\frac{x_i - x}{h}\right)} = \sum_{i=1}^{n} w^\star(x, x_i) y_i,$$

where $w^\star(x, x_i) = \frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}$ are the normalized weights.

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

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

where $w_i \equiv w(x_i, x)$ are the kernel weights defined just above, normalized to sum to one. Derive a concise (matrix) form of the weight vector $\hat{\boldsymbol{a}}$, and by extension, the local function estimate $\hat{f}(x)$ at the target value x. 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.

The problem can be rephrased as a weighted least squares problem in which we are trying to estimate the $(D + 1)$-dimensional vector $\hat{\boldsymbol{a}}_x$, so that $g_x(x_1, \ldots, x_n; \hat{\boldsymbol{a}}) = \hat{f}(x)$.

Let us define the matrix R_x, so that $[R_x]_{ij} = (x_j - x)^{j-1}$, i.e. the $n \times (D + 1)$ matrix

$$R_x = \begin{pmatrix} 1 & x_1 - x & \ldots & (x_1 - x)^D \\ 1 & x_2 - x & \ldots & (x_2 - x)^D \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x & \ldots & (x_n - x)^D \end{pmatrix}.$$

Then we can rewrite the minimization problem as

$$\begin{aligned} \hat{\boldsymbol{a}}_x &= \underset{\mathbb{R}^{D+1}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \{ y_i - g_x(x_i; \boldsymbol{a}) \}^2 \\ &= \underset{\mathbb{R}^{D+1}}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \{ y_i - a_0 - \sum_{k=1}^{D} a_j (x_i - x)^k \}^2 \\ &= \underset{\mathbb{R}^{D+1}}{\operatorname{argmin}} (\boldsymbol{Y} - R_x \boldsymbol{a})^T W (\boldsymbol{Y} - R_x \boldsymbol{a}). \end{aligned}$$

This problem is analogous to the weighted least squares we have already analysed several times, thus the solution is

$$\hat{\boldsymbol{a}}_x = (R_x^T W R_x)^{-1} R_x^T W \boldsymbol{Y},$$

and the corresponding estimated function at the target point x is

$$\hat{f}(x) = g_x(x; \hat{\boldsymbol{a}}_x) = \hat{a}_{x0} = \boldsymbol{e}_1^T \hat{\boldsymbol{a}}_x,$$

where $\boldsymbol{e}_1 = (1, 0, \ldots, 0)^T$. Therefore, for each target point x we have to solve a WLS problem, and the corresponding smoothed fitted value is $\hat{y} = Hy$ where H is the first row of the matrix $(R_x^T W R_x)^{-1} R_x^T W$.

(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

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

We can explicitly write the solution of the WLS problem in the case $D = 1$. In fact, let us denote for brevity $k_i = K(\frac{x-x_i}{h})$. Then we have

$$\begin{aligned} \hat{\boldsymbol{a}}_x &= \left[\begin{pmatrix} 1 & \dots & 1 \\ x_1 - x & \dots & x_n - x \end{pmatrix} \begin{pmatrix} k_1 & & \\ & \ddots & \\ & & k_n \end{pmatrix} \begin{pmatrix} 1 & x_1 - x \\ \vdots & \vdots \\ 1 & x_n - x \end{pmatrix}\right]^{-1} \\ &\quad \cdot \begin{pmatrix} 1 & \dots & 1 \\ x_1 - x & \dots & x_n - x \end{pmatrix} \begin{pmatrix} k_1 & & \\ & \ddots & \\ & & k_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^n k_i & \sum_{i=1}^n k_i(x_i - x) \\ \sum_{i=1}^n k_i(x_i - x) & \sum_{i=1}^n k_i(x_i - x)^2 \end{pmatrix}^{-1} \begin{pmatrix} k_1 & \dots & k_n \\ k_1(x_1 - x) & \dots & k_n(x_n - x) \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\ &\propto \begin{pmatrix} \sum_{i=1}^n k_i(x_i - x)^2 & -\sum_{i=1}^n k_i(x_i - x) \\ -\sum_{i=1}^n k_i(x_i - x) & \sum_{i=1}^n k_i \end{pmatrix} \begin{pmatrix} k_1 & \dots & k_n \\ k_1(x_1 - x) & \dots & k_n(x_n - x) \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \end{aligned}$$

and therefore, considering only the first element of $\hat{\boldsymbol{a}}_x$,

$$\begin{aligned} \hat{a}_{x0} &\propto \begin{pmatrix} k_1[s_2(x) - (x_1 - x)s_1(x)] & \dots & k_n[s_2(x) - (x_n - x)s_1(x)] \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\ &= \sum_{i=1}^n w_i(x) y_i, \end{aligned}$$

where $w_i(x) = k_i[s_2(x) - (x_i - x)s_1(x)]$, which is the desired result apart from the proportionality constant (which is simply give by the sum of the weights).

(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.*

(D) *We don't know the residual variance, but we can estimate it. A basic fact is that if $\boldsymbol{X}$ is a random vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ, then for any symmetric matrix Q of appropriate dimension, the quadratic form $\boldsymbol{X}^T Q \boldsymbol{X}$ has expectation*

$$E(\boldsymbol{X}^T Q \boldsymbol{X}) = tr(Q\Sigma) + \boldsymbol{\mu}^T Q \boldsymbol{\mu}\,.$$

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

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

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

Let us start by expanding out the term at the numerator. Since $\mathrm{E}[\boldsymbol{y}] = f(\boldsymbol{x})$ and $\mathrm{Cov}[\boldsymbol{y}] = \sigma^2 \mathcal{I}_n$, we can write

$$\begin{aligned}
\mathrm{E}[||\boldsymbol{r}||_2^2] &= \mathrm{E}[(\boldsymbol{y} - H\boldsymbol{y})^T (\boldsymbol{y} - H\boldsymbol{y})] \\
&= \mathrm{E}[\boldsymbol{y}^Y \boldsymbol{y} + \boldsymbol{y}^T H^T H \boldsymbol{y} - 2\boldsymbol{y}^T H^T \boldsymbol{y}] \\
&= tr(\mathcal{I}_n \sigma^2 \mathcal{I}_n) + f(\boldsymbol{x})^T \mathcal{I}_n f(\boldsymbol{x}) + tr(H^T H \sigma^2 \mathcal{I}_n) + f(\boldsymbol{x})^T H^T H f(\boldsymbol{x}) \\
&\quad - 2\left[tr(H^T \sigma^2 \mathcal{I}_n) + f(\boldsymbol{x})^T H^T f(\boldsymbol{x})\right] \\
&= \sigma^2 \left[n + tr(H^T H) - 2tr(H)\right] + \left[f(\boldsymbol{x}) - Hf(\boldsymbol{x})\right]^T \left[f(\boldsymbol{x}) - Hf(\boldsymbol{x})\right].
\end{aligned}$$

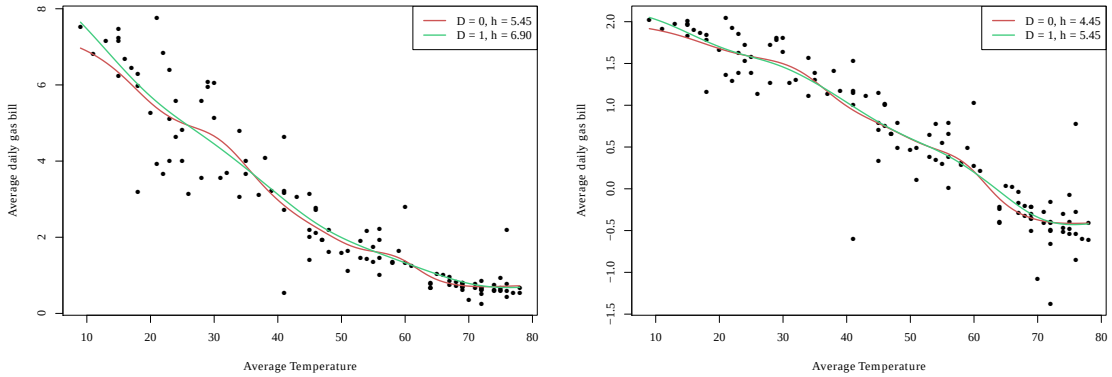
Therefore

$$\mathrm{E}[\hat{\sigma}^2] = \sigma^2 + \frac{\left[f(\boldsymbol{x}) - Hf(\boldsymbol{x})\right]^T \left[f(\boldsymbol{x}) - Hf(\boldsymbol{x})\right]}{n + tr(H^T H) - 2tr(H)},$$

which is nearly unbiased for large n if $f(\boldsymbol{x}) \approx Hf(\boldsymbol{x})$, that is, if the "true" and unknown function $f(x)$ generating the data is similar to its smoothed version, which implies that it is regular enough.

(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.*

The R function is illustrated in Listing **??**. It has been implemented in the general case, and the user can specify the choice of the polynomial degree D. In the particular cases $D = 0$ and $D = 1$, we obtained the results displayed in Figure 4. As one can see, the local linear smoother seems to outperform the local constant smoother on the edges, where the latter suffers of the edge bias problem previously discussed.

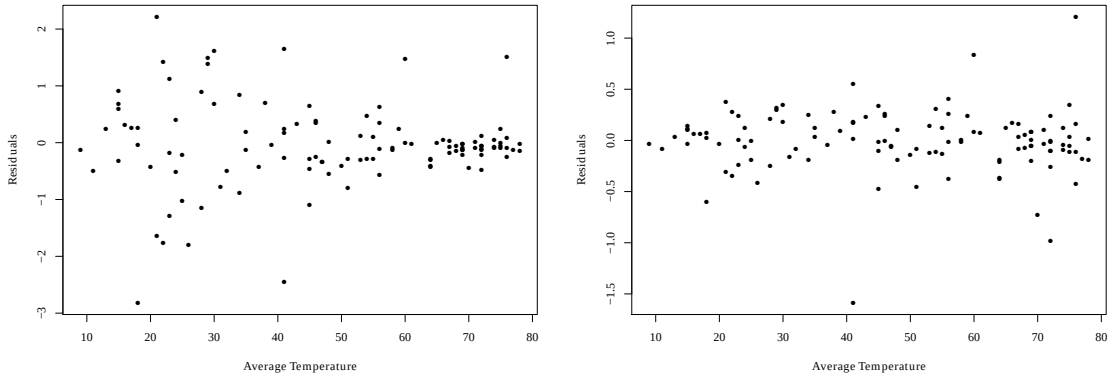


(a) Original data *(b) Log-transformed data*

Figure 4

(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?*

No, as one can see in Figure 5a, the residuals do not seem to have the same variance. A variance stabilizing transformation, such as the logarithm, seems to improve the behaviour of the residuals, see Figure 5b.



(a) Original data *(b) Log-transformed data*

Figure 5: Residuals of the model applied to both the original data and the log-transformed data.

(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.*

Using the estimated variance in part (D), we can build pointwise confidence intervals for the smoothed curve. In fact, we can use

$$\hat{y} \pm \hat{\sigma} \cdot z_{0.975}.$$

In Figure 6 the result of this procedure is illustrated.

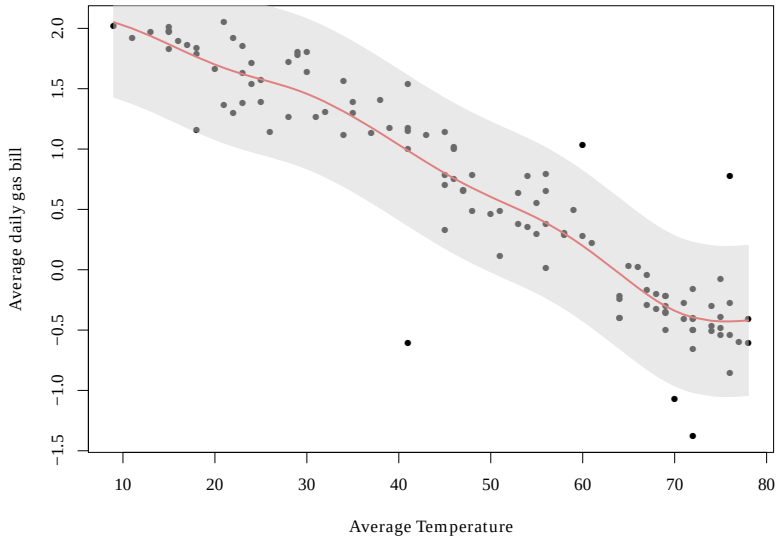


Figure 6: Local linear smoother for the utilities data. The optimal value of the bandwidth has been chosen via LOOCV. The gray shaded area represents the 95% confidence intervals.

Problem 5. 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 $x_1, \dots, x_N \in \mathcal{X}$, the random vector $[f(x_1), \dots, f(x_N)]^T$ 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 of all finite-dimensional marginals of the process, in the sense that

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

for all $x_1, x_2 \in \mathcal{X}$. More generally, the random vector $[f(x_1), \dots, f(x_N)]^T$ 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.

(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.

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 squared exponential 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 role of the parameters for the two covariance function families is the following:

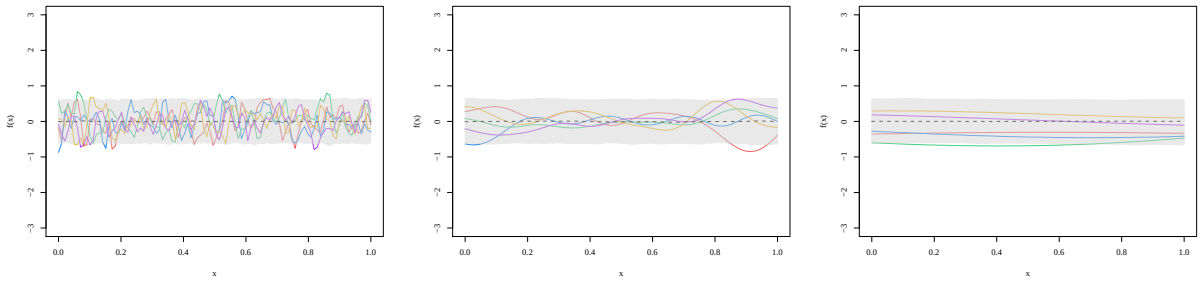
- b controls how much distant points are correlated. Small values of b produce a non correlated process (more noisy), large values of b produce a process with high correlation and, subsequently, smoother functions;
- τ_1^2 is the scaling factor that controls the amplitude of the covariance matrix. Larger values will amplify the values of the process;
- τ_2^2 is an additive scaling factor that controls the amplitude of the variances, that is, how noisy each point is.

As one can remark in Figure 7 and Figure 8, the functions sampled with Matern covariance function tend to be less smooth than the ones with squared exponential covariance function.

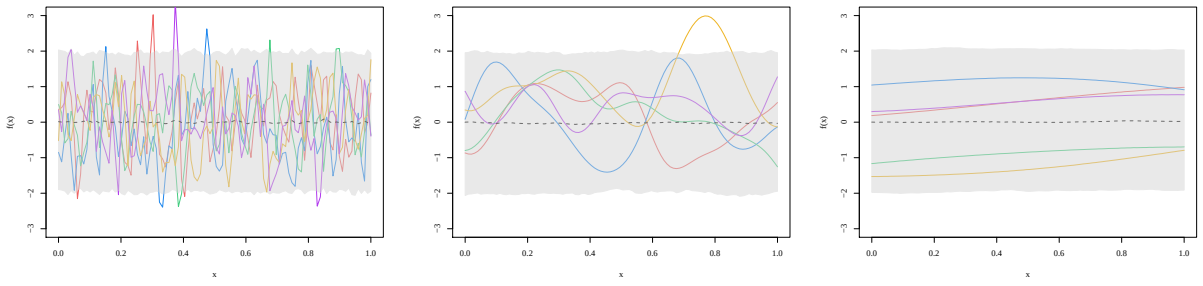
(B) *Suppose you observe the value of a Gaussian process $f \sim GP(m, C)$ at points $x_1, \ldots, x_N$. What is the conditional distribution of the value of the process at some new point $x^\star$? For the sake of notational ease simply write the value of the (i, j) element of the covariance matrix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

(C) *Prove the following lemma.*

Lemma 1. *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.*

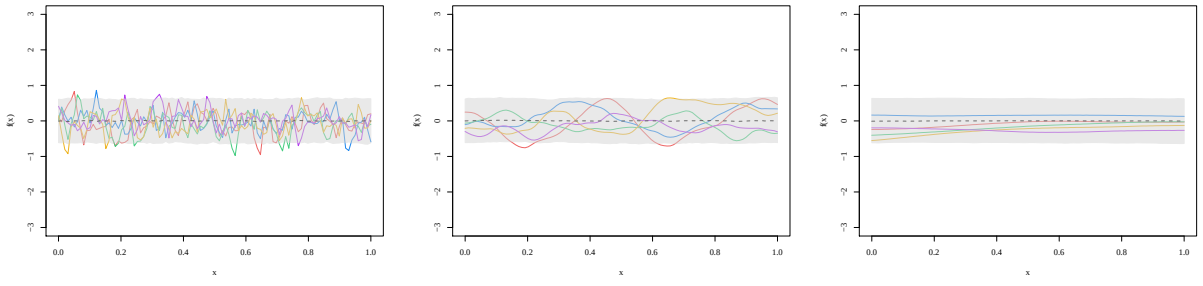


(a) $b = 0.01, \tau_1^2 = 0.1, \tau_2^2 = 10^{-6}$ *(b)* $b = 0.1, \tau_1^2 = 0.1, \tau_2^2 = 10^{-6}$ *(c)* $b = 1, \tau_1^2 = 0.1, \tau_2^2 = 10^{-6}$

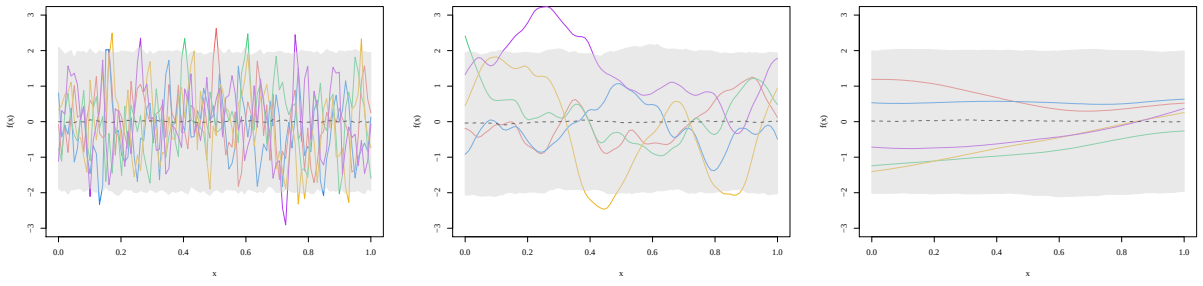


(d) $b = 0.01, \tau_1^2 = 1, \tau_2^2 = 10^{-6}$ *(e)* $b = 0.1, \tau_1^2 = 1, \tau_2^2 = 10^{-6}$ *(f)* $b = 1, \tau_1^2 = 1, \tau_2^2 = 10^{-6}$

Figure 7: Sample of 5 *functions from a Gaussian process with squared exponential covariance function. The mean* $m(x) = 0$ *is represented in dashed black line, the* 95% *predictive interval is shadowed in gray.*



(a) $b = 0.01, \tau_1^2 = 0.1, \tau_2^2 = 10^{-6}$ *(b)* $b = 0.1, \tau_1^2 = 0.1, \tau_2^2 = 10^{-6}$ *(c)* $b = 1, \tau_1^2 = 0.1, \tau_2^2 = 10^{-6}$



(d) $b = 0.01, \tau_1^2 = 1, \tau_2^2 = 10^{-6}$ *(e)* $b = 0.1, \tau_1^2 = 1, \tau_2^2 = 10^{-6}$ *(f)* $b = 1, \tau_1^2 = 1, \tau_2^2 = 10^{-6}$

Figure 8: Sample of 5 *functions from a Gaussian process with Matern* $5/2$ *exponential covariance function. The mean* $m(x) = 0$ *is represented in dashed black line, the* 95% *predictive interval is shadowed in gray.*

Problem 6. In nonparametric regression and spatial smoothing

(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.

(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 $f(x^\star)$.*

(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.*

(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.*

(E) *Return to the utilities or ethanol data sets. Fix $\tau_2^2 = 0$, and evaluate the log of the marginal likelihood function $p(y \mid \tau_1^2, b)$ over a discrete 2-d grid of points. If you're getting errors in your code with $\tau_2^2 = 0$, use something very small instead. Use this plot to choose a set of values $(\hat{\tau_1^2}, \hat{b})$ for the hyperparameters. Then use these hyperparameters to compute the posterior mean for f, given y. Comment on any lingering concerns you have with your fitted model.*

(F) *In `weather.csv` you will find data on two variables from 147 weather stations in the American Pacific northwest.*

- *pressure: the difference between the forecasted pressure and the actual pressure reading at that station (in Pascals)*
- *temperature: the difference between the forecasted temperature and the actual temperature reading at that station (in Celsius)*

There are also latitude and longitude coordinates of each station. Fit a Gaussian process model for each of the temperature and pressure variables. Choose hyperparameters appropriately. Visualize your fitted functions (both the posterior mean and posterior standard deviation) on a regular grid using something like a contour plot or color image. Read up on the `image`, `filled.contour`, or `contourplot` functions in R. An important consideration: is Euclidean distance the appropriate measure to go into the covariance function? Or do we need separate length scales for the two dimensions, i.e.

$$d^2(x, z) = \frac{(x_1 - z_1)^2}{b_1^2} + \frac{(x_2 - z_2)^2}{b_2^2}\,.$$

Justify your reasoning for using Euclidean distance or this "nonisotropic" distance.

Appendix A

R code