Max Planck Institute for Gravitational Physics
IMPRS Lecture Series

Making sense of data: introduction to statistics for gravitational wave astronomy

Problem Sheet 4: Advanced topics in statistics

---

1. (a) The mean is

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 50.2.$$

The autocovariance coefficient at lag k is

$$c_k = \frac{1}{n-k-1}\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x}).$$

We obtain

$$c_0 = 1.803, \qquad c_1 = -0.431, \qquad c_2 = 0.133, \qquad c_3 = 0.056.$$

The autocorrelation function at lag k is $r_k = c_k/c_0$, which gives

$$\hat{r}_1 = -0.239, \qquad \hat{r}_2 = 0.0739, \qquad \hat{r}_3 = 0.0308.$$

Note that in the literature you might see the correlation coefficients defined instead as

$$c_k = \frac{1}{n}\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x}).$$

This is a biased estimator of the covariance coefficients, but it has certain nice properties. Using this definition we obtain

$$c_0 = 1.623, \qquad c_1 = -0.345, \qquad c_2 = 0.093, \qquad c_3 = 0.033.$$

Conclusions below are not altered by using the alternative definition.

(b) The correlogram for these coefficients is shown in Figure 1.

(c) i. Under the white noise hypothesis, the individual coefficients are approximately Normal with variance $1/n = 0.1$. We set a threshold of $|r_k| > 1.96/\sqrt{n} = 0.620$ for a 95% significance test of a single coefficient. The number of coefficients, out of the 3 we have calculated, that exceed the threshold under the null hypothesis, follows a $\mathrm{Bin}(3, 0.05)$ distribution. The probability that the number of coefficients exceeding the threshold is $k \geq 1$ is 0.086, and for $k \geq 2$ it is 0.012. The threshold is therefore 2. In this case 0 coefficients exceed the threshold so we have no evidence to reject the white noise hypothesis.

The largest magnitude coefficient is $|r_1| = 0.239$. This corresponds to a p-value of 0.450 in the $\mathrm{N}(0, 0.1)$ distribution. The probability of one or more successes in a $\mathrm{Bin}(3, 0.450)$ distribution is 0.976, so this is the p-value of the test on this data, i.e., it is highly insignificant.

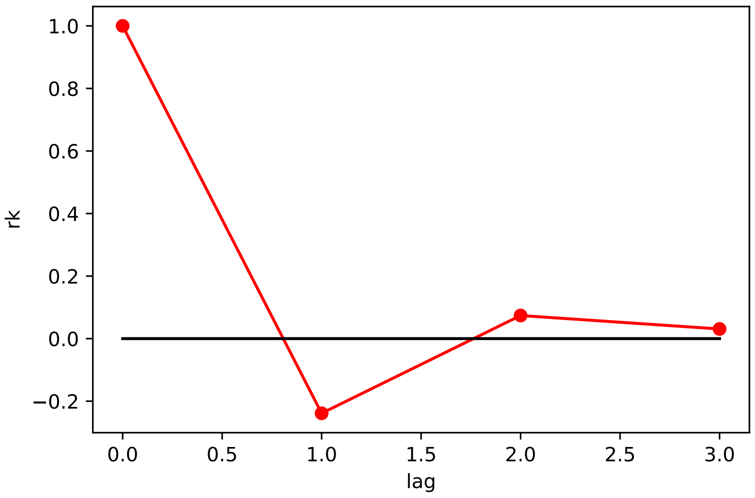

Figure 1: Correlogram for temperature data.

ii. The Ljung-Box test statistic is

$$Q = n(n+2)\sum_{h=1}^{m} \hat{r}_h^2/(n-h).$$

For this data set we compute $Q = 0.859$. This should be compared to a χ_3^2 distribution, for which the upper 95% point is $Q = 7.815$. Again there is no evidence to reject the white noise hypothesis. The p-value in this case is 0.835.

(d) In this case the estimated autocorrelation coefficients are

$$\hat{r}_1 = 0.315, \qquad \hat{r}_2 = 0.375, \qquad \hat{r}_3 = 0.264.$$

The new value of the Ljung-Box test statistic is $Q = 19.8$, which is above the threshold and so we now reject the white noise hypothesis for the full series. In this case the p-value is 0.0002, so there is strong evidence that the series is not white noise.

2. (a) The least squares estimator is given by

$$\hat{\alpha}_1 = \text{argmin}\left(\sum_{i=2}^{n}(x_i - \alpha_1 x_{i-1})^2\right).$$

Differentiating with respect to α_1 and setting the derivative to zero we obtain the equation

$$\sum_{i=2}^{n} x_{i-1}(x_i - \hat{\alpha}_1 x_{i-1}) = 0 \qquad \Rightarrow \qquad \hat{\alpha}_1 = \frac{\sum_{i=1}^{n-1} x_i x_{i+1}}{\sum_{i=1}^{n-1} x_i^2}.$$

(b) The autocorrelation coefficient estimator, $\hat{r}_1$, assuming the mean is zero, is given by

$$\hat{r}_1 = \frac{n}{n-1}\frac{\sum_{i=1}^{n-1} x_i x_{i+1}}{\sum_{i=1}^{n} x_i^2}.$$

There is a difference in the prefactor, but this is close to 1 and it is exactly 1 when using the alternative version of the autocorrelation function estimator mentioned in the previous question (using $1/n$ instead of $1/(n-k-1)$ as the prefactor, or $1/n$ instead of $1/(n-k)$ in the case that the mean is known). There is also a difference in that the sum in the denominator includes x_n^2, while it does not for $\hat{\alpha}_1$. This correction will be small for large numbers of samples n. In practice, $\hat{r}_1$ is often used as the estimator of α_1.

(c) Using the above estimator, $\hat{\alpha}_1$, we obtain for this data

$$\hat{\alpha}_1 = 0.624.$$

The best-fit AR(1) model is thus

$$X_t = 0.624 X_{t-1} + w_t.$$

3. (a) Using the usual backshift operator B this model can be written as

$$\phi(B)X_t = \theta(B)w_t, \qquad \text{where } \phi(B) = 1{+}0.3B{-}0.1B^2, \qquad \theta(B) = 1{+}0.1B{-}0.2B^2.$$

The two polynomials can be factorized

$$\phi(B) = 0.1(5-B)(2+B), \qquad \theta(B) = 0.1(5-2B)(2+B).$$

The roots of $\phi(B)$ are at $B = 5$ and $B = -1$, while those of $\theta(B)$ are at $B = 5/2$ and $B = -2$. All of these roots lie outside the unit circle, so the process is both invertible and causal. However, the polynomials share a common root at $B = -2$ and so the process is not regular.

(b) Cancelling the common root we see that the process is equivalent to

$$X_t - 0.2X_{t-1} = w_t - 0.4w_{t-1}, \qquad \tilde{\phi}(B)X_t = \tilde{\theta}(B),$$

where $\tilde{\phi}(B) = 1 - 0.2B$ and $\tilde{\theta}(B) = 1 - 0.4B$, The only root of $\tilde{\phi}(B)$ is at $B = 5$ and the only root of $\tilde{\theta}(B)$ is at $B = 5/2$. These are both outside the unit circle so the process is regular as required.

(c) Formally we can write

$$\begin{aligned} X_t &= (1-0.2B)^{-1}(1-0.4B)w_t = (1+0.2B+0.04B^2+0.008B^3+\cdots)(1-0.4B)w_t \\ &= (1-0.2B-0.04B^2-0.008B^3+\cdots)w_t \\ &= w_t - 0.2w_{t-1} - 0.04w_{t-2} - 0.008w_{t-3} + \cdots = \sum_{i=0}^{\infty} \pi_i w_{t-i}, \end{aligned}$$

with $\pi_0 = 1$, $\pi_1 = -0.2$, $\pi_2 = -0.04$ and $\pi_3 = -0.008$. Similarly

$$\begin{aligned} w_t &= (1-0.4B)^{-1}(1-0.2B)X_t = (1+0.4B+0.16B^2+0.064B^3+\cdots)(1-0.2B)X_t \\ &= (1+0.2B+0.08B^2+0.032B^3+\cdots)X_t \\ &= X_t + 0.2X_{t-1} + 0.08X_{t-2} + 0.032X_{t-3} + \cdots = \sum_{i=0}^{\infty} \psi_i X_{t-i}, \end{aligned}$$

where $\psi_0 = 1$, $\psi_1 = 0.2$, $\psi_2 = 0.08$ and $\psi_3 = 0.032$.

4. (a) The model can be written as

$$(1-B)^3 Y_t = Z_t - \theta Z_{t-1} = (1-\theta B)Z_t \tag{1}$$

where B is the usual backshift operator defined such that $BY_t = Y_{t-1}$. We recognise this as an ARIMA(0,3,1) model. ARIMA processes with $d \neq 0$ are never stationary and so this is not a stationary time series.

(b) We can write, formally,

$$Z_t = (1-\theta B)^{-1}(1-B)^3 Y_t = \Pi(B)Y_t.$$

If the process is invertible then we can find $\Pi(B)$ via expanding the above expression

$$\begin{aligned}
Z_t &= (1-B)^3(1+\theta B+\theta^2B^2+\theta^3B^3+\cdots)Y_t \\
&= (1-3B+3B^2-B^3)(1+\theta B+\theta^2B^2+\theta^3B^3+\cdots)Y_t \\
&= [(1+\theta B+\theta^2B^2+\theta^3B^3+\cdots)-(3B+3\theta B^2+3\theta^2B^3+\cdots) \\
&\qquad +(3B^2+3\theta B^3+\cdots)-(B^3+\cdots)]Y_t \\
&= Y_t+(\theta-3)Y_{t-1}+(\theta^2-3\theta+3)Y_{t-2}+(\theta^3-3\theta^2+3\theta-1)Y_{t-3}+\cdots
\end{aligned}$$

So we deduce $\pi_0 = 1$, $\pi_1 = \theta - 3$, $\pi_2 = (\theta^2 - 3\theta + 3)$ and $\pi_3 = (\theta^3 - 3\theta^2 + 3\theta - 1)$.

(c) The local trend model is defined by the equations

$$Y_t = a_t + \epsilon_t, \qquad a_t = a_{t-1} + \eta_t.$$

Simple manipulations give us

$$Y_t - Y_{t-1} = \epsilon_t + \eta_t - \epsilon_{t-1}$$

which looks quite similar to an ARIMA(0,1,1) model of the general form

$$Y_t - Y_{t-1} = \zeta_t - \theta\zeta_{t-1}.$$

We need to find a suitable specification of θ and σ_ζ^2 such that these two models are equivalent. Both the models take the form

$$Y_t - Y_{t-1} = \xi_t$$

where ξ_t is a sequence of zero mean Gaussian random variables. A set of Gaussian random variables is completely and uniquely characterised by its mean and two-point function (covariance). Therefore, all we need to do is find θ and σ_ζ^2 such that the ARIMA model has the same covariance properties as the original series.

For the original series we have variance

$$\mathrm{var}(\xi_t) = \mathrm{var}(\epsilon_t + \eta_t - \epsilon_{t-1}) = 2\sigma^2 + \sigma_\eta^2$$

and covariances

$$\mathrm{cov}(\xi_t, \xi_{t-k}) = \mathrm{cov}(\epsilon_t + \eta_t - \epsilon_{t-1}, \epsilon_{t-1} + \eta_{t-1} - \epsilon_{t-2})I(k=1) = -\sigma^2 I(k=1).$$

For the ARIMA model we have

$$\mathrm{var}(\xi_t) = (1+\theta^2)\sigma_\zeta^2$$

and

$$\mathrm{cov}(\xi_t,\xi_{t-k}) = \mathrm{var}(\zeta_t - \theta\zeta_{t-1}, \zeta_{t-1} - \theta\zeta_{t-2})I(k=1) = -\theta\sigma_\zeta^2 I(k=1).$$

Hence we need to solve

$$2\sigma^2 + \sigma_\eta^2 = (1+\theta^2)\sigma_\zeta^2, \qquad \sigma^2 = \theta\sigma_\zeta^2$$

which yield the equation

$$\theta + 1/\theta = 2 + \sigma_\eta^2/\sigma^2.$$

Since $\sigma_\eta^2/\sigma^2 > 0$ this has two solutions, one with $\theta < 1$ and one with $\theta > 1$. So, we can take the solution with $|\theta| < 1$ and obtain an invertible ARIMA process that generates the local trend model.

(d) Subsitituting the definitions of a_{n+1} and B_{n+1} into

$$Y_{n+1}^n = a_{n+1} + b_{n+1}$$

we obtain

$$\begin{aligned}
Y_{n+1}^n &= \alpha Y_n + (1-\alpha)Y_n^{n-1} + \gamma(a_{n+1} - a_n) + (1-\gamma)b_n \\
&= \alpha Y_n + (1-\alpha)(a_n + b_n) + \gamma[\alpha Y_n + (1-\alpha)(a_n+b_n)] - \gamma a_n + (1-\gamma)b_n \\
&= \alpha(1+\gamma)Y_n + (1-\alpha-\alpha\gamma)a_n + (2-\alpha-\alpha\gamma)b_n \\
&= \alpha(1+\gamma)Y_n + (2-\alpha\alpha\gamma)(a_n+b_n) - a_n.
\end{aligned}$$

We now note that $a_n + b_n = Y_n^{n-1}$ and $a_n = \alpha Y_{n-1} + (1-\alpha)Y_{n-1}^{n-2}$. Hence

$$Y_{n+1}^n = \alpha(1+\gamma)Y_n - \alpha Y_{n-1} + (2-\alpha-\alpha\gamma)Y_n^{n-1} - (1-\alpha)Y_{n-1}^{n-2}$$

as required.

5. (a) NW estimator:

$$\hat{f}_h(x) = \frac{\sum_i Y_i K_h(x_i - x)}{\sum_j K_h(x_j - x)}, \qquad K_h(x) = \frac{1}{h}K\left(\frac{x}{h}\right).$$

From the lecture notes, we have

$$\begin{aligned}
b(x) = \mathbb{E}\widehat{f}(x) - f(x) &= \sum_{i=1}^n w_i(x)[f(x_i) - f(x)] \quad \text{[Taylor Expansion]} \\
&\approx \sum_{i=1}^n w_i(x)\left[f(x) + f'(x)(x_i - x) + f''(x)\frac{(x_i-x)^2}{2} - f(x)\right] \\
&= \sum_{i=1}^n \frac{K_h(x_i - x)}{\sum_{j=1}^n K_h(x_j - x)}\left[f'(x)(x_i - x) + f''(x)\frac{(x_i-x)^2}{2}\right] \\
&\approx \frac{1}{n}\left[f'(x)\sum_{i=1}^n (x_i - x)K_h(x_i - x) + f''(x)\sum_{i=1}^n K_h(x_i - x)\frac{(x_i-x)^2}{2}\right] \\
&\approx f'(x)\int_0^1 (z-x)K_h(z-x)dz + f''(x)\int_0^1 K_h(z-x)\frac{(z-x)^2}{2}dz \\
&\approx f'(x)h\int_{-x/h}^{(1-x)/h} K(v)vdv + f''(x)\frac{h^2}{2}\int_{-x/h}^{(1-x)/h} K(v)v^2dv \\
&\approx f'(x)h\int_{-\infty}^{\infty} K(v)vdv + f''(x)\frac{h^2}{2}\int_{-\infty}^{\infty} K(v)v^2dv \\
&= \frac{\mu_2(K)h^2}{2}f''(x)
\end{aligned}$$

using $\frac{1}{n}\sum_{i=1}^n g(x_i) \approx \int_0^1 g(z)dz$ which implies $\frac{1}{n}\sum_{i=1}^n K_h(x_i - x) \approx 1$, $x/h \to +\infty$ and $(1-x)/h \to +\infty$ since $x \in (0,1)$.

Similarly:

$$\begin{aligned}
v(x) &= \sigma^2 \sum_{i=1}^n [w_i(x)]^2 = \sigma^2 \sum_{i=1}^n \frac{[K_h(x_i - x)]^2}{\left[\sum_{j=1}^n K_h(x_j - x)\right]^2} \\
&\approx_{\left\{\frac{1}{n}\sum_{i=1}^n K_h(x_i - x) \approx 1\right\}} \frac{\sigma^2}{n^2} \sum_{i=1}^n [K_h(x_i - x)]^2 \\
\left\{\tfrac{1}{n}\textstyle\sum_{i=1}^n \to \int_0^1\right\} &\approx \frac{\sigma^2}{n} \int_0^1 [K_h(z-x)]^2 dz = \frac{\sigma^2}{nh} \int_0^1 \left[K\left(\frac{z-x}{h}\right)\right]^2 d\left(\frac{z-x}{h}\right) \\
\left\{v = \tfrac{z-x}{h}\right\} &= \frac{\sigma^2}{nh} \int_{-x/h}^{(1-x)/h} \left[K\left(v\right)\right]^2 dv \approx \frac{\sigma^2}{nh} \int_{-\infty}^{\infty} \left[K\left(v\right)\right]^2 dv \\
&= \frac{\sigma^2}{nh} ||K||_2^2.
\end{aligned}$$

(b) The mean square error of an estimator $\hat{f}$ is $\mathbb{E}((\hat{f} - \mathbb{E}(\hat{f}))^2) = v(x) + b^2(x)$. From the previous results

$$\text{AMSE} \approx \frac{\sigma^2}{nh} ||K||_2^2 + \frac{\mu_2(K)^2 h^4}{4} (f''(x))^2.$$

Differentiating with respect to h and setting to zero we find the optimal bandwidth

$$h_{\text{opt}} = \left(\frac{\sigma^2 ||K||_2^2}{n(f''(x))^2 \mu_2(K)^2} \right)^{\frac{1}{5}}.$$

Substituting back into the AMSE we have

$$\text{AMSE}_{\text{opt}} = C \left(\sqrt{\mu_2(K)} ||K||_2^2 \right)^{\frac{4}{5}}$$

where

$$C = \frac{5\sigma^{\frac{8}{5}} (f''(x))^{\frac{2}{5}}}{4n^{\frac{4}{5}}}.$$

(c) i. We require $\int K(x)\mathrm{d}x = 1$, hence

$$1 = 2A \int_0^{\sqrt{5}} (1 - x^2/5)\mathrm{d}x = 2A[x - x^3/15]_0^{\sqrt{5}} = 4A\sqrt{5}/3 \quad \Rightarrow \quad A = 3/(4\sqrt{5}).$$

We have $\int xK(x)\mathrm{d}x = 0$ by inspection and

$$\int x^2 K(x)\mathrm{d}x = 2A[x^3/3 - x^5/25]_0^{\sqrt{5}} = 4A\sqrt{5}/3 = 5 \neq 0$$

so the order is 2.

ii. The distribution is Normal, $\hat{f}_h(x) \sim N(f(x) + b(x), v(x))$, with $f(x) + b(x) = \sum w_i(x) f(x_i)$ and $v(x) = \sigma^2 \sum w_i^2(x)$. For this kernel

$$||K||_2^2 = 2A^2 \int_0^{\sqrt{5}} (1 - x^2/5)^2 \mathrm{d}x = 2A^2[x - 2x^3/15 + x^5/125]_0^{\sqrt{5}} = 16\sqrt{5}A^2/15 = \frac{3}{5\sqrt{5}} = 0.268.$$

The variance is therefore approximately

$$v(x) \approx \frac{0.4^2}{3} \times 0.671 = 0.0143.$$

Assuming the bias is negligible, a confidence interval is therefore

$$|f(0.2) - 1.2| \leq 1.96\sqrt{0.036} = 0.234 \quad \Rightarrow \quad f(0.2) \in [0.965, 1.434].$$

As 1.5 does not lie in this confidence interval, so we reject the null hypothesis at the 5% significance level.

(d) i. The Kernel is of order 2, therefore

$$0 = \int_{-\infty}^{\infty} xK(x)\mathrm{d}x = AbD^2 \quad \Rightarrow \quad b = 0$$

since if $A = 0$ or $D = 0$, the kernel function is identically zero. We find A from

$$1 = \int_{-\infty}^{\infty} K(x)\mathrm{d}x = 2A(D + cD^3/3).$$

ii. Suppose $\mu_2(\tilde{K}) = \int_{-\infty}^{\infty} x^2 K(x)\mathrm{d}x$, then

$$\int_{-\infty}^{\infty} x^2 \tilde{K}_h(x)\mathrm{d}x = h^2 \int_{-\infty}^{\infty} u^2 \tilde{K}(u)\mathrm{d}u = h^2 \mu_2(\tilde{K})$$

where we have made the substitution $u = x/h$. Hence setting $h = 1/\sqrt{\mu_2(\tilde{K})}$ ensures $\mu_2(\tilde{K}_h) = 1$. Imposing this constraint on $K(x)$ gives

$$1 = \int_{-\infty}^{\infty} x^2 K(x)\mathrm{d}x = 2A(D^3/3 + cD^5/5).$$

iii. From the result in part (b), the AMSE is proportional to $\mu_2(K)^{\frac{2}{5}}||K||_2^{\frac{8}{5}}$. Since we can impose the constraint that $\mu_2(K) = 1$ from the previous result, this reduces to $||K||_2^{\frac{8}{5}}$. Hence we want to choose the parameters to minimize

$$||K||_2^2 = \int_{-\infty}^{\infty} K^2(x)\mathrm{d}x = 2A^2(D + 2cD^3/3 + c^2D^5/5).$$

From the previous constraints we have

$$\begin{aligned} \frac{1}{2A} &= D + \frac{cD^3}{3} = \frac{D^3}{3} + \frac{cD^5}{5} \\ \Rightarrow \quad c &= \frac{5(D^2-3)}{D^2(5-3D^2)} \\ A &= \frac{3(3D^2-5)}{8D^3}. \end{aligned}$$

Substituting into the expression above gives

$$\begin{aligned} ||K||_2^2 &= \frac{3}{32D^6}\left(3D(3D^2-5)^2 + 10D(3-D^2)(3D^2-5) + 15D(D^2-3)^2\right) \\ &= \frac{3}{8}\left(\frac{3}{D} - \frac{10}{D^3} + \frac{15}{D^5}\right). \end{aligned}$$

As required. Differentiating with respect to D and setting it to 0 gives

$$0 = -\frac{9}{8D^6}\left(D^2 - 5\right)^2 \quad \Rightarrow \quad D^2 = 5$$

so we recover the Epanechnikov kernel.

6. (a) Differentiation of the expression with respect to θ_0 and θ_1 gives

$$\begin{aligned} 0 &= -2\sum_{i=1}^{n}\left(Y_i - \hat{\theta}_0 - \hat{\theta}_1\frac{(x_i - x)}{h}\right)K_h(x_i - x) \\ 0 &= -2\sum_{i=1}^{n}\left(\frac{x_i - x}{h}\right)\left(Y_i - \hat{\theta}_0 - \hat{\theta}_1\frac{(x_i - x)}{h}\right)K_h(x_i - x). \end{aligned}$$

Using the definitions in the question we obtain the simultaneous equations

$$\begin{aligned} 0 &= S_y - \hat{\theta}_0 S - \hat{\theta}_1 S_x \\ 0 &= S_{xy} - \hat{\theta}_0 S_x - \hat{\theta}_1 S_{xx}. \end{aligned}$$

Taking appropriate linear combinations gives the expressions in the question.

(b) For large n we can approximate $\frac{1}{n}\sum_i f(x_i) = \int_0^1 f(x)\mathrm{d}x$. The various terms in the expression for $\hat{\theta}_1$ can therefore be approximated by

$$\begin{aligned} S &\approx n\int_{-\infty}^{\infty} K(u)\mathrm{d}u = n \\ S_x &\approx n\int_{-\infty}^{\infty} uK(u)\mathrm{d}u = 0 \\ S_{xx} &\approx n\int_{-\infty}^{\infty} u^2K(u)\mathrm{d}u = \mu_2(K). \end{aligned}$$

Putting this together we have

$$\hat{f}'(x) = \frac{\hat{\theta}_1(x)}{h} \approx \frac{nS_{xy}}{n^2\mu_2(K)} = \frac{1}{nh\mu_2(K)}\sum_{i=1}^{n} Y_i\left(\frac{x_1 - x}{h}\right)K_h(x_i - x),$$

as required.

(c) The variance of $\hat{g}(x)$ is

$$\mathrm{var}(\hat{g}) = \sum l_i^2(x)\mathrm{var}(Y_i) = \sigma^2\sum l_i^2(x) = \sigma^2||\mathbf{l}(x)||^2.$$

(x) As required. For the LP(1) estimator of f' we have

$$l_i(x) = \frac{1}{nh\mu_2(K)}\left(\frac{x_1 - x}{h}\right)K_h(x_i - x).$$

Hence

$$\begin{aligned} \mathrm{var}(\hat{f}'(x)) &= \frac{\sigma^2}{n^2h^2\mu_2^2(K)}\sum_{i=1}^{n}\left(\frac{x_1 - x}{h}\right)^2 K_h(x_i - x)^2 \\ &= \frac{\sigma^2}{nh^3\mu_2^2(K)}\left[\frac{1}{nh}\sum_{i=1}^{n}\left(\frac{x_1 - x}{h}\right)^2 K^2\left(\frac{x_1 - x}{h}\right)\right] \\ &\approx \frac{\sigma^2\nu_2(K)}{nh^3\mu_2^2(K)} \\ \text{where } \nu_2(K) &= \int_{-\infty}^{\infty} u^2K^2(u)\mathrm{d}u. \end{aligned}$$

(d) i. Firstly we note

$$||\mathbf{l}(x)||^2 = \left(\frac{1}{nh\mu_2(K)}\right)^2 \sum_{i=1}^{n} \left(\frac{x_1 - x}{h}\right)^2 K_h^2(x_i - x) \approx \left(\frac{1}{nh\mu_2(K)}\right)^2 \frac{n}{h}\nu_2(K).$$

Hence

$$T_i(x) \approx \sqrt{\frac{h}{n\nu_2(K)}} \left(\frac{x_1 - x}{h}\right) K_h(x_i - x) = \sqrt{\frac{1}{nh\nu_2(K)}} G\left(\frac{x_i - x}{h}\right),$$

where $G(u) = uK(u)$. We then find

$$T_i'(x) \approx -\sqrt{\frac{1}{nh^3\nu_2(K)}} G'\left(\frac{x_i - x}{h}\right)$$

and hence

$$||\mathbf{T}'(x)||^2 = \frac{1}{nh^3\nu_2(K)} \sum_{i=1}^{n} (G'\left(\frac{x_i - x}{h}\right))^2 \approx \frac{1}{h^2\nu_2(K)} ||G'||_2^2 = \left(\frac{||G'||_2}{h||G||_2}\right)^2.$$

This is independent of x and so we have

$$\kappa_0 = \int_a^b ||\mathbf{T}'(x)||\mathrm{d}x = (b - a)||\mathbf{T}'(x)|| = \left(\frac{b-a}{h}\right) \frac{||G'||_2}{||G||_2} = \left(\frac{b-a}{h}\right) \frac{||xK' + K||}{||xK||}.$$

As required

ii. To construct the asymptotic confidence we need to compute the variance, and κ_0. For the former we need

$$\nu_2(K) = ||xK||_2^2 = 2\int_0^1 x^2(1-x)^2 \mathrm{d}x = 2[1/3 - 1/2 + 1/5] = 1/15$$

and

$$\mu_2(K) = 2\int_0^1 x^2(1-x)\mathrm{d}x = 2[1/3 - 1/4] = 1/6.$$

Hence we compute

$$\mathrm{var}(\hat{f}') = \frac{0.01 \times 36}{200 \times 0.1^3 \times 15} = 0.12.$$

We then evaluate κ_0, for which we need

$$||K' + xK||_2^2 = 2\int_0^1 (1-2x)^2 \mathrm{d}x = 2[1 - 2 + 4/3] = 2/3$$

hence we find

$$\kappa_0 = \frac{1}{0.1}\sqrt{10} =, \quad \Rightarrow \quad c_\alpha = 3.257.$$

The asymptotic confidence band is

$$|g(x) - \hat{g}_h^{LP(1)}(x)| \leq 1.128.$$

If the function is linear then its derivative must be constant. A constant can only fit within the confidence band if there exists C such that

$$\max_{x\in[0,1]} \hat{g}_h^{LP(1)} - 1.128 \leq C \leq \min_{x\in[0,1]} \hat{g}_h^{LP(1)} + 1.128.$$

In this case we need

$$2.137 \leq C \leq 2.065$$

which is not possible so we reject the null hypothesis.

7. (a) i. Substituting

$$g(x) = \sum_{j=1}^{N} \beta_j h_j(x)$$

into the penalised least squares estimator we find that the smoothing spline is solved by

$$\widehat{\beta} = \arg\min_{\beta\in\mathbb{R}^N} \left\{ \sum_{i=1}^{N} \left[\sum_{j=1}^{N} \beta_j h_j(x_i) \right]^2 - 2\sum_{i=1}^{N} \left[\sum_{j=1}^{N} \beta_j h_j(x_i) \right] Y_i + \lambda \int \left[\sum_{j=1}^{N} \beta_j h_j''(x) \right]^2 dx \right\}$$

$$= \arg\min_{\beta\in\mathbb{R}^N} \left\{ \beta^T H^T H \beta - 2Y^T \mathbf{H}^T \beta + \lambda \beta^T \Omega \beta \right\},$$

where $N \times N$ matrix H has entries $H_{ij} = h_j(x_i)$, $i = 1, \ldots, N$, $j = 1, \ldots, N$, and $N \times N$ matrix Ω has elements $\Omega_{j\ell} = \int h_j''(x) h_\ell''(x) dx$, $j, \ell = 1, \ldots, N$. Differentiation with respect to β gives

$$2\left(\mathbf{H}^T\mathbf{H} + \lambda\mathbf{\Omega}\right)\hat{\beta} - 2\mathbf{H}^T\mathbf{Y} = 0$$

which can be rearranged to give the quoted solution.

ii. The basis functions in this case are

$$h_1(x) \;=\; 1, \qquad h_2(x) = x, \qquad h_3(x) = (x - 1/2)_+^3 - 2(x-1)_+^3 + (x - 3/2)_+^3.$$

The corresponding matrices are

$$\mathbf{H} = \begin{pmatrix} 1 & 1/2 & 0 \\ 1 & 1 & 1/8 \\ 1 & 3/2 & 3/4 \end{pmatrix}, \qquad \mathbf{\Omega} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 99 \end{pmatrix}$$

(b) Requiring the estimator to pass through all the knots is equivalent to

$$\sum_j \beta_j h_j(x_i) = Y_i \quad \forall\, i.$$

This can be written as

$$\mathbf{H}\beta = \mathbf{Y} \quad \Rightarrow \quad \beta = \mathbf{H}^{-1}\mathbf{Y}.$$

The $\lambda = 0$ limit of the smoothing spline has

$$\hat{\beta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{Y} = \mathbf{H}^{-1}(\mathbf{H}^T)^{-1}\mathbf{H}^T\mathbf{Y} = \mathbf{H}^{-1}\mathbf{Y}$$

and so the solutions agree. Setting $\lambda = 0$ puts no weight on the smoothness penalty and so forces the smoothing spline to go through all the points.

(c) i. Because $h_0''(x) = h_1''(x) = 0$, the first two columns and rows of Ω are all zero. Writing

$$\mathbf{H}^T\mathbf{H} + \lambda\Omega = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

where $\mathbf{A}_{11}$ is a 2×2 matrix, in the limit $\lambda \to \infty$ we have $\mathbf{A}_{22} \to \infty$ while all other matrices remain finite. Hence, using the formula in the hint

$$\left(\mathbf{H}^T\mathbf{H} + \lambda\Omega\right)^{-1} \to \begin{pmatrix} (\mathbf{A}_{11})^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

We see that $\hat{\beta}_j \to 0$ for $j \geq 3$ as required. The matrix

$$\mathbf{A}_{11} = \begin{pmatrix} n & \sum x_i \\ \sum x_j & \sum x_j^2 \end{pmatrix}$$

and the first two elements of $\mathbf{H}^T\mathbf{Y}$ are $\sum y_j$ and $\sum x_j y_j$. Hence

$$\begin{aligned} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} &= \frac{1}{n\sum x_j^2 - (\sum x_j)^2} \begin{pmatrix} \sum x_j^2 & -\sum x_j \\ -\sum x_j & n \end{pmatrix} \begin{pmatrix} \sum y_j \\ \sum x_j y_j \end{pmatrix} \\ &= \frac{1}{n\sum (x_j - \bar{x})^2} \begin{pmatrix} n\bar{Y}\sum x_j^2 - n\bar{x}\sum x_j y_j \\ n\sum (x_j - \bar{x}) y_j \end{pmatrix}. \end{aligned}$$

ii. We see immediately that $\hat{\beta}_2$ agrees with $\hat{\alpha}_2$. For $\hat{\beta}_1$ we write

$$\begin{aligned} n\bar{Y}\sum x_j^2 - n\bar{x}\sum x_j y_j &= n\bar{Y}\sum (x_j - \bar{x})^2 + n^2\bar{Y}\bar{x}^2 - n\bar{x}\sum x_j y_j \\ &= n\bar{Y}\sum (x_j - \bar{x})^2 - n\bar{x}\sum (x_j - \bar{x}) y_j \end{aligned}$$

and hence this agrees with $\hat{\alpha}_1$. In the limit $\lambda \to \infty$ we are placing all the weight on smoothness and not the data. The smoothest function (for a f'' penalty) is a linear function. Then we find the best fit linear function to the data, which is this linear least squares estimator.

(d) On a regular grid, with N points in each dimension, the backfitting estimate involves fitting splines to the average of the observations in each direction separately, and using a smoothing parameter λ/N. First we subtract the mean value from all observations $\hat{\alpha} = 27/9 = 3$. Then, we fit in the x_1 direction, averaging over the repeated measurements at each x_1 value. We need to fit a smoothing spline to

$$(1/2, -4/3), \quad (1, 1), \quad (3/2, 1/3)$$

with smoothing parameter $\lambda/3 = 1/99$. The $\mathbf{H}$ and $\mathbf{\Omega}$ matrices are as found in part (a)(ii). Hence we compute

$$\mathbf{A} = \mathbf{H}^T\mathbf{H} + \lambda\Omega = \begin{pmatrix} 3 & 3 & 7/8 \\ 3 & 7/2 & 5/4 \\ 7/8 & 5/4 & 101/64 \end{pmatrix},$$

and

$$\mathbf{b}_1 = \mathbf{H}^T\mathbf{Y} = \begin{pmatrix} 1 & 1 & 1 \\ 1/2 & 1 & 3/2 \\ 0 & 1/8 & 3/4 \end{pmatrix} \begin{pmatrix} -4/3 \\ 1 \\ 1/3 \end{pmatrix} = \begin{pmatrix} 0 \\ 5/6 \\ 3/8 \end{pmatrix}$$

and $\hat{\beta}_1 = \mathbf{A}^{-1}\mathbf{b}_1$. Fitting in the x_2 direction we need to fit the data

$$(1/2, -1), \quad (1, 0), \quad (3/2, 1).$$

The $\mathbf{A}$ matrix is unchanged, but we now have

$$\mathbf{b}_2 = \mathbf{H}^T\mathbf{Y} = \begin{pmatrix} 1 & 1 & 1 \\ 1/2 & 1 & 3/2 \\ 0 & 1/8 & 3/4 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 3/4 \end{pmatrix}$$

and $\hat{\beta}_2 = \mathbf{A}^{-1}\mathbf{b}_2$. The final solution is

$$\hat{f}(x_1, x_2) = \hat{\alpha} + \sum_{j=1}^{3} \hat{\beta}_{1j} h_j(x_1) + \sum_{j=1}^{3} \hat{\beta}_{2j} h_j(x_2).$$

8. (a) The Cascade algorithm

i. Set $c_{Jk} = Y_{k+1}$ for $k = 0, 1, .., 2^J - 1$, set $j = J - 1$;

ii. Set

$$c_{jk} = \sum_{m \in \mathbb{Z}} h_m c_{j+1,2k+m}, \quad d_{jk} = \sum_{m \in \mathbb{Z}} g_m c_{j+1,2k+m};$$

iii. if $j = 0$ stop; else set $j := j - 1$ and repeat step 2.

In this case we find $c_{00} = 7.15/\sqrt{2} = 5.06$, $d_{00} = -2.85/\sqrt{2} = -2.02$, $d_{10} = -0.35$, $d_{11} = -0.8$, $d_{20} = -0.8/\sqrt{2} = -0.566$, $d_{21} = -0.7/\sqrt{2} = -0.495$, $d_{22} = -0.6/\sqrt{2} = -0.424$ and $d_{23} = 0.4/\sqrt{2} = 0.283$.

(b) To test the hypothesis that the function is constant we need to test whether there is a change point of any of the Haar wavelets in the interval $(0, 0.5)$. The wavelets that have change points in $(0.5, 1.0)$ are ψ_{11}, ψ_{22} and ψ_{23}. Therefore the equivalent hypothesis is

$$w_{11} = w_{22} = w_{23} = 0$$

The test statistic is $T = (d_{11}^2 + d_{22}^2 + d_{23}^2)/\sigma^2$ which follows a χ_3^2 distribution. In this case, we find $T = 40.0$ which is greater than $\chi_3^2(5\%) = 7.815$. So we reject the hypothesis that f is constant.

(c) The projection estimator is

$$\hat{f}_2(x) = \frac{1}{2\sqrt{2}} \left(c_{00}\phi(x) + d_{00}\psi(x) + \sqrt{2}d_{10}\psi(2x) + \sqrt{2}d_{11}\psi(2x - 1) \right)$$

At $x = 0.4$ c_{00}, d_{00} and d_{10} contribute and (asymptotically) these are uncorrelated and each has variance σ^2, therefore the combined variance is $(1 + 1 + 2)\sigma^2/n = \sigma^2/2 = 0.01125$. Evaluation of the estimator gives $\hat{f}_2(0.4) = \frac{1}{2\sqrt{2}}(c_{00} + d_{00} - \sqrt{2}d_{10}) = 1.2498$. A point wise 95% confidence interval is $1.2498 \pm 1.96\sqrt{0.01125} = 1.2498 \pm 0.2100$ and so $f(0.4) \in [1.0398, 1.4598]$.

(d) The universal threshold is

$$\lambda = \sigma\sqrt{2 \log n} = 0.306.$$

We set coefficients smaller than this value to zero, which means we set $d_{23} = 0$ only.

(e) i. To prove the first property

$$\begin{aligned} 1 &= \int \phi(x)dx = \sum_{k \in Z} h_k \sqrt{2} \int \phi(2x - k)dx = \sum_{k \in Z} h_k 2^{-1/2} \int \phi(v)dv \\ &= \frac{1}{\sqrt{2}} \sum_{k \in Z} h_k. \end{aligned}$$

To prove the second property

$$\begin{aligned} \delta_{0l} &= \int \phi(x)\phi(x - l)dx = 2 \int \left[\sum_{k \in Z} h_k \phi(2x - k) \sum_{m \in Z} h_m \phi(2x - m - 2l) \right]^2 dx \\ &= \sum_{k,m} h_k h_m \int \phi(2x - k)\phi(2x - m - 2l)d(2x) \\ &= \sum_{k,m} h_k h_m \delta_{k,m+2l} = \sum_k h_k h_{k-2l}. \end{aligned}$$

ii. To prove the first property

$$\begin{aligned} 0 &= \int \phi(x)\psi(x-m)dx = 2\sum_{k,l\in Z} h_l g_k \int \phi(2x-l)\phi(2x-2m-k)dx \\ &= 2\sum_{k,l\in Z} h_l g_k \delta_{l,2m+k} = \sum_{k\in Z} g_k h_{k+2m}. \end{aligned}$$

To prove the second property

$$\begin{aligned} \delta_{0m} &= \int \psi(x)\psi(x-l)dx = 2\int \left[\sum_{k\in Z} g_k\phi(2x-k)\sum_{m\in Z} g_m\phi(2x-m-2l)\right]^2 dx \\ &= \sum_{k,m} g_k g_m \int \phi(2x-k)\phi(2x-m-2l)d(2x) \\ &= \sum_{k,m} g_k g_m \delta_{k,m+2l} = \sum_k g_k g_{k-2l}. \end{aligned}$$

Setting $g_k = (-1)^k h_{1-k}$ we have

$$\begin{aligned} \sum_k g_k h_{k+2m} &= \sum_k (-1)^k h_{1-k} h_{k+2m} = \sum_{k'} (-1)^{1-2m-k'} h_{k'+2m} h_{1-k'} = -\sum_k g_k h_{k+2m} \\ &\Rightarrow \sum_k g_k h_{k+2m} = 0 \end{aligned}$$

where the intermediate step follows from setting $k'+2m = 1-k$. Similarly

$$\sum_k g_k g_{k-2l} = \sum_k (-1)^k h_{1-k}(-1)^{k-2l} h_{1-k+2l} = \sum_{k'} h_{k'} h_{k'-2l} = \delta_{0l}$$

where the intermediate step follows from setting $k' = 1-k$.

iii. With four unknown coefficients the relationships between the scaling coefficients provide three constraints

$$h_0 + h_1 + h_2 + h_3 = \sqrt{2} \tag{2}$$

$$h_0^2 + h_1^2 + h_2^2 + h_3^2 = 1 \tag{3}$$

$$h_0 h_2 + h_1 h_3 = 0. \tag{4}$$

The suggested substitution ensures Eq. (3) is satisfied. Using Eq. (4) we find

$$\begin{aligned} \cos\beta\sin\beta\cos\alpha\cos\gamma + \cos\beta\sin\beta\sin\alpha\sin\gamma &= 0 \\ \Rightarrow \cos(\alpha-\gamma) &= 0 \\ \Rightarrow \alpha-\gamma &= \pi/2 \text{ or } -\pi/2. \end{aligned}$$

Note that there is an alternative solution $\beta = 0$, but for this $h_2 = h_3 = 0$ and it is the Haar wavelet basis. Note also that the two alternative solutions yield the same final expressions for the h_i's, so we use $\alpha-\gamma = \pi/2$ in the following. Using Eq. (2) we now find

$$\begin{aligned} \cos\beta(\cos\alpha+\sin\alpha) + \sin\beta(\cos\gamma+\sin\gamma) &= \sqrt{2} \\ \cos\beta(\cos\alpha+\sin\alpha) + \sin\beta(\sin\alpha-\cos\alpha) &= \sqrt{2} \\ \cos(\alpha-\beta) + \sin(\alpha-\beta) &= \sqrt{2} \\ \Rightarrow \alpha-\beta = \pi/4 \end{aligned}$$

where the last step follows from the fact that the maximum of $\cos x + \sin x$ is $\sqrt{2}$ at $x = \pi/4$.

iv. We first verify $\sum_k g_k = 0$ as suggested in the hint:

$$\begin{aligned}
\sum_k g_k &= -h_0 + h_1 - h_2 + h_3 = \cos\beta(\sin\alpha - \cos\alpha) + \sin\beta(\sin\gamma - \cos\gamma) \\
&= \cos\beta(\sin\alpha - \cos\alpha) - \sin\beta(\sin\alpha + \cos\alpha) = \sin(\alpha - \beta) - \cos(\alpha - \beta) = 0.
\end{aligned}$$

Using $\int x\psi(x)\mathrm{d}x = 0$ we have

$$\begin{aligned}
0 &= \sqrt{2}\sum_k g_k \int x\phi(2x-k)\mathrm{d}x = 1/(2\sqrt{2})\sum_k g_k \int (u+k)\phi(u)\mathrm{d}u \\
\Rightarrow 0 &= \left[\int u\phi(u)\mathrm{d}u\right]\sum_k g_k + \left[\int \phi(u)\mathrm{d}u\right]\sum_k k g_k = \left[\int \phi(u)\mathrm{d}u\right]\sum_k k g_k \\
\Rightarrow \sum_k k g_k &= 0.
\end{aligned}$$

The additional constraint is $-2h_3 + h_2 - h_0 = 0$ from which

$$\begin{aligned}
h_2 - h_0 &= \sin\beta\cos\gamma - \cos\alpha\cos\beta = \sin\beta\sin\alpha - \cos\alpha\cos\beta \\
&= -\cos(\alpha+\beta) = -1/\sqrt{2}(\cos 2\alpha + \sin 2\alpha) \\
&= 2h_3 = 2\sin\beta\sin\gamma = -2\sin\beta\cos\alpha = \sqrt{2}\cos\alpha(\cos\alpha - \sin\alpha) \\
&= 1/\sqrt{2}(1 + \cos 2\alpha - \sin 2\alpha) \\
\Rightarrow 1 + 2\cos 2\alpha &= 0 \quad \Rightarrow \quad \cos 2\alpha = -1 \quad \Rightarrow \quad \alpha = -\pi/6 \text{ or } \pi/3.
\end{aligned}$$

9. While this question can be done by hand, it is much easier to do it using computer software to manipulate matrices, such as Mathematica or python.

(a) The first step is to estimate the hyperparameters of the model. The question suggested using a square exponential covariance function

$$k(x_1, x_2) = A\exp\left[-\frac{1}{2\sigma^2}(x_1 - x_2)^2\right].$$

The normalisation parameter A was not explicitly given in the question, but including it gives the model greater flexibility. We will consider two cases below, in which we either fix $A = 1$ or treat it as a hyperparameter to be constrained by the training data.

From the covariance function we construct the covariance matrix for the training data

$$K_{ij} = A\exp\left[-\frac{0.01}{2\sigma^2}(i-j)^2\right],$$

in which we use the fact that we have a regular design with $x_i = 0.1(i-1)$. We then determine the optimal hyperparameters by maximizing the likelihood for the training data, $\{y_i\}$. This is most easily accomplished by minimizing minus twice the log-likelihood, which up to constant terms is

$$\mathcal{L} = \sum_{i,j} y_i K_{ij}^{-1} y_j + \log(\det[\mathbf{K}]).$$

The optimal choice of hyperparameters can be found using a minimization routine such as scipy.optimize.minimize in python. In the case where we fix $A = 1$ we find the optimal choice $\sigma = 0.144$. Allowing both A and σ to vary we find the optimal choice is $A = 2.41$ and $\sigma = 0.156$.

We denote the kernel function and training data covariance matrix corresponding to these optimized hyperparameters by $k_{\text{opt}}(x_1, x_2)$ and $\mathbf{K}_{\text{opt}}$ respectively. The value of the Gaussian Process at a new set of points $\{z_i\}$ are described by a multi-variate normal distribution

$$p(\mathbf{y}) \propto \exp\left[-\frac{1}{2}(\mathbf{y}-\mu)^T \mathbf{\Sigma}^{-1}(\mathbf{y}-\mu)\right]$$

where $\mathbf{y}^T = \{y(z_i)\}$ and the mean and covariance matrix are given by

$$\begin{aligned}\mu_i &= \sum_{j,k} [\mathbf{K}_*]_{ji}\, [\mathbf{K}_{\text{opt}}^{-1}]_{jk}\tilde{y}_k \\ \Sigma_{ij} &= [\mathbf{K}_{**}]_{ij} - \sum_{k,l} [\mathbf{K}_*]_{ki}[\mathbf{K}^{-1}]_{kl}[\mathbf{K}_*]_{lj} \\ \text{where } [\mathbf{K}_*]_{ij} &= k_{\text{opt}}(x_i, z_j), \qquad [\mathbf{K}_{**}]_{ij} = k_{\text{opt}}(z_i, z_j),\end{aligned}$$

and $\tilde{y}_k$ is the observed value of the function at x_k.

(b) We now construct the GP approximant at the three new points, $\mathbf{z}^T = \{0.15, 0.45, 0.75\}$. Plugging these values into the above expressions, for the case where we fixed the hyperparameter $A = 1$, we obtain the mean and covariance matrix

$$\begin{aligned}\mu &= (-2.167, -2.324, -1.386)^T \\ \Sigma &= \begin{pmatrix} 1.435 \times 10^{-4} & -5.752 \times 10^{-5} & 4.166 \times 10^{-5} \\ -5.752 \times 10^{-5} & 3.856 \times 10^{-5} & -4.009 \times 10^{-5} \\ 4.166 \times 10^{-5} & -4.009 \times 10^{-5} & 6.771 \times 10^{-5} \end{pmatrix}\end{aligned}$$

and for the case where we optimize both A and σ using the training data we obtain

$$\begin{aligned}\mu &= (-2.172, -2.322, -1.391)^T \\ \Sigma &= \begin{pmatrix} 1.120 \times 10^{-4} & -4.060 \times 10^{-5} & 3.522 \times 10^{-5} \\ -4.060 \times 10^{-5} & 2.223 \times 10^{-5} & -2.615 \times 10^{-5} \\ 3.522 \times 10^{-5} & -2.615 \times 10^{-5} & 4.532 \times 10^{-5} \end{pmatrix}.\end{aligned}$$

We note that the covariance is particularly small. Taking the square root of the diagonal elements we estimate uncertainties in the values at the three points of $(0.012, 0.006, 0.008)$ or $(0.011, 0.005, 0.007)$ in the first and second case respectively. The reason for this is that we have constructed this Gaussian Process assuming that the training data contains no error. In fact this is not the case. The data was generated by adding $N(0, 0.15^2)$ errors to the quadratic function $5x^2 - 3x - 2$. As we are not accounting for that error the GP is underestimating the uncertainties. In Figure 2 we show the data, the GP estimates at the requested points, with their uncertainties, and the true function used to generate the data. We see that the value at $x = 0.15$ is not being well estimated, with the true value lying several standard deviations away from the GP estimate.

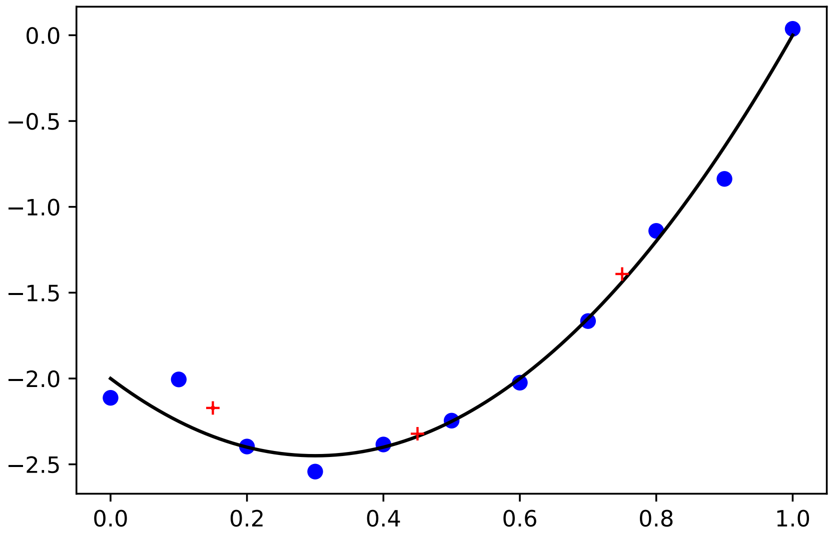

Figure 2: Red crosses show the estimated mean values of the function at the target points, with the vertical tics indicating the 1-σ uncertainties. The blue points are the observed training data points and the black curve indicates the true quadratic function that was used to generate the data.

Errors can be accounted for by modifying the $\mathbf{K}$ matrix to include the training data error terms

$$K_{ij} = A \exp\left[-\frac{0.01}{2\sigma^2}(i-j)^2\right] + \sigma_d^2 \delta_{ij}.$$

In this case we know that $\sigma_d = 0.15$ and so could use that directly. However, an alternative is to treat this parameter as an additional hyperparameter to optimize over the training data. Doing this we obtain optimal values of $A = 4.17$, $\sigma = 0.682$ and $\sigma_d = 0.127$. We note that the optimal covariance function is different, and much broader than before, but the optimal value of σ_d is pretty close to the true value used to generate the data.

Note that we include σ_d in $\mathbf{K}$ only and not in $\mathbf{K}_*$ or $\mathbf{K}_{**}$. Including it in the latter two matrices would be equivalent to attempting to predict the value of a future noisy measurement of the function, rather than the true value of the function. If we are interested in the latter, we account for noise in the training data only.

Using the optimized covariance function including the training data error term we find new values for the mean and covariance matrix at the requested points of

$$\begin{aligned}
\mu &= (-2.292, -2.356, -1.434)^T \\
\Sigma &= \begin{pmatrix} 4.248 \times 10^{-3} & 1.323 \times 10^{-3} & -8.568 \times 10^{-4} \\ 1.323 \times 10^{-3} & 3.777 \times 10^{-3} & 1.315 \times 10^{-3} \\ -8.568 \times 10^{-4} & 1.315 \times 10^{-3} & 4.257 \times 10^{-3} \end{pmatrix}.
\end{aligned}$$

In particular we see that the square roots of the diagonal elements of the covariance matrix are $(0.065, 0.061, 0.065)$, approximately an order of magnitude

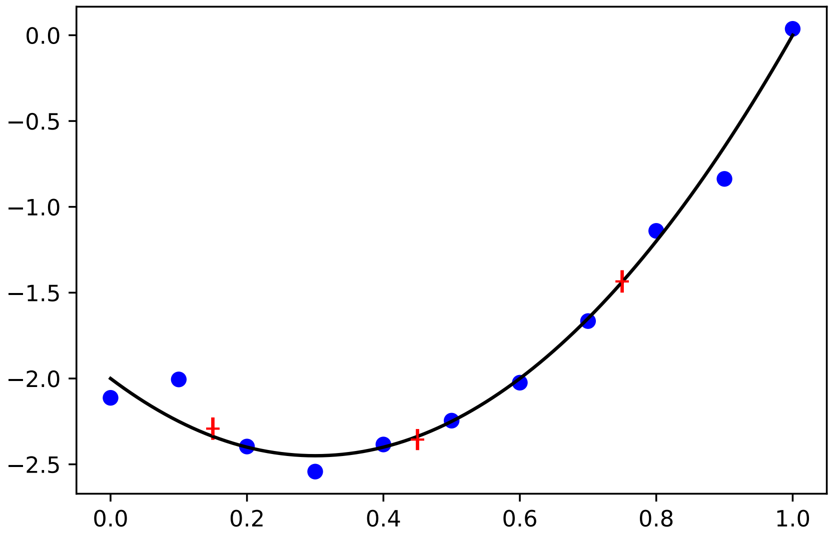

Figure 3: As Figure 2 but now for the Gaussian Process that was fit allowing for uncertainties in the training data.

larger than before. In Figure 3 we compare the new prediction from the GP model to the true function and now find that at all three points we are within the $1-\sigma$ uncertainty estimated from the GP model.

10. (a) We suppose that at iteration n there are $\{n_i\}$ diners at table i, and n_i is drawn from a multinomial distribution with probability q_i. The next ($n+1$'th) diner sits at Table i with probability $n_i/(a+n)$ or is assigned to a new table with probability $a/(a+n)$. The probability that the $n+1$'th diner sits at Table i is therefore

$$\begin{aligned} q &= \sum_{j=0}^{n} p(n_i = j)\frac{j}{(a+n)} = \frac{1}{(a+n)}\sum_{j=0}^{n} j\frac{n!}{j!(n-j)!}q_i^j(1-q_i)^{n-j} \\ &= \frac{nq_i}{(a+n)}\sum_{j=0}^{n}\frac{(n-1)!}{(j-1)!(n-1-(j-1)!}q_i^{j-1}(1-q_i)^{n-1-(j-1)} \\ &= \frac{nq_i}{(a+n)}\sum_{k=0}^{n-1}\frac{(n-1)!}{k!(n-1-k)!}q_i^k(1-q_i)^{n-1-k} \\ &= \frac{nq_i}{(a+n)} \to q_i \qquad \text{as } n \to \infty, \end{aligned}$$

so asymptotically there is a steady state distribution, as required. The exact distribution (a draw from a Dirichlet process) depends on the sequence of draws that are made, since previous events are reinforced by the algorithm. This same argument cannot be used to give the exact probability distribution (the proof that it is a DP is complicated), but it can be used to compute the expected number of diners (or the expected probability weight assigned to a given interval by the DP). This is the probability assigned to the interval by the base distribution. We will see the same feature in the stick-breaking construction in the next part of the question.

(b) The locations of the point masses, $\{U_l\}$, and their weights, determined from $\{V_l\}$, are independent. The probability that any one of the point masses is in the subset B_i is $H_0(B_i)$ since the point masses are drawn from H_0. Hence asymptotically the expected probability assigned to B_i is

$$\begin{aligned}
p(B_i) = H_0(B_i)\mathbb{E}\left(\sum_{l=1}^{\infty} p_l\right) &= H_0(B_i)\sum_{l=0}^{\infty}\frac{1}{1+M_H}\left(\frac{M_H}{1+M_H}\right)^l \\
&= H_0(B_i)\frac{1}{1+M_H}\left(1-\frac{M_H}{1+M_H}\right)^{-1} = H_0(B_i)
\end{aligned}$$

as required. the last line follows from summing the geometric progression using standard results. If the probabilities are drawn from a Dirichlet process, then the probability assigned to B_i and its complement $\bar{B}_i$ should be drawn from a Dirichlet distribution

$$\begin{aligned}
(p(B_i), p(\bar{B}_i)) &= (p(B_i), 1-p(B_i)) \\
&\sim \mathrm{Dir}(aH_0(B_i), aH_0(\bar{B}_i))) = \mathrm{Dir}(aH_0(B_i), a(1-H_0(B_i))).
\end{aligned}$$

The expectation value of a particular component, x_j, drawn from a Dirichlet distribution with parameter vector $\vec{\alpha}$ is $\alpha_j/\sum_k \alpha_k$ and in this case we deduce

$$\mathbb{E}(p(B_i)) = \frac{aH_0(B_i)}{aH_0(B_i) + a(1-H_0(B_i))} = H_0(B_i)$$

and so the result in this question (and the similar result that can be derived for the Chinese restaurant process) is as expected.