BIOST 527 - HOMEWORK 3

KATIE WOOD

Problem 1

Letting $p = 1$, we have

$$\text{MSE}(\hat{g}(x_0), g^*(x_0)) = \text{Bias}(\hat{g}(x_0))^2 + \text{Var}(\hat{g}(x_0))$$

by the Bias-Variance decomposition. Investigating first the bias, we find

$$\begin{aligned}\text{Bias}(\hat{g}(x_0)) &= \mathbb{E}(\hat{g}(x_0)) - g^*(x_0) \\ &= \mathbb{E}\left(\frac{1}{k}\sum_{j\in\mathcal{N}_k(x_0)} Y_j\right) - g^*(x_0)\end{aligned}$$

Plugging in our model, $Y_i = g^*(X_i) + \epsilon_i$, gives

$$\begin{aligned}&= \mathbb{E}\left(\frac{1}{k}\sum_{j\in\mathcal{N}_k(x_0)} (g^*(X_j) + \epsilon_j)\right) - g^*(x_0) \\ &= \frac{1}{k}\sum_{j\in\mathcal{N}_k(x_0)} \left(\mathbb{E}\left(g^*(X_j)\right) + \mathbb{E}\left(\epsilon_j\right)\right) - g^*(x_0) \\ &= \frac{1}{k}\sum_{j\in\mathcal{N}_k(x_0)} (g^*(X_j)) - g^*(x_0)\end{aligned}$$

since $\mathbb{E}(\epsilon_i) = 0$ by assumption. Using the hint, we note that kNN is a linear smoother and can be expressed in the form

$$\hat{g}(x_0) = \sum_i w_{0,i} Y_i$$

where

$$\sum_i w_{0,i} = 1$$

The weights $w_{0,i}$ are given by

$$w_{0,i} = \begin{cases} \frac{1}{k}, & i \in \mathcal{N}_k(x_0) \\ 0, & \text{otherwise} \end{cases}$$

Returning to our equation for the bias, we may now write

$$\begin{aligned}\text{Bias}(\hat{g}(x_0)) &= \frac{1}{k}\sum_{j\in\mathcal{N}_k(x_0)} (g^*(X_j)) - \frac{1}{k}\sum_{j\in\mathcal{N}_k(x_0)} (g^*(x_0)) \\ &= \frac{1}{k}\sum_{j\in\mathcal{N}_k(x_0)} (g^*(X_j) - g^*(x_0))\end{aligned}$$

Since g^* is Lipschitz continuous, we use the bound in (2) to obtain

$$\begin{aligned}\text{Bias}(\hat{g}(x_0)) &\leq \frac{1}{k} \sum_{j \in \mathcal{N}_k(x_0)} B\,|X_j - x_0| \\ &= B \sup_{j \in \mathcal{N}_k(x_0)} (|X_j - x_0|) \sum_{j \in \mathcal{N}_k(x_0)} \frac{1}{k}\end{aligned}$$

Owing to the uniform fixed design, $\sup_{j \in \mathcal{N}_k(x_0)} (|X_j - x_0|) = \frac{k}{2n}$. Thus,

$$\begin{aligned}\text{Bias}(\hat{g}(x_0)) &\leq B\frac{k}{2n} \sum_{j \in \mathcal{N}_k(x_0)} \frac{1}{k} \\ &= B\frac{k}{2n} \cdot k \cdot \frac{1}{k} \\ &= B\frac{k}{2n} \\ &= C\frac{k}{n}\end{aligned}$$

We have found the bias for $p = 1$. We proceed to find the variance:

$$\begin{aligned}\text{Var}(\hat{g}(x_0)) &= \text{Var}\left(\frac{1}{k} \sum_{j \in \mathcal{N}_k(x_0)} Y_j\right) \\ &= \frac{1}{k^2} \sum_{j \in \mathcal{N}_k(x_0)} \text{Var}\left(g^*(x_0) + \epsilon_i\right) \\ &= \frac{1}{k^2} \sum_{j \in \mathcal{N}_k(x_0)} \text{Var}\left(\epsilon_i\right)\end{aligned}$$

Since $\text{Var}\left(\epsilon_i\right) = \sigma^2$ by assumption, we have

$$\begin{aligned}&= \frac{1}{k^2} \sum_{j \in \mathcal{N}_k(x_0)} \sigma^2 \\ &= \frac{1}{k^2} \cdot k \cdot \sigma^2 \\ &= \frac{\sigma^2}{k}\end{aligned}$$

Combining this result with the bias, we have

$$\text{MSE}(\hat{g}(x_0), g^*(x_0)) \leq C^2\frac{k^2}{n^2} + \frac{\sigma^2}{k}$$

Differentiating the expression on the right-hand side with respect to k, we find

$$\frac{\partial}{\partial k}\left(C^2\frac{k^2}{n^2} + \frac{\sigma^2}{k}\right) = 2k\frac{C^2}{n^2} - \frac{\sigma^2}{k^2}$$

Then if we set

$$2k\frac{C^2}{n^2} - \frac{\sigma^2}{k^2} = 0$$

we find that the optimal k^* is

$$k^* = C^* n^{2/3}$$

With this k^*, the MSE becomes

$$\mathrm{MSE}(\hat{g}(x_0), g^*(x_0)) \leq C^{**} n^{-2/3}$$

so the rate of convergence is $\wr_p(n^{-2/3})$.

Next, we let $p = 2$ and investigate again the bias. The proof is identical until the Lipschitz bound:

$$\begin{aligned}
\mathrm{Bias}(\hat{g}(x_0)) &\leq \frac{1}{k} \sum_{j \in \mathcal{N}_k(x_0)} B\sqrt{|X_{j1} - x_{01}|^2 + |X_{j2} - x_{02}|^2} \\
&= B \sup_{j \in \mathcal{N}_k(x_0)} \sqrt{|X_{j1} - x_{01}|^2 + |X_{j2} - x_{02}|^2} \sum_{j \in \mathcal{N}_k(x_0)} \frac{1}{k} \\
&= B \sup_{j \in \mathcal{N}_k(x_0)} \sqrt{|X_{j1} - x_{01}|^2 + |X_{j2} - x_{02}|^2} \\
&\leq B \sup_{j \in \mathcal{N}_k(x_0)} \left(|X_{j1} - x_{01}|^2 + |X_{j2} - x_{02}|^2\right) \\
&= B\left(\left(\frac{k}{2n}\right)^2 + \left(\frac{k}{2n}\right)^2\right) \\
&= C\frac{k^2}{n^2}
\end{aligned}$$

Now we find the variance. It is identical to the $p = 1$ case:

$$\mathrm{Var}(\hat{g}(x_0)) = \frac{\sigma^2}{k}$$

Combining these two results, we find:

$$\mathrm{MSE}(\hat{g}(x_0), g^*(x_0)) \leq C^2\frac{k^4}{n^4} + \frac{\sigma^2}{k}$$

Differentiating the expression on the right-hand side with respect to k, we find

$$\frac{\partial}{\partial k}\left(C^2\frac{k^4}{n^4} + \frac{\sigma^2}{k}\right) = 4k^3\frac{C^2}{n^4} - \frac{\sigma^2}{k^2}$$

Then if we set

$$4k^3\frac{C^2}{n^4} - \frac{\sigma^2}{k^2} = 0$$

we find that the optimal k^* is

$$k^* = C^* n^{4/5}$$

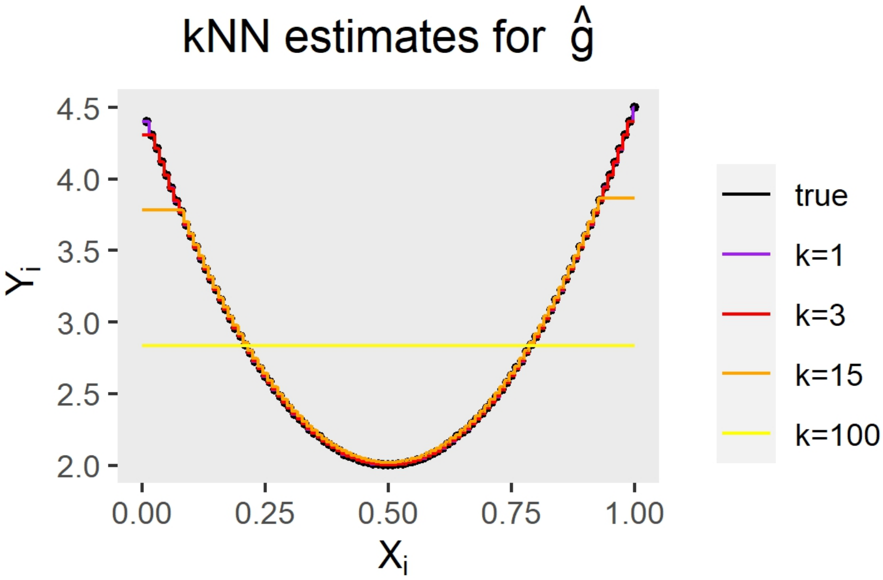
With this k^*, the MSE becomes

$$\mathrm{MSE}(\hat{g}(x_0), g^*(x_0)) \leq C^{**} n^{-4/5}$$

so the rate of convergence is $\wr_p(n^{-4/5})$.

Comparing the results for $p = 1$ and $p = 2$, we find that the rate of convergence for $p = 2$ is faster. Intuitively, this may be because the distance to the furthest nearest neighbor increases more slowly for $p = 2$, so the estimate provided by the k nearest neighbors in 2 dimensions is more accurate.

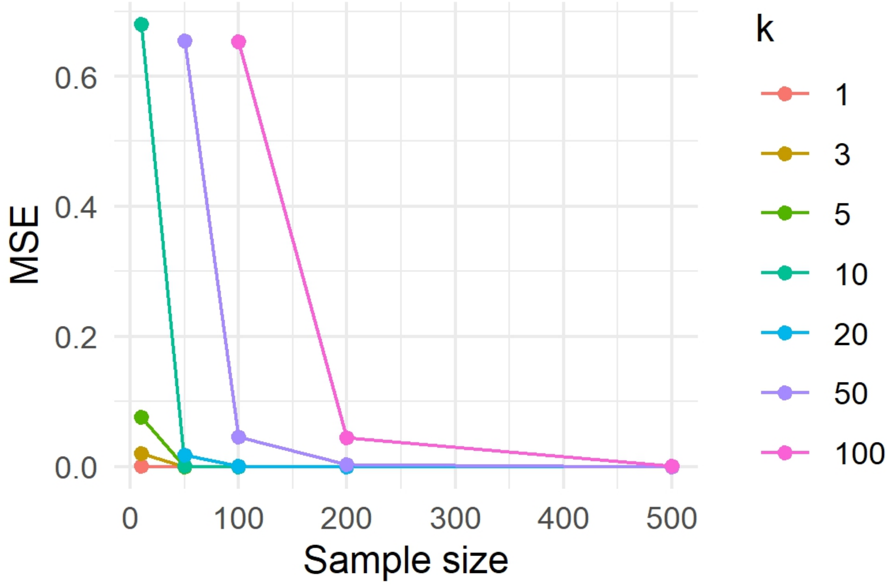
Problem 2



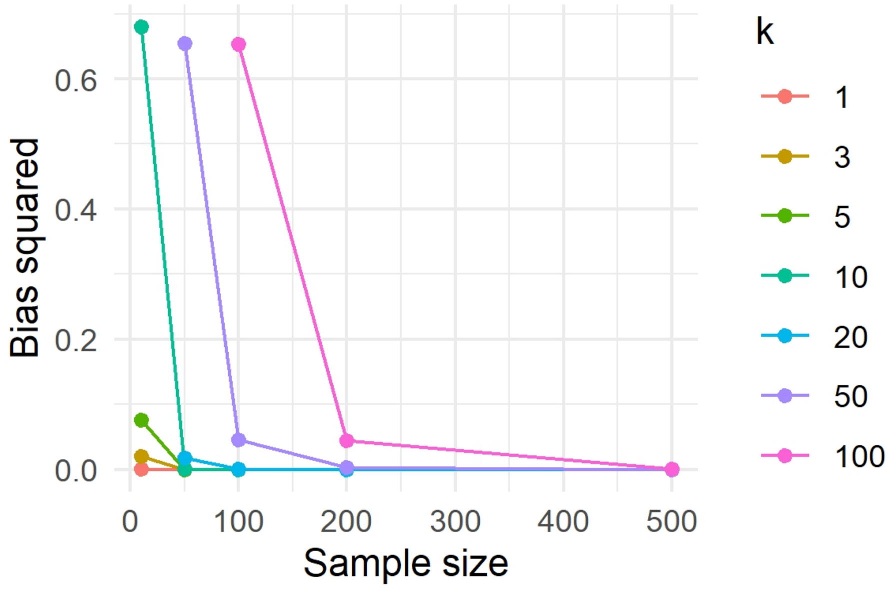
We observe that the bias of the kNN estimator increases with k. This effect is most noticeable at the boundaries, but is also present near the local minimum at the center of the plot. When all $k = 100$ nearest neighbors are considered, the estimate reduces to a simple average over the domain. Lastly, on a dense enough grid of points (here, 1000 points were used) all of the kNN estimators show small "wiggles" due to the small gap between consecutive X_i.

Sample size	k	MSE	Bias squared	Variance
10	1	6.319798e-04	6.260697e-04	5.915976e-06
10	3	2.006160e-02	2.005966e-02	1.935755e-06
10	5	7.561438e-02	7.561328e-02	1.106429e-06
10	10	6.805862e-01	6.805856e-01	5.971704e-07
50	1	8.557124e-05	7.949653e-05	6.080792e-06
50	3	1.900107e-04	1.879276e-04	2.085217e-06
50	5	2.100109e-06	9.047941e-07	1.196512e-06
50	10	1.092080e-03	1.091496e-03	5.844116e-07
50	20	1.769661e-02	1.769630e-02	3.055736e-07
50	50	6.545041e-01	6.545040e-01	1.145761e-07
100	1	6.511704e-06	1.953971e-08	6.498663e-06
100	3	2.560648e-06	4.645466e-07	2.098200e-06
100	5	5.164771e-06	3.947087e-06	1.218903e-06
100	10	1.281291e-05	1.218957e-05	6.239599e-07
100	20	1.483131e-03	1.482804e-03	3.272566e-07
100	50	4.559357e-02	4.559344e-02	1.227612e-07
100	100	6.536979e-01	6.536979e-01	5.985449e-08
200	1	6.621591e-06	7.894136e-09	6.620317e-06
200	3	2.362963e-06	3.210684e-08	2.333189e-06
200	5	1.589574e-06	2.306486e-07	1.360286e-06
200	10	7.764520e-07	1.476779e-07	6.294035e-07
200	20	3.485307e-05	3.454710e-05	3.062749e-07
200	50	2.984768e-03	2.984642e-03	1.255681e-07
200	100	4.446850e-02	4.446844e-02	6.216798e-08
500	1	5.897302e-06	5.654153e-09	5.897546e-06
500	3	2.063244e-06	6.459351e-09	2.058843e-06
500	5	1.282841e-06	9.495363e-09	1.274620e-06
500	10	2.391114e-06	1.733037e-06	6.587358e-07
500	20	4.160008e-07	1.000397e-07	3.162774e-07
500	50	5.369290e-05	5.357489e-05	1.181318e-07
500	100	1.178000e-03	1.177940e-03	6.008199e-08

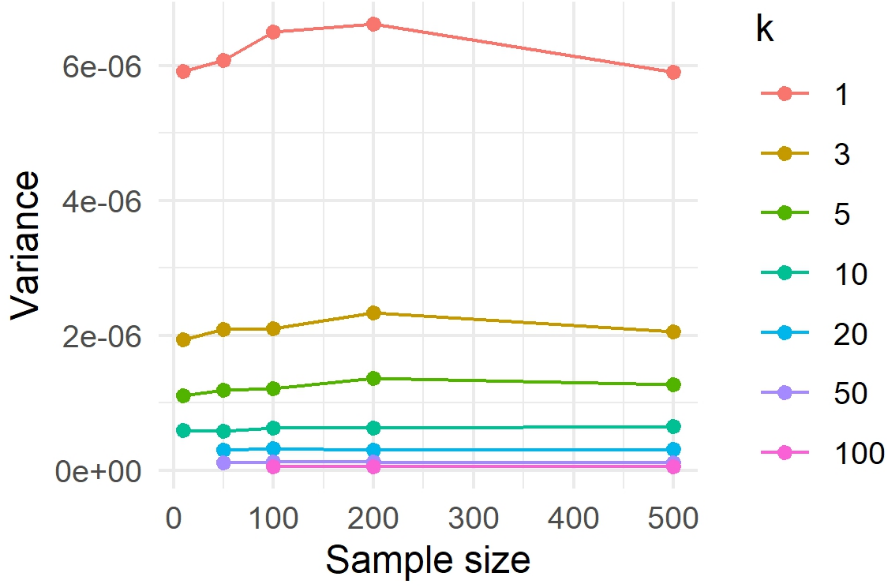
TABLE 1. Above, we show the approximate bias squared, variance, and MSE at $x_0 = 0.45$. Each row represents 1000 predictions generated from our model.



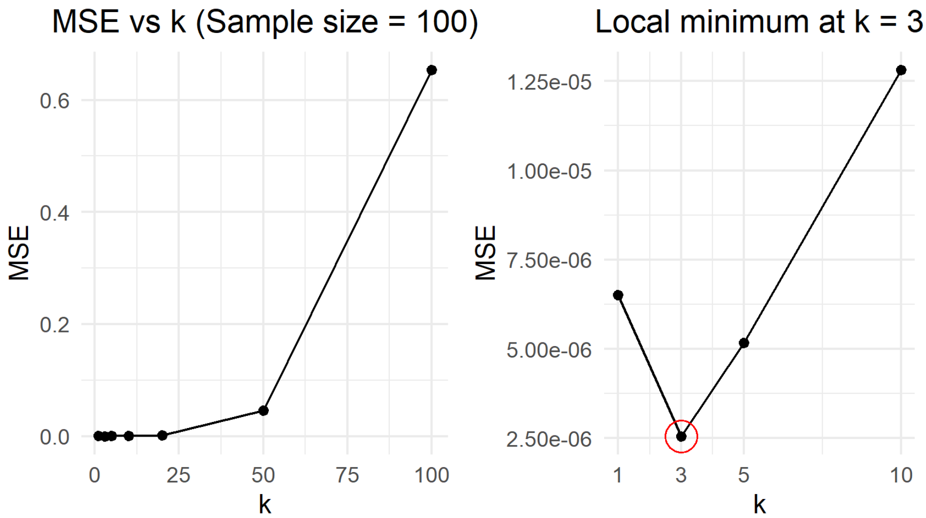
Above we see that for all values of k, the MSE tends to 0 as n becomes large (this is the expected behaviour). Additionally, the MSE at any given sample size appears to be larger for larger k, though it is difficult to tell this exactly for $k = 1, 3, 5$.



As expected, we observe that bias (at a given sample size) increases with increasing k. Also, since the bias squared is much larger than the variance in this problem, the MSE is approximately equal to the bias squared. This explains why the above plot nearly replicates that of the MSE.



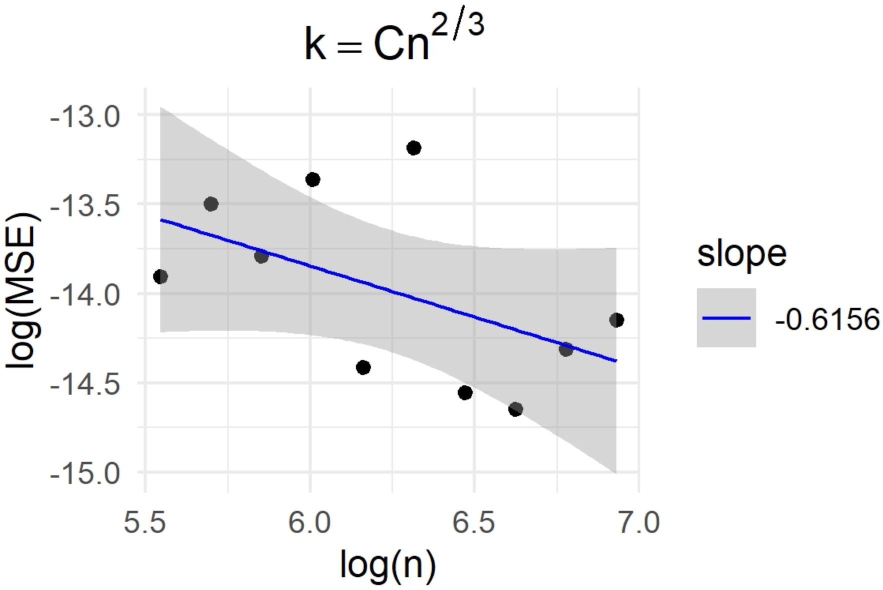
As expected, we find that the variance decreases with increasing k, as shown in the above plot. Also, the variance is largely unaffected by the sample size, which is consistent with our finding in Problem 1 that the variance term equals σ^2/k (i.e., does not depend on n). It is not easy to tell, since the k values are not evenly spaced, but it also appears that the variance is decreasing at a decreasing rate, consistent with the $1/k$ behaviour.



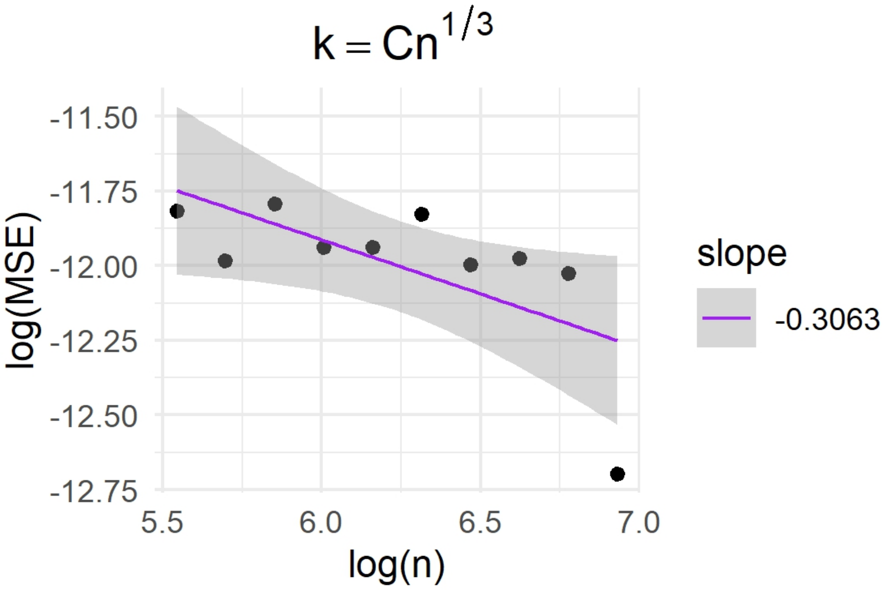
Above left, we observe what appears to be MSE increasing (monotonically) with k. However, zooming in to $k < 10$ enables us to detect the local minimum at $k = 3$. This suggests that $k = 3$ is close to the optimal k for this problem.

Sample size	MSE
256	9.155436e-07
298	1.370692e-06
348	1.027616e-06
406	1.574341e-06
474	5.499503e-07
552	1.877884e-06
645	4.778374e-07
752	4.359224e-07
877	6.090865e-07
1024	7.162666e-07

TABLE 2. Approximate MSE at $x_0 = 0.45$ with $k = 0.2n^{2/3}$

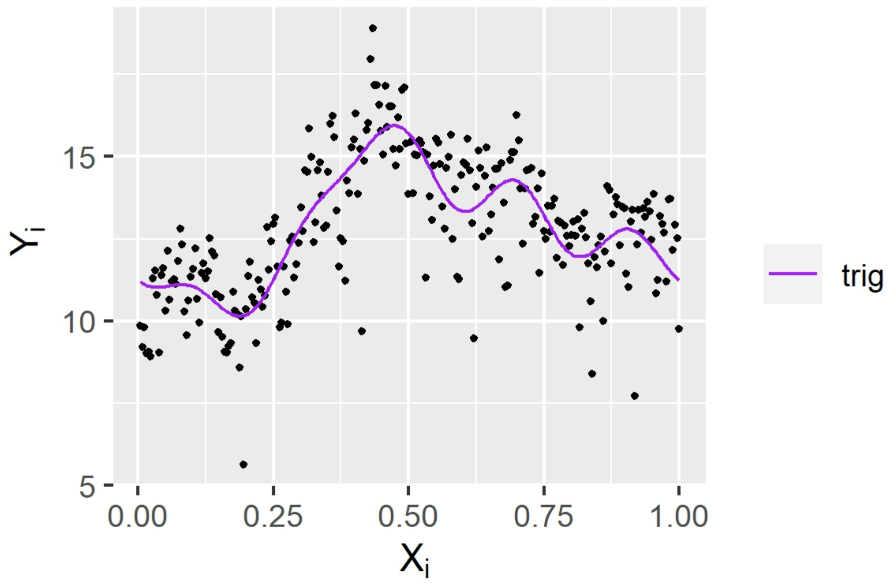


Considering the theoretical results from Problem 1, we expect the above plot to look like a straight line with slope equal to $-2/3$ (since $\log(n^{-2/3}) = -(2/3)\log(n)$. Indeed, when we estimate the slope of the line of best fit through the points, it is approximately -0.62. However, there is a fair bit of noise, with 4 out of 10 data points falling more than one standard error away from the line of best fit. Thus, the result above is encouraging, but should be taken with a grain of salt.

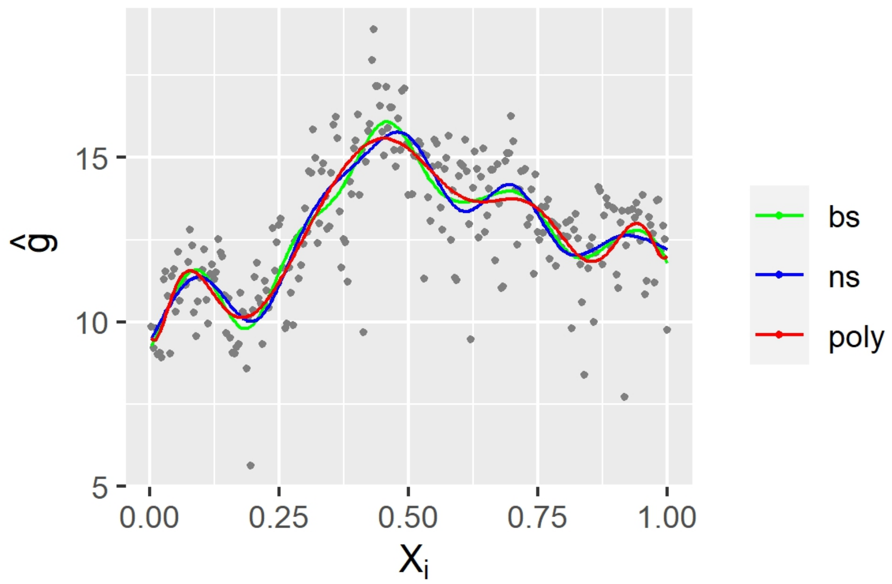


Now, we expect the variance to dominate the asymptotic behaviour of the MSE (since with this choice of k, the variance decays like $n^{-1/3}$, while the bias squared decays like $n^{-4/3}$). Thus, the overall rate of convergence of the MSE becomes $\wr_p(n^{-1/3})$. We then expect that the above plot will show a linear relationship with slope $-1/3$. Indeed, the slope of the line of best fit we find to be approximately -0.31. Like the $k = Cn^{2/3}$ plot, however, there is noise, albeit less so, that cautions us to view this result with a grain of salt.

Problem 3



Above, we visualize the orthogonal projection estimator from the trigonometric basis, which we have applied to the data from the first log-periodogram. By visual inspection, we determined that a truncation parameter of $M = 10$ gave a close to optimal fit, without over-fitting. We observe that $\hat{g}^{OP}$ captures what appear to be three peaks in the log-periodogram (near $x = 0.5, 0.75, 1$). At the same time, the basis struggles a bit more to capture the fourth (left-most) peak near $x = 0$. This is likely because this left-most peak represents a higher frequency than the current truncation parameter allows us to capture.



We now compare the results obtained from the polynomial, natural splines, and B-splines bases. Using visual inspection, we determined truncation parameters of $M = 13$, 10, and 13 respectively, slightly higher than for the trigonometric basis. All three bases appear to perform similarly, however the most satisfactory fit seems to be that of the natural splines. The natural splines fit clearly captures all four peaks in the log-periodogram, with little noise interfering in between peaks. The B-splines basis also captures the four peaks, but displays somewhat more wiggles in between. The polynomial basis does the least well to capture the four peaks, particularly struggling with the third one (near $x = 0.75$). This could be because the polynomial basis requires higher degree to capture repeated changes of concavity, particularly over a small interval.