Problem sheet 2

1. Consider data \((x_i, y_i), i = 1, 2, \ldots, n\) and, for \(x \in \mathbb{R}\), let \(\hat{m}(x)\) be a local linear smoother with kernel \(K(x)\). Let \(\theta_x = (\theta_{x,1}, \theta_{x,2})\) be the parameters of a linear regression
\[
f_x(z) = \theta_{x,1} + \theta_{x,2}(z - x)
\]
fitted to the data around \(x\) by weighted linear regression with weight function \(w_i(x) = K((x - x_i)/h)\). Here \(h > 0\) is a fixed smoothing parameter.

Show that
\[
\hat{m}(x) = \sum_{j=1}^{n} \frac{\tilde{w}_j(x)}{\sum_{k=1}^{n} \tilde{w}_k(x)} y_j
\]
where
\[
\tilde{w}_j(x) = w_j(x)(T_{n,2}(x) - (x_j - x)T_{n,1}(x))
\]
and
\[
T_{n,1} = \sum_{k=1}^{n} w_k(x)(x_k - x), \quad T_{n,2} = \sum_{k=1}^{n} w_k(x)(x_k - x)^2
\]
so that it is not necessary to fit the regression separately at each point.

2. We are given \(n + 1\) data points \((x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\) and we want to find a smooth function \(y = f(x)\) to represent the trend in these points. We define a smooth function to be one that is continuous and has continuous first and second derivatives.

Suppose that \(\lambda > 0\) is a fixed constant. If \(f\) is smooth and minimises
\[
\sum_{j=0}^{n} (y_j - f(x_j))^2 + \lambda \int_{-\infty}^{\infty} [f''(x)]^2 dx, \quad \lambda > 0
\]
then it is a cubic spline.

(a) Suppose there are just three data points: \((-1, y_0), (0, y_1), (1, y_2)\). Find the best cubic spline according to the criterion in (a). Describe how the fitted curve depends on \(\lambda\).

(b) Suppose that the \(y\)-values are known to be independent and normally distributed with variance \(\sigma^2\). How should that information influence your choice of the tuning parameter \(\lambda\)? Suppose that the variance \(\sigma^2\) is unknown. How can \(\lambda\) be chosen now?

3. We are given \(n\) data points \((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\) with equally spaced \(x\)-values, i.e. \(x_1 = 1, x_2 = 2, \ldots, x_n = n\). The proposal is to smooth the data by fitting a local linear regression with window count \(m = 3\). In other words, the smoothed value \(\hat{y}_k\) will be obtained by fitting a linear regression through the points \((k - 1, y_{k-1}), (k, y_k), (k + 1, y_{k+1})\) for \(k = 2, 3, \ldots, n - 1\). At the endpoints, the regression through \((1, y_1), (2, y_2), (3, y_3)\) is used to get \(\hat{y}_1 = \hat{m}(x_1)\), and similarly for \(\hat{y}_n = \hat{m}(x_n)\). At the conclusion of this calculation, the vector of smoothed values \(\hat{y} = (\hat{m}(x_1), \ldots, \hat{m}(x_n))\), can be expressed as a linear transformation of \(y = (y_1, \ldots, y_n)\), the original \(Y\)-vector, i.e. \(\hat{Y} = SY\).
(a) What are the elements of the matrix $S$?

(b) Suppose that the $y$-values are independent each with the same variance $\sigma^2$. What is the variance of $\hat{y}_i$, $i = 1, 2, \ldots, n$?

(c) Give the form of the leave-one-out cross-validation estimator $\hat{\text{MSE}}(h)$ of the mean squared error

$$E((Y - \hat{m}_h(x))^2),$$

where the expectation is with respect to new observations $Y$ and $x$ chosen randomly within $x_1, \ldots, x_n$. Give the form of the generalized cross-validation version (GCV) and comment on the difference between the two for large sample sizes $n$.

(d) What would happen if you smoothed the smoothed values, i.e. what would a plot of the $k$th row of $S^2$ and the $k$th row of $S^3$ look like?

(e) How is this smoothing method related to ‘kernel smoothing’?

4. Suppose that for a sample of size $n$ there are observations $(x_i, Y_i)$, $i = 1, \ldots, n$ where $x_i = i$ are the design points and $Y_i$ are the corresponding measured response values. A kernel smoother is applied to the data with kernel $K$ of the form

$$K(x, \tilde{x}) = \max\left\{0, 1 - \frac{(x - \tilde{x})^2}{h}\right\}$$

where $h$ is the bandwidth. The fitted values are denoted by $\hat{Y}_i, i = 1, \ldots, n$. Let $Y$ and $\hat{Y}$ represent the vectors of response values and fitted values respectively. The kernel smoother is a linear smoother and thus $\hat{Y} = SY$ for some $n \times n$ matrix $S$.

(a) Suppose that $n = 7$. Find the smoothing matrix $S$ for the bandwidth $h = 2$. What are the degrees of freedom $\nu = \text{trace}(S)$ for a bandwidth of 2? What are the degrees of freedom for sample size $n > 7$ for a bandwidth of 2? What are the degrees of freedom for sample size $n > 7$ for a bandwidth of 3?

(b) Suppose that for a sample of size $n = 1000$, the residual sum of squares is $RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$. Using a bandwidth of 2, $RSS = 950$, and for a bandwidth of 3, $RSS = 1020$. Using this information, how would you decide which of the two bandwidths provides the better fit to the data?

(c) Suppose that the true model for the underlying data is $Y_i = m(x_i) + \epsilon_i$ where $m(x) = \alpha x^2 + \beta x + \gamma$ and errors $\epsilon_i, i = 1, \ldots, n$ are i.i.d with $\epsilon_i \sim N(0, \sigma^2)$ for some $\sigma^2 > 0$. Denote the estimated values of a kernel smoother by $\hat{Y}_i = \hat{m}(x_i)$. Define and calculate the bias and variance of the kernel estimator at the design points $x_i = i$, where $2 < i < n - 2$ for a bandwidth $h = 2$. What is the expected quadratic loss, $E[(\hat{m}(x_i) - m(x_i))^2]$, at these design points?