Lecturer: Geoff Nicholls

Lecture 5: Smoothing and non-parametric regression

Notes and Problem sheets are available at

http://www.stats.ox.ac.uk/~nicholls/CompStats/(L1-7)
Degrees of freedom of a linear smoother

The residual sum of squares, $RSS = \|Y - SY\|^2$ is natural measure of the goodness of fit, and the basis for a simple estimate of the variance of $Y - m(x)$, which is often of interest.

The degrees-of-freedom $df$ of a smoother $S$ is also of interest. We will see a bias-variance trade off reflected in the balance between smoother complexity, $df$, and goodness of fit, $RSS$. The two are linked in the variance.

Since $E(Y) = m(x)$, the variance-covariance matrix for $Y$ is

$$\text{var}(Y) = E((Y - m(x))(Y - m(x))^T)$$

and I assume $\text{var}(Y) = \sigma^2 I_n$ with $I_n$ the $n \times n$ identity.
First, if $v$ is a random vector, and $A$ a commensurate matrix,

$$E(v^T Av) = E(v)^T AE(v) + \text{tr}(A \text{var}(v)),$$

(note $E(v^T Av) = E(\text{tr}(v^T Av))$ and permute the trace). Hence

$$E(RSS) = E(\| (I - S) Y \|^2)$$

$$= E(Y^T (I - S)^T (I - S) Y))$$

$$= m(x)^T (I - S)^T (I - S) m(x) + \sigma^2 \text{tr}( (I - S)^T (I - S))$$

$$= \| (I - S) m(x) \|^2 + \sigma^2 (\text{tr}(S^T S) - 2 \text{tr}(S) + n)$$

$$= \| (I - S) m(x) \|^2 + \sigma^2 (n - df).$$

where I am defining the degrees of freedom of the smoother as

$$df = 2 \text{tr}(S) - \text{tr}(S^T S).$$

As $df$ goes up, $RSS$ goes down.
Why this definition for $df$? First if $S$ is just linear regression then $S^T S = S$ so $df = \text{tr}(S)$ and so we have $p$ DOF as expected. Also, for regression, $m(x) = X\theta$ and $(I - S)X\theta = 0$ and so $\|(I - S)m(x)\| = 0$. The result above reduces to

$$\sigma^2 = E(RSS)/(n - df)$$

with $df = p$, a familiar result.

The bias in our estimates $m(x)$ is

$$E(\hat{Y} - m(x)) = Sm(x) - m(x),$$

so $\|(I - S)m(x)\|^2$ is the sum of the squared biases. The bias $\|(I - S)m(x)\|$ of our smoother will in general be small compared to $\sigma^2$ so

$$RSS/(n - df)$$

will estimate the variance of $Y$ with little bias.
Exercise: the mean square error at $x$ is

$$MSE(\hat{m}(x)) = E((\hat{m}(x) - m(x))^2).$$

The summed MSE, MSSE say, at $x_1, \ldots, x_n$ is

$$MSSE(\hat{m}(x)) = \sum_{i=1}^{n} E((\hat{m}(x_i) - m(x_i))^2).$$

Show that if $\text{var}(Y) = \sigma^2 I$ then

$$MSSE(\hat{m}(x)) = \|(I - S)m(x)\|^2 + \sigma^2 \text{tr}(S^T S).$$
The linear smoother Zoo

Two basic approaches: local regression estimators, and penalised estimators.

Local regression includes Kernel estimators. The kernel restricts the points that contribute to estimation of $\hat{Y}_i$ to just those with $x$-values close to $x_i$.

Penalised estimators take a large set of basis functions (so $p$ is typically very large). They minimise the residual sum of squares, adding a penalty term to avoid over-fitting.
Local regression estimators: Kernel estimators

Let \( K(x) \) be a continuous bounded symmetric probability density.

Example: Epanechnikov Kernel, \( K(x) = 3(1 - x^2)\mathbb{1}(|x| < 1)/4 \).
Example: Normal Kernel, \( K(x) = \exp(-x^2/2)/\sqrt{2\pi} \).

Let \( w_{i,j} = K((x_i - x_j)/h) \) (controls contribution if \( Y_j \) to \( \hat{Y}_i \)).
Let \( w_i(x) = K((x - x_i)/h) \) at generic \( x \).

The bandwidth \( h \) sets the smoothing scale.
Example: the Nadaraya-Watson kernel estimator,

\[ \hat{m}(x) = \sum_{j=1}^{n} \frac{w_j(x)}{\sum_{k=1}^{n} w_k(x)} Y_j. \]

In this case

\[ \hat{m}(x_i) = \sum_{j=1}^{n} \frac{w_{i,j}}{\sum_{k=1}^{n} w_{i,k}} Y_j \]

If \( S_{i,j} = \frac{w_{i,j}}{\sum_{k=1}^{n} w_{i,k}} \) for \( i, j \) in \( 1, \ldots, n \) then \( S \) is a function of \( X \) alone, and \( \hat{Y} = SY \), so we have a linear smoother.

Weights are normalised. If \( Y_i = c, \ i = 1, \ldots, n \) then \( \hat{m}(x) = c \).

It follows that \( \sum_j S_{i,j} = 1 \).
The NW kernel estimator solves a least squares problem.

Claim: If at each $x$, $\hat{m}(x)$ satisfies

$$\hat{m}(x) = \arg \min_a \sum_{i=1}^n w_i(x)(Y_i - a)^2$$

then

$$\hat{m}(x) = \sum_{j=1}^n \frac{w_j(x)}{\sum_{k=1}^n w_k(x)} Y_j.$$  

Proof: solve $dC/da = 0$ for $a$ with $C(a) = \sum_{i=1}^n w_i(x)(Y_i - a)^2$.

Notice we solve a least squares problem at each $x$-value.
Example: smoothing the CMB spectral power distribution. 
\( \hat{m}(x) \) depends on bandwidth \( h \) as 
\[ w_i(x) = K((x - x_i)/h). \]
Local weighted regression from constant to linear local smoothers
Weighted regression

Claim: If $Y \sim N(X \theta; \Sigma)$, with $\Sigma^{-1} = W$ known, then $\hat{\theta} = (X^T W X)^{-1} X^T W Y$ and $\hat{Y} = SY$ with

$$S = X^T (X^T W X)^{-1} X^T W.$$

Proof: factorise $W = L^T L$. The least-squares estimate (LSE) for $\theta$ (the MLE here) minimises

$$(Y - X \theta)^T W (Y - X \theta) = ||LY - LX \theta||^2.$$

This is just regression with $Y$ replaced by $LY$ and $X$ by $LX$ so the LSE is

$$\hat{\theta} = ((LX)^T L X)^{-1} (LX)^T (LY),$$

and so $\hat{\theta} = (X^T W X)^{-1} X^T W Y$ etc.
Local Weighted regression

Regressing $Y$ against $x$ using local weighted data around $x$. Let $X_x = [X_{x,i}]_{i=1}^n$ be the matrix with rows $X_{x,i} = (1, x_i - x)$. Let

$$
\hat{\theta}_x = \arg \min_{\theta_x} \sum_{i=1}^n \left(y_i - \theta_{x,1} - \theta_{x,2}(x_i - x)\right)^2 K((x_i - x)/h)
$$

If $W_x = \text{diag}(w_1(x), \ldots, w_n(x))$ with $w_i(x) = K((x_i - x)/h)$,

$$
\hat{\theta}_x = \arg \min_{\theta_x} (Y - X_x \theta_x)^T W_x (Y - X_x \theta_x).
$$

This weighted regression is minimised by $\hat{\theta}_x = P_x Y$ where

$$
P_x = (X_x^T W_x X_x)^{-1} X_x^T W_x.
$$
This gives a regression line

\[ \hat{Y}_{x,z} = \hat{\theta}_{x,1} + \hat{\theta}_{x,2}(z - x) \]

that is a function of \( z \) fitting the data around \( x \).

The fitted value at \( x \) itself is at \( z = x \), yielding the intercept,

\[ \hat{m}(x) = \hat{\theta}_{x,1}. \]

That is

\[ \hat{m}(x) = (1, 0)P_x Y \]

or in other words the inner product of the top row \([P_x]_1,:) \) with \( Y \). At observation points

\[ \hat{Y}_i = [P_{x_i}]_1,:)Y \]

so \( \hat{Y} = SY \) with \( S_{i,:} = [P_{x_i}]_1,:) \).
The idea can be generalised directly to local polynomial regressions, taking for example

\[ X_{x,i} = (1, x_i - x, (x_i - x)^2/2, \ldots, (x_i - x)^p/p!) \]

and

\[ \hat{\theta}_x = \arg\min_{\theta_x} \sum_{i=1}^{n} (y_i - X_{x,i}\theta_x)^2 K((x_i - x)/h) \]

The local design matrix at \( x \) is \( X_x = [X_{x,i}]_{i=1}^{n} \), now \( n \times (p+1) \). The calculation is unchanged. Again, the local prediction is \( \hat{m}(x) = (1, 0, \ldots, 0)\hat{\theta}_x = \hat{\theta}_{x,1} \), since \( (1, 0, \ldots, 0) \) is the covariate vector for a prediction at \( x \), so

\[ \hat{m}(x) = (1, 0, \ldots, 0)P_xY \]

with \( P_x = (X_x^TW_xX_x)^{-1}X_x^TW_x \) as before and \( W_x \) as above.
Remark 1). If $Y_i = 1$, $i = 1, \ldots, n$ then $\hat{m}(x) = 1$ at all $x^*$ so we smooth constant data to a constant function.

Define $\ell_j(x) = [Px]_{1,j}$. Since $\hat{\theta}_{x,1} = \sum_{j=1}^{n} [Px]_{1,j}Y_j$ we have

$$\sum_{j=1}^{n} \ell_j(x) = 1$$

for local polynomial smoothers.

Remark 2) Convergence [outside syllabus]. If $\int |K(x)| \, dx < \infty$, $\lim_{|x| \to \infty} xK(x) = 0$, and $E(Y^2) < \infty$ with $h_n \to 0$ and $nh_n \to \infty$ when $n \to \infty$,

$$\hat{m}(x) \xrightarrow{P} m(x)$$

[Hardle, "Applied Non-Parametric Regression", CUP, (1990)]

*local residuals are zero when $\hat{\theta} = (1, 0, \ldots, 0)$ and $\hat{m}(x) = \hat{\theta}_{x,1}$
Boundary bias: How does the bias depend on the data design over $x$? If $\ell_i(x) = [P_x]_{1,i}$ then

$$\hat{m}(x) = \sum_i \ell_i(x)Y_i$$

with $E(Y_i) = m(x_i)$ and $m(x)$ smooth (say). Taylor expanding,

$$m(x_i) = m(x) + m^{(1)}(x)(x_i - x) + \frac{m^{(2)}(x)}{2}(x_i - x)^2 + \ldots,$$

about $x$, and noting $\sum_i \ell_i = 1$, we have

$$E(\hat{m}(x)) = \sum_i \ell_i(x)m(x_i)$$

$$= m(x) + m^{(1)}(x)\sum_i (x_i - x)\ell_i(x) +$$

$$\frac{m^{(2)}(x)}{2} \sum_i (x_i - x)^2 \ell_i(x) + \ldots$$

Consider $|x_i - x|$ of $O(h)$ as $K((x - x_i)/h) \simeq 0$ for $|x_i - x| \gg h$. 
The bias, $b(x) = E(\hat{m}(x)) - m(x)$, in the NW smoother is

$$b(x) = m^{(1)}(x) \sum_i (x_i - x) \ell_i(x) + \frac{m^{(2)}(x)}{2} \sum_i (x_i - x)^2 \ell_i(x) + R.$$  

The first term is

$$m'(x) \sum_i \frac{(x_i - x) K \left(\frac{x-x_i}{h}\right)}{\sum_k K \left(\frac{x-x_k}{h}\right)}$$

This is zero if $x_i$ are symmetric around $x$ but not zero at the boundary (all $x_i - x$ same sign).

Suppose $m(x)$ was actually linear. Despite using constant-only local linear regression, NW has bias $b(x) \simeq 0$ at interior point $x$. 
Claim: let $a = (a_1, ..., a_p)$. In LP regression of order $p$,

$$\sum_i \ell_i(x)[a_1(x_i - x) + \ldots + \frac{a_p}{p!}(x_i - x)^p] = 0$$

for all $a \in \mathbb{R}^p$. Proof: if $m(x)$ is polynomial order $p$ (same order as fit) and we observe $p + 1$ noise-free data points then $\hat{m}(x) = m(x)$. Then $m(x) = E(\hat{m}(x)) = \sum_i \ell_i(x)m(x_i)$ gives

$$\sum_i \ell_i(x)(m(x_i) - m(x)) = 0,$$

using $\sum_i \ell_i = 1$. Since $m(x)$ is polynomial order $p$,

$$m(x_i) = m(x) + m^{(1)}(x)(x_i - x) + \ldots + \frac{m^{(p)}}{p!}(x_i - x)^p,$$

we have, for all choices of the coefficients $m^{(j)}(x), j = 1, ..., p$,

$$\sum_i \ell_i(x)[m^{(1)}(x)(x_i - x) + \ldots + \frac{m^{(p)}(x)}{p!}(x_i - x)^p] = 0.$$

as $\ell_i(x), i = 1, ..., n$ depends on $x$ and $x_i, i = 1, ..., n$, not $m(x)$. 
Claim: in LP regression of order $p$ the bias is

$$E(\hat{m}(x)) - m(x) = \frac{m^{(p+1)}(x)}{(p+1)!} \sum_j (x_j - x)^{p+1} \ell_j(x) + R$$

with $R$ dominated by terms of order $h^{p+2}$.

Proof: in general $m(x)$ is not polynomial of finite order so

$$m(x_i) = m(x) + m^{(1)}(x)(x_i - x) + \ldots + \frac{m^{(p)}(x)}{p!}(x_i - x)^p$$

$$+ \frac{m^{(p+1)}(x)}{(p+1)!}(x_i - x)^{p+1} + \ldots,$$

and hence, using the previous result,

$$\sum_j \ell_j(x)m(x_j) = m(x) + \frac{m^{(p+1)}(x)}{(p+1)!} \sum_j (x_j - x)^{p+1} \ell_j(x) + \ldots$$

The claim follows as $E(\hat{m}(x)) = \sum_j \ell_j(x)m(x_j)$. 
Why not choose \( p = 20\)?

If \( Y_i = m(x) + \sigma \varepsilon_i \) with \( \varepsilon_i \sim \mathcal{N}(0, 1) \), the variance of the linear smoother, \( \hat{m}(x) = \sum_j \ell_j(x) Y_i \), is,

\[
\text{Var}(\hat{m}(x)) = \sigma^2 \sum_j \ell_j^2(x) = \sigma^2 \| \ell(x) \|_2^2
\]

and \( \| \ell(x) \|_2 \) tends to be large if \( p \) is large. In practice, \( p = 1 \) is a good choice.