1 Basic concept

1.1 Bias-variance decomposition

Define $e(x) = f(x) - \hat{f}(x)$ and $p(e)$ as its pdf.

$$\text{MSE} = E(e^2) = \int e^2 pde = (\int epde)^2 + (\int e^2 pde - (\int epde)^2) = E(e)^2 + Var(e) \tag{1}$$

1.2 A simple example

(A) Y follows binomal distribution $B(n, \pi_h)$. $E(Y) = n\pi_h$, $Var(Y) = n\pi_h(1 - \pi_h)$. One can estimate $f(0)$ by

$$f(0)_{\text{estimate}} = \frac{Y}{nh} \tag{2}$$

(B)

$$\begin{aligned} \pi_h &\approx hf(0) + \frac{f''(0)}{2}\int x^2 dx \\ &= hf(0) + \frac{f''(0)h^3}{24} \end{aligned} \tag{3}$$

Choosing h so that both $h \ll 1$ and $\pi_h \ll 1$ are true, we have

$$\begin{aligned} \text{MSE}(0) &= (E(\hat{f}(0)) - f(0))^2 + Var(\hat{f}(0)) \\ &\approx (\frac{\pi_h}{h} - f(0))^2 + \frac{\pi_h}{nh^2} \\ &= (\frac{f''(0)}{24})^2 h^4 + \frac{1}{nh}(f(0) + \frac{f''(0)h^2}{24}) \\ &\approx (\frac{f''(0)}{24})^2 h^4 + \frac{1}{nh} f(0) \end{aligned} \tag{4}$$

(C)

$$\frac{\partial \text{MSE}(0)}{\partial h} = 4Ah^3 - \frac{f(0)}{nh^2} \tag{5}$$

In order to minimize the MSE, $h = (\frac{f(0)}{4An})^{\frac{1}{5}}$

2 Curve Fitting by linear smoothing

(A)

$$y_{\text{estimate}} = \beta_{\text{estimate}} x$$

$$\begin{aligned}
&= (\vec{x}^T\vec{x})^{-1}\vec{x}^T\vec{y}x \\
&= \frac{\sum_i x_i y_i}{\sum_i x_i^2}x \\
&= \sum_i \frac{x_i x}{\vec{x}^2} y_i
\end{aligned} \tag{6}$$

So we have $\omega_i = \frac{x_i x}{\vec{x}^2}$

(B) See "linear smoothing\linear smoothing.r" and "linear smoothing\weight.r".

3 Cross validation

(A) See "linear smoothing\weight.r"

(B) See "linear smoothing\testmodel.r"

(C)

4 Local polynomial regression

(A) Define matrix R where $R_{ij}(\vec{x}) = (x_i - x)^{j-1}$. Then $g(x_i, a) = (R\vec{a})_i$. The cost function can be written as

$$\sum_{i=1}^{n} \omega_i (y_i - (R\vec{a})_i)^2 \tag{7}$$

Optimizing this with $\vec{a}$ is equivalent to linear regression on $f(\vec{x}) = \vec{x}^T\vec{a}$ given observed data $X = R$ and $\vec{y}$ and weighted cost function. The answer is

$$\vec{a} = (R^T\Omega R)^{-1}R^T\Omega\vec{y} \tag{8}$$

$$\Omega = \mathrm{diag}(\omega_1, \omega_2, \cdots) \tag{9}$$

The estimate of $f(x)$ is just $g(x, a) = a_0$.

(B) Call the old weight function ω and the new ones γ

$$\begin{aligned}
(R^T\Omega R)^{-1} &= \begin{pmatrix} \sum_i \omega_i & \sum_i R_{i2}\omega_i \\ \sum_i R_{i2}\omega_i & \sum_i R_{i2}^2\omega_i \end{pmatrix}^{-1} \\
&= \frac{1}{(\sum_i \omega_i)(\sum_i R_{i2}^2\omega_i) - (\sum_i R_{i2}\omega_i)^2} \begin{pmatrix} \sum_i R_{i2}^2\omega_i & -\sum_i R_{i2}\omega_i \\ -\sum_i R_{i2}\omega_i & \sum_i \omega_i \end{pmatrix}
\end{aligned} \tag{10}$$

$$R^T\Omega\vec{y} = \begin{pmatrix} \sum_i \omega_i y_i \\ \sum_i R_{i2}\omega_i y_i \end{pmatrix} \tag{11}$$

$$\begin{aligned} f(x) = a_0 &= \frac{(\sum_i R_{i2}^2\omega_i)(\sum_i \omega_i y_i) - (\sum_i R_{i2}\omega_i)(\sum_i R_{i2}\omega_i y_i)}{(\sum_i \omega_i)(\sum_i R_{i2}^2\omega_i) - (\sum_i R_{i2}\omega_i)^2} \\ &= \sum_i \frac{((\sum_j R_{j2}^2\omega_j) - (\sum_j R_{j2}\omega_j)R_{i2})\omega_i}{\sum_k((\sum_j R_{j2}^2\omega_j) - (\sum_j R_{j2}\omega_j)R_{k2})\omega_k} y_i \end{aligned} \tag{12}$$

Define weight function $\gamma(x_i, x) = ((\sum_j R_{j2}^2\omega_j) - (\sum_j R_{j2}\omega_j)R_{i2})\omega_i$, the above can be written as a weighted sum of y_i. Further evaluate γ:

$$\begin{aligned} \gamma(x_i, x) &= ((\sum_j R_{j2}^2\omega_j) - (\sum_j R_{j2}\omega_j)R_{i2})\omega_i \\ &= \frac{1}{h^2}((\sum_j (x_j - x)^2 K_j) - (\sum_j (x_j - x)K_j)(x_i - x))K_i \\ &= \frac{1}{h^2}(s_2 - s_1(x_i - x))K_i \end{aligned} \tag{13}$$

Here we used K_i to denote $K(\frac{x-x_i}{h})$. Since the function will be normalized so we can ignore $\frac{1}{h^2}$.

(C) From $y = f(x) + \epsilon$

$$\text{Var}(\vec{y}) = \sigma^2 \tag{14}$$

$$\text{Mean}(\vec{y}) = f(\vec{x}) \tag{15}$$

$$\begin{aligned} \text{Mean}(a_0) &= \sum_i \gamma_i f(x_i) \\ \text{Var}(a_0) &= \text{Tr}\text{Var}(H\vec{y}) \\ &= \sum_i |\gamma(x_i, x)|^2\sigma^2 \end{aligned} \tag{16}$$

(D) Define $\vec{\mu} = \text{Mean}(\vec{y}) = f(\vec{x})$

$$\begin{aligned} E(\sigma^2) &\propto E((\vec{y} - H\vec{y})^T(\vec{y} - H\vec{y})) \\ &= E(\vec{y}^T\vec{y}) - 2E(\vec{y}^T H\vec{y}) + E(\vec{y}^T H^2\vec{y}) \\ &= E(\vec{y}^T\vec{y}) - 2 * (\vec{\mu}^T H\vec{\mu} + \sigma^2\text{Tr}(H)) + \vec{\mu}^T(H^2)\vec{\mu} + \sigma^2\text{Tr}(H^2) \end{aligned} \tag{17}$$

Apply trace trick to the first terms:

$$E(\vec{y}^T\vec{y}) = \text{Tr}(E(\vec{y}\vec{y}^T)) = \text{Tr}(\sigma^2 + \vec{\mu}\vec{\mu}^T) = n\sigma^2 + \vec{\mu}^T\vec{\mu} \tag{18}$$

$$E(\sigma^2) = \sigma^2 + \frac{\vec{\mu}^T(1-H)^2\vec{\mu}}{n - 2\text{Tr}(H) + \text{Tr}(H^2)} \tag{19}$$

(E) See "local polynomial regression\utilities analysis.r" and "local polynomial regression\local linear estimator.r"

(F) See "local polynomial regression\utilities analysis.r"

(G) See "local polynomial regression\utilities analysis.r"

5 Gaussian Process

(A) See "gaussian process\comparing gaussian"

(B) Define C_{ij} to be element of covariance matrix computed by covariance function, from the joint distribution

$$p(f(x_1), f(x_2), \cdots, f(x_n), f(x^*)) \propto e^{-\frac{1}{2}\sum_{i,j} C^{-1}_{ij}(f(x_i)-m(x_i))(f(x_j)-m(x_j))} \tag{20}$$

we have

$$\begin{aligned} p(f(x^*)|f(x_1), f(x_2), \cdots, f(x_n)) &\propto e^{-\frac{1}{2}(C^{-1}_{**}(f(x^*)-m(x^*))^2+2(f(x^*)-m(x^*))\sum_i C_{*i}(f(x_i)-m(x_i)))} \\ &\propto e^{-\frac{C^{-1}_{**}}{2}(f(x^*)^2+2f(x^*)(\sum_i \frac{C^{-1}_{*i}}{C^{-1}_{**}}(f(x_i)-m(x_i))-m(x^*))} \end{aligned} \tag{21}$$

So we have $(f^*|\mathbf{f}) \sim N(\mathbf{m} - \frac{1}{(C^{-1})_{**}}(C^{-1})^T_*(\mathbf{f}-\mathbf{m}), \frac{1}{(C^{-1})_{**}})$

We have split matrix C and C^{-1} as

$$\begin{pmatrix} C_{**} & C^T_* \\ C_* & C_0 \end{pmatrix} \tag{22}$$

We can write the element of C^{-1} explicitly

$$(C^{-1})_{**} = (C_{**} - C^T_* C_0^{-1} C_*)^{-1}$$

$$\frac{(C^{-1})^T_*}{(C^{-1})_{**}} = -C^T_* C_0^{-1}$$

C_{**} is the covariance function evaluated at 0.

(C) For any linear combination of $(\vec{y}, \vec{\theta})$: $(\vec{a}^T\vec{y} + \vec{b}^T\vec{\theta})$. Compute the moment generating function

$$E(e^{t\vec{a}^T\vec{y}+t\vec{b}^T\vec{\theta}}) = \int e^{t\vec{a}^T\vec{y}+t\vec{b}^T\vec{\theta}} p(\vec{y}|\vec{\theta})p(\vec{\theta})d\vec{y}d\vec{\theta} \tag{23}$$

Change the variable being integrated

$$\vec{y} \rightarrow \vec{y}' = \vec{y} - R\vec{\theta} \tag{24}$$

$$\vec{\theta} \rightarrow \vec{\theta}' = \vec{\theta} - \vec{m} \tag{25}$$

The Jacobian of this transformation is 1. The logrithmic pdf of $e^{t\vec{a}^T\vec{y}+t\vec{b}^T\vec{\theta}}$ is:

$$t\vec{a}^T\vec{y}' + t(\vec{b}^T + \vec{a}^T R)\vec{\theta}' + t(\vec{b}^T + \vec{a}^T R)\vec{m} - \frac{\vec{y}^{T\prime}\Sigma^{-1}\vec{y}'}{2} - \frac{\vec{\theta}^{T\prime}V^{-1}\vec{\theta}'}{2} + \cdots \tag{26}$$

Complete the square for $\vec{y}'$,

$$-\frac{\vec{y}^{T\prime}\Sigma^{-1}\vec{y}'}{2} + t\vec{a}^T\vec{y}' = -\frac{(\vec{y}' - t\Sigma\vec{a})^T\Sigma^{-1}(\vec{y}' - t\Sigma\vec{a})}{2} + \frac{\vec{a}^T\Sigma\vec{a}}{2}t^2 \tag{27}$$

and similarly for $\vec{\theta}'$. The t-dependent part after the the integrating is

$$\exp\left(\frac{\vec{a}^T\Sigma\vec{a} + (\vec{b}+R^T\vec{a})^T V(\vec{b}+R^T\vec{a})}{2}t^2 + (\vec{b}^T + R^T\vec{a})t\right) \tag{28}$$

The coefficient of t^2 is positive definite. And the constant multiplied to this exponential must be 1 because $M(0) = E(1) = 1$.

6 In nonparametric regression and spatial smoothing

(A)

$$p(\vec{y}, f(\vec{x})) = p(\vec{y}|f(\vec{x}))p(f(\vec{x})) \propto e^{-\frac{|\vec{y}-f(\vec{x})|^2}{2\sigma^2}} e^{-\frac{f(\vec{x})^T C^{-1} f(\vec{x})}{2}} \tag{29}$$

$$\begin{aligned} p(f(\vec{x})|\vec{y}) &\propto p(\vec{y}, f(\vec{x}))|_{f(\vec{x})} \\ &\propto e^{-\frac{f(\vec{x})^T(C_0^{-1}+\frac{1}{\sigma^2})f(\vec{x})}{2} + \frac{\vec{y}^T f(\vec{x})}{\sigma^2}} \\ &\propto e^{-\frac{(f(\vec{x})-(C_0^{-1}\sigma^2+1)^{-1}\vec{y})^T(C_0^{-1}+\frac{1}{\sigma^2})(f(\vec{x})-(C_0^{-1}\sigma^2+1)^{-1}\vec{y})}{2}} \end{aligned} \tag{30}$$

Therefore $\mathbf{f}|\mathbf{y} \sim N\left((C_0^{-1}\sigma^2+1)^{-1}\mathbf{y}, \sigma^2(C_0^{-1}\sigma^2+1)^{-1})\right)$
Using Woodbury identity's special case where both matrix are full rank.

$$(A^{-1}+B^{-1})^{-1} = A - A(A+B)^{-1}A$$

we have

$$(C_0^{-1}\sigma^2+1)^{-1} = \frac{C_0}{\sigma^2}\left(1-(C_0+\sigma^2)^{-1}C_0\right) = C_0(C_0+\sigma^2)^{-1}$$

or

$$(C_0^{-1}\sigma^2+1)^{-1} = 1 - (\frac{C_0}{\sigma^2}+1)^{-1}$$

(B) From previous question $E(f^*|\mathbf{f}) = C_*^T C_0^{-1}\mathbf{f}$ and $E(\mathbf{f}|\mathbf{y}) = (C_0^{-1}\sigma^2+1)^{-1}\mathbf{y}$. We have

$$\begin{aligned} E(f(x^*)|\vec{y}) &= \int E(f(x^*)|f(\vec{x}))p(f(\vec{x})|\vec{y})df(\vec{x}) \\ &= \int C_*^T C_0^{-1}\mathbf{f}p(f(\vec{x})|\vec{y})df(\vec{x}) \\ &= C_*^T C_0^{-1}E(\mathbf{f}|\mathbf{y}) = W^T\mathbf{y} \end{aligned} \tag{31}$$

Where $W^T = C_*^T C_0^{-1}(C_0^{-1}\sigma^2+1)^{-1} = C_*^T(C_0+\sigma^2)^{-1}$.

To compute variance, first compute $E(f(x^*)^2)$, using $Var(x) = E(x^2) - E(x)^2$ and the fact x_i are independent of each other

$$\begin{aligned} E(f^{*2}|\mathbf{y}) &= \int E(f(x^*)^2|\vec{x})p(f(\vec{x})|\vec{y})df(\vec{x}) \\ &= \int (Var(f^*|\mathbf{f}) + (C_*^T C_0^{-1}\mathbf{f})^2)p(f(\vec{x})|\vec{y})df(\vec{x}) \\ &= Var(f^*|\mathbf{f}) + C_*^T C_0^{-1} E C_0^{-1} C_* \end{aligned} \tag{32}$$

Where

$$\begin{aligned} E &= E(\mathbf{f}\mathbf{f}^T|\mathbf{y}) \\ &= Cov(\mathbf{f}|\mathbf{y}) + E(\mathbf{f}|\mathbf{y})E(\mathbf{f}|\mathbf{y})^T \end{aligned} \tag{33}$$

so

$$\begin{aligned} E(f^{*2}|\mathbf{y}) &= Var(f^*|\mathbf{f}) + C_*^T C_0^{-1}(Cov(\mathbf{f}|\mathbf{y}) + E(\mathbf{f}|\mathbf{y})E(\mathbf{f}|\mathbf{y})^T)C_0^{-1}C_* \\ &= Var(f^*|\mathbf{f}) + C_*^T C_0^{-1} Cov(\mathbf{f}|\mathbf{y}) C_0^{-1} C_* + E(f^*|\mathbf{y})^2 \\ &= C_{**} - C_*^T C_0^{-1} C_* + C_*^T C_0^{-1} C_0 \left(1 - (C_0 + \sigma^2)^{-1} C_0\right) C_0^{-1} C_* + E(f^*|\mathbf{y})^2 \\ &= C_{**} - C_*^T (C_0 + \sigma^2)^{-1} C_* + E(f^*|\mathbf{y})^2 \end{aligned} \tag{34}$$

So

$$Var(f^*|\mathbf{y}) = C_{**} - C_*^T (C_0 + \sigma^2)^{-1} C_*$$

(C) See "non-parametric regression and spatial smoothing\utilities analysis non-parametric regression.r".

(D) $\mathbf{y} = \mathbf{f} + \epsilon$. Given grid $(x_1, x_2 \cdots)$, $\mathbf{y}$ is the sum of two independent multivariate normal distribution. $\mathbf{y} \sim N(0, C_0 + \sigma^2)$

(E) See "non-parametric regression and spatial smoothing\utilities analysis non-parametric regression.r".

(F) See "non-parametric regression and spatial smoothing\weather.r".