Data Fitting by a Spline

By

Kozo ICHIDA* Fujiichi YOSHIMOTO** and Takeshi KIYONO*

(Received March 19, 1973)

Abstract

This paper presents an algorithm for the computation of data-fitting by a cubic spline. The normal equation derived from the algorithm is not ill-conditioned. If it is assumed that the data are subjected to some independently distributed error about their trend, the regression theory provides the desired estimate of variance, like the expression that is known when we use the orthogonal polynomials.

1. Introduction

Approximating a function, whose values at a sequence of points are generally known only empirically and are subject to inherent errors, is a problem that has been studied. The standard polynomials in least squares data-fitting have a difficulty in solving the normal equation, because it becomes quite ill-conditioned as the degree of polynomial increases. Orthogonal polynomials avoid this difficulty, but there is a case in which they are not satisfactory.

Recently, data-fitting using spline functions has been studied. Sometimes the spline functions seem more adequate than orthogonal polynomials. In [3], the spline function is represented as

$$S_m(x)=a_0+a_1x+\cdots+a_mx^m+\sum_{i=1}^{J} c_i(x-\hat{x}_i)_+{}^m, \qquad (1)$$

where $\hat{x}_i$'s are a set of joints. But eq.(1) is ill-conditioned and cumbersome to evaluate for the large m and J.

Here, a new algorithm is given in which the normal equation is not ill-conditioned.

* Department of Information Science
** Akashi Technical College (Mr. Yoshimoto was Studying at Kyoto University as a visiting researcher from April 1970 to March 1971)

2. Representation of a Spline Function

Let $S(x)$ be a cubic spline function having the equidistant knots $x^{(i)}(i=0, 1,\dots, n)$ and satisfying the equation

$$S(x^{(i)})=y^{(i)} \qquad (i=0, 1, \dots, n).$$

In the interval $[x^{(i-1)}\leqq x\leqq x^{(i)}]$

$$S(x)=M_{i-1}\left\{\frac{(x^{(i)}-x)^3}{6h}-\frac{x^{(i)}-x}{6}h\right\}+M_i\left\{\frac{(x-x^{(i-1)})^3}{6h}-\frac{x-x^{(i-1)}}{6}h\right\}$$
$$+\frac{x^{(i)}-x}{h}y^{(i-1)}+\frac{x-x^{(i-1)}}{h}y^{(i)}, \tag{2}$$

where $h=x^{(i)}-x^{(i-1)}$[1]. From the continuity of $S'(x)$ at $x^{(i)}$, we get the following equation

$$\begin{bmatrix} 2 & \frac{1}{2} & 0 & & & 0 \\ \frac{1}{2} & 2 & \frac{1}{2} & & & \\ & \ddots & \ddots & \ddots & & \\ & & & \frac{1}{2} & 2 & \frac{1}{2} \\ 0 & & & 0 & \frac{1}{2} & 2 \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-2} \\ M_{n-1} \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-2} \\ d_{n-1} \end{bmatrix}, \tag{3}$$

where

$$d_i=3(y^{(i+1)}-2y^{(i)}+y^{(i-1)})/h^2, \tag{4}$$

and we set $M_0=M_n=0$.

Let D_n be the following determinant

$$D_n=\begin{vmatrix} 2 & \frac{1}{2} & 0 & & & 0 \\ \frac{1}{2} & 2 & \frac{1}{2} & & & \\ & \ddots & \ddots & \ddots & & \\ & & & \frac{1}{2} & 2 & \frac{1}{2} \\ 0 & & & 0 & \frac{1}{2} & 2 \end{vmatrix}. \tag{5}$$

From eq. (5) we get

$$D_n=\left\{\left(1+\frac{\sqrt{3}}{2}\right)^{n+1}-\left(1-\frac{\sqrt{3}}{2}\right)^{n+1}\right\}\Big/\sqrt{3}\ . \tag{6}$$

Writing eq. (3) as

$$AM=\boldsymbol{d}, \tag{7}$$

the element of A^{-1} (inverse matrix of A) is:

$$\left.\begin{aligned} A_{i,j}^{-1}&=\frac{(-1)^{i+j}D_{i-1}D_{n-1-j}}{2^{j-i}D_{n-1}} \quad (1\leqq i\leqq j\leqq n-1),\\ A_{i,j}^{-1}&=\frac{(-1)^{i+j}D_{j-1}D_{n-1-i}}{2^{i-j}D_{n-1}} \quad (1\leqq j\leqq i\leqq n-1),\end{aligned}\right\} \tag{8}$$

and $|A|=D_{n-1}$.

From eqs. (7) and (8), we have

$$M_i=\sum_{j=1}^{n-1}A_{i,j}^{-1}d_j=\sum_{j=1}^{n-1}A_{i,j}^{-1}\frac{y^{(j+1)}-2y^{(j)}+y^{(j-1)}}{h^2/3}\equiv\sum_{j=0}^{n}a_{i,j}y^{(j)}, \tag{9}$$

where

$$\left.\begin{aligned} a_{i,j}&=3(A_{i,j+1}^{-1}-2A_{i,j}^{-1}+A_{i,j-1}^{-1})/h^2,\\ A_{i,-1}^{-1}&=A_{i,0}^{-1}=A_{i,n}^{-1}=A_{i,n+1}^{-1}=0.\end{aligned}\right\} \tag{10}$$

Substitution of M_i of eq. (9) into eq. (2) gives

$$\begin{aligned} S(x)=&\sum_{j=0}^{n}(b_i(x)a_{i-1,j}+\beta_i(x)a_{i,j})y^{(j)}\\ &+\frac{x^{(i)}-x}{h}y^{(i-1)}+\frac{x-x^{(i-1)}}{h}y^{(i)}, \end{aligned} \tag{11}$$

where

$$\left.\begin{aligned} b_i(x)&=\frac{(x^{(i)}-x)^3}{6h}-\frac{x^{(i)}-x}{6}h,\\ \beta_i(x)&=\frac{(x-x^{(i-1)})^3}{6h}-\frac{x-x^{(i-1)}}{6}h.\end{aligned}\right\} \tag{12}$$

3. Data-fitting by the Spline

Let $x_k(k=1, 2,\dots, N)$ be given values of an independent real variable x. Suppose that corresponding to each value x_k we have observed value f_k which generally will

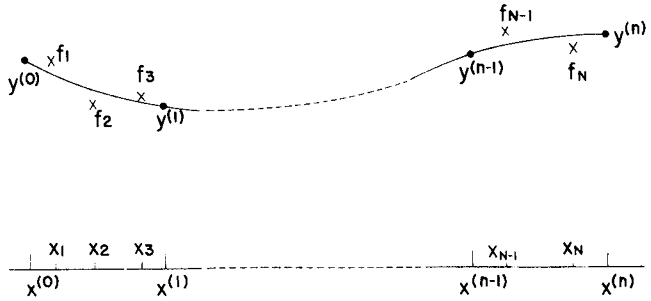

Fig. 1. Representation of data fitting by a spline function.

be in error (Fig. 1). We use a spline function (11) for the least squares approximation. The length of each interval $x_k - x_{k-1}$ may be different. We suppose $N > n+1$, and $x^{(0)} \leqq x_k \leqq x^{(n)}$ $(k=1, 2, \ldots, N)$. Let minimum and maximum x_k for $x^{(i-1)} \leqq x_k \leqq x^{(i)}$ be x_{p_i} and x_{q_i} respectively. Then the sum of the squares of the residual becomes

$$F(y^{(0)}, y^{(1)}, \ldots, y^{(n)}) = \sum_{i=1}^{n} \sum_{k=p_i}^{q_i} \{S(x_k) - f_k\}^2. \tag{13}$$

Differentiating $F(y^{(0)}, y^{(1)}, \ldots, y^{(n)})$ with $y^{(i)}$ $(i=0, 1, \ldots, n)$ and setting to zero,

$$\frac{\partial F(y^{(0)}, y^{(1)}, \ldots, y^{(n)})}{\partial y^{(i)}} = 0 \quad (i=0, 1, \ldots, n), \tag{14}$$

we get the normal equation

$$\begin{bmatrix} t_{0,0} & t_{0,1} & \cdots\cdots & t_{0,n} \\ t_{1,0} & t_{1,1} & \cdots\cdots & t_{1,n} \\ \vdots & \vdots & & \vdots \\ t_{n,0} & t_{n,1} & \cdots\cdots & t_{n,n} \end{bmatrix} \begin{bmatrix} y^{(0)} \\ y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} = \begin{bmatrix} g_0 \\ g_1 \\ \vdots \\ g_n \end{bmatrix}. \tag{15}$$

The matrix $\{t_{m,r}\}$ is symmetric and its element $t_{m,r}$ is expressed as follows.

$$t_{m,m} = V_{m,m} + W_{m,m} + C_{m,m} + U_{m,m} + Z_{m,m} + E_m + R_m, \tag{16}$$

$$t_{m+1,m} = V_{m+1,m} + W_{m+1,m} + C_{m+1,m} + U_{m+1,m} + Z_{m+1,m} + B_{m+1,m}, \tag{17}$$

$$t_{m-1,m} = V_{m-1,m} + W_{m-1,m} + C_{m-1,m} + U_{m-1,m} + Z_{m-1,m} + J_{m-1,m}, \tag{18}$$

$$t_{r,m} = V_{r,m} + W_{r,m} + C_{r,m} + U_{r,m} + Z_{r,m} \quad (r \geqq m+2 \quad \text{or} \quad m \geqq r+2), \tag{19}$$

where

$$V_{r,m} = \sum_{i=1}^{n} \sum_{k=p_i}^{q_i} \{b_i(x_k) a_{i-1,m} + \beta_i(x_k) a_{i,m}\} \{b_i(x_k) a_{i-1,r} + \beta_i(x_k) a_{i,r}\}, \tag{20}$$

$$W_{r,m} = \sum_{k=p_{m+1}}^{q_{m+1}} \{(x^{(m+1)} - x_k)/h\} \{b_{m+1}(x_k) a_{m,r} + \beta_{m+1}(x_k) a_{m+1,r}\}, \tag{21}$$

$$C_{r,m} = \sum_{k=p_{r+1}}^{q_{r+1}} \{(x^{(r+1)} - x_k)/h\} \{b_{r+1}(x_k) a_{r,m} + \beta_{r+1}(x_k) a_{r+1,m}\}, \tag{22}$$

$$U_{r,m} = \sum_{k=p_m}^{q_m} \{(x_k - x^{(m-1)})/h\} \{b_m(x_k) a_{m-1,r} + \beta_m(x_k) a_{m,r}\}, \tag{23}$$

$$Z_{r,m} = \sum_{k=p_r}^{q_r} \{(x_k - x^{(r-1)})/h\} \{b_r(x_k) a_{r-1,m} + \beta_r(x_k) a_{r,m}\}, \tag{24}$$

$$E_r = \sum_{k=p_{r+1}}^{q_{r+1}} \{(x^{(r+1)} - x_k)/h\}^2, \tag{25}$$

$$R_r = \sum_{k=p_r}^{q_r} \{(x_k - x^{(r-1)})/h\}^2, \tag{26}$$

$$B_{r,m}=\sum_{k=p_r}^{q_r}\{(x^{(r)}-x_k)/h\}\{(x_k-x^{(m)})/h\}, \tag{27}$$

$$J_{r,m}=\sum_{k=p_m}^{q_m}\{(x^{(m)}-x_k)/h\}\{(x_k-x^{(r)})/h\}, \tag{28}$$
$$W_{r,n}=C_{n,m}=U_{r,0}=Z_{0,m}=0,$$
$$R_0=E_n=0$$
$$(0\leqq r,\ m\leqq n).$$

Moreover if we set

$$G_r=\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}f_k\{b_i(x_k)a_{i-1,r}+\beta_i(x_k)a_{i,r}\}, \tag{29}$$

$$H_r=\sum_{k=p_{r+1}}^{q_{r+1}}f_k\{(x^{(r+1)}-x_k)/h\},\quad H_n=0, \tag{30}$$

$$Q_r=\sum_{k=p_r}^{q_r}f_k\{(x_k-x^{(r-1)})/h\},\quad Q_0=0, \tag{31}$$

then we have

$$g_r=G_r+H_r+Q_r. \tag{32}$$

The matrix $\{t_{r,m}\}$ has an element that becomes smaller in absolute as it is apart from the diagonal, and is not ill-conditioned.

4. Unbiased Estimator of Variance of Error

Suppose the f_k in eq.(13) is expressed for $[x^{(i-1)}\leqq x_k\leqq x^{(i)}]$ $(i=1, 2, \ldots, n)$ as

$$f_k=S(x_k)+\varepsilon_k=\sum_{j=0}^{n}\{b_i(x_k)a_{i-1,j}+\beta_i(x_k)\,a_{i,j}\}\,y^{(j)}$$
$$+\frac{x^{(i)}-x_k}{h}y^{(i-1)}+\frac{x_k-x^{(i-1)}}{h}y^{(i)}+\varepsilon_k, \tag{33}$$

ε_k is an error that has an expectation and variance

$$\left.\begin{aligned}&E[\varepsilon_k]=0,\\&V[\varepsilon_k]=\sigma^2.\end{aligned}\right\} \tag{34}$$

The estimators of the regression coefficients $y^{(0)}, y^{(1)}, \ldots, y^{(n)}$ are written as $\hat{y}^{(0)}, \hat{y}^{(1)}, \ldots, \hat{y}^{(n)}$. With these estimators, the sum of the squares of the residual is

$$S=\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}\Big[f_k-\Big\{\sum_{j=0}^{n}(b_i(x_k)a_{i-1,j}+\beta_i(x_k)a_{i,j})\hat{y}^{(j)}$$
$$+\frac{x^{(i)}-x_k}{h}\hat{y}^{(i-1)}+\frac{x_k-x^{(i-1)}}{h}\hat{y}^{(i)}\Big\}\Big]^2. \tag{35}$$

Determining $\hat{y}^{(0)}, \hat{y}^{(1)}, \ldots, \hat{y}^{(n)}$ to minimize S, we get

$$\begin{bmatrix} t_{0,0} & t_{0,1} & \cdots & t_{0,n} \\ t_{1,0} & t_{1,1} & \cdots & t_{1,n} \\ \vdots & \vdots & & \vdots \\ t_{n,0} & t_{n,1} & \cdots & t_{n,n} \end{bmatrix} \begin{bmatrix} \hat{y}^{(0)} \\ \hat{y}^{(1)} \\ \vdots \\ \hat{y}^{(n)} \end{bmatrix} = \begin{bmatrix} g_0 \\ g_1 \\ \vdots \\ g_n \end{bmatrix}. \tag{36}$$

The solution $\hat{y}^{(i)}$ $(i=0, 1, \dots, n)$ of eq. (36) gives the estimator of the regression coefficients $y^{(i)}$. We define column vectors $\hat{\mathbf{y}}$ and $\mathbf{y}$ of which $\hat{y}^{(i)}$ and $y^{(i)}$ $(i=0, 1, \dots, n)$ are respective elements. Then eq. (36) is written as

$$T\hat{\mathbf{y}}=\mathbf{g}, \tag{37}$$

where $T=\{t_{r,m}\}$. From eq. (34) the expectation of f_k for $[x^{(i-1)} \leqq x_k \leqq x^{(i)}]$ is

$$E[f_k]=\sum_{j=0}^{n}\{b_i(x_k)a_{i-1,j}+\beta_i(x_k)a_{i,j}\}\, y^{(j)} + \frac{x^{(i)}-x_k}{h}y^{(i-1)}+\frac{x_k-x^{(i-1)}}{h}y^{(i)}. \tag{38}$$

We now have the following lemma.

Lemma 1

$$E[g_r]=\sum_{m=0}^{n} t_{r,m}y^{(m)} \quad (0 \leqq r \leqq n). \tag{39}$$

Proof.

From eqs. (29)∼(32),

$$E[g_0]=\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}\{b_i(x_k)a_{i-1,0}+\beta_i(x_k)a_{i,0}\}\, E[f_k]+\sum_{k=p_1}^{q_1}\frac{x^{(1)}-x_k}{h}E[f_k]$$

$$=\sum_{m=0}^{n}V_{0,m}y^{(m)}+\sum_{m=0}^{n-1}W_{0,m}y^{(m)}+\sum_{m=1}^{n}U_{0,m}y^{(m)}+\sum_{m=0}^{n}C_{0,m}y^{(m)}+E_0y^{(0)} +J_{0,1}y^{(1)}=\sum_{m=0}^{n}t_{0,m}y^{(m)}.$$

For $1 \leqq r \leqq n-1$

$$E[g_r]=\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}\{b_i(x_k)a_{i-1,r}+\beta_i(x_k)a_{i,r}\}\, E[f_k] +\sum_{k=p_{r+1}}^{q_{r+1}}\frac{x^{(r+1)}-x_k}{h}E[f_k]+\sum_{k=p_r}^{q_r}\frac{x_k-x^{(r-1)}}{h}E[f_k]$$

$$=\sum_{m=0}^{n}V_{r,m}y^{(m)}+\sum_{m=0}^{n-1}W_{r,m}y^{(m)}+\sum_{m=1}^{n}U_{r,m}y^{(m)}+\sum_{m=0}^{n}C_{r,m}y^{(m)} +E_ry^{(r)}+J_{r,r+1}y^{(r+1)}+\sum_{m=0}^{n}Z_{r,m}y^{(m)}+B_{r,r-1}y^{(r-1)}+R_ry^{(r)}$$

$$=\sum_{m=0}^{n}t_{r,m}y^{(m)}.$$

$$E[g_n]=\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}\{b_i(x_k)a_{i-1,n}+\beta_i(x_k)a_{i,n}\}E[f_k]+\sum_{k=p_n}^{q_n}\frac{x_k-x^{(n-1)}}{h}E[f_k]$$

$$=\sum_{m=0}^{n}V_{n,m}y^{(m)}+\sum_{m=0}^{n-1}W_{n,m}y^{(m)}+\sum_{m=1}^{n}U_{n,m}y^{(m)}+\sum_{m=0}^{n}Z_{n,m}y^{(m)}$$

$$+B_{n,n-1}y^{(n-1)}+R_ny^{(n)}$$

$$=\sum_{m=0}^{n}t_{n,m}y^{(m)}.$$

Writing eq. (39) as

$$T\boldsymbol{y}=E[\boldsymbol{g}], \tag{40}$$

we have

$$\boldsymbol{y}=T^{-1}E[\boldsymbol{g}]. \tag{41}$$

From eq. (37), we get

$$\hat{\boldsymbol{y}}=T^{-1}\boldsymbol{g}. \tag{42}$$

The expectation of eq. (42) become

$$E[\hat{\boldsymbol{y}}]=E[T^{-1}\boldsymbol{g}]=T^{-1}E[\boldsymbol{g}]=\boldsymbol{y}, \tag{43}$$

i.e.,

$$E[\hat{y}^{(r)}]=y^{(r)} \quad (r=0, 1, \dots, n). \tag{44}$$

Lemma 2

$$Cov[g_r, g_m]=t_{r,m}\sigma^2 \quad (0\leqq r,\ m\leqq n). \tag{45}$$

Proof.

For $1\leqq r,\ m\leqq n-1$

$$Cov[g_r, g_m]=E\{[g_r-E[g_r]\}\{g_m-E[g_m]\}]$$

$$=E\Bigg[\Bigg\{\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}(b_i(x_k)a_{i-1,r}+\beta_i(x_k)a_{i,r})f_k+\sum_{k=p_{r+1}}^{q_{r+1}}\frac{x^{(r+1)}-x_k}{h}f_k$$

$$+\sum_{k=p_r}^{q_r}\frac{x_k-x^{(r-1)}}{h}f_k-\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}(b_i(x_k)a_{i-1,r}+\beta_i(x_k)a_{i,r})(f_k-\varepsilon_k)$$

$$-\sum_{k=p_{r+1}}^{q_{r+1}}\frac{x^{(r+1)}-x_k}{h}(f_k-\varepsilon_k)-\sum_{k=p_r}^{q_r}\frac{x_k-x^{(r-1)}}{h}(f_k-\varepsilon_k)\Bigg\}$$

$$\cdot\Bigg\{\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}(b_i(x_k)a_{i-1,m}+\beta_i(x_k)a_{i,m})f_k+\sum_{k=p_{m+1}}^{q_{m+1}}\frac{x^{(m+1)}-x_k}{h}f_k$$

$$+\sum_{k=p_m}^{q_m}\frac{x_k-x^{(m-1)}}{h}f_k-\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}(b_i(x_k)a_{i-1,m}+\beta_i(x_k)a_{i,m})(f_k-\varepsilon_k)$$

$$-\sum_{k=p_{m+1}}^{q_{m+1}}\frac{x^{(m+1)}-x_k}{h}(f_k-\varepsilon_k)-\sum_{k=p_m}^{q_m}\frac{x_k-x^{(m-1)}}{h}(f_k-\varepsilon_k)\Bigg\}\Bigg]$$

$$=E\Big[\Big\{\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}(b_i(x_k)a_{i-1,r}+\beta_i(x_k)a_{i,r})\varepsilon_k+\sum_{k=p_{r+1}}^{q_{r+1}}\frac{x^{(r+1)}-x_k}{h}\varepsilon_k$$

$$+\sum_{k=p_r}^{q_r}\frac{x_k-x^{(r-1)}}{h}\varepsilon_k\Big\}\Big\{\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}(b_i(x_k)a_{i-1,m}+\beta_i(x_k)a_{i,m})\varepsilon_k$$

$$+\sum_{k=p_{m+1}}^{q_{m+1}}\frac{x^{(m+1)}-x_k}{h}\varepsilon_k+\sum_{k=p_m}^{q_m}\frac{x_k-x^{(m-1)}}{h}\varepsilon_k\Big\}\Big].$$

i) $r\neq m,\ r\neq m+1,\ r\neq m-1$

$$Cov[g_r, g_m]=E\Big[\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}(b_i(x_k)a_{i-1,r}+\beta_i(x_k)a_{i,r})\,(b_i(x_k)a_{i-1,m}$$

$$+\beta_i(x_k)a_{i,m})\varepsilon_k^2\Big]+E\Big[\sum_{k=p_{r+1}}^{q_{r+1}}\frac{x^{(r+1)}-x_k}{h}(b_{r+1}(x_k)a_{r,m}$$

$$+\beta_{r+1}(x_k)\,a_{r+1,m})\varepsilon_k^2\Big]+E\Big[\sum_{k=p_r}^{q_r}\frac{x_k-x^{(r-1)}}{h}(b_r(x_k)a_{r-1,m}$$

$$+\beta_r(x_k)a_{r,m})\varepsilon_k^2\Big]+E\Big[\sum_{k=p_{m+1}}^{q_{m+1}}\frac{x^{(m+1)}-x_k}{h}(b_{m+1}(x_k)a_{m,r}$$

$$+\beta_{m+1}(x_k)a_{m-1,r})\varepsilon_k^2\Big]+E\Big[\sum_{k=p_m}^{q_m}\frac{x_k-x^{(m-1)}}{h}(b_m(x_k)a_{m-1,r}$$

$$+\beta_m(x_k)a_{m,r})\varepsilon_k^2\Big]$$

$$=(V_{r,m}+W_{r,m}+U_{r,m}+Z_{r,m}+C_{r,m})\sigma^2=t_{r,m}\sigma^2.$$

ii) $r=m$

$$Cov[g_m, g_m]=(V_{m,m}+W_{m,m}+U_{m,m}+Z_{m,m}+C_{m,m}+E_m+R_m)\sigma^2$$
$$=t_{m,m}\sigma^2.$$

iii) $r=m+1$

$$Cov[g_{m+1}, g_m]=(V_{m+1,m}+W_{m+1,m}+U_{m+1,m}+Z_{m+1,m}+C_{m+1,m}$$
$$+B_{m+1,m})\sigma^2=t_{m+1,m}\sigma^2.$$

iv) $r=m-1$

$$Cov[g_{m-1}, g_m]=(V_{m-1,m}+W_{m-1,m}+U_{m-1,m}+Z_{m-1,m}+C_{m-1,m}$$
$$+J_{m-1,m})\sigma^2=t_{m-1,m}\sigma^2.$$

For $r=0,\ 1\leqq m\leqq n-1$

$$Cov[g_0, g_m]=E\Big[\Big\{\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}(b_i(x_k)a_{i-1,0}+\beta_i(x_k)a_{i,0})\varepsilon_k$$

$$+\sum_{k=p_1}^{q_1}\frac{x^{(1)}-x_k}{h}\varepsilon_k\Big\}\Big\{\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}(b_i(x_k)a_{i-1,m}+\beta_i(x_k)a_{i,m})\varepsilon_k$$

$$+\sum_{k=p_{m+1}}^{q_{m+1}}\frac{x^{(m+1)}-x_k}{h}\varepsilon_k+\sum_{k=p_m}^{q_m}\frac{x_k-x^{(m-1)}}{h}\varepsilon_k\Big\}\Big].$$

i) $m \neq 1$

$$Cov[g_0, g_m]=(V_{0,m}+W_{0,m}+U_{0,m}+C_{0,m})\sigma^2=t_{0,m}\sigma^2.$$

ii) $m=1$

$$Cov[g_0, g_1]=(V_{0,1}+W_{0,1}+U_{0,1}+C_{0,1}+J_{0,1})\sigma^2=t_{0,1}\sigma^2.$$

For $r=m=0$

$$Cov[g_0, g_0]=E\Big[\Big\{\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}(b_i(x_k)a_{i-1,0}+\beta_i(x_k)a_{i,0})\varepsilon_k$$
$$+\sum_{k=p_1}^{q_1}\frac{x^{(1)}-x_k}{h}\varepsilon_k\Big\}^2\Big]=(V_{0,0}+W_{0,0}+C_{0,0}+E_{0,0})\sigma^2$$
$$=t_{0,0}\sigma^2.$$

For $r=0$, $m=n$

$$Cov[g_0, g_n]=E\Big[\Big\{\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}(b_i(x_k)a_{i,0}+\beta_i(x_k)a_{i-1,0})\varepsilon_k$$
$$+\sum_{k=p_1}^{q_1}\frac{x^{(1)}-x_k}{h}\varepsilon_k\Big\}\Big\{\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}(b_i(x_k)a_{i-1,n}+\beta_i(x_k)a_{i,n})\varepsilon_k$$
$$+\sum_{k=p_n}^{q_n}\frac{x_k-x^{(n-1)}}{h}\varepsilon_k\Big\}\Big]=(V_{0,n}+C_{0,n}+U_{0,n})\sigma^2$$
$$=t_{0,n}\sigma^2.$$

For $m=0$, $1\leqq r\leqq n-1$ and $m=0$, $r=n$ we have the result in eq. (45) similarly.

From lemma 1 and 2, we have the following theorem.

Theorem: Let S_m be the minimum of the sum of the squares of the residual. Then

$$S_e=\frac{S_m}{N-n-1} \tag{46}$$

is an unbiased estimator of σ^2.

Proof:

From eqs. (42) (44) and lemma 2, the covariance matrix becomes

$$\begin{aligned}\{Cov[\hat{y}^{(r)},\hat{y}^{(m)}]\} &= \{[(\hat{y}^{(r)}-y^{(r)})(\hat{y}^{(m)}-y^{(m)})]\}\\ &=[(\hat{\boldsymbol{y}}-\boldsymbol{y})(\hat{\boldsymbol{y}}-\boldsymbol{y})^t]\\ &=E[T^{-1}(\boldsymbol{g}-E[\boldsymbol{g}])(\boldsymbol{g}-E[\boldsymbol{g}])^tT^{-1}]\\ &=T^{-1}E[(\boldsymbol{g}-E[\boldsymbol{g}])(\boldsymbol{g}-E[\boldsymbol{g}])^t]T^{-1}\\ &=T^{-1}(T\sigma^2)T^{-1}=\{d_{r,m}\}\sigma^2,\end{aligned}$$

where $\{d_{r,m}\}=D=T^{-1}$.

The minimum of the sum of the squares of the residual becomes

$$S_m=\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}\Big[f_k-\Big\{\sum_{j=0}^{n}(b_i(x_k)a_{i-1,j}+\beta_i(x_k)a_{i,j})\hat{y}^{(j)}+\frac{x^{(i)}-x_k}{h}\hat{y}^{(i-1)}+\frac{x_k-x^{(i-1)}}{h}\hat{y}^{(i)}\Big\}\Big]^2$$

$$=\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}f_k^2-2\sum_{j=0}^{n}G_j\hat{y}^{(j)}-2\sum_{j=0}^{n-1}H_j\hat{y}^{(j)}-2\sum_{j=1}^{n}Q_j\hat{y}^{(j)}$$

$$+\sum_{j=0}^{n}\sum_{m=0}^{n}V_{m,j}\hat{y}^{(m)}\hat{y}^{(j)}+2\sum_{j=0}^{n}\sum_{i=0}^{n-1}C_{i,j}\hat{y}^{(i)}\hat{y}^{(j)}$$

$$+2\sum_{j=0}^{n}\sum_{i=1}^{n}Z_{i,j}\hat{y}^{(i)}\hat{y}^{(j)}+\sum_{i=0}^{n-1}E_i(\hat{y}^{(i)})^2+2\sum_{i=0}^{n-1}J_{i,i+1}\hat{y}^{(i)}\hat{y}^{(i+1)}$$

$$+\sum_{i=1}^{n}R_i(\hat{y}^{(i)})^2$$

$$=\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}f_k^2-2\sum_{r=0}^{n}g_r\hat{y}^{(r)}+\sum_{r=0}^{n}g_r\hat{y}^{(r)}$$

$$=\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}f_k^2-\sum_{r=0}^{n}g_r\hat{y}^{(r)}. \tag{47}$$

The expectation of the first term of the righthand side in eq. (47) is

$$E\Big[\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}f_k^2\Big]=\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}[V[f_k]+(E[f_k])^2]$$

$$=\sum_{i=1}^{n}\sum_{k=p_i}^{q_i}\Big[\sigma^2+\Big\{\sum_{j=0}^{n}(b_i(x_k)a_{i-1,j}+\beta_i(x_k)a_{i,j})y^{(j)}+\frac{x^{(i)}-x_k}{h}y^{(i-1)}+\frac{x_k-x^{(i-1)}}{h}y^{(i)}\Big\}^2\Big]$$

$$=N\sigma^2+\sum_{r=0}^{n}\sum_{j=0}^{n}V_{r,j}y^{(r)}y^{(j)}+\sum_{r=0}^{n}\sum_{j=1}^{n}W_{r,i-1}y^{(r)}y^{(i-1)}$$

$$+\sum_{r=0}^{n}\sum_{i=1}^{n}U_{r,i}y^{(r)}y^{(i)}+\sum_{i=1}^{n}\sum_{j=0}^{n}C_{i-1,j}y^{(i-1)}y^{(j)}$$

$$+\sum_{i=1}^{n}\sum_{j=0}^{n}Z_{i,j}y^{(i)}y^{(j)}+\sum_{i=1}^{n}E_{i-1}(y^{(i-1)})^2+\sum_{i=1}^{n}R_i(y^{(i)})^2$$

$$+2\sum_{i=1}^{n}B_{i,i-1}y^{(i)}y^{(i-1)}$$

$$=N\sigma^2+\sum_{r=0}^{n}\sum_{m=0}^{n}t_{r,m}y^{(r)}y^{(m)}.$$

The expectation of the second term is

$$E\Big[\sum_{r=0}^{n}\hat{y}^{(r)}g_r\Big]=E[\hat{\mathbf{y}}^t\mathbf{g}]=E[\hat{\mathbf{y}}^tT\hat{\mathbf{y}}]$$

$$=E\Big[\sum_{r=0}^{n}\sum_{m=0}^{n}\hat{y}^{(r)}t_{r,m}\hat{y}^{(m)}\Big]=\sum_{r=0}^{n}\sum_{m=0}^{n}t_{r,m}E[\hat{y}^{(r)}\hat{y}^{(m)}]$$

$$=\sum_{r=0}^{n}\sum_{m=0}^{n}t_{r,m}\{Cov[\hat{y}^{(r)},\hat{y}^{(m)}]+y^{(r)}y^{(m)}\}$$

$$=\sum_{r=0}^{n}\sum_{m=0}^{n}t_{r,m}\{d_{r,m}\sigma^2+y^{(r)}y^{(m)}\}$$

$$=(n+1)\sigma^2+\sum_{r=0}^{n}\sum_{m=0}^{n}t_{r,m}y^{(r)}y^{(m)} \quad (\because D=T^{-1}).$$

Therefore we have

$$E[S_m]=E\Big[\sum_{i=0}^{n}\sum_{k=p_i}^{q_i}f_k^2\Big]-E\Big[\sum_{r=0}^{n}y^{(r)}g_r\Big]$$
$$=N\sigma^2-(n+1)\sigma^2=(N-n-1)\sigma^2,$$

which complete the proof.

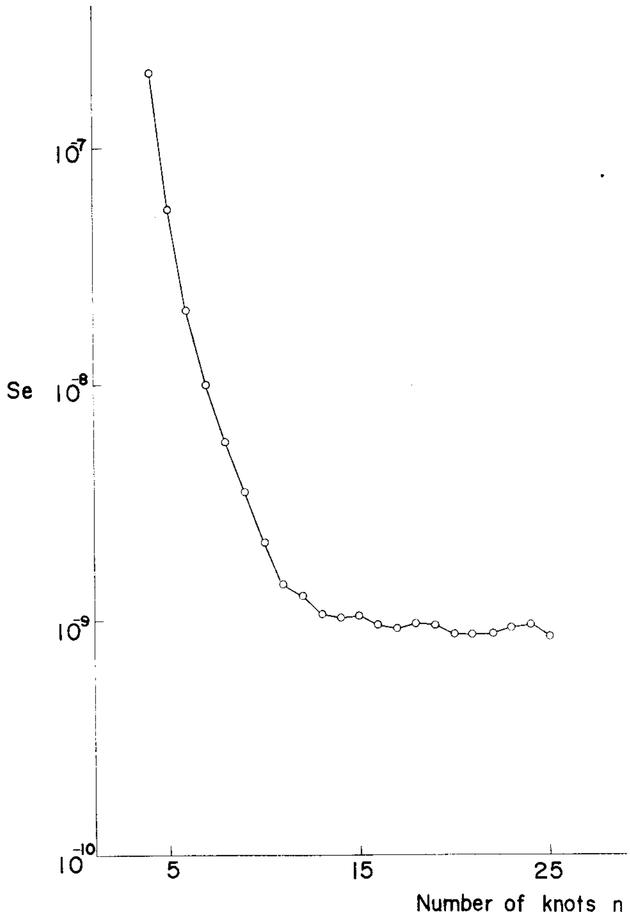

Fig. 2. Graph of Se of example 1.

In practice, since we do not yet know the appropriate n, we would solve the normal equation (37) for $n=1, 2,\ldots,$ and compute S_e, to continue as long as S_e decreases significantly with an increasing n. As soon as a value of n is reached after which no significant decrease occurs in S_e, then we use this n for the desired least squares approximation.

5. Numerical Examples

Example 1 The data used for this example are values of $\sin x$ rounded to 4 decimal places, for $x_k=0°(1°)40°$. In Fig. 2 we plot S_e against $n(x^{(0)}=0°, x^{(n)}=40°)$, and for $n=14$ $S(x)$ is drawn in Fig. 3.

Example 2 This example is due to Powell [2]. The data is given by

$$f_k=\frac{1}{0.01+(x_k-0.3)^2}+\varepsilon_k$$

for $x_k=0.005(0.01)0.995$. The ε_k is an error which is independent for each x_k. Its expectation and variance are 0 and 4 respectively. The figure of Se is plotted in Fig. 4 ($x^{(0)}=0$, $x^{(n)}=1.0$) and $S(x)$ is drawn for $n=11$ in Fig. 5.

These numerical examples were done with FACOM 230-60 of Data Processing Center of Kyoto University.

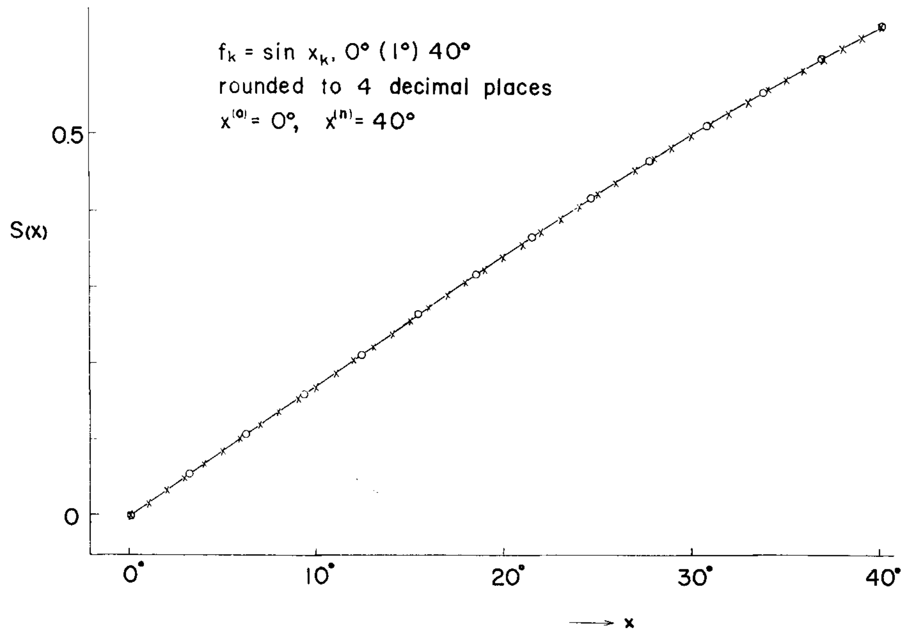

Fig. 3. Computed result of example 1.

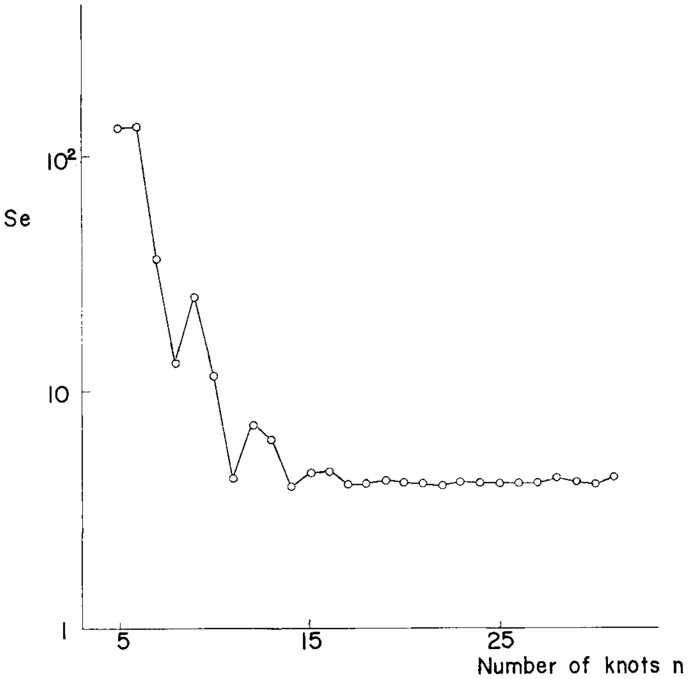

Fig. 4. Graph of *Se* of example 2.

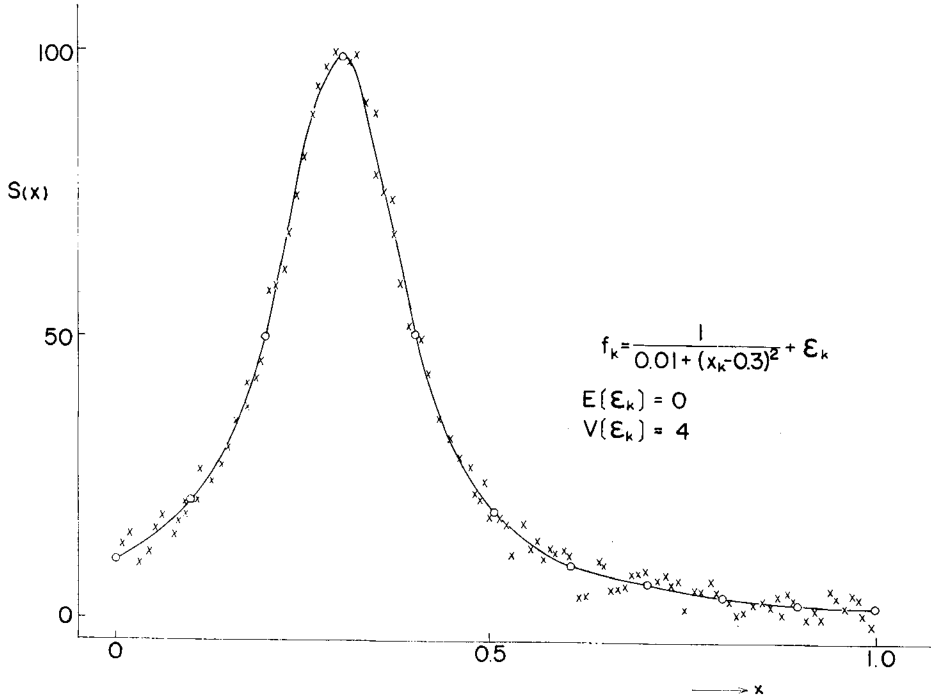

Fig. 5. Computed result of example 2.

References

[1] J. H. Ahlberg et al: The Theory of Splines and Their Applications, Academic Press, New York, 1967.
[2] M. J. D. Powell: Curve Fitting by Splines in One Variable, Numerical Approximation to Functions and Data, J. G. Hayes, ed., Athlone Press, London, 1970, pp. 65–83.
[3] L. B. Smith: The Use of Man- machine Interaction in Data-fitting Problems, Technical Report No. CS 131, Stanford University, 1969.