Curve Fitting by a One-Pass Method With a Piecewise Cubic Polynomial

KOZO ICHIDA and TAKESHI KIYONO
Kyoto University, Japan
FUJIICHI YOSHIMOTO
Akashi Technical College, Japan

A method is described for fitting a piecewise cubic polynomial to a sequence of data by a one-pass method. The polynomial pieces are calculated as the data is scanned only once from left to right. The algorithm is shown to be stable when a piecewise cubic polynomial is used which is continuous with its first derivative, while it becomes unstable if the polynomial is made continuous up to the second derivative. The knots of the approximating function are determined successively using a criterion by Powell.

Key Words and Phrases. curve fitting, least squares, one-pass method, piecewise cubic polynomial, stability
CR Categories: 5 19

1. INTRODUCTION

Piecewise polynomials have become more widely used lately in least-squares data fitting since they are more flexible than simple polynomials. Most of the methods assume that the beginning and the end of the data are fixed. However, for on-line systems, it is frequently necessary to follow and fit the data without knowing its end beforehand. The one-pass method, in which the polynomial pieces are calculated as the data is scanned only once from left to right, is useful for such cases. Rice [3, 4] considered an orthogonalization process for the one-pass method in which the least-squares approximation to the data is required to have an error within a prescribed bound. However, it is difficult to determine the tolerance when the magnitude of the error is not known beforehand.

In this paper we discuss curve fitting by a one-pass method with a piecewise cubic polynomial. For the one-pass method to be feasible, the algorithm must be stable, as is the step-by-step method for solving ordinary differential equations. The one-pass algorithm described here is shown to be stable when the fitting function is continuous only up to the first derivative, while it becomes unstable if the function is continuous up to the second derivative.

Authors' addresses: K. Ichida and T. Kiyono, Department of Information Science, Kyoto University, Kyoto, Japan; F. Yoshimoto, Department of General Studies, Akashi Technical College, Uozumicho, Akashi, Japan.

The knots of the approximating function are determined successively using a criterion given by Powell [2] to extract almost all the information from the data. To meet the requirement of data compression and smoothness of approximation, the number of knots is made as small as possible while still satisfying the criterion. Furthermore, a device is introduced to reduce the amount of computation. Here, it is not necessary that the magnitude of error be known in advance.

2. LEAST-SQUARES FITTING WITH A PIECEWISE CUBIC POLYNOMIAL

Suppose we have the data expressed by

$$F_k = f(x_k) + \epsilon_k \quad (k = 1, 2, \ldots), \tag{1}$$

where $f(x)$ is an unknown function, ϵ_k is an error having a mean of zero and variance σ^2 (less than ∞), and x_k is the data point. The errors ϵ_k are assumed mutually independent. First we consider least-squares fitting with a piecewise cubic polynomial which is continuous with its first derivative. Let us refer to this as "one-pass method 1."

A piecewise cubic polynomial which can be used for one-pass method 1 is given in [1, p. 12]. However, since the form displayed there is not convenient for examining stability, we employ the following function. In the interval $I = [s, t]$ the approximating function is in the form

$$S(x) = y_0 + m_0(x - s) + a(x - s)^2 + b(x - s)^3, \tag{2}$$

where s and t denote knots, and $y_0 = S(s)$ and $m_0 = S'(s)$ are values of $S(x)$ and of its derivative at s, respectively. The constants a and b are determined by the least-squares calculation. From eq. (2), the condition for continuity of $S(x)$ and its derivative at t is

$$\begin{aligned} y &\equiv S(t) = y_0 + m_0 h + ah^2 + bh^3, \\ m &\equiv S'(t) = m_0 + 2ah + 3bh^2, \end{aligned} \tag{3}$$

where $h = t - s$.

We consider the following expression for the sum of the squares of the residuals (see Figure 1):

$$Q = \sum_{x_k \in \Delta I + I} [S(x_k) - F_k]^2 + w \sum_{x_k \in \Delta I_1 + I_1} [S_1(x_k) - F_k]^2. \tag{4}$$

In this equation w is a weighting factor, and

$$S_1(x) = y + m(x - t) + a_1(x - t)^2 + b_1(x - t)^3 \tag{5}$$

is a tentatively calculated function in $I_1 = [t, t_1]$ (with t_1 chosen so that there is the same number of data in each interval I and I_1). a_1 and b_1 are constants. In eq. (4) we consider future and past data in the calculation, where the data in the interval I_1 are the future data and those in ΔI are the past data. Without the future data the approximating function becomes unstable as fitting proceeds. This situation results from the fact that we have determined m, the derivative at the right end of the interval I, from the data in the interval alone. [This corresponds to the case of $w = 0$ in Fig. 2 (see Section 3)]. The data in ΔI and ΔI_1 in eq. (4) con-

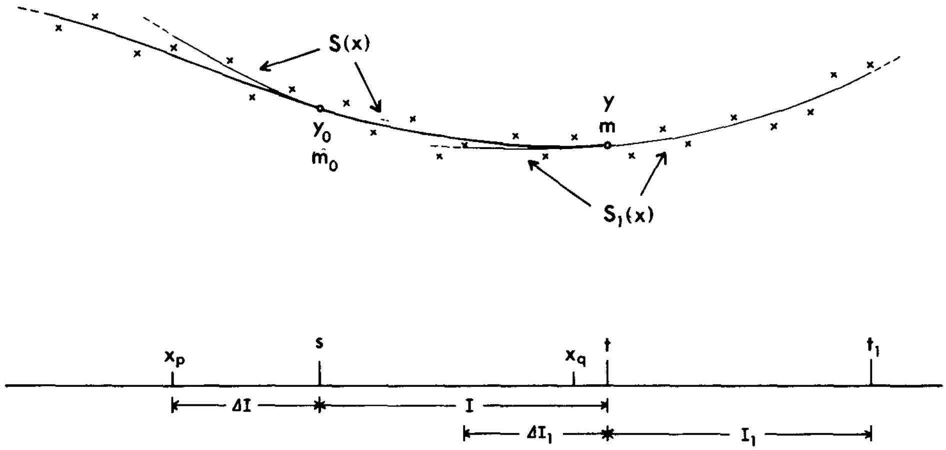

Fig. 1. Least-squares fitting with a piecewise cubic polynomial using future and past data; s and t denote the knots, and y and m are the values of the approximating function and its derivative at t, respectively.

tribute to the smoothness of the approximating function. We can differentiate Q in eq. (4) with respect to the parameters a, b, a_1, and b_1, then set the result equal to zero to obtain the normal equation, and thus determine a and b, and then y and m by eq. (3).

Some applications use the smoother cubic spline where the second derivative is also continuous. Let us refer to the one-pass method with a cubic spline as "one-pass method 2." In this case, the approximating function is expressed in the interval I as

$$R(x) = y_0 + m_0(x - s) + \tfrac{1}{2}M_0(x - s)^2 + a(x - s)^3, \tag{6}$$

where $y_0 = R(s)$, $m_0 = R'(s)$, $M_0 = R''(s)$, and a is a constant. $R_1(x)$, which corresponds to $S_1(x)$, is given as

$$R_1(x) = y + m(x - t) + \tfrac{1}{2}M(x - t)^2 + a_1(x - t)^3. \tag{7}$$

The conditions of continuity of the approximating function and its first and second derivatives are

$$\begin{aligned} y &\equiv R(t) = y_0 + m_0h + \tfrac{1}{2}M_0h^2 + ah^3, \\ m &\equiv R'(t) = m_0 + M_0h + 3ah^2, \\ M &\equiv R''(t) = M_0 + 6ah. \end{aligned} \tag{8}$$

The sum of the squares of the residuals is

$$Q = \sum_{x_k \in \Delta I + I} [R(x_k) - F_k]^2 + w \sum_{x_k \in \Delta I_1 + I_1} [R_1(x_k) - F_k]^2, \tag{9}$$

where the suffixes are the same as in eq. (4). We can differentiate eq. (9) with respect to a and a_1, set the result equal to zero, and thus determine a. The quantities y, m, and M are determined by eq. (8).

3. STABILITY

In order to investigate the stability of the methods, we need to know the relation between y, m and y_0, m_0 for one-pass method 1 or that between y, m, M and y_0, m_0, M_0 for one-pass method 2. These relations are found from eqs. (3) or (8) mentioned in Section 2. In general, the obtained difference equations have variable coefficients and their stability cannot be determined easily.

Hence we consider the ideal case where the knots and data points are at equal intervals and the data are dense enough so that the summations can be replaced by integrals. Then, for one-pass method 1, eq. (4) becomes

$$Q = \int_{s-\alpha h}^{t} [S(x) - F(x)]^2 \, dx + w \int_{t-\alpha h}^{t_1} [S_1(x) - F(x)]^2 \, dx, \tag{10}$$

where $[s - \alpha h, s]$ and $[t - \alpha h, t]$ denote ΔI and ΔI_1, respectively. We obtain the normal equation as before and substitute the parameters a and b into eq. (3) to obtain the difference equation

$$\begin{bmatrix} y \\ m \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12}h \\ c_{21}/h & c_{22} \end{bmatrix} \begin{bmatrix} y_0 \\ m_0 \end{bmatrix} + \begin{bmatrix} u \\ v \end{bmatrix}, \tag{11}$$

where c_{ij} $(i, j = 1, 2)$ are constants that depend on w and α but not on h; and u and v are independent of y_0 or m_0.

The stability of the system (11) is determined by the eigenvalues of this system or by the roots λ_1 and λ_2 of the quadratic equation

$$\lambda^2 - (c_{11} + c_{22})\lambda + c_{11}c_{22} - c_{12}c_{21} = 0. \tag{12}$$

These eigenvalues do not depend on h. Figure 2 is a graph of $\max_{1 \le i \le 2} | \lambda_i |$ versus w, with parameter α. This graph shows that the system (11) is stable if $w > 0.2$. However a smaller magnitude of $\max | \lambda_i |$ does not always give a better fit. It was found by numerical experiments that $w = 1$ and $\alpha = 0.5$ give satisfactory results.

For one-pass method 2, eq. (9) becomes

$$Q = \int_{s-\alpha h}^{t} [R(x) - F(x)]^2 \, dx + w \int_{t-\alpha h}^{t_1} [R_1(x) - F(x)]^2 \, dx. \tag{13}$$

We obtain the normal equation as before and substitute a into eq. (8) to obtain the difference equation

$$\begin{bmatrix} y \\ m \\ M \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12}h & d_{13}h^2 \\ d_{21}/h & d_{22} & d_{23}h \\ d_{31}/h^2 & d_{32}/h & d_{33} \end{bmatrix} \begin{bmatrix} y_0 \\ m_0 \\ M_0 \end{bmatrix} + \begin{bmatrix} u' \\ v' \\ w' \end{bmatrix}, \tag{14}$$

where d_{ij} $(i, j = 1, 2, 3)$ are constants depending on w and α but not on h; and u', v', and w' are independent of y_0, m_0, or M_0. Again the eigenvalues λ_i $(i = 1, 2, 3)$ of this system do not depend on h. The graph of $\max_{1 \le i \le 3} | \lambda_i |$ versus w, with α as the parameter, is shown in Figure 3. The value of $\max | \lambda_i |$ decreases with increasing w. However, it is much greater than unity (for $w > 5$ $\max | \lambda_i |$ is nearly constant), and the system (14) is unstable. Accordingly, one-pass method 2 for cubic splines is not suitable for curve fitting.

For the case where the knots and data points are equally spaced but the sum-

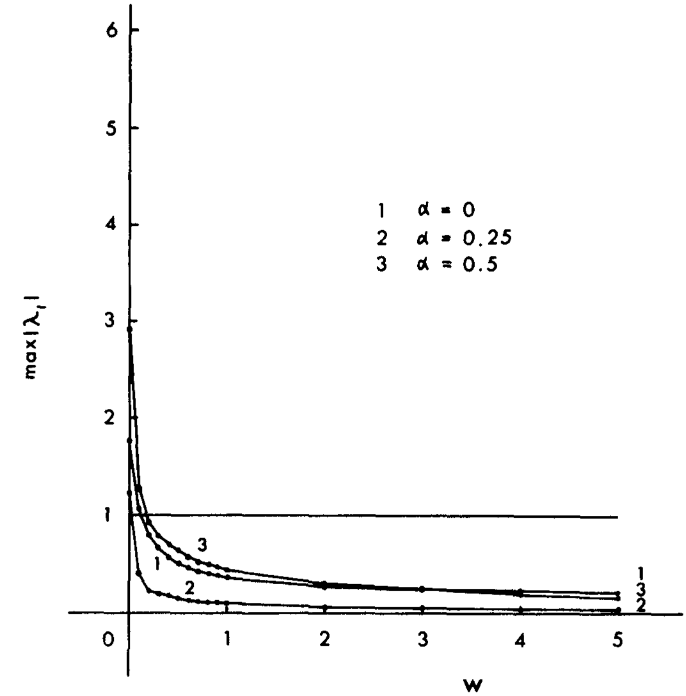

Fig. 2. Graph of the maximum of the absolute of the eigenvalues of eq. (11) versus w, with parameter α.

mations cannot be replaced by integrals, figures similar to Figures 2 and 3 can be obtained. In this case the systems corresponding to eqs. (11) and (14) become constant-coefficient systems, and the stability is similar to that for the continuous case above.

In other cases, eqs. (11) and (14) become variable-coefficient systems, and the stability cannot be determined from the eigenvalues alone. However, several numerical experiments have shown that the graphs of max $|\lambda_i|$ are not much different from those of Figures 2 or 3 (the eigenvalues are different in each interval), and it appears that one-pass method 1 is stable, while one-pass method 2 is unstable for these cases.

4. CRITERION FOR FITTING AND DETERMINATION OF KNOTS

It is necessary to have some criterion for fitting in order to determine the locations of knots. One may proceed as follows: (1) make h as large as possible within the limit that the sum of the squares of the residuals be less than a prescribed bound, and (2) use an unbiased estimator of variance of errors. Both these methods require that the magnitude of error be known in advance. In one-pass methods, however, the fitting has to be done automatically, and therefore the above criteria are not adequate. Rice [3] introduced a method that made use of the error curve, but it is quite complicated. We use the criterion of Powell [2], called the "trend" for data fitting, where knots are inserted successively until a satisfactory fit is

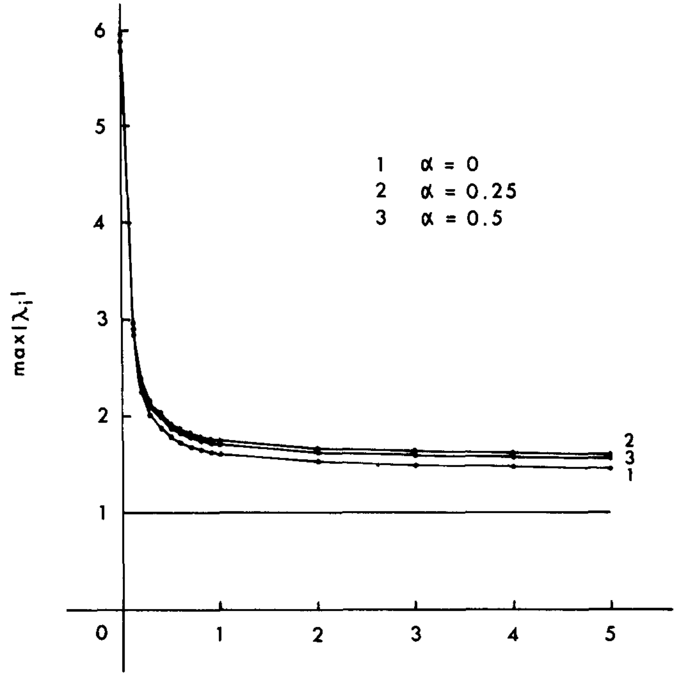

Fig. 3. Graph of the maximum of the absolute of the eigenvalues of eq. (14) versus w, with parameter α.

obtained. An advantage of this criterion is that it is applicable even if the magnitude of error is unknown; essentially, it attempts to extract all the information from the data.

To understand this criterion, consider the interval $\Delta I + I$. The situation is shown in Figure 1. The residuals r_k are

$$r_k = S(x_k) - F_k \quad (k = p, \ldots, q). \tag{15}$$

The objective of the trend criterion is to increase the accuracy of the fit until the residuals behave as random variables, in which case we assume that the piecewise cubic polynomial is a good approximation to the underlying smooth function $f(x)$. The criterion is expressed as follows [2]:

$$T \geq \sqrt{(q-p)\sum_{k=p}^{q} r_k^2/(q-p+1)}, \tag{16}$$

where

$$T = \sum_{k=p+1}^{q} r_{k-1} r_k \,. \tag{17}$$

If inequality (16) is satisfied, we infer that there is a trend in this interval, which means the fitting is not satisfactory. Otherwise we search for the most adequate approximation among the fittings that do not satisfy the criterion.

For a one-pass method it is necessary that there exist an h which has no trend. Therefore we replace expression (16) by the stronger test

$$T \geq \sum_{k=p}^{q} r_k^2 / \sqrt{(q - p)} \equiv U. \tag{18}$$

It is easily seen that the inequality $T \leq U$ in eq. (18) always holds when $q - p \leq 2$ [2]; thus we can make sure that there exists an h which has no trend. Usually $(q - p)$ is much larger than 2, and there is little difference between eqs. (16) and (18).

The knot t is determined so that h is as large as possible, using the trend criterion. Since this meets the requirement of extracting all of the information from the data, we accept the approximation on this interval. Thus we compute the least-squares approximation with eq. (4), increasing t as long as there is no trend.

To reduce the amount of calculation required, we apply the following scheme. Let μ be the number of data points in I. We seek the maximum of μ so that there will be no trend in the interval $\Delta I + I$ or in the next tentative interval, $\Delta I_1 + I_1$. Let the initial estimate of μ be $\mu^{(0)}$, and let the minimum variation of μ be unity. For the initial interval, $\mu^{(0)}$ is arbitrary; for other intervals, $\mu^{(0)}$ is taken to be the value of μ accepted in the previous interval. Note that we must require $\mu \geq 2$ if the normal equation is to have a solution. First we set $\mu = \mu^{(0)}$ and calculate the approximating function to see if there is a trend or not. Then we look for a suitable value of μ by use of the method of bisection for solving an equation, as follows. (Figure 4 illustrates the two cases where [i] there is not or [ii] there is a trend at $\mu^{(0)}$.) In eqs. (19) through (21) the constant K is set to be $+1$ or -1 depending on whether we are dealing with case [i] or [ii], respectively.

In case [i] we continue to increase μ by

$$\mu^{(j)} = \mu^{(j-1)} + K2^{j-1} \quad (j = 1, 2, \ldots, \theta_1), \tag{19}$$

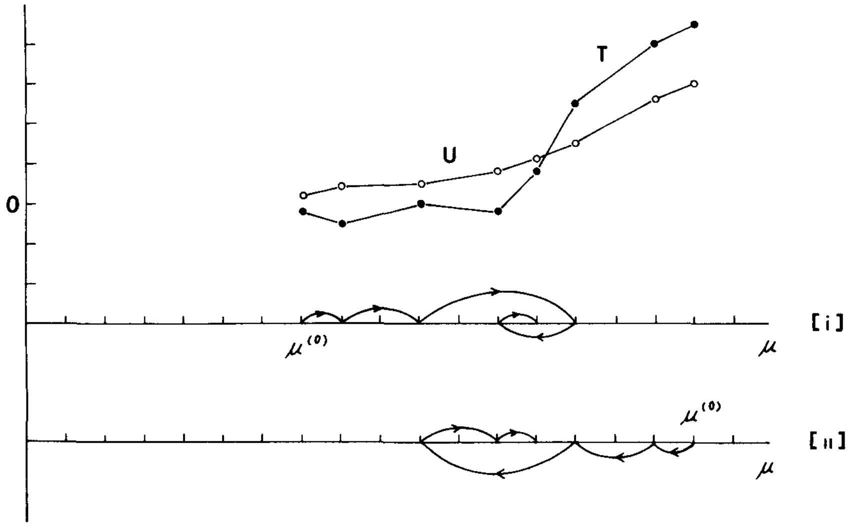

Fig 4. Determination of the location of a knot by the method of bisection using the trend criterion; μ denotes the number of data points in I, and $\mu^{(0)}$ is the initial estimate of μ.

until there is a trend at $j = \theta_1$. Next we continue to decrease μ by

$$\begin{aligned}\mu^{(0)} &\equiv \mu^{(\theta_1)}, \\ \mu^{(j)} &= \mu^{(j-1)} - K2^{(\theta_1-1)}/2^j \quad (j = 1, 2, \ldots, \theta_2),\end{aligned} \tag{20}$$

until there is no trend at $j = \theta_2$. Again we increase μ until there is a trend by

$$\begin{aligned}\mu^{(0)} &\equiv \mu^{(\theta_2)}, \\ \mu^{(j)} &= \mu^{(j-1)} + K[2^{(\theta_1-1)}/2^{(\theta_2)}]/2^j \quad (j = 1, 2, \ldots, \theta_3).\end{aligned} \tag{21}$$

Continuing in this manner, we alternately increase and decrease μ L times until the change in μ is 1 or 0. This gives the desired value of μ, namely, the largest one possible which is not a trend. Experience indicates that L is usually less than or equal to 3. Case [ii] is treated similarly.

This algorithm can be used for irregular data as well as for regular data. One must be careful in choosing α, the proportion of past data. If it is too large, μ becomes too small. If it is too small, smoothness of approximation is lost. Numerical experiments indicate that $\alpha \simeq 0.5$ is satisfactory. The knot t is taken to be the midpoint of the interval between data points where a trend first appears.

5. ALGORITHM

The actual algorithm is summarized as follows.

Step 1. We calculate y and m at the first knot $s = x^{(0)}$ (see Figure 5). We set $x^{(0)}$ as the coordinate that we wish to begin fitting, and fit a cubic polynomial $P(x)$ to data (x_k, F_k), $k = 1, 2, \ldots, n$. The sum of the squares of the residuals is

$$Q = \sum_{k=1}^{n} [P(x_k) - F_k]^2. \tag{22}$$

By computing the parameters of $P(x)$ which minimize eq. (22), we obtain $y = P(x^{(0)})$, $m = P'(x^{(0)})$. The number of data points n in eq. (22) is determined using the trend criterion, which we express in the form

$$\sum_{k=2}^{n} r_{k-1} r_k \geq \beta \sqrt{(n-1)Q/n}, \tag{23}$$

where r_k is a residual and β is a constant. We look for the maximum value of n that does not satisfy eq. (23); $\beta \simeq 0.5$ seems satisfactory.

Step 2. We calculate the fitting function for the first interval $[s, t] = [x^{(0)}, x^{(1)}]$. Since no past data is available in this interval, we set $\Delta I = \Delta I_1 = \phi$ in eq. (4).

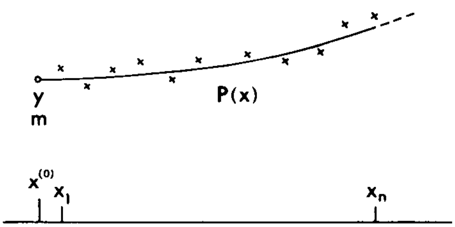

Fig. 5. Calculation of y and m at the first knot using a cubic polynomial $P(x)$.

The values of y and m are calculated from eq. (3). The knot $x^{(1)}$ is determined so that there is no trend in I or I_1, using the criterion (18).

Step 3. We calculate y and m for each interval after the first interval. The knot is determined so that there is no trend in $\Delta I + I$ or in $\Delta I_1 + I_1$, using the criterion (18).

Step 4. If t_1 is greater than or equal to the last data point x_N, we set t_1 as the coordinate that we wish to finish fitting. Then t is taken so that there are nearly the same number of data points in I and I_1, and $S_1(x)$ is accepted as the approximating function of the last interval. In this case we express the trend criterion as

$$T \geq \gamma U, \tag{24}$$

where γ is a constant. Since we have no data points for $x > x_N$, we cannot increase t arbitrarily. Instead, we increase the number of past data points, while at the same time satisfying eq. (24) to keep the approximation smooth. Numerical experiments show that $\gamma \simeq 0.5$ is satisfactory.

In the actual calculations involved in one-pass methods, a shift of the origin is desirable for accuracy of computation.

6. NUMERICAL EXAMPLE

Consider the data

$$F_k = [0.01 + (x_k - 0.3)^2]^{-1} + [0.015 + (x_k - 1.2)^2]^{-1} + \epsilon_k, \tag{25}$$

where ϵ_k is the normally distributed error with mean 0 and variance 1. The number of data points is taken to be 200. As an example of equally spaced data, assume that x_k are given as 0.005(0.01)1.995. Figure 6 shows the result of fitting this data by

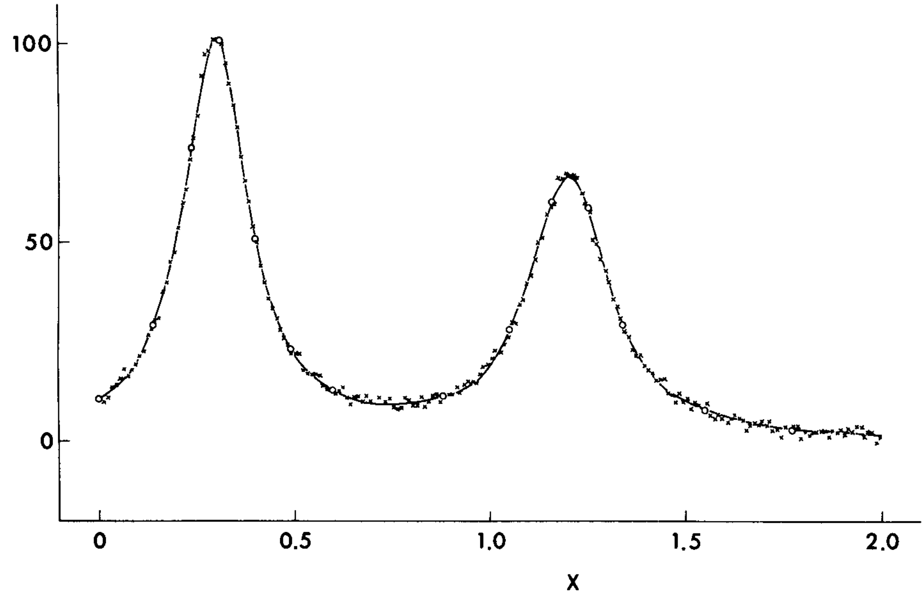

Fig. 6. Result of fitting of the example by one-pass method 1 for equally spaced data; the small circles show the joining points of the polynomial pieces.

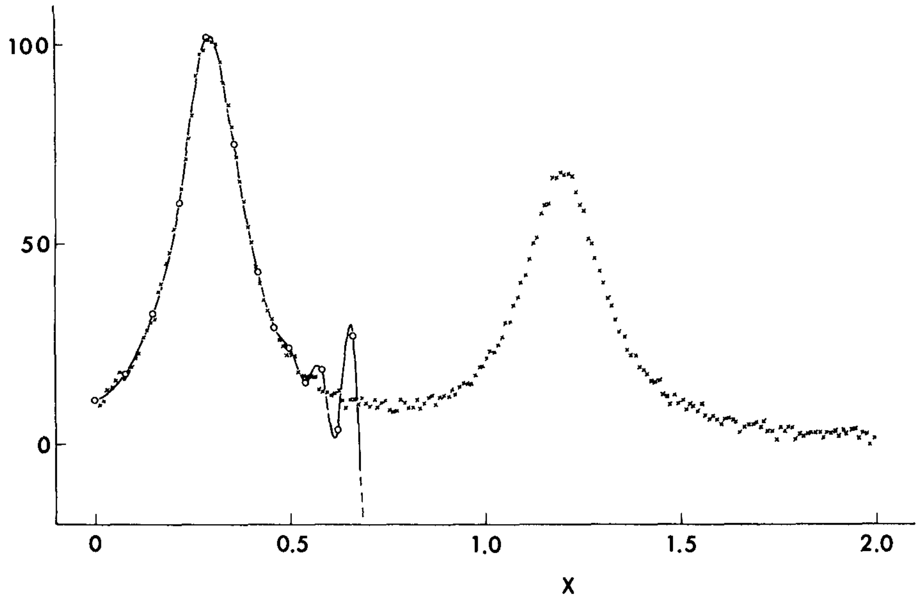

Fig. 7. Result of fitting of the example by one-pass method 2 for equally spaced data; the small circles show the joining points of the polynomial pieces.

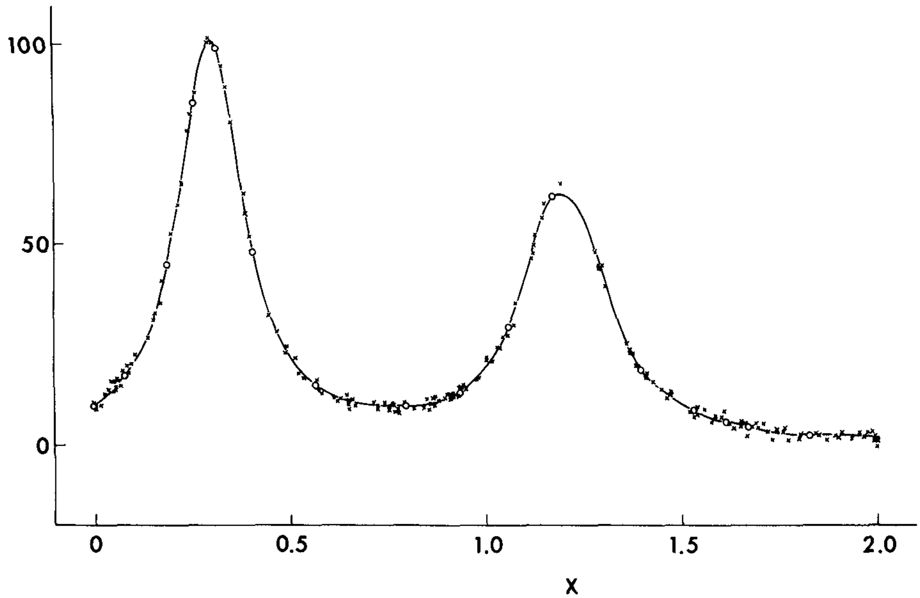

Fig. 8. Result of fitting of the example by one-pass method 1 for irregularly spaced data; the small circles show the joining points of the polynomial pieces.

using one-pass method 1, and Figure 7 shows the result of fitting the data by using one-pass method 2. Note that in the latter case the system is unstable and oscillates heavily as the fitting proceeds.

As an example of irregularly spaced data, assume that x_k are given by uniformly distributed random numbers in [0, 2], so that the coordinates of the data are totally random. Figure 8 shows the result of fitting the data by using one-pass method 1. For one-pass method 2, a result similar to that shown in Figure 7 is obtained. In all these calculations we set $\alpha = 0.5$ as the proportion of past data; for the other parameters we set $\beta = \gamma = 0.5$ (see Section 5).

7. CONCLUSION

A one-pass method for curve fitting has been described in which the data is scanned only once from left to right. A piecewise cubic polynomial is used for the approximating function. By examining the stability of the method, it is seen that the method is stable if the polynomial is C^1, while it is unstable if the polynomial is C^2. Further study is necessary for one-pass methods with piecewise polynomials of degree other than three, and with piecewise polynomials having a different degree of continuity at each knot.

REFERENCES

1. AHLBERG, J.H., NILSON, E.N., AND WALSH, J.L. *The Theory of Splines and Their Applications.* Academic Press, London, 1967.
2. POWELL, M.J.D. Curve fitting by splines in one variable. In *Numerical Approximation to Functions and Data,* J.G. Hayes, Ed., Athlone Press, London, 1970, pp. 65–83.
3. RICE, J.R. General purpose curve fitting. In *Approximation Theory,* A. Talbot, Ed., Academic Press, London, 1970, pp. 191–204
4. RICE, J.R. Running orthogonalization. *J. Approx. Theory 4* (1971), 332–338.

Received May 1974; revised October 1976