An Adaptive Algorithm for Fitting with Splines

T. M. Tao, A. T. Watson
Department of Chemical Engineering
Texas A&M University
College Station, TX 77843

In many situations, a parametric representation of discrete measured data may be desired. These include data smoothing and prediction and estimation of derivatives or integrals for rate or parameter estimation (Churchill, 1979). The accurate estimation of derivatives is particularly difficult since small measurement errors tend to be amplified by the process of differentiation (Anderson and Bloomfield, 1974). It is important to have a numerical method which is capable of generating accurate representations of response functions from measured data.

Splines have been widely used to approximate response functions from measured data (Klaus and Van Ness, 1967; Guttmann, 1975; Read and Dunfield, 1972; Wold, 1971). A spline is defined as a collection of many small piecewise polynomial segments with every two consecutive pieces joining smoothly at the knots; the function as well as some of its derivatives are required to be continuous across these knots. The precision of spline approximations depends mainly on the choice of the number of knots and their locations. It might be expected that we can approximate the true (but unknown) response function as closely as possible by adding more knots. For measured data, however, this is not necessarily true. As the number of knots increases, deviations between spline estimates and measurements will decrease accordingly. However, as the dimension of the spline approaches the number of data, the accuracy of the function approximation decreases since the function will tend to reflect the errors associated with the data measurement. Thus, we must limit the number of knots used in approximations to avoid the problem of overfitting (Draper and Smith, 1981). For a specified number of knots, the accuracy of spline estimates is also influenced by the distributions of knots (de Boor, 1978).

While the selection of knots (both the number of knots and their respective locations) may be vital to the success of spline approximations, they are usually chosen by a trial and error process. Solutions so obtained are heavily dependent on experience, and they may not provide the optimal choice. Therefore, it is desirable to choose the knots in a specific and, hopefully, optimal manner. For a fixed number of knots, de Boor and Rice (1968) and Jupp (1978) introduced a variable-knots approach wherein the positions of knots are chosen to minimize a least-squares objective function. While the variable-knots approach is promising in reducing the approximation error, studies showed that the solution is sensitive to the initial knot locations; a local minimum rather than a global minimum may be obtained if this initial guess is far from the expected optimal solution (de Boor and Rice, 1968; Jupp, 1978).

Recently, we have developed a new adaptive variable-knots algorithm (Tao, 1987). Both the number of knots and their locations are determined by the algorithm. It is not necessary to provide an initial guess of the number of knots and their respective locations. A systematic procedure is introduced to select them automatically; trial and error is no longer needed to proceed with spline approximation. If the number of knots is specified and fixed during computations, a unique global optimal solution is expected. Moreover, no prior experience or knowledge of splines is required for the usage of this algorithm.

Spline Approximations

It is well recognized that spline functions can be represented efficiently through the use of some linearly independent basic spline elements called B-splines (de Boor, 1978; Schumaker, 1981). Let the domain of interest be $a \le x \le b$; it is further divided into $k + 1$ subintervals with locations of the k interior joints or knots being denoted as $t = [t_i \mid i = 1, \ldots, k; a < t_i < b]$. Then, a spline function $S(x)$ of order m defined on this domain may be represented as a linear combination of a finite number of mth order B-splines:

$$S(x) = \sum_{i=1}^{n} C_i B_i^m(x, t) \tag{1}$$

where C_i are the representation coefficients. B-splines, $B_i^m(x, t)$, are determined uniquely by the order of spline and the knots. The degree of continuity of $S(x)$ at these knots is controlled by a multiplicity vector $e = [e_i \mid i = 1, \ldots, k; 1 \le e_i < m]$; the func-

Correspondence concerning this paper should be addressed to A. T. Watson.

tion and its first $m - e_i - 1$ derivatives are continuous across the ith knot. The dimensionality or the number of linearly independent B-splines defined on t is determined by $n = m + \sum_{i=1}^{k} e_i$. Further details on the theory and implementation of B-splines may be found elsewhere (de Boor, 1978; Cox, 1972; Schumaker, 1981).

$S(x)$ is linear in these representation coefficients C_i, while the knots t_i are related to $S(x)$ nonlinearly. For a given set of data $[(x, f)_i | i = 1, \ldots, np;\ a \le x \le b]$, the approximating spline function is determined by choosing the unknown coefficients and knots. This may be done by interpolation, whereby the spline estimates $S(x_i)$ represent the data f_i exactly, or by smoothing, wherein the unknowns are chosen to minimize some error function (generally the least-squares function). When dealing with measured data, the latter approach is preferred from a statistical view point.

Fixed-knots algorithm

If the number of knots and their locations are specified and fixed during calculations, the remaining unknown coefficients $C = [C_i | i = 1, \ldots, k]$ are then chosen as those that minimize the sum of squared residuals (i.e., a least-squares objective function):

$$F(C) = [f - GC]^T [f - GC] \tag{2}$$

This is a linear least-squares problem wherein elements of the $np \times n$ matrix G are given by $G_{ij} = B_j^m(x_i, t)$. The residuals could be weighted to reflect relative accuracies in measuring these data, as in weighted least-squares (Draper and Smith, 1981); here, we assume equal weight for convenience.

Solution of Eq. 2 can be obtained by solving the well-known normal equations. Nevertheless, the explicit formation of the coefficient matrix of the normal equations (i.e., G^TG) can cause the loss of twice as many significant figures as is necessary (Gill et al., 1981); consequently, an orthogonal decomposition method which does not require formation of the matrix G^TG is preferred for accuracy and numerical stability (Lawson and Hanson, 1974). The fixed-knots algorithm is computationally straightforward. The accuracy of spline estimates depends mainly on the proper placement of knots.

Variable-knots algorithm

It is known that the precision of spline approximations improves greatly if the knots are also free parameters (de Boor and Rice, 1968; Jupp, 1978; and Brannigan, 1981). In other words, we look for C and t that minimize the following objective function:

$$F(C, t) = [f - GC]^T [f - GC] \tag{3}$$

This is a highly nonlinear least-squares problem and can only be solved iteratively. The two sets of unknowns (i.e., C and t) are disjointed; C can be determined alone by the fixed-knots approximation once t is specified. Thus, we consider $C(t)$ and write Eq. 3 as:

$$F(t) = [f - GC(t)]^T [f - GC(t)] \tag{4}$$

With a fixed number of knots, de Boor and Rice (1968) presented an interesting variable-knots algorithm for least-squares fitting by cubic splines. In that approach, each knot is, in turn, varied so as to minimize the sum of squared residuals as a function of this knot. This process is continued until termination criteria are satisfied. To improve the efficiency, Jupp (1978) and Brannigan (1981) used the Gauss-Newton method to solve Eq. 4. In those previous studies, the number of knots used to construct spline functions and an initial distribution of these knots is specified in advance. While the variable-knots approach is promising in improving the accuracy of spline approximations, some serious problems for the variable-knots approximations have been the existence of (potentially) many stationary points of the nonlinear least-squares objective function and the dependence of a final solution on the initial guess of knot locations (de Boor and Rice, 1968; Jupp, 1978). Additionally, it is not an easy task to determine the appropriate dimensionality, or the number of knots, of a model for an arbitrary set of data (Akaike, 1974; Schwarz, 1978). Most people choose the number of knots and their initial locations based on intuition or personal experience.

Adaptive variable-knots algorithm

We have developed a systematic procedure which selects the appropriate number of knots as well as their optimal locations automatically for approximating an arbitrary set of discrete measurements (Tao, 1987). Beginning with no knots, we increase the number of knots one at a time. At the kth iteration, the Levenberg-Marquardt method (Fletcher, 1971) is used to solve Eq. 4. Once the optimal solution for the k interior knots are found, a local sum of squared residuals SSR_j^k and a knot addition index KAI_j^k for those data that fall within the jth subinterval, $t_{j-1} \le x \le t_j$, are calculated according to:

$$\mathrm{SSR}_j^k = \sum_{i=1}^{n_j} [f_i - S(x_i)]^2 \quad t_{j-1} \le x_i \le t_j;\ t_0 = a,\ t_{k+1} = b \tag{5}$$

$$\mathrm{KAI}_j^k = \frac{\mathrm{SSR}_j}{\mathrm{TSSR}^k} + \frac{n_j}{np} \quad j = 1, \ldots, k+1 \tag{6}$$

where n_j is the number of data points in the jth subinterval and TSSR^k, the total sum of squared residuals for the kth iteration, is defined as the sum of all the $k + 1$ local sum of squared residuals. At the end of the kth iteration, an extra knot will be added to the middle of the subinterval having the largest KAI. This extra knot, in addition to the k optimal knots obtained from current iteration, will serve as the starting points for the next $(k + 1)$th iteration. This procedure is continued until either one of the following criteria is met:

$$\left(\frac{\mathrm{TSSR}}{np - n}\right)^k > \left(\frac{\mathrm{TSSR}}{np - n}\right)^{k-1}$$

$$\frac{\mathrm{TSSR}^0 - \mathrm{TSSR}^{k-1}}{\mathrm{TSSR}^0 - \mathrm{TSSR}^k} > 0.95 \tag{7}$$

where TSSR^0 refers to the TSSR for the no-knots case. The first criterion is a statistical measure. The appearance of overfitting is indicated by an increasing error variance. The second criterion is a measure of the cost we are willing to pay to achieve some specified degree of accuracy. In other words, the iterations terminate when the improvement of precision of the fit due to the addition of one more knot is insignificant.

As pointed out by de Boor and Rice (1968) and Jupp (1978),

an appropriate initial guess of the distribution of knots is vital to the success of variable-knots approximations. Many initial guesses may result in local, rather than global, minima. Our algorithm tends to avoid nonglobal minima. Instead of relying on the solution of a single large minimization problem from a chosen initial guess for the knot locations, our algorithm generates an initial guess that is close to the global solution by solving a sequence of smaller problems. Global minima can often be readily determined for minimization problems with few unknowns. Since at each step only a single knot is added, the optimal knot distribution for the previous problem often serves as a good initial guess for obtaining the global solution for the next problem. Thus, at each step we expect to obtain the global solution for that corresponding number of knots.

Results and Discussion

First, we consider the correlation of titanium heat data. Without explaining explicitly the physical nature of these measurements, de Boor and Rice (1968) presented a variable-knots approximation with cubic splines for a set of 49 data values that express a thermal property of titanium. As they reported, these data contain a significant amount of noise, or measurement error, and are not well approximated with polynomials. They used five interior knots to construct the spline approximations. For the purpose of comparison, we also used five interior knots to fit the titanium heat data. Our results, as well as those reported in previous studies, are summarized in Table 1 wherein the locations of knots and the sum of squared residuals (SSR) for corresponding spline approximations are presented.

In Case 1, equally-spaced knots are used. Equally-spaced knots are the simplest choice for data fitting. However, this case has the largest SSR among all the tested cases. Starting with equally-spaced knots, de Boor and Rice (1968) varied the locations of knots to improve the accuracy of approximation; their results are reported as Case 2. A significant reduction in the SSR is observed; the SSR is less than 5% of the corresponding value for Case 1. This result demonstrates the value of variable-knots approximation. Starting with another initial distribution of knots t^0 = (725, 850, 910, 975, 1,040, 1,075), de Boor and Rice (1968) repeated the calculation, and the results are presented as Case 3. The SSR is almost a tenth that of Case 2. Evidently, a local minimum is obtained for Case 2. Results of Cases 2 and 3 reveal the dependence of a final solution on the initial distribution of knots and the difficulty of locating the optimal locations of a given number of knots.

Jupp (1978) drew a similar conclusion from his findings. He showed that there exist four stationary points (i.e., one interior optimum, two local minima, and one saddle point) within the domain of interest. The interior optimum is obtained only when the starting point is chosen close to the expected optimal solution and it is reported as Case 4. Even for people who are familiar with spline approximations, there is generally no assurance that the provided initial guess is close enough to the optimal solution; consequently, a local minimum may be determined (e.g., Case 2).

As stated above, our algorithm chooses the knots automatically; initial guesses for the distribution of knots are generated in a way that tends to avoid convergence to a nonglobal optimum. Our results are shown as Case 5. This solution was obtained via a systematic scheme rather than a trial and error process, and it is slightly better than the global solution as reported by Jupp (1978) (i.e., Case 4). The small difference between Cases 4 and 5 may be due to the round-off error arising from the variable transformation proposed by Jupp (1978). The optimal spline fit together with the data are plotted in Figure 1.

For the second example, we consider the estimation of partial molar enthalpies from discrete measurements of excess enthalpies. This requires the estimation of the derivatives of measured data. Partial molar properties are important thermodynamic functions used to predict or describe behaviors of liquid mixtures. Let H^e denote the excess enthalpy measured on mixing components 1 and 2. Then, the partial molar enthalpies, $\overline{H}_1$ and $\overline{H}_2$, are calculated by (Smith and Van Ness, 1975):

$$\overline{H}_1 = H^e + (1 - y)\left(\frac{dH^e}{dy}\right)_{T,P}$$

$$\overline{H}_2 = H^e - y\left(\frac{dH^e}{dy}\right)_{T,P} \tag{8}$$

where y is the mole fraction of Component 1. Derivatives of H^e with respect to compositions at constant temperature and pressure are needed to evaluate $\overline{H}_1$ and $\overline{H}_2$. For a given set of measurements of H^e, a spline function is first fitted to these raw data, and then derivatives calculated directly from the resulting spline function are used to estimate $\overline{H}_1$ and $\overline{H}_2$ via Eq. 8.

In Figure 2, measurements of H^e for a binary liquid mixture of pentafluorobenzene, Component 1, and n-propylbenzene, Component 2, are plotted against a mole fraction of pentafluorobenzene (Skillerne de Bristowe and Stubley, 1973). Only ten data points are available. This measured excess enthalpy is an

Table 1. Knots for Variable-Knots Cubic Splines

	Case				
Knots	1	2 de Boor and Rice	3 de Boor and Rice	4 Jupp	5
t_1	675.0	755.3	835.3	835.9	832.4
t_2	755.0	839.6	876.6	876.4	876.2
t_3	835.0	877.1	902.5	898.2	899.6
t_4	915.0	896.2	910.5	916.3	915.2
t_5	995.0	910.2	977.9	974.0	976.0
SSR	1.5235	0.0618	0.0072	0.0061	0.0059

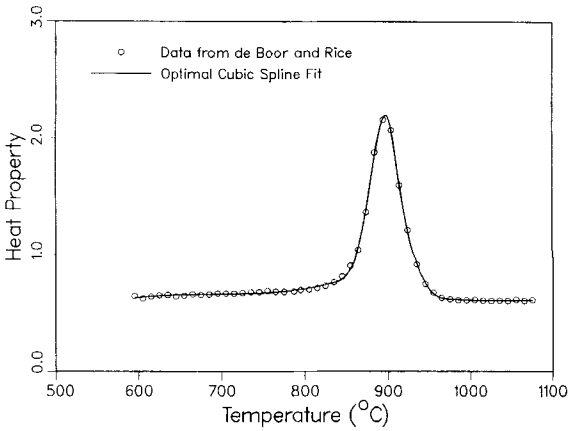

Figure 1. Heat property of titanium.

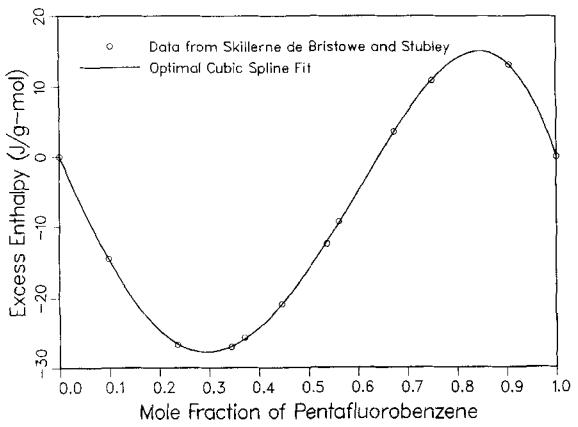

Figure 2. Excess enthalpy for mixtures of pentafluorobenzene with *n*-propylbenzene at 328 K.

S-shaped function with zero values at $y = 0$ and $y = 1$. Within the domain of interest, there are two extremum points (i.e., one maximum and one minimum) and one inflection point. We must approximate the response function as well as derivatives from these limited number of observations. The optimal cubic spline approximation for the excess enthalpy calculated by the adaptive variable-knots algorithm is also shown in Figure 2. The algorithm chose three interior knots at locations $t^* = (0.237, 0.577, 0.84)$. In Table 2, values of excess enthalpy and partial molar enthalpies estimated from the optimal spline approximation are compared with those reported by Skillerne de Bristowe and Stubley (1973). The partial molar enthalpies estimated from our variable-knots algorithm are very close to the published values. This example demonstrates that the adaptive variable-knots algorithm is also suitable for the estimation of derivatives from discrete measurements.

Notation

B_i^m = the ith B-spline or order m
C = B-spline representation coefficients
e = multiplicity vector
f = measured response function
F = objective function
G = coefficient matrix
H^e = excess enthalpy
$\bar{H}$ = partial molar enthalpy
k = number of interior knots
KAI = knot-addition index
m = order of spline function
n = dimension of spline space
n_j = number of data in the jth interval
np = number of measurements
S = spline approximating function
SSR = sum of squared residuals
t = locations of knots or time
TSSR = total sum of squared residuals
VAR = estimated variance
x = independent variable
y = mole fraction

Subscripts

T,P = at constant temperature and pressure

Superscript

T = transpose of matrix
$*$ = optimal solution

Table 2. Partial Molar Enthalpies for Mixtures of Pentafluorobenzene with *n*-Propylbenzene

Composition	Optimal Spline Fit			Skillerne de Bristowe and Stubley Results		
x	He	$\bar{H}_1$	$\bar{H}_2$	He	$\bar{H}_1$	$\bar{H}_2$
0.1	−14.7	−128.8	−2.0	−15.2	−127.0	−2.7
0.2	−24.6	−79.1	−11.0	−24.5	−73.9	−12.2
0.3	−27.8	−24.6	−29.1	−27.5	−25.9	−28.1
0.4	−24.3	12.5	−48.8	−24.3	11.2	−48.0
0.5	−16.1	33.7	−65.9	−16.2	34.1	−66.4
0.6	−5.0	42.4	−76.2	−5.0	42.0	−75.6
0.7	6.4	37.0	−65.1	6.3	37.0	−65.3
0.8	14.0	22.6	−20.3	14.1	23.2	−22.5
0.9	13.5	7.4	68.4	13.8	7.4	68.2
SSR		0.210			0.701	
VAR		0.030			0.078	

Literature Cited

Akaike, H., "A New Look at the Statistical Model Identification," *IEEE Trans. Automat. Contr.*, **AC-19**(6), 716 (1974).

Anderssen, R. S., and P. Bloomfield, "A Time Series Approach to Numerical Differentiation," *Technometrics*, **16**(1), 69 (1974).

Brannigan, M., "An Adaptive Piecewise Polynomial Curve Fitting Procedure for Data Analysis," *Comm. Statist.-Theor. Meth.*, **A10**(18), 1823 (1981).

Churchill, S. W., *The Interpretation and Use of Rate Data: The Rate Concept*, Hemisphere Publishing, Washington (1979).

Cox, M. G., "The Numerical Evaluation of B-splines," *Inst. Maths. Applics.*, **10**, 134 (1972).

de Boor, C., and J. R. Rice, "Least Squares Cubic Splines Approximation II-Variables Knots," CSD TR 21, Purdue Univ., Lafayette, IN (1968).

de Boor, C., *A Practical Guide to Splines*, Springer-Verlag, New York (1978).

Draper, N., and H. Smith, *Applied Regression Analysis*, 2nd ed., John Wiley and Sons, New York (1981).

Fletcher, R., "A Modified Marquardt Subroutine for Nonlinear Least Squares," Report R6799, AERE, Harwell (1971).

Gill, P. E., W. Murray, and M. H. Wright, *Practical Optimization*, Academic Press, New York (1981).

Guttmann, A. J., "Analysis of Experimental Specific Heat Data Near Critical Temperature," *J. Phys. C.: Solid State Phys.*, **8**, 921 (1975).

Jupp, D. L. B., "Approximation of Data by Splines with Free Knots," *SIAM J. Numer. Anal.*, **15**, 328 (1978).

Klaus, R. L., and H. C. Van Ness, "An Extension of the Spline-Fit Technique and Application to Thermodynamic Data," *AIChE J.*, **13**(6), 1132 (Nov., 1967).

Lawson, C. L., and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, NJ (1974).

Read, J. F., and L. G. Dunfield, "Determination of Reaction Rates by the Use of Cubic Spline Interpolation," *J. Chem. Phys.*, **57**, 2178 (1972).

Schumaker, L. L., *Spline Functions: Basic Theory*, John Wiley and Sons, New York (1981).

Schwarz, G., "Estimating the Dimension of a Model," *Ann. Statist.*, **6**, 461 (1978).

Skillerne de Bristowe, B. J. and D. Stubley, "Excess Enthalpies of Mixtures of Pentafluorobenzene with Some Mono-Alkyl Benzenes," *J. Chem. Thermody.* **5**, 121 (1973).

Smith, J. M. and H. C. Van Ness, *Introduction to Chemical Engineering Thermodynamics*, McGraw-Hill, New York (1975).

Tao, Teh-Ming, "Novel Applications of B-splines to Engineering Problems," PhD Dissertation, Texas A&M University, College Station, TX (1987).

Wold, S., "Analysis of Kinetic Data by Means of Spline Functions," *Chemica Scripta*, **1**, 97 (1971).

Manuscript received Oct. 6, 1987, and revision received Apr. 21, 1988.