A Data-Fitting Algorithm with Shape-Preserving Features

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

Data (or curve) fitting refers to the process of obtaining a parametric representation of discrete data. We consider here the univariate data-fitting problem, in which a dependent variable (or response function) is taken as a function of a single independent variable. A variety of numerical functions have been used for data fitting. It is now recognized that polynomial splines possess a number of properties that make them generally the most appropriate choice for data fitting. Simply stated, any function and all of its derivatives can be represented arbitrarily accurately with a spline, provided sufficient numbers of knots are used (Schumaker, 1981). Approximating splines can be developed by interpolation, whereby the spline is required to pass through each data point, or by smoothing, whereby the spline coefficients are chosen to minimize some error functional (such as a least-squares performance index). The smoothing approach is to be preferred from a statistical viewpoint. A suitable data-fitting procedure should provide more accurate estimates for values of the response function than the measurements themselves.

In many situations, certain features about the graph of the true (but unknown) response function may be known. It is often desirable that the approximating function retain those features. There are two common situations. The first is one for which the response function is known to be a monotonic increasing (or decreasing) function of the independent variable. The second common situation is one for which the function is known to be convex (or concave). Standard data-fitting procedures might produce approximating splines that do not retain features known to exist for the true function, even though a fairly accurate representation may be obtained. This can be particularly troublesome when derivatives of the response function are desired, such as for rate or parameter estimation (Landis and Nilson, 1962; de Nevers, 1966; Churchill, 1979; Tao and Watson, 1984a, b).

Here we present an algorithm for constructing spline approximations which, if desired, satisfy the requirement that the approximating function, or any of its derivatives, be monotonic increasing (or decreasing) over the entire range of the independent variable, or over any portion of that range. Note that a function whose first derivative is monotonic increasing is a convex function. Thus, the two common data-fitting problems discussed previously—construction of monotonic increasing or convex approximating functions—are solved as special cases. This process is carried out by minimizing a least-squares objective function subject to linear inequality constraints on the spline coefficients. Our algorithm is fully automatic, and the (global) minimum value of the sum of squared residuals is always obtained. We demonstrate the algorithm with examples for the two common data-fitting problems previously cited.

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

Function Approximation with Shape Constraints

A spline function defined on $\{x|a \le x \le b\}$ can be written as (de Boor, 1978; Schumaker, 1981):

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

For a selected order and partition (i.e., knot locations and multiplicities), an approximating spline is normally calculated by determining the spline coefficients that minimize a least-squares objective function (de Boor, 1978):

$$J = (f - GC)^T W (f - GC) \tag{2}$$

Elements of the $N \times n$ matrix G are given by $G_{ij} = B_j^M(x_i)$. The problem we consider is that of obtaining the approximating spline that minimizes a least-squares objective function, subject to the constraint that the function, or one of its derivatives, be monotonic increasing. The alternative case, that the function or one of its derivatives be monotonic decreasing, is handled simply by changes in sign in the constraint equations. With a restriction on the order of the approximating spline, the problem can be solved using quadratic programming.

We consider the requirement that the ℓth derivative of $S(x)$ be monotonic increasing with x (note that the zeroth derivative refers to the function itself). $S^{\ell}(x)$ will be monotonic increasing if the derivative $S^{\ell+1}(x)$ is nonnegative for all values of x. Note that if $S(x)$ is a spline of order M, $S^{\ell+1}(x)$ is a spline of order $M - (\ell + 1)$. If we restrict the order of the approximating spline to be $\ell + 2 \le M \le \ell + 3$, its $(\ell + 1)$th derivative will be either linear or a uniform value within each subinterval $[t_i, t_{i+1}]$. Consequently, we can ensure that $S^{\ell+1}(x)$ is nonnegative for all values of x by requiring that it be nonnegative at only a finite number of locations. Specifically, with knots $\{t_i | a < t_i < b, i = 1, 2, \ldots, k\}$ and multiplicity $\{m_i | 1 \le m_i < M, i = 1, 2, \ldots, k\}$, satisfaction of the following constraints ensures that $S^{\ell}(x)$ is monotonic increasing:

$$S^{\ell+1}(t_i) \ge 0 \quad \text{when } \ell \le M - m_i - 2$$

or

$$S^{\ell+1}(t_i^+) \ge 0 \quad \text{when } \ell > M - m_i - 2$$
$$S^{\ell+1}(t_i^-) \ge 0 \qquad (3)$$

When ℓ exceeds $M - m_i - 2$, $S^{\ell+1}(x)$ may be discontinuous across t_i. Therefore, two constraints need be imposed at location t_i, as is specified in Eq. 3. The least-squares approximating spline that preserves the requirement that its ℓth derivative be monotonic increasing is now determined by calculating the coefficient vector C that minimizes Eq. 2 subject to the inequality constraints given in Eq. 3. This is a numerical minimization problem that has a specific structure. The objective function is quadratic in C. Constraints given by Eq. 3 are linear in C. The minimization of a quadratic objective function subject to linear inequality constraints is a quadratic programming problem; it can be effectively solved by any of a number of quadratic programming algorithms (Gill et al., 1981). There are specially designed methods when, as is the case here, the objective function is the least-squares function (Stoer, 1971). The important characteristic of the solution is that the exact minimum can be determined in a finite number of steps. Furthermore, the solution is unique provided G has full rank; this condition is guaranteed when the knots are placed so that each B-spline is sampled by a distinct data point (Jupp, 1978). In the examples that follow, we used a computer code published elsewhere (Hanson and Haskell, 1982) to solve the minimization problem.

Results and Discussion

We illustrate the data-fitting algorithm developed here by fitting discrete thermodynamic data. First, we consider the equilibrium melting temperature of a zinc-indium bimetallic solution or alloy. Experimentally determined values of the melting temperature at various compositions (Rhines and Grobe, 1944) are plotted in Figure 1. The melting points for pure indium and zinc are 155 and 419.4°C, respectively. Beginning with pure indium, the melting temperature decreases as small amounts of zinc are added. The eutectic temperature is reached at weight fraction [Zn] = 2.73%. A sharp increase in the melting temperature is observed as more zinc is added. The rate of increase of the melting temperature is much smaller for zinc compositions greater than [Zn] = 25%. The function corresponding to zinc compositions in excess of the eutectic is believed to be monotonic increasing. Similar results on this zinc-indium bimetallic system have been reported by several other investigators (Svivbely and Selis, 1953).

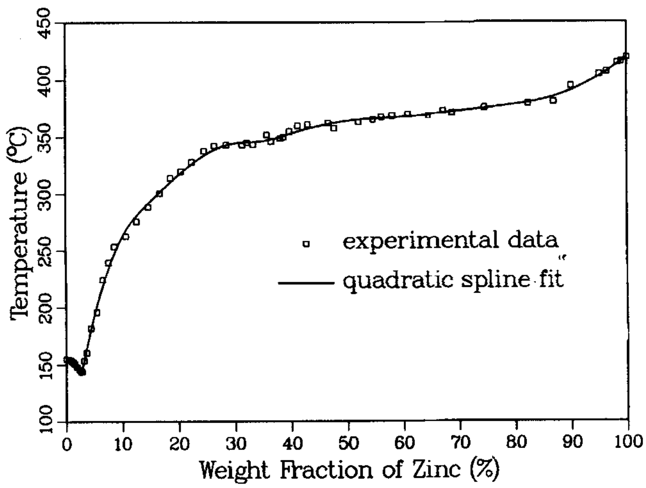

Figure 1. Melting points of Zn-In alloy.

We fitted a quadratic spline to the data using the knot locations listed in Table 1. We allowed the first derivative to be discontinuous at the eutectic in order to represent the sharp corner indicated there by the data. This was accomplished by assigning the multiplicity m_i a value of 2 at that location. Simple knots (unit multiplicty) are used elsewhere. Values of the fitted function and its first derivative at the knot locations are listed in Table 1. When constraints are not used, the first derivative has a negative value at the knot located at [Zn] = 32.2%. This indicates that the first derivative is negative for some range of composition, which is contrary to our understanding of the system.

The monotonic increasing approximating spline that minimizes the least-squares objective function was determined by including inequality constraints on the first derivative of the function. The values of the first derivative at the knots, Table 1, illustrate that the function is monotonic increasing for compositions greater than the eutectic. Note that the function values at the knots actually differ little from the corresponding values for the unconstrained case. This illustrates that the precision of the fit need not be compromised significantly in determining a fitted function with the desired monotonicity property. The fitted function is plotted with the observed data in Figure 1.

Table 1. Function and Derivative Values for Example 1

Inner Knot Location [Zn]%	Spline without Constraints		Spline with Shape Constraints	
	f	f'	f	f'
1.23	152.2	−5.25	152.1	−4.70
2.73^-	142.9	−7.10	142.8	−7.69
2.73^+	142.9	25.9	142.8	26.2
5.42	201.1	17.4	201.1	17.2
12.6	281.1	5.00	281.2	5.19
22.3	327.0	4.40	326.5	4.07
32.2	346.6	−0.428	346.5	0.0
38.6	351.0	1.80	351.7	1.62
54.5	366.9	0.204	366.6	0.258
82.4	380.3	0.760	380.4	0.731

The second example is the fitting of equilibrium boiling temperature of tetrahydrofuran-water at atmospheric pressure (Hirata et al., 1975). The boiling temperature is believed to be a convex function of composition. The appearance of unwarranted inflection points in fitting these data with cubic splines has been recognized previously (Taitel and Tamir, 1983). They proposed a method for obtaining a fit that is free from unwarranted inflection points; however, their solution may not be optimal, in that it may not provide the minimum value for the sum of squared residuals.

Dierckx (1980) has reported a constrained minimization algorithm for constructing a convex or concave approximating cubic spline. While he does obtain the minimum value for the sum of squared residuals, a tedious and time-consuming Theil-Van de Panne procedure is used to solve the quadratic programming problem. This procedure does not take advantage of the special structure of the least-squares minimization problem.

In Table 2 we show the values of the second derivative of the approximating function at each of the knot locations. We have used the same knot locations as previous investigators (Taitel and Tamir, 1983). When no constraints are used (case 1), the second derivative has negative values at two knot locations. This illustrates that the first derivative for the approximating function is not monotonic increasing as desired. When inequality constraints are included to require that the second derivative be nonnegative, a convex approximating function is obtained. The values of the second derivative at the knots, and the corresponding sum of squared residuals, are listed as case 2 in Table 2. Note that there is a single active constraint, i.e., constraint that is satisfied as an equality. The fitted function and observed data are plotted in Figure 2.

The solution to this problem reported by Taitel and Tamir is evidently not the convex approximating spline that minimizes the least-squares objective function. We have reproduced their solution by using equality constraints to require that the second derivative be zero at the first and third interior knots. Results of this fit are listed as case 3 in Table 2. While the first derivative is monotonic increasing, as desired, a larger value of the sum of squared residuals is obtained than for the global optimum represented by case 2. This example illustrates the value of quadratic programming methods for determining the global minimum of the objective function.

Table 2. Second Derivative and Sum of Squared Residuals Values for Example 2

Knot Locations	Case		
	1	2	3
$t_0 = 0$	81,400	80,400	81,600
$t_1 = 0.05$	−124	305	0.0
$t_2 = 0.09$	73.6	26.7	31.0
$t_3 = 0.52$	−27.9	0.0	0.0
$t_4 = 0.85$	46.7	5.5	3.9
$t_5 = 1$	465	602	608
Sum of squared residuals	0.434	0.839	0.876
Inequality constraints	No	Yes	No
Equality constraints (knots)	No	No	1,3

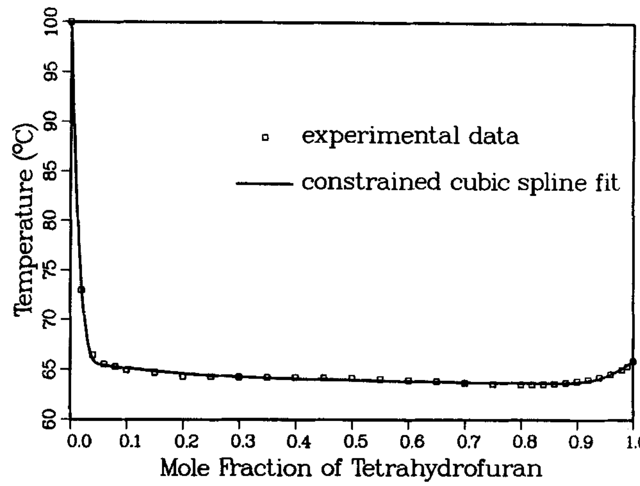

Figure 2. Boiling temperature of tetrahydrofuran-water mixture.

Conclusions

We have developed an efficient and versatile algorithm for approximating univariable data with polynomial splines. We use a *B*-spline formulation so that the order of the spline and the degree of continuity at the knots can be specified conveniently. The user may require that the approximating spline, or any of its derivatives, be monotonic increasing (or decreasing) over any range of the independent variable. The data-fitting algorithm is fully automatic and always determines the least-squares optimum fit. We demonstrate the algorithm with two situations that are commonly encountered in data-fitting problems: the fitting of monotonic increasing functions and convex functions.

Notation

a, b = end points for independent variable
$B_i^M(x)$ = ith *B*-spline of order M
C_i = coefficients in *B*-spline representation
f_i = value of dependent variable at x_i
$G_{ij} = B_j^M(x_i)$
J = objective function
k = number of knots
m_i = multiplicity of ith inner knot
M = order of spline function
n = dimension of spline
N = number of experimental observations
$S(x)$ = polynomial spline
t_i = partition or location of knots
W = weighting matrix
x = independent variable

Superscripts

ℓ = ℓth derivative
$+$ = approach the point from the right
$-$ = approach the point from the left

Literature Cited

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

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

de Nevers, N., "Rate Data and Derivatives," *AIChE J.*, **12**(6), 1110 (1966).

Dierckx, P., "An Algorithm for Cubic Spline Fitting with Convexity Constraints," *Computing* **24**, 349 (1980).

Gill, P. E., W. Murray, and M. H. Wright, *Practical Optimization,* Academic Press, New York (1981).
Hanson, R. J., and K. H. Haskell, "Two Algorithms for the Linearly Constrained Least-Squares Problem," *ACM Trans. Math Softw.,* **8**(3), 323 (1982).
Hirate, M., S. Ohe, and K. Naguhama, *Computer-Aided Data Book of Vapor-Liquid Equilibria,* Kodansha Ltd., Elsevier (1975).
Jupp, D. L. B., "Approximation to Data by Splines with Free Knots," *SIAM J. Numer. Anal.,* **15,** 328 (1978).
Landis, F. and E. N. Nilson, "The Determination of Thermodynamic Properties by Direct Differentiation Technique," *Progress in International Research on Thermodynamics and Transport Properties,* Academic Press, New York, 218 (1962).
Rhines, F. N., and Grobe, A. H., "Constitution of the System Indium-Zinc," *Trans. AIME,* **156,** 253 (1944).
Svivbely, W. J., and Selis, S. M., "A Thermodynamic Study of the Zinc-Indium System," *J. Am. Chem. Soc.,* **75,** 1532 (1953).
Schumaker, L. L., *Spline Functions: Basic Theory,* Wiley, New York (1981).
Stoer, J., "On the Numerical Solution of Constrained Least-Squares Problems," *SIAM J. Numer Anal.,* **8**(2), 382 (1971).
Taitel, Y., and A. Tamir, "Avoiding Unwarranted Inflection Points in Fitting of Data," *AIChE J,* **29**(1), 153 (1983).
Tao, T. M., and A. T. Watson, "Accuracy of JBN Estimates of Relative Permeability. 1. Error Analysis," *Soc. Pet. Eng. J.,* **24**(2), 209 (1984a).
———, "Accuracy of JBN Estimates of Relative Permeability. II. Algorithms," *Soc. Pet. Eng. J.,* **24**(2), 215 (1984b).

Manuscript received Jan. 20, 1987, and revision received April 13, 1987.