Two Methods of Selecting Smoothing Splines Applied to Fermentation Process Data

Nina F. Thornhill
Dept. of Computer Science

Mauro Manela and John A. Campbell
Dept. of Electronic and Electrical Engineering

Karl M. Stone
Advanced Centre for Biochemical Engineering
University College London, Gower Street, London WC1E 6BT, England

Two methods for generating smoothing splines are compared and applied to data from a fed-batch fermentation process. One method chose both the degree of the spline and its parameters by minimizing the generalized cross validation (GCV) function using a genetic algorithm (GA). The other method adjusted the smoothing spline to a specified chi-square goodness-of-fit, requiring prior knowledge of the measurement variability. The GCV/GA method led to excellent results with all the fermentation data records. The goodness-of-fit method gave a family of spline fits; splines with a low percentage fit extracted trends from the data, while for general use a 50% fit appeared satisfactory. The goodness-of-fit method executed more quickly than the GCV/GA method, but the GCV/GA method was more generally applicable as it chose both the degree of the spline and the amount of smoothing automatically.

Introduction

A typical requirement in the analysis of process data with spline functions is smoothing and interpolating a set of noisy measurements for the capture of real features that are not easy to see in a sparse data record, using the smoothest spline with acceptable accuracy. This work concerns nonsteady data trajectories from penicillin fermentations that include process discontinuities and process disturbances. It investigates methods for adjusting the parameters that determine the compromise between the accuracy and the smoothness of the spline; these were a smoothness factor, the degree of the spline, and the number of basis functions in the composite spline.

A spline is a piecewise continuous function consisting of polynomial segments that can provide an approximation to an unknown function represented only by noisy discrete measured data. The polynomial segments join at the points called knots, where all the derivatives, but the highest, are continuous. The third derivative of a degree three (cubic) spline, for example, is discontinuous at the knots. A common method for constructing spline functions is to use a linear combination of simple spline functions, such as the bell-shaped B-splines (Perry et al., 1984), each weighted with a coefficient determined by fitting the composite spline to the discrete data.

Many authors have recognized that the amount of smoothing of the spline approximation should be controlled to avoid the spline following the random errors in the measurements. Reinsch (1967) proposed an adjustable smoothness factor (denoted by p) which controlled a balance between the fidelity of the spline to the measurements on the one hand, and its roughness indicated by the values of the higher derivatives on the other. Dierckx (1975) published a FORTRAN algorithm to generate a spline with minimum third derivative discontinuities subject to a fidelity constraint (denoted by S) on the permitted deviation of the spline from the measurements. Both the number of knots (and hence the number of B-splines) and the smoothness factor, p, were adjusted in the search for a spline with the desired value of S. The fidelity constraint was specified by the user, although the relationship between the specified fidelity and the quality of the visual fit was far from intuitive. Reinsch (1967) noted that the standard deviations of the measurements affected the fidelity demanded of the spline, and the

Correspondence concerning this article should be addressed to N. F. Thornhill.

specified fidelity constraint therefore ought to depend on the standard deviations as well as on the number of points in the data set.

Craven and Wahba (1979) found an optimum choice of smoothness factor by generalized cross validation (GCV), without a fidelity constraint. One way to visualize the GCV concept is that it optimizes the ability of a set of splines fitted to all except one data point to predict the omitted values. These authors considered the case where the number of basis functions was the same as that of data points, while Pope and Gadh (1988) applied the GCV approach to the case of dense data by reducing the number of basis functions. Procedures for calculation of the coefficients of the basis functions in a smoothing spline have been optimized by Hutchinson and de Hoog (1985), while FORTRAN routines for generalized cross validation computations have been published by Bates et al. (1987).

The use of smoothing splines, especially cubic splines, is widespread in the analysis of process and laboratory data (Klaus and van Ness, 1967; Dunfield and Read, 1972; Tao and Watson, 1988; Steemson and White, 1988). Oner et al. (1986) reported the use of cubic spline functions in the analysis of data from anaerobic fermentation processes. In their work, measurements of several chemical components from the same fermentation were smoothed simultaneously, each with a fidelity constraint and smoothness factor such that the smoothed data sets satisfied the available electron balance for the biochemical reactions involving the measured components. Buono et al. (1986) compared splines smoothed using electron and carbon balance closure with the GCV technique for choosing the smoothness factor and commented that the balance method had the advantage that it used all the fermentation measurements and not just those of the data set to be smoothed. The balance method was, however, applicable only when there were measurements of all the relevant chemical species.

We also have been using spline functions with fermentation data, working with a pilot-scale industrial penicillin fermentation and measuring a limited number of components (the dry weight concentration, penicillin concentration, and carbon dioxide evolution rate). The work has led to two developments in procedures for fitting smoothing splines to individual data sets which have been illustrated using the penicillin fermentation data. In one development, the fidelity constraint was treated as a random chi-square variate. When the measurement variability had been determined experimentally we calculated the appropriate value for S taking into account the number of data points, the number of knots, the known measurement standard deviations, and a chi-square goodness-of-fit specified by the user.

When the measurement variability was unknown, measurements were instead smoothed by a spline with the combination of smoothness factor and number of knots which minimized the GCV function. The GCV minimization was a nonlinear problem with a large search space that combined searching in discrete and continuous dimensions (the type of search space discussed by Goldberg, 1989) and the problem also suffered from numerical instabilities for certain combinations of parameters due to the existence of ill-conditioned matrices (de Hoog and Hutchinson, 1987). It belongs to the class of constrained optimization problems since there was a maximum permitted number of knots for a given size of data set. This is exactly the type of problem where a genetic algorithm (GA) is a candidate for the task. Genetic algorithms are population-based so, for example, the performance is not degraded by an occasional ill-conditioned matrix.

It is standard practice in the applications literature to use cubic spline functions for data smoothing. Exploiting the flexibility of the GA has enabled us to incorporate the degree of the spline polynomials as another dimension in the GCV search space and hence to search for both the structure of the model and its parameters at the same time. This provided a bonus in several cases where a quadratic, rather than a cubic, spline yielded better results. The alternative goodness-of-fit procedure was restricted to cubic splines since there was no obvious method for comparing the qualities of splines with the same goodness-of-fit but of different degree.

The next section presents the theoretical and experimental methods, comprising the procedures for fitting spline functions by the above techniques, an outline of the GA, and the experimental methods for the penicillin fermentations. The derivations of the B-spline coefficients and the associated cost functions are outlined in the Appendix. The third section illustrates the search for the required degree of smoothing in a detailed study of measurements from one of the penicillin fermentations. It illustrates that smoothing splines at the GCV minimum have visual appeal to a fermentation technologist compared with ones which are away from the GCV minimum and also shows the effect of different goodness-of-fit selections. It also comments on some interesting results from additional fermentations and assesses the computational load. Finally, the techniques are compared in terms of efficiency, and appropriate uses for each are recommended.

Theory and Methods

Theory of smoothing splines

A spline function is defined over a set of basis functions $b_{j,\lambda}(x)$ and we will follow Dierckx (1975) in using normalized B-spline functions. Perry et al. (1984), and Cox (1972) indicate how B-splines may be evaluated. Given a set of data points $\{x_i,y_i\}$, $i=1, 2, \ldots m$, and knots t_j, $j=\lambda+1, \lambda+2, \ldots n-\lambda$ in the range $[x_1, x_m]$, a spline function of degree λ with knots t_j, $j=1, 2, \ldots n$ may be represented as a linear combination of a finite number of λth degree B-splines:

$$s_\lambda(x) = \sum_{j=1}^{n-\lambda-1} c_j b_{j,\lambda}(x) = \underline{b}_\lambda(x)^T \underline{c}, \tag{1}$$

$$\underline{b}_\lambda(x)^T = \{ b_{1,\lambda}(x) \quad b_{2,\lambda}(x) \quad \ldots \quad b_{n-\lambda-1,\lambda}(x) \} \tag{2}$$

where c_j are the representation coefficients and the B-spline $b_{j,\lambda}(x)$ is determined uniquely by the degree of the spline and the knots t_j to $t_{j+\lambda+1}$. We shall refer to $\underline{c}$ as the vector of coefficients and to $\underline{sx} = \boldsymbol{B}\underline{c}$ as the discrete samples from the spline function $s_\lambda(x)$ at $x_1, x_2, \ldots, x_m$, where the ith row of matrix $\boldsymbol{B}$ is $\underline{b}_\lambda(x_i)^T$. Following Dierckx (1975), the roughness is measured by the vector of discontinuities in the highest derivative of the spline, which we shall refer to as $\underline{rc} = \boldsymbol{R}\underline{c}$. The discontinuity jump of $s_\lambda(x)$ at t_k is:

$$rc_{\{k-(\lambda+1)\}} = \left(\frac{d^\lambda s_\lambda(t_k)}{dx^\lambda} \right)^+ - \left(\frac{d^\lambda s_\lambda(t_k)}{dx^\lambda} \right)^- \tag{3}$$

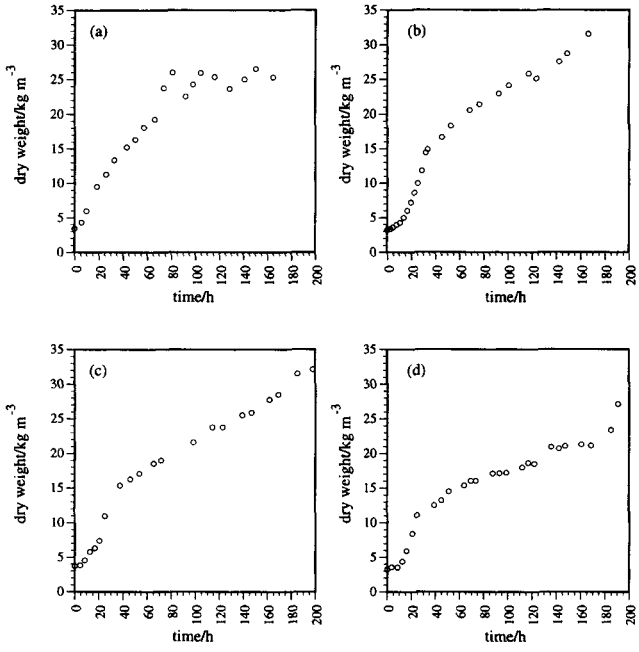

Figure 1. Dry weight concentration measurements for several penicillin fermentations.

(a) Oscillations caused by variations in the precursor feed; (b) fermentation with a temporary interruption in growth because of a blocked feed pump; (c) fermentation with no unusual features; (d) fermentation receiving excess nutrients toward the end.

The superscripts $^+$ and $^-$ indicate the value of the λth derivative either side of the knot t_k. Only discontinuities at the knots inside the range $[x_1, x_m]$ are important and hence, for example, for a cubic spline the jth column $\underline{r}_j$ of matrix $\boldsymbol{R}$ is:

$$\underline{r}_j = \{r_{5,j}\ r_{6,j}\ r_{7,j}\ \ldots\ r_{n-4,j}\}^T \tag{4}$$

$$r_{k,j} = \left(\frac{d^3 b_{j,3}(t_k)}{dx^3}\right)^+ - \left(\frac{d^3 b_{j,3}(t_k)}{dx^3}\right)^- \tag{5}$$

Throughout this article we assume that measurements $\underline{y}$ are generated by an underlying function corrupted by additive random noise. For a given spline with samples $\underline{sx}$ fitted to measurements, the weighted residuals are given by $\underline{f} = \boldsymbol{W}^{1/2}\underline{e}$, where $\underline{e} = \underline{sx} - \underline{y}$ and the weighting matrix, $\boldsymbol{W}$, would be the inverse of the variance matrix of the random noise sequence if this were known. To account for the roughness of the spline and any departure from the measurements Dierckx (1975) introduced a cost function given by:

$$J(n, \underline{c}, p) = p(\underline{f}^T\underline{f}) + (\underline{rc}^T\underline{rc}) = p\|\underline{f}\|^2 + \|\underline{rc}\|^2 \tag{6}$$

The factor p controls the tradeoff between the roughness of the fit and its accuracy. Once the number of knots, their positions, and a value for p have been chosen, the minimization of J, which is a quadratic function of the coefficient vector $\underline{c}$ is straightforward using the procedure of the Appendix. The positions of the knots should reflect the distribution of the measurements (Dierckx, 1975), while the choice of p and n is a nonlinear problem; the solution of that problem is the primary focus of this article.

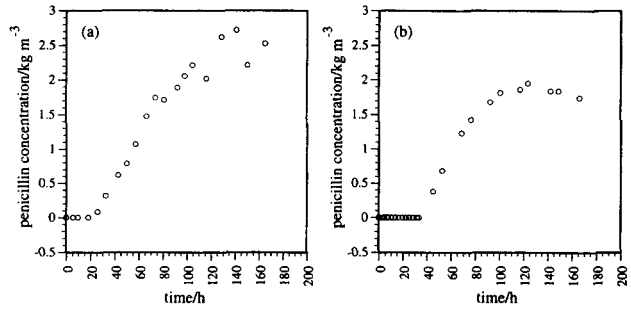

Figure 2. Penicillin concentration measurements corresponding with the fermentations in Figures 1a and 1b.

Note that the measured concentration is always nonnegative.

Craven and Wahba (1979) showed that the GCV function, $\|\underline{z}\|^2$, is the sum of squares of the residuals, f_i, weighted by the sensitivity of each residual to variations in the corresponding experimental measurement:

$$\|\underline{z}\|^2 = \sum_{i=1}^{m}\left(\frac{f_i}{\dfrac{\partial f_i}{\partial y_i}}\right)^2 \tag{7}$$

In this methodology, the matrix $\boldsymbol{W}$ need reflect only the relative weightings for the measurements; absolute values of the measurement error variances are not required. The Appendix gives the expression for the numerical calculation of $\|\underline{z}\|^2$. Our method (Manela et al., 1993) used a GA to search for a good combination of n and p. For each candidate pair of parameters, n_a and p_a, a straightforward minimization of $J(n_a, \underline{c}, p_a)$ gave the coefficients for an arbitrary spline $\underline{sx}_a$, which was used to compute the GVC function for that particular set of parameters. The GA searched for the combination that minimized the GCV function.

Reinsch (1967) chose the smoothness factor, p, using a fidelity constraint such that $\|\underline{f}\|^2$ achieved a specified value, S. He suggested empirically that the value of S should be in the range $m \pm \sqrt{2m}$, but this is unsatisfactory from a practical point of view as there is no obvious means of choosing a specific value within that range other than by trial and error. Moreover, the best choice depends also on the number of knots in use. When the variance of the measurement noise is known the $\boldsymbol{W}$ matrix is approximately equal to the inverse of the covariance matrix of the error sequence $\underline{e}$. In that case (Press et al., 1986), $\|\underline{f}\|^2$ is a random chi-square variate with $\{m - [n - (\lambda + 1) - 1]\}$ degrees of freedom. We have extended the smoothness factor concept to take account of the chi-square distribution of $\|\underline{f}\|^2$. The approach is appropriate when an independent experimental assessment of the standard deviation of the measurement error is available to calculate the variances for the $\boldsymbol{W}$ matrix. The user must specify the goodness-of-fit which is translated to a value of the $P(\chi^2|\nu)$ probability function. For a 50% fit, for example, $P(\chi^2|\nu)$ would be 0.5 and the value of S set to the corresponding χ^2 value using a numerical inversion of the chi-square function (Press et al., 1986). A way to visualize a 50% spline fit is that in cases where $\nu \approx m$ the 50% fit lies within one standard deviation of about half the measurements.

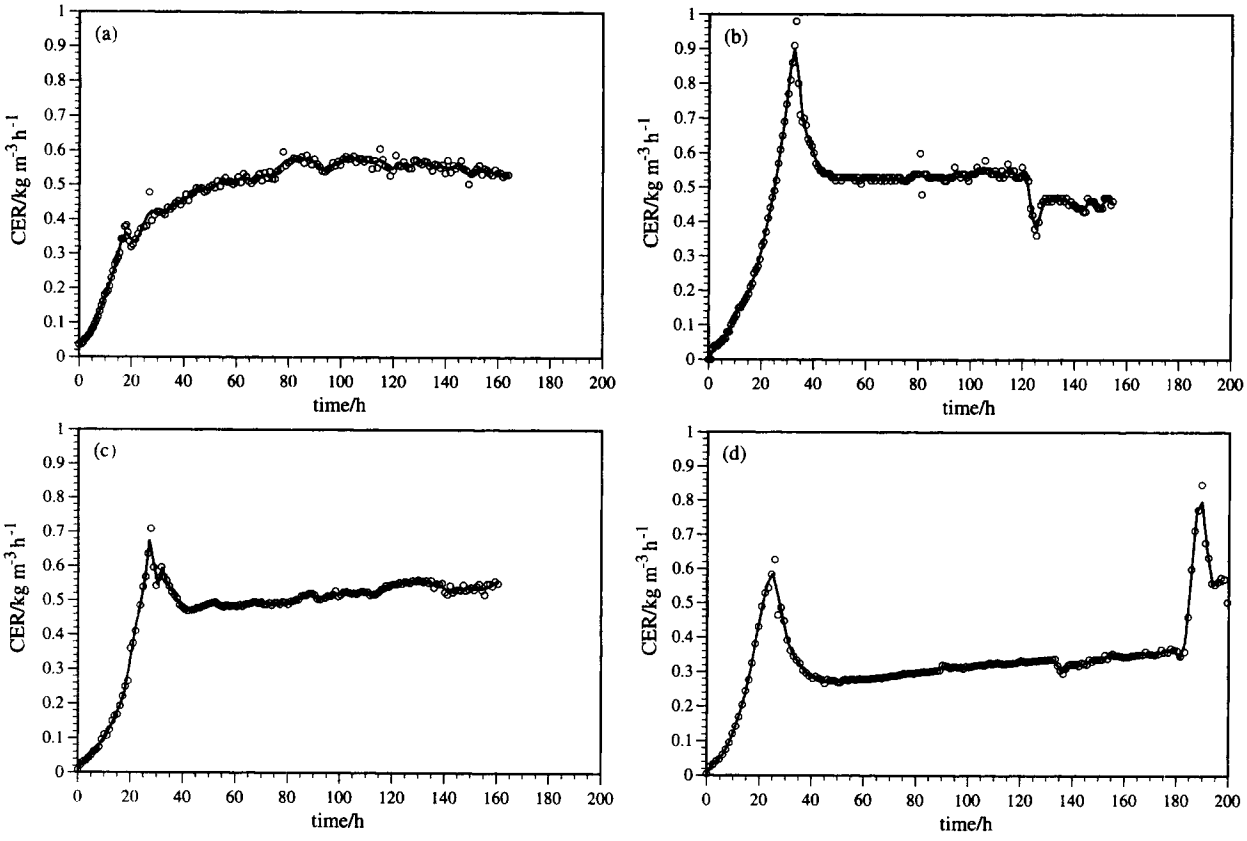

Figure 3. Carbon dioxide evolution rate (CER) records fitted with GCV/GA splines.

The GCV/GA splines recognized real events such as the peaks at the end of batch growth phase and the fault at 120 hours in Figure 3b.

Genetic algorithm

Genetic algorithms belong to a class of algorithms introduced by Holland (1975) that are loosely based on biological mechanisms of natural selection and which aim to converge to a desired optimum solution by iterating through generations of individual candidate solutions. This section explains the means by which the population in the GCV/GA application evolved and how the candidate solutions were encoded.

In a standard GA methodology (Goldberg, 1989), the population consists of individual candidate solutions encoded as binary strings. A new population is derived from the current population by a processing step involving a probabilistic selection procedure (in which the probability of selection is weighted toward the "fitter" individuals) combined with a controlled amount of crossover (in which two binary strings swap random substrings) and mutation (where binary digits are changed). In our application, candidates from the current population with lower GCV values were chosen preferentially at the selection stage.

A candidate solution was a binary string comprising three binary substrings. The parameters to be encoded were λ, n, and p, where p was a continuous variable, while n and λ were

Table 1. Values of the GCV Function for Dry Weight Data in Figure 1a for Cubic Splines

	No. of Knots							
log(p)	10	12	14	16	18	20	22	24
−10.00	42.24	42.15	42.07	42.05	41.87	41.74	41.67	41.67
−8.80	40.90	40.21	39.75	39.66	39.26	39.21*	39.29	39.32
−7.60	42.67	44.34	47.09	49.67	50.53	53.89	57.75	58.40
−6.40	78.42	83.16	80.26	89.01	88.18	88.97	92.55	94.07
−5.20	96.25	88.93	106.4	114.0	116.3	122.0	121.2	121.0
−4.00	97.91	76.01	236.6	115.8	120.9	112.5	95.07	41.87
−2.80	98.01	74.55	287.0	1,486	1,469	598.1	191.6	21.11
−1.60	98.02	74.45	291.2	9,334	7,859	1,881	1,142	12.55*
−0.40	98.02	74.44*	291.5	11,050	9,020	2,080	1,729	15.32
0.80	98.02	74.44*	291.6	11,210	9,104	2,205	2,212	15.63
2.00	98.02	74.44*	291.6	11,220	9,110	2,330	9,216	15.65
∞	98.02	74.44*	291.6	11,220	9,110	2,351	411,300	NaN

* Local minima

Table 2. Values of the GCV Function for Penicillin Concentration Data in Figure 2a for Cubic Splines.

	No. of Knots							
log(p)	10	12	14	16	18	20	22	24
−10.00	1.057	1.050	1.046	1.046	1.033	1.027	1.026	1.026
−8.80	1.004	0.932	0.903	0.906	0.830	0.809	0.813	0.815
−7.60	0.818	0.756*	0.757	0.804	0.781*	0.809	0.869	0.903
−6.40	0.942	1.917	1.664	2.090	2.185	2.201	2.427	2.872
−5.20	1.098	6.168	7.449	6.630	6.539	6.587	6.583	8.729
−4.00	1.115	7.651	17.11	10.10	10.22	10.31	10.91	11.05
−2.80	1.116	7.776	19.13	4.717	1.570	0.463	0.308*	3.562
−1.60	1.116	7.784	19.28	0.856*	6.675	23.02	56.56	1.890
−0.40	1.116	7.784	19.29	2.164	10.73	32.63	83.42	1.675
0.80	1.116	7.784	19.29	3.151	11.25	36.81	96.45	1.659
2.00	1.116	7.784	19.29	3.241	11.29	42.10	336.6	1.658*
∞	1.116	7.784	19.29	3.247	11.29	42.99	6,408	NaN

* Local minima.

integers. Since the practical range for p was large we encoded its logarithm using ten binary bits within the interval $(-4, +7)$, a decision that will be discussed further in the Application section. Not all combinations of λ and n were permitted because of the following constraint (Dierckx, 1975):

$$3\lambda + 1 \le n \le m + k + 1,$$

but a geometrical transform (Manela et al., 1993) in which the permitted combination resided in either a triangular or trapezoidal region in the parameter space overcame the constraint. The permitted values of λ and n were expressed by auxiliary polar coordinates which were encoded as binary strings.

Let the candidate solutions in the population be π_l, $l = 1$ to N. Each π_l corresponds to a unique spline fit with a corresponding GCV value of $\|\underline{z}\|_l^2$. The purpose of the GA was to find the π_l corresponding to the spline with the minimum GCV function. The population size was either 30 of 50 and each "trial" consisted of the evaluation of a single new candidate solution. The initial population was chosen randomly, and although after only 300 trials the GA solutions had begun to cluster in the region of the GCV minimum, we allowed it to run for 1,000 trials.

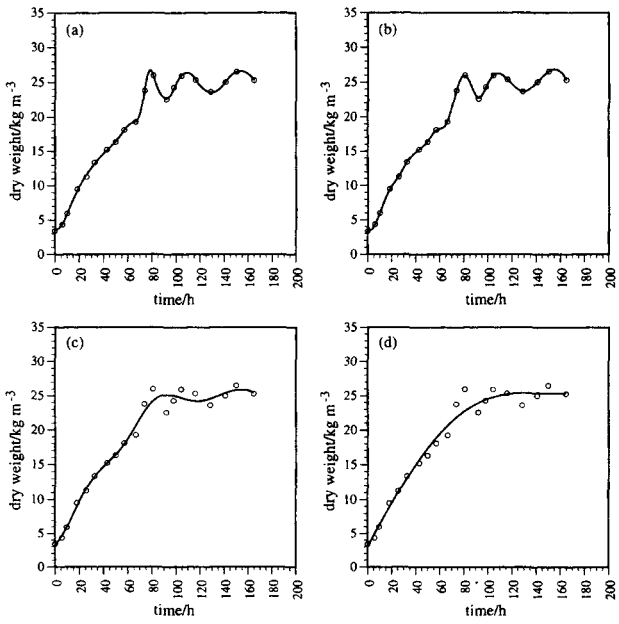

Figure 4. Spline fits to the dry weight data shown in Figure 1a.

(a) Spline with the best overall GCV value according to the GA optimization routine, a quadratic spline with 16 knots; (b) best cubic spline, with 24 knots; (c) 12-knot cubic spline from the local GCV minimum at the bottom lefthand corner of Table 1; (d) from the local GCV minimum at the top of Table 1 with 20 knots. The oscillations in these data are a real effect, but the splines away from the global GCV minimum do not follow the oscillations reliably.

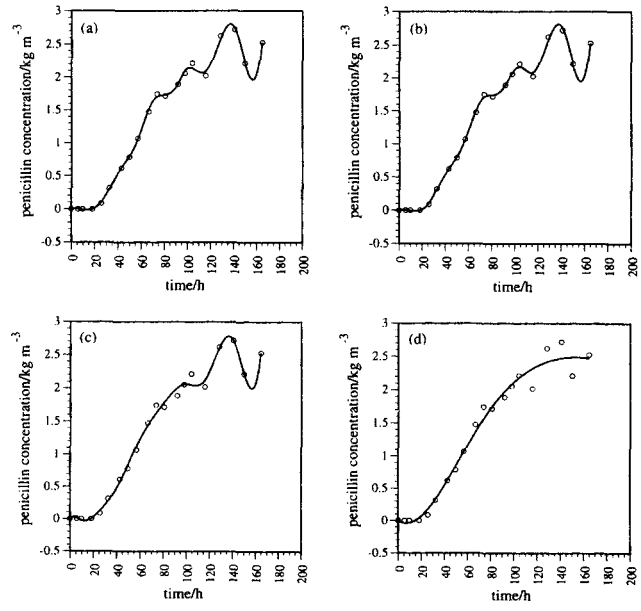

Figure 5. Spline fits to the penicillin concentration data in Figure 2a.

(a) Spline with the best overall GCV value according to the GA optimization routine, a cubic spline with 19 knots; (b) from the local GCV minimum on the righthand side of Table 2 with 22 knots; (c) from the local GCV minimum with 16 knots in the middle of Table 2; (d) very smooth spline from the top lefthand corner of Table 2. Some of the fitted splines have negative values near the start.

Fermentation methods

The penicillin fermentations involved the aerobic growth of a mould, *P. chrysogenum P2*, on a semidefined medium containing corn steep liquor. After a period of rapid and almost exponential growth, the initial charge of nutrients became exhausted and the growth was subsequently limited by a glucose feed, giving almost linear growth for the remainder of the fermentation. Penicillin G production began during the linear growth in response to a feed of the precursor phenyl acetic acid (PAA) and reached up to 4 $kg \cdot m^{-3}$ by the end. Laboratory assays monitored biomass and penicillin concentrations of samples withdrawn from the fermenter, and on-line measurement of the carbon dioxide evolution rate (CER) using a mass spectrometer gave an indication of the culture's metabolic activity. The dry weight and penicillin concentration records for several fermentations are presented in Figures 1 and 2. Figure 3 shows the CER records.

Fermentation one (Figures 1a, 2a, and 3a) had oscillations in its growth caused by intermittent boosts to the PAA addition rate to compensate for underfeeding. Fermentation number two (Figures 1b, 2b, and 3b) had a temporary interruption to its growth at 120 hours because of an interruption in the glucose feed. Fermentations three (Figures 1c and 3c) and four (Figures 1d and 3d) had no unusual features although fermentation four was the subject of an experiment in which the glucose feed rate increased at the end, giving a second period of exponential growth. The features in these data records challenge the capability of the spline fitting procedures to extract the real features from random noise in the measurements.

Application of the Spline Fitting Procedures

This section illustrates the application of the GCV/GA minimization procedure and the goodness-of-fit procedure to the dry weight, penicillin concentration and CER data records from the penicillin fermentations. Spline fitting procedures using selected dry weight and penicillin concentrations measurements are detailed, and then performances of the two procedures on additional data sets are compared.

The dry weight and penicillin concentration measurements from Figures 1a and 2a posed problems for a smoothing procedure because they were sparse and had oscillations defined by only a few datum points. Tables 1 and 2 show the GCV function of cubic splines for those biomass and penicillin data records. These GCV functions have features such as elongated valleys and local minima; the regions where there are local minima are shaded in the tables, while Figures 4 and 5 plot the smoothing splines corresponding with different local minima in the tables, some of which are clearly inadequate representations of the data. Very small values of p were always unsatisfactory, and the numerical experiments indicated that the interval $(-4, +7)$ for $\log(p)$ in the GA was convenient, although there was no special theoretical reasoning behind the choice. The GA successfully ignored the local minima, and its convergence was also unaffected by the occasional individual for which the spline computation became numerically unstable because of coincidental singularities in the matrices. In each case, the GA returned a solution near the global minimum, such as a cubic spline with 24 knots and a p value of 0.0643 as the best GCV cubic spline fit for the dry weight data.

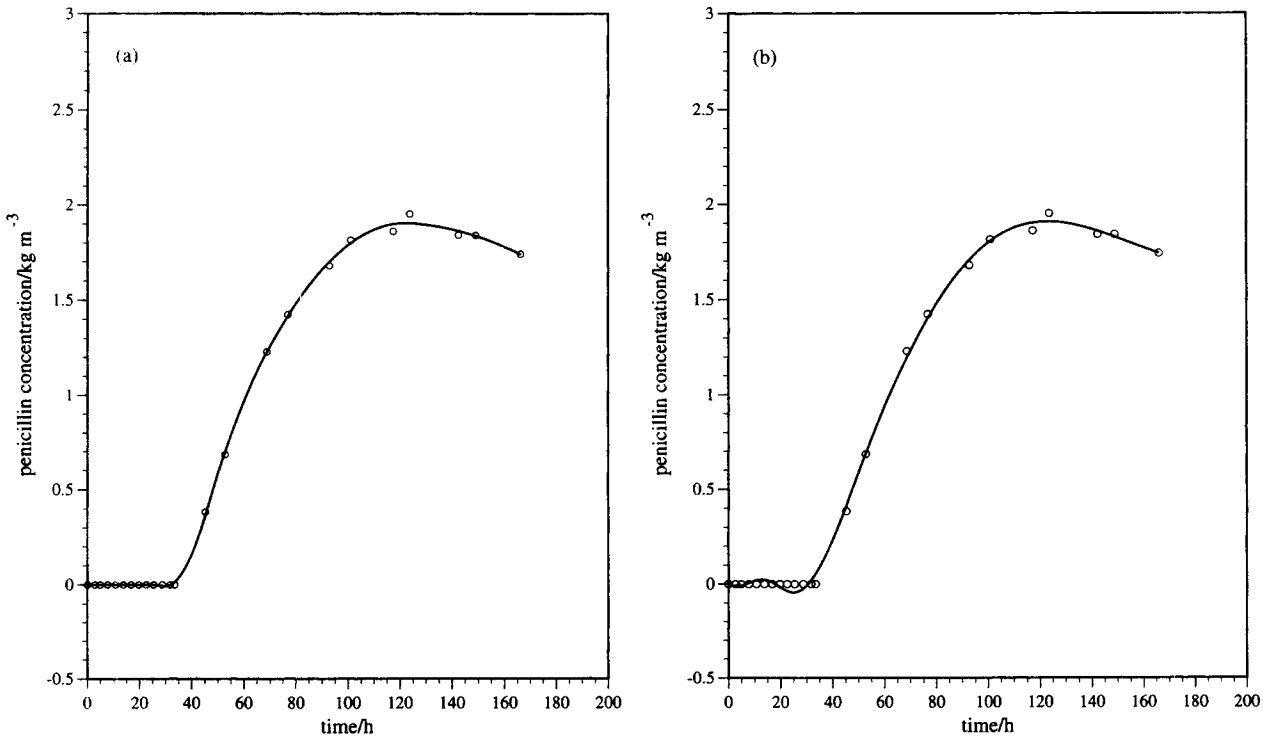

Figure 6. Cubic vs. quadratic splines for penicillin concentration data in Figure 2b.

(a) Quadratic spline with the overall GCV minimum, as assessed by the GA algorithm; (b) cubic spline with minimum GCV value. There is a discontinuity at about 35 hours when penicillin production begins; the quadratic spline with its discontinuous second derivative shows higher fidelity in the region of the "corner" than the cubic spline.

Table 3. Values of $\|\underline{f}\|^2$ for Dry Weight Data of Figure 1a for Which the Measurement Standard Deviation was $\pm 0.6\ kg \cdot m^{-3}$

	No. of Knots							
log(p)	17	18	19	20	21	22	23	24
−8	71.21	70.99	70.63	70.46	70.56	70.39	70.49	70.32
−7	66.14	65.76	65.19	64.82	65.18	64.80	64.11	63.77
−6	56.11	55.50	55.60	55.05	55.16	54.58	53.67	53.13
−5	43.68	42.66	42.51	41.34	41.85	40.67	40.10	38.88
−4	32.61	30.82	27.61	25.51	24.07	21.76	21.62	19.30
−3	30.13	24.61	10.76	8.675	6.819	5.139	6.119	4.433
−2	29.77	19.27	3.058*	1.678	2.747	1.225	1.964	0.439
−1	29.76	18.74	2.694	1.269	2.392	0.878	1.483	0.011
0	29.76	18.73	2.689	1.264	2.383	0.869	1.469	0.002
1	29.76	18.73	2.689	1.264	2.383	0.869	1.469	0.001
2	29.76	18.73	2.689	1.264	2.383	0.869	1.469	0.001
∞	29.76	18.73	2.689	1.264	2.383	0.869	1.469	0.001

* With the combination of parameters for the smoothest 50% goodness-of-fit spline.

When the GA search was extended to include also the degree of the spline polynomial it returned the quadratic spline with 16 knots in Figure 4a with a GA value that was only 30% of the GCV value for Figure 4b. While visually there was little to choose between the best quadratic and best cubic splines in this instance, the penicillin data of Figure 2b illustrates a major benefit that arose from using the GA to select simultaneously the structure and parameters of the model. The data record showed a discontinuity at about 35 hours when penicillin production began. The GA selected the quadratic spline of Figure 6a as the best spline of any degree compared with the best of the cubic splines in Figure 6b. The quadratic spline with a

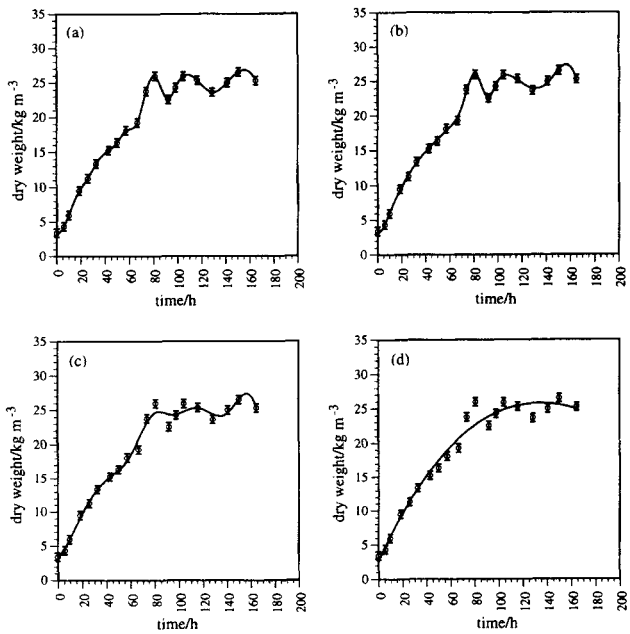

Figure 7. Cubic spline smoothing of dry weight data characterized by percentage fit.

(a) 95% fit; (b) a 50% fit (of these, the 50% fit makes more use of the measurement uncertainty and provides a smoother representation); (c) low-quality fit (<0.5%) straying significantly outside the error bars; (d) 0% fit extracting the overall trend of the data by ignoring all the detailed features.

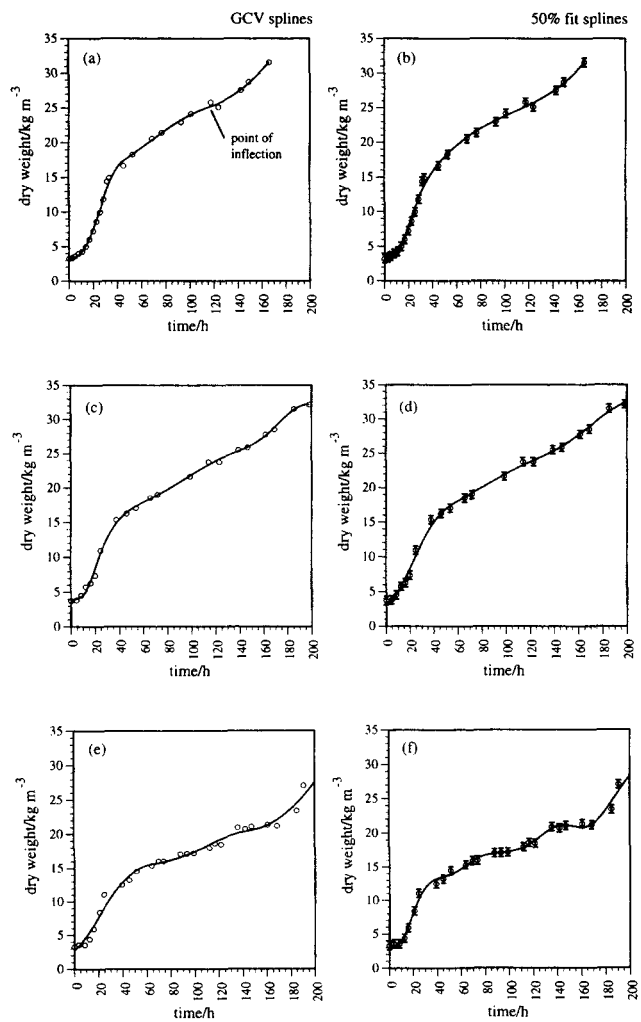

Figure 8. Comparison of GCV/GA and 50% spline fits for three additional fermentations.

(a), (b) Fermentation with temporary interruption to its growth at 120 hours. The GCV spline reflected this reduction in growth rate with a point of inflection, but the 50% spline did not because the error bars indicate the feature (which is real) could be accounted for by measurement noise. (a), (e) Quadratic splines; (c) a quartic spline. All the 50% fits are cubic splines.

Table 4. Computational Loads Associated with the GCV/GA and Goodness-of-Fit Methods of Selecting Smoothing Splines

Spline Fitting Procedure (Tested on Sets of 20 to 180 Measurements)	Evaluations of $\|\underline{z}\|^2$	Evaluations of $\|\underline{f}\|^2$	Minimizations of J	Timing Example Using Figure 1a Data
GCV/GA	300, minimum 500–600, typical	none	300, minimum 500–600, typical	for 1,000 GCV/GA trials: SUN 4: 30 s PC286: 12 min (estimated)
Goodness-of-Fit	none	$\sim m/2$	$\sim m/2$	SUN 4: <1 s PC286: 15 s

discontinuous second derivative showed higher fidelity to the data in the region of the "corner" than the cubic spline. Moreover, the quadratic spline, unlike the cubic spline, avoided giving impossible values for the penicillin concentration which cannot, of course, become negative.

The goodness-of-fit method selects a cubic spline for which $\|\underline{f}\|^2 \leq \chi^2_{\alpha,\nu}$. One can, however, use the goodness-of-fit method only when the standard deviations of the measurements are known. Table 3 tabulates $\|\underline{f}\|^2$ values for the dry weight data of Figure 1a, for which the measurement standard deviation has an upper bound of $\pm 0.6\ \mathrm{kg \cdot m^{-3}}$ (Stone et al., 1992), and highlights the cell giving the best 50% spline. The value of $\chi^2_{0.5,\nu}$ for this data record with 19 knots was 5.347, and following Dierckx (1975) we considered the smoothest spline to be the one with the smallest value of p and n which met the goodness-of-fit constraint. Figure 7 shows cubic spline fits for dry weight data with various percentage accuracies. The fits at 0.5% and 0% (a single cubic polynomial) resembled GCV splines which in Figures 4c and 4d were considered inadequate. This time, however,the infidelity to the data was manipulated by the user to extract trends from the data; thus, the splines in Figures 7c and 7d were fit for their purpose. The standard deviation of the penicillin concentration measurement was unknown, and the goodness-of-fit procedure was not applied to that data record.

Figure 8 compares the GCV/GA splines with 50% cubic splines for several sets of dry weight data. The figure suggests that the 50% cubic splines reflect most but not all, of the features captured by the GCV/GA splines. The fermentation in Figures 8a and 8b, for example, had a temporary interruption in its growth at 120 hours because of a blocked pump. The GCV/GA spline in Figure 8a reflected the reduction in growth rate with a point of inflection in the trajectory, but the 50% spline did not because the error bars indicated that the feature (which is real) could statistically be accounted for by measurement noise. The point of inflection is indicated with a pointer in Figure 8a.

On-line CER measurements were more numerous than the dry weight or penicillin concentration measurements. Figure 3 shows GCV/GA spline fits for four CER fermentation records. The peaks in most of the records between 20 and 30 hours occurred because of the abrupt change in the metabolism of the culture when the initial nutrients were exhausted and substrate-limited growth began. The GCV/GA splines handled this discontinuity well. These splines also reflected the features that have already been identified in the corresponding dry weight trajectories. Figure 3b shows that the interruption of the glucose feed at 120 hours was accompanied by a dip in the carbon dioxide production rate, while Figure 3d was the fermentation with an induced second rapid growth phase at the end of the run.

Table 4 indicates the computational load involved with each of the spline fitting procedures, and it invites a couple of comments. First, the number of evaluations of the GCV function in the GCV/GA procedure is not heavily dependent on the size of the data record that is being smoothed, whereas the goodness-of-fit search is size-dependent. The GCV/GA algorithm, though, is more computationally intensive because an evaluation of $\|\underline{z}\|^2$ involves more operations than the computation of $\|\underline{f}\|^2$. As the example execution times indicate, the timing depends on the available computing power, although there is some preliminary evidence to suggest that a nonrandom initial choice of population will reduce the number of trials in cases where there exists some prior knowledge about the shape of the GCV function in different parts of the search space.

Conclusions

Two procedures for finding a smoothing spline for a set of measurements are described, where the adjusted parameters were the number of knots and the smoothness factor. When the standard deviation of the measurement error was unknown, a spline that minimized the generalized cross validation function gave a good fit. When we knew the measurement error,

Table 5. Observations about the GCV/GA and Goodness-of-Fit Methods of Selecting Smoothing Splines with Guidance for Their Application

Spline Fitting Procedure	Applicability	Automation	Processing Requirements
GCV/GA	Generally applicable	Fully automated selection of the degree and parameters of a smoothing spline	Best with a powerful processor, if timing is a concern, such as Sun Sparcstation
goodness-of-fit	Applicable when measurement errors variances are known	User controls the smoothness by specifying the goodness-of-fit. Restricted to cubic splines	Satisfactory on low-power processor, such as IBM PC AT 80286/7

the spline could alternatively be fitted using a chi-square goodness-of-fit criterion, which gave the user control over the fit. In particular, a low goodness-of-fit spline extracted an overall trend from the data trajectory.

The GCV/GA splines and the 50% goodness-of-fit splines recognized real events in data from several penicillin fermentations while rejecting outliers and noisy points. The ability of the GCV/GA procedure to automatically select the degree of the spline, as well as its parameters, was a pleasing outcome of the work that illustrated the advantage of GCV/GA splines over the traditional engineering limitation to cubic spline smoothing.

The penicillin fermentation data records were time-varying and discontinuous, and the off-line records were sparse and contained features characterized by few points. Because of their success with the penicillin data we suggest that the spline fitting procedures outlined in this article could be applied to other nonsteady processes, including those which have discontinuous behavior. Table 5 summarizes our observations about the GCV/GA and the goodness-of-fit techniques and will help guide the choice of technique applicable to a particular smoothing problem.

Acknowledgment

This work has been supported by Capes (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), Brazil, by the SERC Advanced Centre for Biochemical Engineering, University College London and by the Centre for Process Systems Engineering at Imperial College of Science, Technology and Medicine.

Notation

$b_{j,\lambda}(x)$ = jth B-spline function of degree λ
$\underline{b}_\lambda(x_i)^T$ = row vector of B-splines, $j=1$ to $n-\lambda-1$, evaluated at x_i
$\boldsymbol{B}$ = matrix whose ith row is $\underline{b}_\lambda(x_i)^T$
$\underline{c}$ = vector of B-spline coefficients
$\underline{e}$ = vector of residuals from the spline fit
$\underline{f}$ = vector of weighted residuals from the spline fit
i = index for data points, $i=1$ to m
j = index for knots, $j=1$ to n
$J(n,\underline{c},p)$ = cost function
k = index for interior knots, $k=\lambda+2$ to $n-\lambda-1$
l = index for candidate solutions in the GA, $l=1$ to N
m = number of data points
n = number of knots
N = number of candidate solutions in the GA population
p = smoothness factor
$r_{k,j}$ = discontinuity jump in the highest derivative of the jth B-spline at knot k
$\underline{r}_j$ = column vector of $r_{k,j}$ for $k=\lambda+2$ to $n-\lambda-1$
$\boldsymbol{R}$ = matrix whose jth column is $\underline{r}_j$
$\underline{rc}$ = vector of discontinuities of the composite spline
S = fidelity constraint
$s_\lambda(x)$ = composite spline of degree λ fitted to experimental data y_i
$\underline{sx}$ = vector of samples from the spline $s_\lambda(x)$
t_j = abscissas of knots
$\boldsymbol{W}$ = weighting matrix
x_i = abscissas of experimental data points
y_i = ordinates of experimental data points
$\underline{y}$ = vector whose elements are y_i
$\|\underline{z}\|^2$ = GCV function

Greek letters

α = $(1-\alpha)$ is the chi-square goodness of fit
$\chi^2_{\alpha,\nu}$ = value of χ^2 for which the probability function $P(\chi^2|\nu)$ has the value α
λ = degree of the spline
ν = chi-square degrees of freedom
π_l = binary string representing a candidate solution of the GA problem

Literature Cited

Bates, D. M., M. J. Lindstrom, G. Wahba, and B. S. Yandell, "GCVPACK—Routines for Generalised Cross Validation," *Commun. Statist. Simula.*, **16**, 263 (1987).

Buono, M. A., S. S. Yang, and L. E. Erickson, "Comparison of Two Methods of Selecting Smoothing Spline Functions for Estimation of Specific Rates in Fermentations," *Chem. Eng. Commun.*, **45**, 145 (1986).

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

Craven, P., and G. Wahba, "Smoothing Noisy Data with Spline Functions," *Numer. Math.*, **31**, 377 (1979).

de Hoog, F. R., and M. F. Hutchinson, "An Efficient Method for Calculating Smoothing Splines using Orthogonal Transformations," *Numer. Math.*, **50**, 311 (1987).

Dierckx, P., "An Algorithm for Smoothing, Differentiation and Integration of Experimental Data using Spline Functions," *J. of Comput. and Appl. Math.*, **1**, 165 (1975).

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

Goldberg, D. E., *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison Wesley, Reading, MA (1989).

Holland, J. H., *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor (1975).

Hutchinson, M. F., and F. R. de Hoog, "Smoothing Noisy Data with Spline Functions," *Numer. Math.*, **47**, 99 (1985).

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

Manela, M., N. F. Thornhill, and J. A. Campbell, "Fitting Spline Functions to Noisy Data using a Genetic Algorithm," *Dept. of Computer Science Research Note RN/93/5*, Univ. College London (1993).

Oner, M. D., L. E. Erikson, and S. S. Yang, "Utilisation of Spline Functions for Smoothing Fermentation Data and for Estimation of Specific Rates," *Biotechnol. Bioeng.*, **28**, 902 (1986).

Perry, R. H., D. W. Green, and J. O. Maloney, eds., *Perry's Chemical Engineers' Handbook*, 6th ed., McGraw-Hill, New York (1984).

Pope, S. B., and R. Gadh, "Fitting Noisy Data Using Cross-Validated Cubic Smoothing Splines," *Commun. Statist. Simula.*, **17**, 349 (1988).

Press, W. H., B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge (1986).

Reinsch, C. H., "Smoothing by Spline Functions," *Numer. Math.*, **10**, 177 (1967).

Stone, K., F. W. Roche, and N. F. Thornhill, "Dry Weight Measurement of Microbial Biomass and Measurement Variability Analysis," *Biotechnol. Techniques*, **6**, 207 (1992).

Steemson, M. L., and E. T. White, "Numerical Modelling of Steady State Continuous Crystallization Processes Using Piecewise Cubic Spline Functions," *Comput. Chem. Eng.*, **12**, 81 (1988).

Tao, T. M., and A. T. Watson, "An Adaptive Algorithm for Fitting with Splines," *AIChE J.*, **34**, 1722 (1988).

Appendix: Derivation of Spline Coefficients and Related Functions

The derivation of the splines and their associated functions follows Dierckx (1975) and Craven and Wahba (1979). For a given spline with B-spline coefficients $\underline{c}$ and samples $\underline{sx}=\boldsymbol{B}\underline{c}$ fitted to a set of measurements $\underline{y}$ the weighted residuals are $\underline{f}=\boldsymbol{W}^{1/2}(\underline{sx}-\underline{y})$. The vector $\underline{rc}=\boldsymbol{R}\underline{c}$ is the vector of discontinuities at the interior knots, defined by Eqs. 3 to 5. With $\boldsymbol{E}=\boldsymbol{W}^{1/2}\boldsymbol{B}$ and $\underline{ym}=\boldsymbol{W}^{1/2}\underline{y}$, the cost function (Eq. 6) becomes:

$$J(n, \underline{c}, p) = p(E\underline{c} - \underline{ym})^T(E\underline{c} - \underline{ym}) + \underline{c}^T R^T R\underline{c} \qquad [A1]$$

The minimizing value of $\underline{c}$ in Eq. A1 is:

$$\hat{\underline{c}} = \left(E^T E + \frac{R^T R}{p}\right)^{-1} E^T \underline{ym}$$

The smoothing spline is generated from:

$$\underline{sx} = B\hat{\underline{c}} = X\underline{y}, \text{ where } X = W^{-1/2} E \left(E^T E + \frac{R^T R}{p}\right)^{-1} E^T W^{1/2}$$

The sum of squares function, $\|\underline{f}\|^2$, and the GCV function, $\|\underline{z}\|^2$, are given by:

$$\|\underline{f}\|^2 = \underline{y}^T (X - I)^T W (X - I)\underline{y}$$

$$\|\underline{z}\|^2 = \{[\text{diag}(X - I)]^{-1}\underline{f}\}^T \{[\text{diag}(X - I)]^{-1}\underline{f}\}$$

where I is the identity matrix and $\text{diag}(X - I)$ is a diagonal matrix whose elements are the diagonal elements of $(X - I)$. In the $\|\underline{z}\|^2$ computation, W is the identity matrix.

Manuscript received Apr. 28, 1993, and revision received July 6, 1993.