
DEPARTMENT OF STATISTICS

University of Wisconsin
Madison, Wisconsin 53706

TECHNICAL REPORT NO. 379

June 1974

A COMPLETELY AUTOMATIC FRENCH CURVE:

FITTING SPLINE FUNCTIONS BY CROSS VALIDATION

by

Grace Wahba* and Svante Wold**
University of Wisconsin
Madison, Wisconsin

Typist: Bernice R. Weitzel

* Research was supported by the Air Force Office of Scientific Research under Grant No. AFOSR 72-2363A.

** Statistician-in-Residence 1973-74.

Permanent address: Research Group for Chemometrics, Institute of Chemistry,
Umeå University, S 901 87, Umeå, Sweden.

A COMPLETELY AUTOMATIC FRENCH CURVE: FITTING SPLINE FUNCTIONS BY CROSS VALIDATION

G. Wahba

Department of Statistics, University of Wisconsin, Madison

S. Wold

Research Group for Chemometrics,
Institute of Chemistry, Umeå University, Umeå Sweden

ABSTRACT

The cross validation mean square error technique is used to determine the correct degree of smoothing, in fitting smoothing splines to discrete, noisy observations from some unknown smooth function. Monte Carlo results show amazing success in estimating the true smooth function as well as its derivative.

1. INTRODUCTION

We consider the problem of recovering a smooth function when only discrete, noisy measurements of it are available. We are interested in the situation where the particular form of the true function is not known, what is known is that the desired function is smooth. (We will define smoothness shortly). The fitted curves that result from our work look much like what one would expect from an experienced draftsman with a French curve.

Formally, the model is

$$(1.1) \qquad Y(t) = f(t) + \varepsilon(t), \quad t\varepsilon[0,1]$$

where $E\varepsilon(s)\varepsilon(t) = \sigma^2$, $s = t$, $E\varepsilon(s)\varepsilon(t) = 0$ $s \neq t$, and $f\varepsilon W_2^{(2)}$,

$$W_2^{(2)} = \{f: f,\ f' \quad \text{abs.cont.},\ f''\varepsilon \mathcal{L}_2[0,1]\}.$$

The noise variance σ^2 is generally unknown. The smallness of the integral

$$\int_0^1 (f''(t))^2 dt = M$$

is taken as the measure of smoothness of f. If f has a small M associated with it, it will visually appear not too wiggly. Since f is unknown, M is also unknown. $Y(t)$ is observed for $t = t_1, t_2, \ldots, t_n$, $0 \leq t_1 < t_2 < \ldots < t_n \leq 1$.

In this note we explore the use of the cubic smoothing spline for recovering f and also f', and, in particular, demonstrate the use of the cross-validation mean square error (CVMSE) technique for estimating from the data the appropriate degree of smoothing.

2. SMOOTHING SPLINES

Consider the solution to the problem: Find $f\varepsilon W_2^{(2)}$ to

$$\min\left\{\frac{1}{n}\sum_{j=1}^{n}(Y(t_j) - f(t_i))^2 + \lambda\int_0^1 (f''(t))^2\, dt\right\},$$

where λ is a non-negative given real number. The first term is a measure of fidelity to the data, and the second term is λ times the "smoothness" of f. It is well known (see Greville [6], Reinsch [11]) that for fixed $\lambda > 0$, there is a unique solution $f_{n,\lambda}$ in $W_2^{(2)}$. It is a so-called natural cubic spline, possessing the following properties:

i) $f_{n,\lambda}$ is a (possibly different) polynomial of at most degree 3 in each of the intervals $[0,t_1]$, $[t_j, t_{j+1}]$, $j = 1, 2, \ldots, n-1$, $[t_n,1]$,

ii) The polynomial pieces are joined so that f, f' and f''

are continuous,

iii) $f''_{n,\lambda}(0) = f''_{n,\lambda}(1) = 0$, and if $t_1 > 0$, $t_n < 1$, then $f'''_{n,\lambda}(t_1-) = f'''_{n,\lambda}(t_n+) = 0$.

As $\lambda \to \infty$, $f_{n,\lambda}$ becomes increasingly smooth, and the limiting function $f_{n,\infty}$ is the least squares straight line through the data. As $\lambda \to 0$,

$$\frac{1}{n} \sum_{j=1}^{n} (Y(t_j) - f_{n,\lambda}(t_j))^2 \to 0$$

until, in the limit, $f_{n,0}$ passes through the data. ($f_{n,0}$ is the natural cubic spline of interpolation to the data, see Schoenberg [13]).

The use of $f_{n,\lambda}(t)$, $t\varepsilon[0,1]$, as an estimate of $f(t)$, $t\varepsilon[0,1]$, when it is known that $f\varepsilon W_2^{(2)}$, is advantageous because of its nice convergence properties. See, for example Schultz [14] for convergence properties of interpolating splines $(\lambda = 0)$ and their derivatives, when there is no noise, i.e., $\sigma^2 = 0$. When there is noise, then λ must decrease suitably with n to obtain convergence. If this is done,

$$f_{n,\lambda}(t) \xrightarrow{q.m.} f(t)$$

$$f'_{n,\lambda}(t) \xrightarrow{q.m.} f'(t)$$

for all $f\varepsilon W_2^{(2)}$. (See Wahba [16], [17])

Now, consider the problem: Find $f\varepsilon W_2^{(2)}$ to

$$\min \int_0^1 (f''(t))^2 \, dt$$

subject to

$$\frac{1}{n} \sum_{j=1}^{n} (Y(t_j) - f(t_j))^2 \leq S,$$

where S is specified. It is well known [11] that if

$$S \geq \inf_{a,b} \sum_{j=1}^{n} (Y(t_j) - (a + bt_j))^2$$

then there exists a unique $\lambda = \lambda(S)$ such that $f_{n,\lambda}$ is the solution to this problem, and

$$\frac{1}{n} \sum_{j=1}^{n} (Y(t_j) - f_{n,\lambda}(t_j))^2 = S.$$

A computer program is available (Reinsch [11], [12]) which, given λ or S, and the data $\{t_i, Y(t_i)\}$, $i = 1, 2, \ldots, n$, delivers $f_{n,\lambda}(t)$ and $f'_{n,\lambda}(t)$. Reinsch suggests that S be chosen approximately as σ^2. (The S here is $\frac{1}{n}$ times the S in Reinsch's papers.)

3. HOW MUCH SMOOTHING SHOULD THERE BE?

Practically speaking, the choice of λ is critical, if λ is too small, the spline is too wiggly and picks up too much noise (overfit, Fig. 3), if λ is too large, the spline is too smooth and signal is lost (underfit, Fig. 4).

The present authors independently came to the conclusion that, in fact S should be chosen less than σ^2 by a fudge factor k, $0 < k < 1$, defined by

$$S = k\sigma^2.$$

Wold [19] showed by Monte Carlo methods that k in the cases tried should be between .7 and .95. Wahba [16] showed that for f having 3 continuous derivatives with $f^{(iv)} \varepsilon \mathcal{L}_2$, and satisfying certain boundary conditions, that the expected mean square error

$$E \frac{1}{n} \sum_{j=1}^{n} (f_{n,\lambda}(t_j) - f(t_j))^2$$

is minimized by setting

$$(1-k) = c \frac{\theta}{n^{8/9}} (1 + o(1)) \tag{3.1}$$

where c is a constant and

$$\theta = \left[\int_0^1 (f^{(iv)}(t))^2 \, dt \Big/ \sigma^2 \right]^{1/9}.$$

If S is chosen optimally, then favorable q.m. convergence rates are shown to obtain. Of course neither of these solutions tells us how to choose S (equivalently k if σ^2 is known) in practice, since both assume knowledge about f and σ^2 that is not generally available to the experimenter.

4. A PRACTICAL METHOD FOR DETERMINING THE DEGREE OF SMOOTHING

The CVMSE has shown great promise as a criterion for the determination of optimal fit in a number of applications. See Fienberg and Holland [5], Hocking [8], Mosteller and Wallace [10], Stone [15]. A fit giving a minimal CVMSE corresponds to a model representation of the data where the model gives the best prediction (in the least squares sense) of each data point by means of the model and the other data points. Another way to say this is that the minimum CVMSE gives the parameter(s) which maximize the internal consistency of the data set with respect to the applied model. It was therefore natural to try the CVMSE criteria for choosing S to obtain the optimal degree of smoothing. Our procedure goes as follows:

1. Divide the data set into p groups:

$$\begin{array}{ll} \text{Group 1:} & t_1, t_{1+p}, \cdots \\ \text{Group 2:} & t_2, t_{2+p}, \cdots \\ \vdots & \\ \text{Group p:} & t_p, t_{2p}, \cdots \end{array}$$

2. Guess a starting value of S

3. Delete the first group of data. Fit a smoothing spline to the remaining data using Reinsch's program with the S of Step 2. (Data deletion may be done cheaply by manipulating the weights in Reinsch's program). Compute the sum of squared deviations of this smoothing spline from the deleted data points.

4. Delete <u>instead</u> the second group of data. Fit a smoothing spline to the remaining data with the S of Step 2. Compute the sum of squared deviations of the spline from the deleted data points

5. Repeat Step 4. for the 3rd, 4th, ..., pth group of data.

6. Add the sums of squared deviations from steps 3 to 5 and divide by n. This is the CVMSE for S, denoted CV(S).

7. Vary S systematically and repeat steps 3-6 until CV(S) shows a minimum.

5. MONTE CARLO RESULTS

Some preliminary Monte Carlo experiments to test the validity of this procedure have been performed, and the results are extremely encouraging. The first simulation consisted of generating data for $f(t) = \sin t$, t_i equidistant between 0 and π, and

$$Y(t_i) = f(t_i) + \varepsilon_i, \qquad i = 1, 2, \ldots, n$$

where the ε_i are pseudo random numbers, independent and normally distributed with mean 0 and variance σ^2. 10 runs were made for each of the cases $n = 50$, $\sigma^2 = 10^{-6}, 10^{-4}, 10^{-2}, 1$ and $n = 100$, $\sigma^2 = 10^{-6}, 10^{-4}, 10^{-2}, 1$. We chose $p = 10$ A "run" consists of one set of n simulated data points. Figure 1 shows the cross validation mean square error $CV(S) = CV(k\sigma^2)$ plotted as a function of k, for four typical runs. The true mean square error $TR(k)$ and the derivative mean square error $D(k)$, defined by

$$TR(k) = \frac{1}{n} \sum_{j=1}^{n} (f_{n,\lambda}(t_j) - f(t_j))^2$$

$$D(k) = \frac{1}{n} \sum_{j=1}^{n} (f'_{n,\lambda}(t_j) - f'(t_j))^2$$

are also plotted in Figure 1. (Recall that λ is a function of k through the relationship $\lambda = \lambda(S) = \lambda(k\sigma^2)$, in practice, $CV(S))$ is plotted and its minimum determined). We plot CV, TR and D as a function of k for convenient comparison on the same range of abcissae. The ideal k (actually S) minimzes $TR(k)$ which is also not known. Note that in all four cases, two with small σ^2 and two with large, $CV(k)$ follows $TR(k)$ and the minimizing values $\hat{k}$ and k^* respectively are very close. More importantly, note that using $\hat{k}$ leads only to a slightly larger mean square error than the minimum attainable if the ideal k, namely k^*, were known. The following table gives values of the relative inefficiency $TR(\hat{k})/TR(k^*)$ for the four runs of Figure 1.

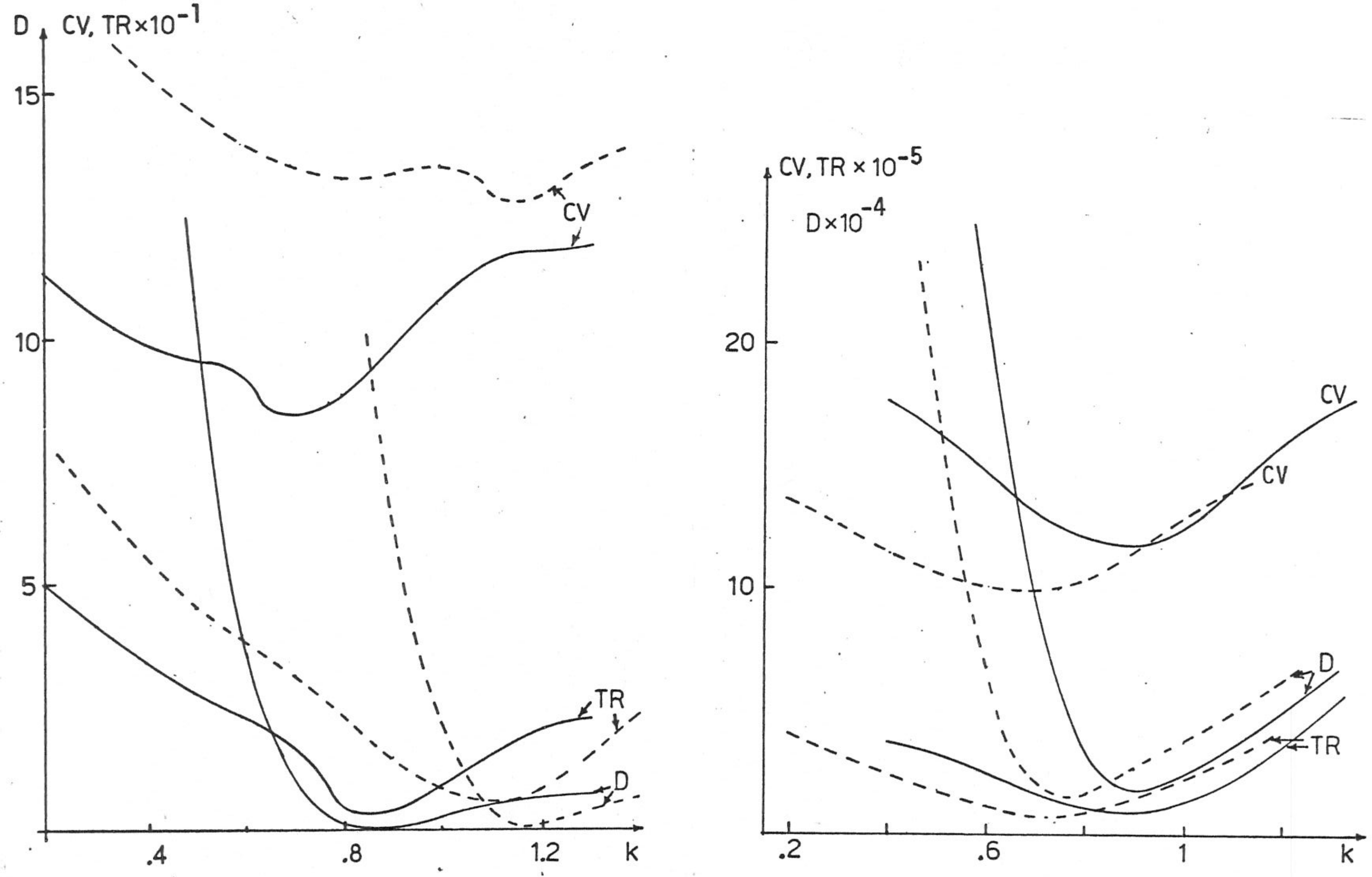

FIG. 1.

The cross validation mean square error (CV), the exact mean square error (TR) and the exact derivative mean square error (D) plotted against k for two runs with $\sigma = 1$ (left picture) and $\sigma = 0{,}01$ (right picture). Runs A, dashed lines, and runs B, solid lines. (Generating function $y = \sin x$; $x = 0{,}2\Pi$; $n = 100$)

TABLE I

The Relative Inefficiency $TR(\hat{k})/TR(k^*)$, Four Cases

σ^2 \ Run	A	B
10^{-4}	1.13	1.03
1	1.14	3.0

These four runs are typical of the other 76 and were selected before the results were examined. Run A with $\sigma^2 = 1$ actually has the 3rd largest relative inefficiency of all 80 cases.

A few cases of $p = 5$ were run but the results were not as good as these chosen here. The classical CVMSE frequently uses $p = n$, which corresponds to deleting one point at a time. This would no doubt give even better results, but is at least $n/10$ times as expensive to run, with the program we are using now.

Table II gives the mean and standard deviation (for each set of ten runs) of $u = TR(\hat{k})/TR(k^*)$.

TABLE II

The Relative Inefficiency $TR(\hat{k})/TR(k^*)$, Means and Standard Deviations of Ten Runs

n \ σ^2	10^{-6}		10^{-4}		10^{-2}		1	
	$\bar{u}$	σ_u	$\bar{u}$	σ_u	$\bar{u}$	σ_u	$\bar{u}$	σ_u
50	1.28	.29	1.23	.37	1.27	.65	2.49	2.0
100	1.07	.09	1.08	.09	1.10	.12	1.59	1.2

We found it remarkable how closely the CVMSE procedure comes to achieving the smallest possible mean square error in each run, over the range of parameters. Note from Figure 1 that $\hat{k}$ is quite a bit different for the 2 replicates for each set of parameters, nevertheless $\hat{k}$ comes close to k^* and $TR(\hat{k})$ comes close to $TR(k^*)$ each time.

Table III gives the mean and standard deviation (for each set of ten runs) of $v = D(\hat{k})/D(k^{**})$. k^{**} is that value of k which minimizes D(k).

TABLE III

The Relative Inefficiency for the Derivative, $D(\hat{k})/D(k^{**})$, Means and Standard Deviations of 10 runs.

n \ σ^2	10^{-6}		10^{-4}		10^{-2}		1	
	$\bar{u}$	σ_u	$\bar{u}$	σ_u	$\bar{u}$	σ_u	$\bar{u}$	σ_u
50	2.37	1.1	1.99	1.4	2.44	3.2	3.84	5.61
100	1.51	.68	1.53	.75	2.19	1.6	2.06	1.14

We believe that these results demonstrate the feasibility of this technique for estimating the derivative.

Figure 2 plots, on a log log scale, the average $(1-\hat{k})$ for each

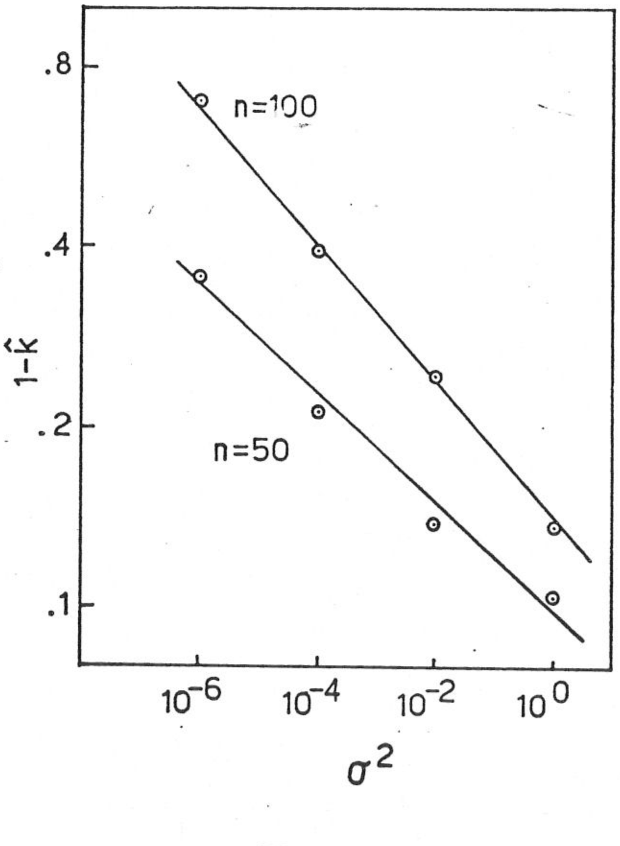

FIG. 2.

Average $(1-\hat{k})$ for 10 runs vs. σ^2.

group of 10 runs, for the four values of σ^2 and two values of n of Table II. The average k^*, and hence $\hat{k}$ can be expected to behave according to Equation (3.1), thus straight lines approximating this data should have a slope of -1/9. The actual slopes are about -.12 and -.09. The distance between the straight lines for n = 100 and n = 50 should be about 8/9 log[100/50] = .27 cycles. It is about .24 cycles.

We conclude this discussion with a visual demonstration of the kind of results one can expect. Figures 3-5 give data generated according to the model of (1.1) with n = 100, σ = .2 and $f(t) = 4.26(e^{-t} - 4e^{-2t} + 3e^{-3t})$. Figure 3 gives a smoothing spline with λ too small, and Figure 4 gives a smoothing spline with λ

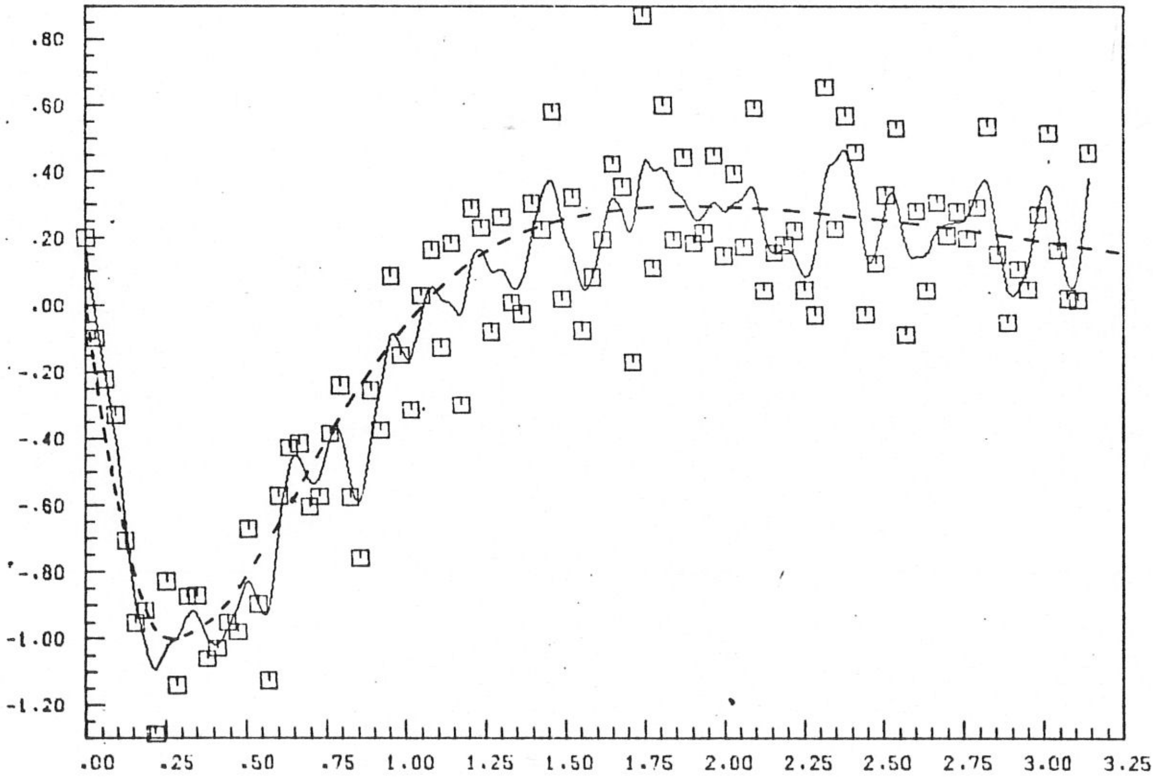

FIG. 3.

Data generated according to the model (1.1) with n = 100, σ = .2 and f(t) = 4.26 (exp(-t) - 4 exp(-2t) + exp(-3t)). Dashed curve is f(t). Solid curve is fitted spline with k = .5σ (too small).

too large. Figure 5 gives the smoothing spline with λ determined by the CVMSE criteria with p = 10. Considering the magnitude of the noise, we considered the result impressive.

Remarks:

1. It remains to prove that the S estimated by CVMSE, converges in an appropriate sense, to the optimal S. We believe a proof will be found. For $\sigma^2 = 10^{-6}$ we have noted a pronounced bias in the estimation such that $k^* - \hat{k}$ is invariably positive. This bias was barely perceptible at $\sigma^2 = 10^{-4}$ and not at all at 10^{-2} or 1. From our present understanding of the relation between CV(λ) and TR(λ) we believe that a theoretical relationship is: For <u>fixed</u> $\sigma^2 > 0$ $\min_\lambda CV(\lambda) \to \min_\lambda TR(\lambda)$ as $n \to \infty$.

2. This technique is related to but not the same as that of Andersen and Bloomfield [1] for obtaining the derivative

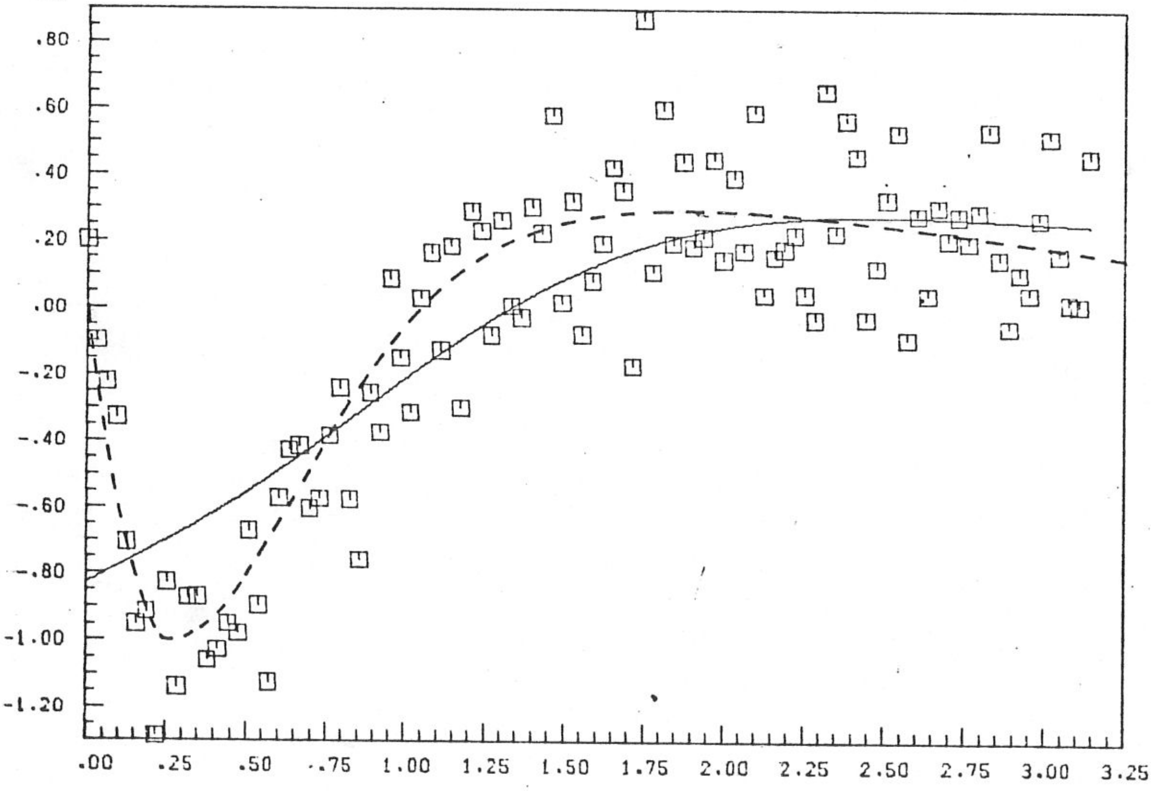

FIG. 4.

Same data as in Fig. 3. Spline (solid curve) is fitted with k = 2σ (too big).

f' from discrete noisy observations. See also Cullum [3]. The relation between the CVMSE technique for determining λ and Anderssen and Bloomfield's method is discussed in [18].

3. This method has great promise for determining the appropriate degree of smoothness for a spectral density estimate, when the spectral density is known to be smooth and bounded above and below. Let $Z(t)$, $t = \ldots, -1, 0, 1, \ldots$, be a zero mean stationary Gaussian process with spectral density f, let $Y(\omega)$ be the periodogram

$$Y(\omega) = \frac{1}{n}\left|\sum_{\tau=1}^{n} Z(\tau)e^{i\tau\omega}\right|^2 .$$

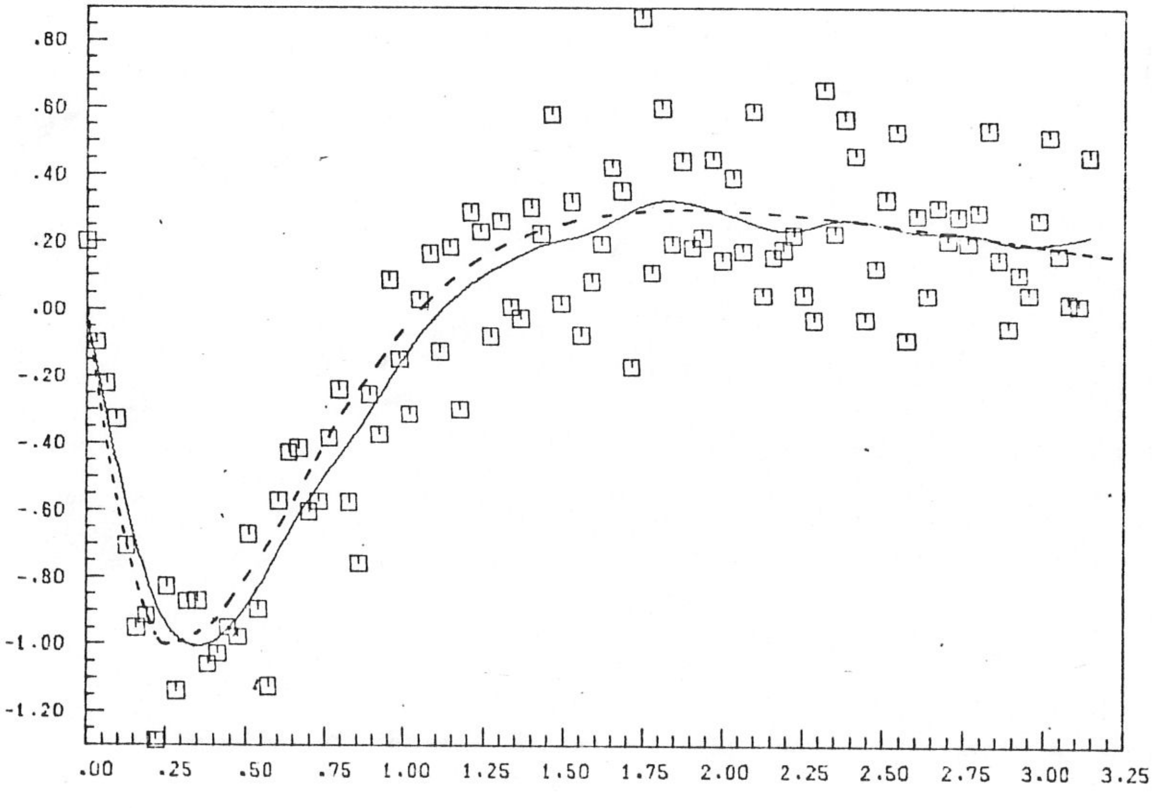

FIG. 5.

Same data as in Fig. 4. Spline (solid curve) is fitted with $k = .99\sigma$ which was the k-value corresponding to the minimum CVMSE.

Then (for n even)

$$Y(t_j) \simeq f(t_j) + \varepsilon_j, \quad t_j = 2\pi j/n, \quad j = 0, 1, \ldots, n/2$$

where the ε_j are approximately independent, zero mean random variables with the standard deviation of ε_j approximately $f(t_j)$ for $j = 1, 2, \ldots, n/2-1$, and $2f(t_j)$ for $j = 0, n/2$. Thus, the data above approximately fits our model, possibly with weights added. Cogburn and Davis [2] give a theoretical procedure for choosing λ which ensures a favorable rate of convergence. We believe the CVMSE is a reasonable estimate of the optimum λ. de Figueiredo and Thompson [4] also studied a smoothing spline spectral density estimate numerically. We would like to see some experimental work on CVMSE and smoothing splines for spectral density estimation. See [18].

4. The problem studied here is a special case of the problem of choosing the parameter λ in the method of regularization for solving noisy linear operator equations. Let

$$Y(t) = (Kf)(t) + \varepsilon(t), \qquad t\varepsilon[0,1]$$

where K is a linear operator on a Hilbert space $\mathcal{H}$ of real valued functions on $[0,1]$ with norm or semi-norm $||\cdot||_{\mathcal{H}}$, and with the property that

$$|Kf(t)| \leq C||f||_{\mathcal{H}}, \qquad t\varepsilon[0,1],$$

where C is a constant. For example, let $\mathcal{H} = W_2^{(2)}$ with the semi-norm $||f||_{\mathcal{H}} = \int_0^1 (f''(t))^2\, dt$ and $(Kf)(t) = \int_0^1 K(t,s)f(s)ds$, $K(t,s)$ continuous. The method of regularization estimates f as the solution $f_{n,\lambda}$ to the problem: Find $f\varepsilon\mathcal{H}$ to

$$\min \frac{1}{n} \sum_{i=1}^{n} (Y(t_i) - f(t_i))^2 + \lambda||f||_{\mathcal{H}}^2.$$

See Wahba [17] for explicit formulae for $f_{n,\lambda}$. The problem of choosing λ has been attacked by a large number of authors (see Wahba [17] for references) but no satisfactory practical solution seems to exist. See Hilgers [7] for some recent numerical experiments. We believe that the CVMSE technique can be shown to be a satisfactory practical solution to the problem of choosing λ in this problem.

5. It is shown in Wahba [17] that $E||f - f_{n,\lambda}||^2_{\mathcal{H}} \xrightarrow{q.m.} 0$ as $n \to \infty$ for $\mathcal{H} = W_2^{(2)}$, if λ is chosen suitably. This entails that if L is a continuous linear functional on $W_2^{(2)}$ (endowed with any of the usual norms) then

$$E|Lf - Lf_{n,\lambda}| \to 0,$$

so that e.g. we estimate $Lf \equiv \int_0^1 f(t)dt$ by $\int_0^1 f_{n,\lambda}(t)dt$. This suggests that we ought to be able to estimate some non-linear functionals like $\max_t f'(t)$ by $\max_t f'_{n,\lambda}(t)$.

6. As is usual and well known in these types of problems, there exists a Bayesian model which gives a smoothing spline as the posterior mean. It is as follows: Let $Y(t) = X(t) + \varepsilon(t)$ where $\varepsilon(t)$ is as before and $X(t)$ is a zero mean Gaussian process of the form

$$X(t) = \theta_1 + \theta_2 t + \int_0^t ds \int_0^s dW(u)$$

where the covariance matrix of θ_1 and θ_2 is $\gamma I_{2\times 2}$ and $W(u)$ is a Wiener process. Thus $X''(t)$, while it doesn't exist, is the formal derivative of a Wiener process (continuous time "white noise"). Let $\hat{X}_\gamma(t)$ be the posterior mean of $X(t)$, given the data

$Y(t_1), \ldots, Y(t_n)$, and let $\hat{X}(t) = \lim_{\gamma\to\infty} \hat{X}_\gamma(t)$. Then, $\hat{X}(t)$, considered as a function of t, is, for some λ, the cubic smoothing spline for the data $Y(t_1), \ldots, Y(t_n)$. (See Kimeldorf and Wahba [9]). Quintic smoothing splines are obtained by letting X''' be white noise. Then "smoothness" means $\int (f'''(t))^2\, dt$ is small. More generally, Tchebychev smoothing splines are obtained when $\sum_{j=0}^{n} a_j X^{(j)}$ is white noise. However, we prefer X'' white noise, because the concommittant definition of smoothness appears to be psychologically valid.

BIBLIOGRAPHY

[1] Anderssen, R. S., and Bloomfield, P. (1974). A time series approach to numerical differentiation. Technometrics 16, 1, 69-75.

[2] Cogburn, R., and Davis, H. T. (1973). Periodic splines and spectral estimation. University of New Mexico, Department of Mathematics and Statistics, Technical Report No. 253 (Rev.).

[3] Cullum, J. (1971). Numerical differentiation and regularization, SIAM J. Numer. Anal. 8, 254-265.

[4] de Figueiredo, R. J. P. and Thompson, J. R. (1972). Power spectral density estimation by spline smoothing in the frequency domain. Proc. Third Symp. on Nonlinear Estimation Theory and its applications, San Diego, California.

[5] Fienberg, S. E., and Holland, P. W. (1972). On the choice of flattening constants for estimating multinomial probabilities, J. Mult. Anal. 2, 127-134.

[6] Greville, T. N. E. (1969). Introduction to spline functions, in "Theory and Application of Spline Functions". T. N. E. Greville, Ed. Academic Press, New York. 1-36.

[7] Hilgers, J. (1973). Non-iterative methods for solving operator equations of the first kind. Univeristy of Wisconsin-Madison, Mathematics Research Center TSR # 1413.

[8] Hocking, R. R. (1972). Criteria for selection of a subset regression: Which one should be used? Technometrics 14, 967-970.

[9] Kimeldorf, G. S., and Wahba, G. (1971). Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. 33, 1, 82-95.

[10] Mosteller, F., and Wallace, D. L. (1963). Inference in an authorship problem. JASA, 58, 302, 275-309.

[11] Reinsch, C. H. (1967). Smoothing by spline functions. Numer. Math. 10, 177-183.

[12] Reinsch, C. H. (1971) Smoothing by spline functions II. Numer. Math. 16, 451-454.

[13] Schoenberg, I. (1967). On spline functions. In "Inequalities", O. Shisha, Ed., Academic Press, New York, 255-291.

[14] Schultz, M. (1970). Error bounds for polynomial spline interpolation, Math. Comp. 24, 111, 507-515.

[15] Stone, M. (1974). Cross-validation and multinomial prediction. Technical Report 35, Department of Statistics, University of Michigan, Ann Arbor.

[16] Wahba, G. (1974). Smoothing noisy data by spline functions, II. University of Wisconsin-Madison, Department of Statistics, Technical Report No. 380.

[17] Wahba, G. (1973). Convergence properties of the method of regularization for noisy linear operator equations. University of Wisconsin-Madison, Mathematics Research Center, TSR No. 1132.

[18] Wahba, G., and Wold, S. (1974). Periodic splines for spectral density estimation: The use of cross validation for determining the degree of smoothing. University of Wisconsin-Madison, Department of Statistics, Technical Report No. 381.

[19] Wold, S. (1974). Spline functions in data analysis. Technometrics 16, 1, 1-11.