Numer. Math. 47, 99-106 (1985)

Numerische Mathematik
© Springer-Verlag 1985

Smoothing Noisy Data with Spline Functions

M.F. Hutchinson and F.R. de Hoog

CSIRO Division of Mathematics and Statistics, GPO Box 1965, Canberra, ACT 2601, Australia

Summary. A procedure for calculating the trace of the influence matrix associated with a polynomial smoothing spline of degree $2m-1$ fitted to n distinct, not necessarily equally spaced or uniformly weighted, data points is presented. The procedure requires order $m^2 n$ operations and therefore permits efficient order $m^2 n$ calculation of statistics associated with a polynomial smoothing spline, including the generalized cross validation. The method is a significant improvement over an existing method which requires order n^3 operations.

Subject Classifications: AMS(MOS): 65D, 65K; CR: G.1.2, G.1.1.

1. Introduction

Since its introduction by Schoenberg [13] and Reinsch [11], the polynomial smoothing spline has provided an attractive way of smoothing noisy data values observed at n distinct points on a finite interval. Craven and Wahba [1] have shown how to choose the degree of smoothing of this spline objectively, both when the amount of noise associated with the data is known, and when it is not known. In the first case, one may minimize the expected mean square error over the data points, and in the second case, one may minimize the generalized cross validation (GCV), a procedure which is asymptotically the same as minimizing the expected mean square error. However, in each case, the function to be minimized involves the trace of the influence matrix associated with the smoothing spline. An existing method lor calculating the trace in the non-equally spaced data point case, as described in Craven and Wahba [1] and implemented by IMSL [10], is expensive, requiring order n^3 operations and approximately n^2 storage locations. Utreras [15] has provided a method for calculating an approximation to the trace in order n operations when the weighting is uniform and the data points are equally spaced. He has indicated in [16] a method for calculating an approximation to the trace in the non-equally spaced case in order n^2 operations.

In this article we provide a method for calculating the trace in the general, not necessarily equally spaced or uniformly weighted case, which requires just order $m^2 n$ operations and order mn storage locations. This overcomes the principal objection raised to the method of Craven and Wahba [1] by Weinert et al. [21] and Wecker and Ansley [20]. A second objection raised in [20] concerning repeated observations is not valid, since these may be taken into account quite rigorously, as apparently realized by Reinsch [11], by taking the mean of each set of repeated observations and setting the relative weight of each data point appropriately.

Our method also provides the diagonal elements of the influence matrix which may be used, as indicated in Wahba [19], to provide confidence intervals for the smoothed data values. The method depends on being able to calculate the central $2m+1$ bands of the inverse of an $(n-m)\times(n-m)$ symmetric, positive definite, $(2m+1)$-banded matrix in order $m^2 n$ operations.

2. Mathematical Preliminaries

A model for which the polynomial smoothing spline is applicable goes as follows. Let $x_1 < \dots < x_n$ be a set of n ordered points on a finite interval and let $y_1, \dots, y_n$ be a corresponding set of noisy observations given by

$$y_i = g(x_i) + \varepsilon_i \qquad (i=1, \dots, n) \tag{2.1}$$

where g is a suitably smooth, but unknown, function and the ε_i are random errors satisfying

$$\begin{aligned} &E(\varepsilon_i)=0, \\ &E(\varepsilon_i \varepsilon_j)=0 \quad \text{for } i \neq j, \\ &E(\varepsilon_i^2)=w_i^2 \sigma^2 \end{aligned} \tag{2.2}$$

where E denotes expectation. The w_i are known positive constants while the value of σ^2 may or may not be known. A polynomial spline function of degree $2m-1$ (m an integer $\geqq 1$) arises (see [3, 11-13]) as the unique real valued function f, with absolutely continuous $(m-1)$-st derivative and square integrable m-th derivative, which minimizes

$$p \sum_{i=1}^{n} \left[\frac{y_i - f(x_i)}{w_i}\right]^2 + \int_{-\infty}^{\infty} (f^{(m)})^2\, dx \tag{2.3}$$

where p is positive. Here $f^{(m)}$ denotes the mth derivative of f and p controls the amount of smoothing of the data. Let f_p denote the function minimizing (2.3). According to [12], (see [17] for a slightly different development involving B-splines) the mth derivative of f_p may be expressed as

$$f_p^{(m)} = \sum_{i=1}^{n-m} c_i M_i \tag{2.4}$$

where M_i are the minimum support splines of Curry and Schoenberg [2]. The coefficients $c = (c_1, \dots, c_{n-m})^T$ and $a = (f_p(x_1), \dots, f_p(x_n))^T$ uniquely determine f_p.

They can be obtained from the linear system

$$(G^T W^2 G + pH)c = pG^T y \tag{2.5}$$

$$a = y - \frac{1}{p} W^2 G c \tag{2.6}$$

where $y=(y_1, \ldots, y_n)^T$, $W=\operatorname{diag}(w_1, \ldots, w_n)$, and H and $G^T W^2 G$ are symmetric, positive definite band matrices of bandwidth $2m-1$ and $2m+1$ respectively. The elements of H are given by

$$h_{ij} = \int_{-\infty}^{\infty} M_i M_j \, dx \tag{2.7}$$

and G is an $(m+1)$-banded, lower triangular, $n \times (n-m)$ matrix with elements in the ith column given by the coefficients of the mth order divided differences based on $x_i, \ldots, x_{i+m}$.

Let the coefficient matrix of (2.5) be denoted by

$$B_p = (G^T W^2 G + pH). \tag{2.8}$$

The *influence matrix* associated with the smoothing spline f_p is the unique $n \times n$ symmetric matrix A_p satisfying

$$a = A_p y. \tag{2.9}$$

From (2.5), (2.6), (2.8) we have

$$y - a = W^2 G B_p^{-1} G^T y \tag{2.10}$$

so that

$$I - A_p = W^2 G B_p^{-1} G^T. \tag{2.11}$$

The total, squared, weighted residual is given by

$$F(p) = \|W^{-1}(I - A_p) y\|^2 = \|WGB_p^{-1} G^T y\|^2 \tag{2.12}$$

where $\|.\|$ denotes the usual L^2-norm in n dimensional Euclidean space.

The algorithm of Reinsch [11, 12] uses repeated rational Cholesky decompositions of the coefficient matrix B_p for different values of p in order to determine the value of p (and the smoothing spline f_p) such that the residual $F(p)=S$, where S is a non-negative number no greater than $F(0)$. Reinsch suggests that S should be approximately $n\sigma^2$ when σ^2 is known, but leaves open the question of how to determine S when σ^2 is unknown. Reinsch's [11] algorithm for the case $m=2$ requires approximately $30n$ operations to calculate $F(p)$, of which only $16n$ need to be performed again to calculate $F(p)$ for each different value of p. Here one operation consists of one multiplication (or division) and one addition (or subtraction).

Wahba [18] has indicated that Reinsch's suggestion when σ^2 is known leads to systematic oversmoothing and Craven and Wahba [1] show that it is preferable to choose p in order to minimize an unbiased estimate of the expected true mean square error given by

$$T_p = \frac{1}{n} \|W^{-1}(I - A_p)y\|^2 - (2\sigma^2/n)\, Tr(I - A_p) + \sigma^2 \tag{2.13}$$

where Tr denotes the trace. Moreover, when σ^2 is unknown, Craven and Wahba [1] show that p may be chosen to minimize the generalized cross validation (GCV) given by

$$V_p = \frac{\frac{1}{n}\|W^{-1}(I - A_p)\, y\|^2}{\left[\frac{1}{n}\, Tr(I - A_p)\right]^2} \tag{2.14}$$

since the minimizer of V_p is asymptotically the same as the minimizer of T_p. Practical minimization of either T_p or V_p therefore requires efficient calculation of $F(p) = \|W^{-1}(I - A_p)y\|^2$ and of $Tr(I - A_p)$.

The algorithm suggested in [1], which has been implemented in subroutine ICSSCV of [10] for the case $m=2$, first calculates the singular value decomposition of $WGH^{-\frac{1}{2}}$, after which $F(p)$ and $Tr(I - A_p)$ may be calculated in approximately $3n$ operations for each value of p. This method avoids the explicit solution of Eq. (2.5) which involves the potentially ill-conditioned matrix B_p. However, the singular value decomposition requires order n^3 operations and approximately n^2 storage locations and is therefore impracticable for large values of n.

Utreras [15] has presented an approximate method for calculating $Tr(I - A_p)$ in $2n$ operations for the special case when the data points are equally spaced and uniformly weighted. We show how to calculate $Tr(I - A_p)$ in the general case, from the rational Cholesky decomposition of B_p in just $(m+1)^2 n$ operations. Since this decomposition of B_p may also be used, as in [11], to calculate $F(p)$, this leads to an efficient order $m^2 n$ algorithm for evaluating, and minimizing, either T_p or V_p.

3. The Main Result

The proposed method depends on the following theorem for obtaining the central bands of the inverse of a banded matrix. Several authors have developed recursive formulae appropriate for this problem, notably [6–9] and [14]. We follow one of the earliest and simplest approaches as described in [7, 14]. Note that the method described in [9] differs from the others in not requiring a Cholesky factorization.

Theorem 3.1. *Let B be a $(2m+1)$-banded, $n \times n$ matrix product of the form*

$$B = U^T D^{-1} U \tag{3.1}$$

where D is a diagonal matrix with positive diagonal elements and U is a real, unit upper triangular matrix of bandwidth $m+1$. Then the central $2m+1$ bands of B^{-1} may be obtained from (3.1) by performing $\frac{1}{3}(m-1)m(m+1) + (n-m)m(m+1)$ operations.

Proof. Let $B^{-1}=(\hat{b}_{ij})_{i,j=1}^{n}$, $U=(u_{ij})_{i,j=1}^{n}$ and $D=\operatorname{diag}(d_1,\dots,d_n)$. Since B^{-1} is symmetric, it is sufficient to calculate the upper $m+1$ bands of B^{-1} whose elements are given by

$$\hat{b}_{i,i+k} \quad (i=1,\dots,n;\ k=0,\dots,\min(m,n-i)) \tag{3.2}$$

Following [7, 14] we have

$$B^{-1}=DU^{-T}+(I-U)B^{-1}, \tag{3.3}$$

which may be easily obtained from (3.1). Since DU^{-T} is lower triangular and U is $(m+1)$-banded, unit upper triangular, this gives rise to the following recurrence formulae for the upper triangular elements of B^{-1},

$$\hat{b}_{i,i+l}=-\sum_{k=1}^{\min(m,n-i)} u_{i,i+k}\hat{b}_{i+k,i+l} \qquad (l>0) \tag{3.4}$$

and

$$\hat{b}_{ii}=d_i-\sum_{l=1}^{\min(m,n-i)} u_{i,i+l}\hat{b}_{i,i+l}. \tag{3.5}$$

For each i formula (3.4) expresses $\hat{b}_{i,i+l}$, for each $l=1,\dots,\min(m,n-i)$, in terms of elements of the ith row of U and previously calculated elements $b_{i+k,i+l}$ $(k>0)$, in $\min(m,n-i)$ operations. Formula (3.5) then expresses $\hat{b}_{ii}$ in terms of elements of the ith rows of D and U, and elements calculated by (3.4), also in $\min(m,n-i)$ operations. In particular, for the first step of the procedure, we have $\hat{b}_{nn}=d_n$. The total number of operations for the whole procedure is then easily seen to be given by

$$\sum_{k=0}^{m-1} k(k+1)+(n-m)m(m+1)=\tfrac{1}{3}(m-1)m(m+1)+(n-m)m(m+1).$$

Remarks. The elements $\hat{b}_{ii}, \hat{b}_{i,i+1},\dots,\hat{b}_{i,i+m}$ may overwrite the storage locations for d_i, $u_{i,i+1},\dots,u_{i,i+m}$ respectively using just one additional storage location, provided that formula (3.5) is used to progressively update d_i to $\hat{b}_{ii}$ as each $\hat{b}_{i,i+l}$ is calculated. It is computationally more straightforward however, to provide m additional storage locations to temporarily store the elements $u_{i,i+1},\dots,u_{i+m}$. The elements of each additional upper band of B^{-1} may be similarly calculated using formula (3.4), each element requiring exactly m operations. The complete upper triangle of B^{-1} may therefore be calculated in $\frac{1}{3}(m-1)m(m+1)+\frac{1}{2}(n-m)m(n+m+1)$ operations. Since every $n\times n$ matrix has bandwidth no greater than $2(n-1)+1$, the method may also be applied to full matrices, giving an operation count of $\frac{1}{3}(n-1)n(n+1)$. This is the same as for the standard method which takes advantage of the triangularity of U (see p. 3.16 of [4]).

We now proceed to the calculation of $Tr(I-A_p)$. Firstly, one may form the rational Cholesky decomposition of the $(n-m)\times(n-m)$ matrix B_p in approximately $\frac{1}{2}(m+1)(m+2)n$ operations, giving

$$B_p=U_p^T D_p^{-1} U_p \tag{3.6}$$

where D_p and U_p satisfy the conditions of Theorem 3.1. The central $2m+1$ bands of B_p^{-1} may therefore be calculated from (3.6) in no more than $m(m+1)n$ operations. Using (2.11) above and an elementary property of the trace, we have

$$Tr(I-A_p)=Tr(G^T W^2 G B_p^{-1}). \tag{3.7}$$

Since the $(n-m)\times(n-m)$ matrix $G^T W^2 G$ is symmetric and $(2m+1)$-banded, it is easy to see that $Tr(I-A_p)$ may be calculated from $G^T W^2 G$ and the central $2m+1$ bands of B_p^{-1} in no more than $(m+1)n$ operations. Thus $Tr(I-A_p)$ may be calculated from (3.6) in no more than $(m+1)^2 n$ operations. The terms of (3.7) may be rearranged to give

$$\begin{aligned} Tr(I-A_p) &= Tr((B_p-pH)B_p^{-1}) \\ &= n-m-p\,Tr(HB_p^{-1}) \end{aligned} \tag{3.8}$$

which may be calculated in n fewer operations since H is $(2m-1)$-banded, but this formula cannot be used in general since all accuracy is lost as p approaches ∞ and $p\,Tr(HB_p^{-1})$ approaches $n-m$. It is however quite accurate for small values of p.

The matrix B_p becomes ill-conditioned for small values of p, or when the data points are very unequally spaced, leading to loss of accuracy in the calculation of the Cholesky decomposition. This problem can be alleviated if, instead of forming the Cholesky decomposition of B_p, one performs a QR factorization of the $(2n-m)\times(n-m)$ matrix

$$Z=\begin{bmatrix} WG \\ p^{\frac{1}{2}} R \end{bmatrix}$$

where $R^T R$ is the Cholesky factorization of H, in the manner described by Eldén [5] at the expense of approximately 4 times as many operations. We will further investigate more accurate ways of calculating $Tr(I-A_p)$ and $F(p)$ elsewhere.

Finally, note that the diagonal elements of A_p, which can be used to provide confidence intervals for the smoothed data values (see [19]), may be calculated using (2.11) and the central $2m+1$ bands of B_p^{-1} in $(m+1)(m+2)n$ operations.

4. Numerical Results

An algorithm based on the algorithm of Reinsch [11] and Theorem 3.1 above, for determining the cubic smoothing spline f_p and its generalized cross validation V_p (or its true mean square error estimate T_p) for each p now goes as follows:

(i) Compute H, $G^T W^2 G$ and $G^T y$.

(ii) Compute the rational Cholesky decomposition of $B_p=(G^T W^2 G+pH)$.

(iii) Compute u from $B_p u=G^T y$ using (i), (ii).

(iv) Compute $v = WGu$ and $F(p) = v^T v$ (see (2.12)).

(v) Compute the central 5 bands of B_p^{-1} using (ii) and Theorem 3.1.

(vi) Compute $Tr(I - A_p) = Tr(G^T W^2 G B_p^{-1})$ using (i), (v).

(vii) Compute V_p (or T_p) using (iv), (vi).

(viii) Compute $a = y - Wv$, $c = pu$ and the remaining coefficients of f_p (see [11]).

If V_p or T_p is to be minimized then step (i) need only be performed once. Steps (ii), ..., (vii) may then be repeated in a global search for the optimal value of p, after which step (viii) may be performed. Steps (i) and (viii) require approximately $21n$ operations while the repeated steps (ii), ..., (vii) require a total of approximately $25n$ operations.

The above algorithm was implemented in double precision on a VAX 750 computer without floating point hardware in standard FORTRAN V. The search method employed to minimize V_p was the same as that used for the approximate method of Utreras [15] as implemented in subroutine ICSSCV of [10]. Average execution times for our algorithm and for the double precision versions of the algorithms of Utreras [15] and Craven and Wahba [1], as implemented in [10], are presented in Table 1. The order n property of our algorithm is clear. Its execution times are almost the same as those for the approximate method of Utreras [15]. They are dramatically less than the times for the original Craven and Wahba [1] algorithm. Times for our procedure when applied to non-equally spaced, non-uniformly weighted data are similar to those for equally spaced, uniformly weighted data as given in Table 1. Source code for this procedure may be obtained from the authors on request.

Table 1. Execution times in seconds for the proposed algorithm and for the algorithms of Utreras [15] and Craven and Wahba [1]

Number of data points	Proposed algorithm	Utreras	Craven and Wahba
50	4	4	179
100	11	11	1,358
200	25	24	10,474
400	53	50	–
800	108	104	–

References

1. Craven, P., Wahba, G.: Smoothing noisy data with spline functions. Numer. Math. **31**, 377–403 (1979)
2. Curry, H.B., Schoenberg, I.J.: On polya frequency functions IV: The fundamental spline functions and their limits. J. Anal. Math. **17**, 71–107 (1966)
3. de Boor, C.: A Practical Guide to Splines. Appl. Math. Sci. vol. 27. New York: Springer 1978
4. Dongarra, J.J., Moler, C.B., Bunch, J.R., Stewart, G.W.: Linpack User's Guide. Philadelphia: Society for Industrial and Applied Mathematics 1979
5. Eldén, L.: An algorithm for the regularization of ill-conditioned banded least squares problems. SIAM J. Sci. Stat. Comput. **5**, 237–254 (1984)

6. Eldén, L.: A note on the computation of the generalized cross-validation function for ill-conditioned least squares problems. BIT **24**, 467–472 (1984)
7. Erisman, A.M., Tinney, W.F.: On computing certain elements of the inverse of a sparse matrix. Commun. ACM **18**, 177–179 (1975)
8. Golub, G.H., Plemmons, R.J.: Large-scale geodetic least-squares adjustment by dissection and orthogonal decomposition. Lin. Alg. Appl. **34**, 3–27 (1980)
9. Haley, S.B.: Solution of band matrix equations by projection-recurrence. Lin. Alg. Appl. **32**, 33–48 (1980)
10. IMSL: Library Reference Manual, edition 9. Houston: IMSL 1982
11. Reinsch, C.H.: Smoothing by spline functions. Numer. Math. **10**, 177–183 (1967)
12. Reinsch, C.H.: Smoothing by spline functions, II. Numer. Math. **16**, 451–454 (1971)
13. Schoenberg, I.J.: Spline functions and the problem of graduation. Proc. Natl. Acad. Sci. USA **52**, 947–950 (1964)
14. Takahashi, K., Fagan, J., Chin, M.-S.: Formation of a sparse bus impedance matrix and its application to short circuit study. Power Industry Computer Applications Conf. Proc. Minneapolis, Minn. **8**, 63–69 (June 4–6, 1973)
15. Utreras, F.: Sur le choix de parametre d'adjustement dans le lissage par fonctions spline. Numer. Math. **34**, 15–28 (1980)
16. Utreras, F.: Optimal smoothing of noisy data using spline functions. SIAM J. Sci. Stat. Comput. **2**, 349–362 (1981)
17. Utreras, F.: Natural spline functions, their associated eigenvalue problem. Numer. Math. **42**, 107–117 (1983)
18. Wahba, G.: Smoothing noisy data with spline functions. Numer. Math. **24**, 383–392 (1975)
19. Wahba, G.: Bayesian "confidence intervals" for the cross-validated smoothing spline. J.R. Stat. Soc., Ser. B **45**, 133–150 (1983)
20. Wecker, W.E., Ansley, C.F.: The signal extraction approach to non linear regression and spline smoothing. J. Am. Stat. Assoc. **78**, 81–89 (1983)
21. Weinert, H.L., Byrd, R.H., Sidhu, G.S.: A stochastic framework for recursive computation of spline functions: Part II, smoothing splines. J. Optimization Theory Appl. **30**, 255–268 (1980)

Received October 25, 1984 / March 18, 1985