ROCKY MOUNTAIN
JOURNAL OF MATHEMATICS
Volume 19, Number 1, Winter 1989

CONTINUOUS APPROXIMATION METHODS FOR THE REGULARIZATION AND SMOOTHING OF INTEGRAL TRANSFORMS

R.P. BENNELL AND J.C. MASON

ABSTRACT. Continuous approximation methods are described for obtaining a numerical solution $f(t)$ to an integral transform $g(s) = \int K(s,t)f(t)dt$, where the given function $g(s)$ may be affected by noise and where the problem may be ill-posed. The approximate solution is expressed in the linear form $f^* = \sum a_j\phi_j$, where a_j are parameters and ϕ_j are certain basis functions; the values of a_j are determined by the minimization of a regularising/smoothing measure, which takes account of both the discrete l_2 error in the integral transform and the continuous L_2 norm of f^* or of one of its derivatives. A Generalized Cross-Validation technique, based on the work of G. Wahba, is used for determining the smoothing parameter, and efficient algorithms are developed for three specific sets of basis functions $\{\phi_j\}$, including a novel algorithm when $\{\phi_j\}$ are chosen to be a set of eigenfunctions. Numerical examples are given to compare the merits of the various algorithms. In the case where the function $g(s)$ is not affected by noise, the established "Method of Truncated Solutions" is adopted and an improved version of this method, based on B-splines, is described and then tested on numerical examples.

1. Introduction. Consider the Fredholm integral equation of the first kind

$$\int_a^b K(s,t)f(t)dt = g(s), \qquad c \le s \le d, \tag{1}$$

where $K(s,t)$ is a given kernel and $g(s)$ is a given function, (the range $[c,d]$ of values of s, does not necessarily coincide with the range of integration $[a,b]$).

We may write (1) in the form of an operator equation

$$Kf = g \quad \text{where } K : F \to G, \tag{2}$$

Received by the editors on October 15, 1986.

which is a well-posed problem provided that it has a unique solution $f \in F$, which depends continuously on $g \in G$. However, it is easily shown [**1**], that such an equation may be ill-posed, and hence a small perturbation in g may result in a large change in f; it is therefore appropriate to adopt a regularisation method to obtain numerical solutions, based on the minimization of a measure of smoothness (see Tikhonov [**2**], Ribiére [**3**]).

In general we shall consider the problem (1) where the function $g(s)$ is not known explicitly but where n measured observations of g are given with a possible "white-noise" contamination. That is, the data are

$$(3) \qquad g(s_i) = \int_a^b K(s_i,t)f(t)dt + \varepsilon(s_i), \quad i = 1,2,\dots,n,$$

where $\varepsilon(s_i) \sim N(0,\sigma)$, $i = 1,2,\dots,n$, (i.e., $\varepsilon(s_i)$ are independent errors with a normal distribution of mean zero and common variance σ^2 (unknown)).

Two important continuous approximation methods have been adopted in the literature for the problem (1). Firstly, Baker et al. [**4**] and Lewis [**5**] have considered a "Method of truncated solutions", for problems where $g(s)$ is exact, based on eigenfunction expansions. This method is discussed in §2 below, where it is modified to adopt B-splines and deal with data in the discrete form. Secondly, Wahba [**6**] has introduced a cross-validation method for coping with the general data (3), based on a reproducing kernel Hilbert space formulation. This method is studied here for the "natural" basis $\{K(s_i,t)\}$ and it is shown to have certain advantages but to be of limited practical applicability. Two other bases are considered in the context of cross-validation, namely B-splines and eigenfunctions. Our use of eigenfunctions in the context of a generalized cross-validation appears to be new.

2. The method of truncated solutions. Baker et al. [**4**] and Lewis [**5**] have studied the "Method of truncated solutions" for obtaining a numerical solution to (1) when the kernel is symmetric, i.e., $K(s,t) = K(t,s)$.

On defining eigenvalues λ_j and corresponding eigenfunctions $\phi_j(s)$,

$j = 1, 2, \ldots$, of the kernel $K(s,t)$ by the relation

$$\int_a^b K(s,t)\phi_j(t)dt = \lambda_j \phi_j(s), \qquad j = 1, 2, \ldots, \tag{4}$$

it is easy to deduce that, for a symmetric kernel, $\{\phi_j(s)\}$ is an orthogonal system which may be made orthonormal by choosing

$$\int_a^b [\phi_j(t)]^2 dt = 1. \tag{5}$$

The first step in the method is to determine, by a least squares or related technique, an approximation $g_m(s)$ to $g(s)$ of the form

$$g_m(s) = \sum_{j=1}^{m} a_j \phi_j(s). \tag{6}$$

The second step is then to determine an approximation $f_m(t)$ to $f(t)$ by solving (1) with g replaced by g_m, namely

$$\int_a^b K(s,t) f_m(t)dt = g_m(s). \tag{7}$$

It immediately follows from (4), (6), (7) that f_m may be determined explicitly in the form

$$f_m(t) = \sum_{j=1}^{m} a_j \lambda_j^{-1} \phi_j(t), \tag{8}$$

where the λ_j are ordered such that $|\lambda_j| > |\lambda_{j+1}|$.

In practice the eigenfunctions $\phi_j(s)$ are not determined exactly, but rather they are fitted by splines. Whereas Lewis [**5**] used a truncated power function basis, we here adopt a B-spline basis. A particular eigenfunction $\phi(t)$, corresponding to an eigenvalue λ, is approximated in the form

$$\phi(t) = \sum_{j=1}^{N} c_j N_{4j}(t),$$

where N_{4j} denotes a normalized cubic B-spline based on the j^{th} knot.

Substitution in (4) (deleting j) gives

$$\sum_{j=1}^{N} c_j \int_a^b K(s,t)N_{4j}(t)dt = \lambda \sum_{j=1}^{N} c_j N_{4j}(s). \tag{9}$$

On approximating the integrals by B-splines using a least squares or comparable procedure,

$$P_j(s) = \int_a^b K(s,t)N_{4j}(t)dt = \sum_{i=1}^{N} b_{ij} N_{4i}(s), \tag{10}$$

and, substituting into (9) and discretizing over N points in s, it follows that

$$\underline{\mathrm{N}}^T B\underline{c} = \lambda \underline{\mathrm{N}}^T \underline{c}, \tag{11}$$

where $\underline{\mathrm{N}} = [N_{4j}(s_k)]$, $B = [b_{ij}]$, $\underline{c} = [c_j]$.

Since B-splines are linearly independent, we obtain

$$B\underline{c} = \lambda\underline{c}, \tag{12}$$

namely a linear algebraic eigenvalue problem for λ, $\underline{c}$.

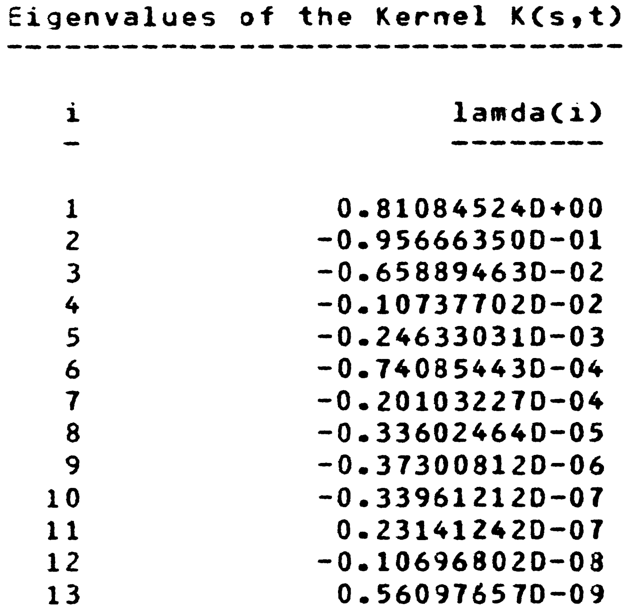
(If we prefer to discretise over more than N points in s, then we must use (11) rather than (12) as our algebraic eigenvalue problem).

A number of techniques were tested by us for determining the coefficients a_j in (6), based on suggestions in [**3**], and the optimal procedure in practice was to minimize

$$\sum_{i=1}^{n} \sum_{j=1}^{m} [a_j \phi_j(s_i) - g(s_i)]^2 \tag{13}$$

where s_i are the data points.

The choice of m, the number of eigenfunctions, is critical in this method. If it is too small, then accuracy is inadequate; if it is too large, then highly oscillatory eigenfunctions corresponding to small eigenvalues may "swamp" the solution.



```
Eigenvalues of the Kernel K(s,t)
--------------------------------

 i                      lamda(i)
 -                      --------

 1                    0.81084524D+00
 2                   -0.95666350D-01
 3                   -0.65889463D-02
 4                   -0.10737702D-02
 5                   -0.24633031D-03
 6                   -0.74085443D-04
 7                   -0.20103227D-04
 8                   -0.33602464D-05
 9                   -0.37300812D-06
10                   -0.33961212D-07
11                    0.23141242D-07
12                   -0.10696802D-08
13                    0.56097657D-09
```

TABLE 1

2.1. *Numerical results.* We have successfully tackled many problems of the form (1), and one such problem is given by

$$K(s,t) = \sqrt{s^2+t^2}, \qquad 0 \leq 2, t \leq 1,\ g(s) = ((1+s^2)^{\frac{3}{2}} - s^3)/3$$

for which the approximate λ_j and ϕ_j are shown in Table 1.

Excellent results were obtained for $N = 13$, and m up to 8. The approximate and true solutions f differed by less than .0004 on $[0,1]$ for $n = 6$ and the errors are shown in Table 2.

However, for $m \geq 9$, substantial errors growing with m appeared in the solution.

In our tests it was observed that good results were normally obtained for equally spaced knots, so that knot choice was not a significant problem. However, for non-exact data in which noise is present, as in problem (3), poor results were obtained.

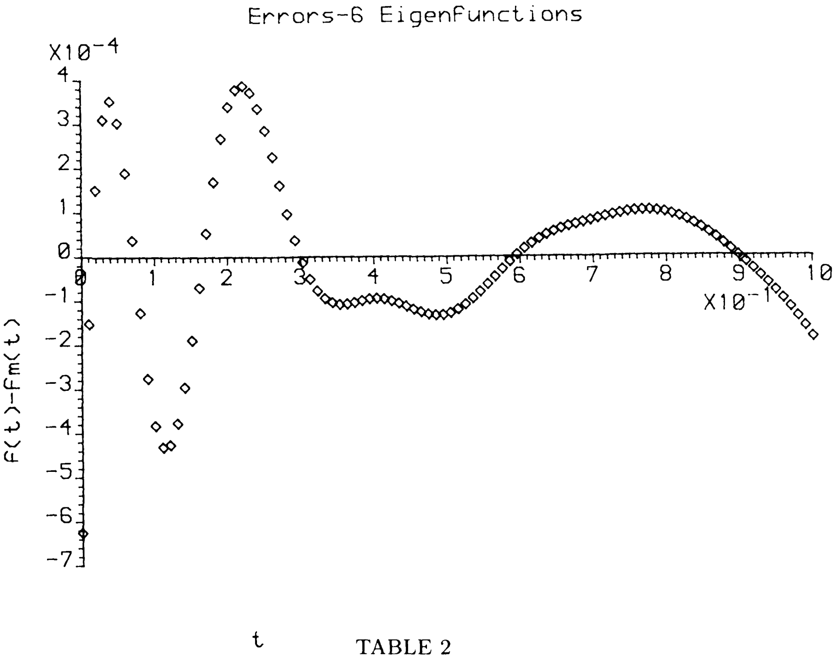


TABLE 2

3. Smoothing/regularisation methods for noisy data. If the problem (1) is solvable but still-posed, then we wish to approximate f by f^*, by minimizing over a specified approximation space, the smoothing measure (see [**2**], [**3**]),

$$I_\lambda[f^*, g] = \frac{1}{n}\sum_{i=1}^{n}(Kf^*(s_i) - g(s_i))^2 + \lambda \int_a^b [f^{*^{(r)}}(t)]^2\,dt, \tag{14}$$

where λ is the smoothing parameter and r is the order of regularization.

In all cases f^* is expressed in the form

$$f^* = \sum a_j\phi_j, \tag{15}$$

where ϕ_j are certain basis functions. In some cases $\{\phi_j\}$ are specified in advance, and define the approximation space. In other cases $\{\phi_j\}$ are deduced by the minimization of (14) over a specified space (e.g., $L_2[a,b]$), typically by using a variational principle.

It is very important to choose the correct value of λ, and so we have adopted the generalized cross validation method of Wahba [**6**] and minimized

$$(16) \qquad V(\lambda) = \frac{1}{n}\|(I - A(\lambda)\underline{g}\|^2 \Big/ \left[\frac{1}{n} \text{ Trace } (I - A(\lambda))\right]^2$$

where $Kf^*(s) = A(\lambda)\underline{g}$, $\underline{g} = [g(s_i)]$.

There are a number of possible choices or deductions of sets of basis functions $\{\phi_j\}$; we discuss the relevant merits of three such sets and give efficient algorithms in each case.

3.1. *Kernel function basis.* First we deduce a basis by minimizing over a specified approximation space, based on a general discussion of Wahba [**6**], and Kimeldorf and Wahba [**7**].

Suppose $f \in L_2[0,1]$ in (1), then by the Riesz representation theorem there exist functions η_s in $L_2[0,1]$ such that

$$Kf(s) = \langle \eta_s, f\rangle = \int_0^1 \eta_s(t)f(t)dt, \qquad s = s_1, \dots, s_n.$$

Thus $\eta_s(t) = K(s,t)$, and it follows that if we minimize (14) over all f^* in $L_2[0,1]$, then f^* takes the form

$$(17) \qquad f^*(t) = \sum_{i=1}^{n} a_i \eta_{s_i}(t) = \sum_{i=1}^{n} a_i K(s_i, t).$$

We have thus "deduced" that the appropriate basis is $\phi_i = K(s_i, t)$. Substituting (15) into (14) and minimizing over a_i, $(i = 1, 2, \dots, n)$ we obtain a system of equations for $\underline{a} = \{a_i\}$,

$$(18) \qquad (Q + n\lambda I)\underline{a} = \underline{g},$$

where $Q_{ij} = \int_0^1 K(s_i,t)K(s_j,t)dt$.

Hence we may deduce that $A(\lambda)$ (in (16)) is given by

$$A(\lambda) = Q(Q + n\lambda I)^{-1},$$

and

$$I - A(\lambda) = I - VD(D + n\lambda I)^{-1}V^T$$

(by eigenvalue decomposition where $Q = VDV^T$).

It follows that, if d_i are the diagonal entries in D, then

$$(19) \qquad V(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{n\lambda z_i}{d_i + n\lambda}\right]^2 \Big/ \left[\lambda \sum_{i=1}^{n}\left[\frac{1}{d_i + n\lambda}\right]^2\right],$$

where $\underline{z} = V^T\underline{g}$.

Clearly the minimization of $V(\lambda)$ can be carried out very simply via this explicit algebraic formula. Having determined the optimal value $\hat{\lambda}$, say, of λ, then a_i are determined by solving (18) for this $\hat{\lambda}$.

3.2. *B-Spline basis.* Following Wahba [**6**], and O'Sullivan and Wahba [**8**], we now exploit the computational efficiency of B-splines. We define k interior knots $x_1, \ldots, x_k$ and place $x_{-3}, \ldots, x_0$ at a and $x_{k+1}, \ldots, x_{k+4}$ at b. Then, using normalized cubic B-splines N_{4j}, as our specification for ϕ_j, i.e.,

$$(20) \qquad f^*(t) = \sum_{j=1}^{k+4} a_j N_{4j}(t),$$

and, on substituting into (14) and minimizing over the space of all approximations of the form (20), we obtain a system of equations for $\underline{a} = \{a_j\}$,

$$(21) \qquad (E^T E + n\lambda B)\underline{a} = E^T\underline{g},$$

where $\underline{a}[a_1, \ldots, a_{k+4}]^T$, $\underline{g} = (g(s_1) \ldots g(s_n))^T$

$$B_{ij} = \int_a^b N_{4i}(t)N_{4j}(t)dt, \; E_{ij} = \int_a^b K(s_i, t)N_{4j}(t)dt.$$

Note that E is a non-square (n by $k + 4$) matrix in this case, and B is not the identity matrix so that (21) is potentially more difficult to solve than was (18) above.

Now $Kf^*(s) = A(\lambda)\underline{g}$, where

$$A(\lambda) = W(W^TW + n\lambda I)^{-1}W^T \quad \text{with } W = EB^{-1/2}.$$

Thus

$$I - A(\lambda) = I - WV(D + n\lambda I)^{-1}V^TW^T,$$

where $W^TW = VDV^T$ (by eigenvector decomposition).

The matrix $B^{1/2}$ may be determined from

$$B = V_1DV_1^T \text{ (by eigenvector decomposition)}$$

in the form

$$B^{1/2} = V_1D^{1/2}V_1^T.$$

We now deduce that

$$(22) \quad V(\lambda) = \frac{n^{-1}\left[\sum_{i=1}^{n} g^2(s_i) - \sum_{i=1}^{k+4} z_i^2(d_i + 2n\lambda)(d_i + n\lambda)^{-2}\right]}{\left[\lambda\sum_{i=1}^{k+4}(d_i + n\lambda)^{-1} + \frac{n-(k+4)}{n}\right]^2},$$

where $\underline{z} = V^TW^T\underline{g}$.

The minimization over λ of $V(\lambda)$ can, in form (22), now be readily computed and, for the optimal $\lambda = \hat{\lambda}$, (21) may be solved for $\underline{a}$.

3.3. *Eigenfunction basis.* Suppose that we now adopt the new (in cross-validation approximation,

$$(23) \qquad f^*(t) = \sum_{j=1}^{m} a_j\phi_j(t),$$

where ϕ_j are approximate eigenfunctions, determined in B-spline form (as in §2)

$$(24) \qquad \phi_j(t) = \sum_{p=1}^{k+4} c_{pj}N_{4p}(t).$$

Substituting (23) and (24) into (14), and minimizing over the space of approximations of the form (23), we obtain a system of equations for $\underline{a} = \{a_j\}$,

$$(25) \qquad (Q^T Q + n\lambda I)\underline{a} = Q^T \underline{g},$$

where Q is the $m \times n$ (non-square) matrix $Q_{ij} = [\lambda_j \phi_j(s_i)]$ and the $n \times n$ identity matrix I results from the orthonormality of $\{\phi_j\}$.

We observe that the problem (25) is noticeably simpler than that (21) obtained for a B-spline basis.

It may be deduced that

$$\begin{aligned} A(\lambda) &= Q(Q^T Q_n \lambda I)^{-1} Q^T \\ &= QV(D + n\lambda I)^{-1} V^T Q^T, \end{aligned}$$

where $Q^T Q = VDV^T$ (by eigenvector decomposition). Hence

$$I - A(\lambda) = I - QV(D + n\lambda I)^{-1} V^T Q^T,$$

which leads to the formula

$$(26) \quad V(\lambda) = \frac{n^{-1}\left[\sum_{i=1}^{n} g^2(s_i) - \sum_{i=1}^{m} (z_i)^2 (d_i + 2n\lambda)(d_i + n\lambda)^{-2}\right]}{\left[\lambda \sum_{i=1}^{m} (d_i + n\lambda)^{-1} + \frac{(n-m)}{n}\right]^2},$$

where $\underline{z} = V^T Q^T \underline{g}$.

The minimization over λ of $V(\lambda)$ in this explicit form is readily carried out, and we note that precisely the same computer code may be used here as for splines (22) but with Q and n replacing W and $k + 4$, respectively. For the optimal $\lambda = \hat{\lambda}$, (25) is then solved for $\underline{a}$.

Numerical results and features for the three distinct Cross-validation methods of 3.1, 3.2, 3.3 will be discussed below.

4. Numerical results for cross-validation methods. We now test the three methods of §3. As might be expected from the similarity of their analysis, we have observed in practice that somewhat comparable results are obtained by using either B-splines or eigenfunctions as a basis. Although the eigenfunctions need to be determined initially, the cross validation algorithm based on them is then a slightly faster one than that for B-splines. The situation with regard to the use of kernel functions is somewhat different. The actual algorithm, based on (17) is certainly simplest of the three cross-validation algorithms, but the use of $K(s_i, t)$ as a basis function can be unfortunate in some cases.

We consider the three algorithms in turn, and first show ill-conditioning is averted by the inclusion of $\lambda \neq 0$ in the first algorithm.

4.1. *Kernel function bases.* Unless otherwise stated, we take the order r of regularization in (15) to be zero. We have found this choice of r satisfactory in the problems we have tested. However, we note that higher values might well be necessary to ensure smoothness in other problems.

Consider the example

$$\int_0^\infty e^{-st} f(t)dt = g(s)$$

and transform this, under $t = x(1-x)^{-1}$ into a finite interval problem

$$\int_0^1 (1-x)^{-2} e^{-sx(1-x)^{-1}} f[x(1-x)^{-1}]dx = g(s).$$

Defining $K(s,x) = (1-x)^{-1} e^{-sx}(1-x)^{-1}$ and $h(x) = (1-x)^{-1}$ $f[x(1-x)^{-1}]$, we obtain

$$\int_0^1 K(s,x)h(x)dx = g(s), \tag{27}$$

where $h(x)$ is in $L_2[0,1]$. The matrix Q (of kernel inner products) has components

$$Q_{ij} = \int_0^1 K(s_i,t)K(s_j,t)dt = \left[\frac{-e^{-(s_i+s_i)t(1-t)^{-1}}}{(s_i+s_j)}\right]_0^1; \tag{28}$$

thus $Q_{ij} = (s_i+s_j)^{-1}$. It follows that, for s_i equally spaced ($s_i \propto i$), [Q] is proportional to the *Hilbert matrix*. Clearly, then, the solution of the linear system (18) for $\lambda = 0$ is extremely ill-conditioned, and, without a smoothing factor the accurate determination of $\underline{a}$ is difficult. However, we have found that the inclusion of even a very small λ regularizes the numerical problem into a well-conditioned one. This is equivalent to the interesting statement that only a very small multiple of an identity matrix needs to be added to the Hilbert matrix to greatly reduce its condition number.

The Hilbert matrix traditionally occurs in least squares polynomial approximation on a continuum, using a power basis, and the present example is an analogue in the context of approximation by kernel basis.

The above clear advantage of the method is, however, immediately balanced by the observation that, in this case,

$$K(s_j, t) = e^{-s_j t}$$

and hence the approximation (17) of $f^*(t)$ to $f(t)$ is a *sum of negative exponentials.* Clearly this is a poor approximation basis unless it so happens that $f(t)$ is itself a rapidly decaying function.

In Table 3, we show the numerical results obtained for

$$(29) \qquad g(s) = 1/(s+1) + e(s), \qquad 0 \leq s \leq 1,$$

where $e(s) = N(0, 0.01)$.

The approximation to the true solution $f(t) = e^{-t}$ is good, and the method has simultaneously achieved both regularization of an ill-conditioned problem and smoothing of the data noise. Noticeably less accurate results were obtained as the noise level was raised.

We also tested problems where the solution $f(t)$ was not exponentially decaying and extremely poor results were sometimes obtained. This confirms that this basis has very limited areas of application.

4.2. *Cubic B-spline basis.* Consider

$$(30) \qquad \int_0^1 (s^2 + t^2)^{1/2} f(t) dt = [(1 + s^2)^{3/2} - s^3]/3,$$

for which the true solution is $f(t) = t$.

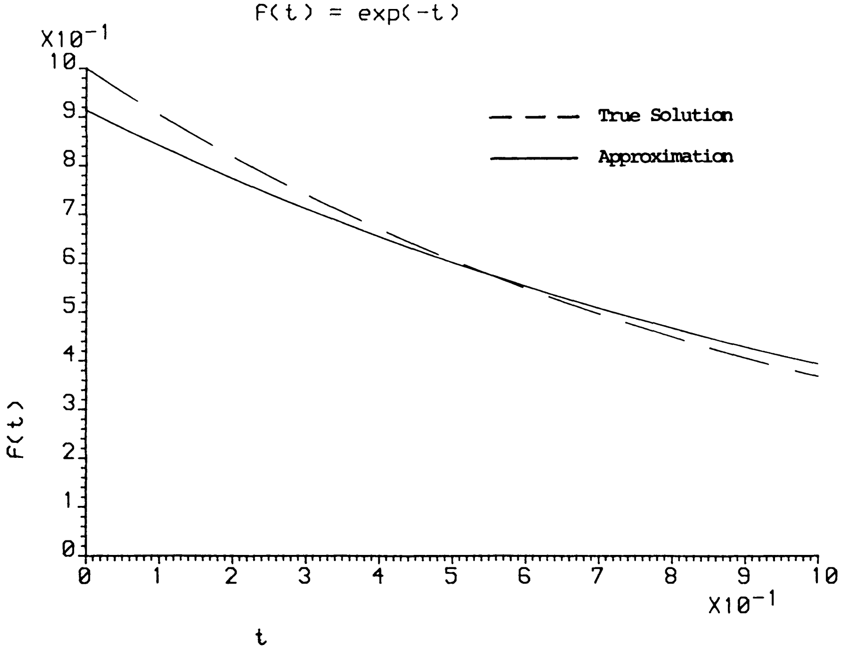


TABLE 3

In Table 4, we show numerical results obtained for the same data noise level as (29), and see that good results were obtained.

We have tested many problems using *b*-splines and have obtained consistently satisfactory results.

4.3. *Eigenfunction basis.* We have considered the problem

$$\int_0^1 K(s,t)f(t)dt = g(s),$$

where

$$K(s,t) = \begin{cases} (1-s)t, & 0 \le t \le s \le 1, \\ (1-t)s, & 0 \le s \le t \le 1, \end{cases}$$

and

$$g(s) = \frac{s}{12}(1 - 2s^2 + s^3),$$

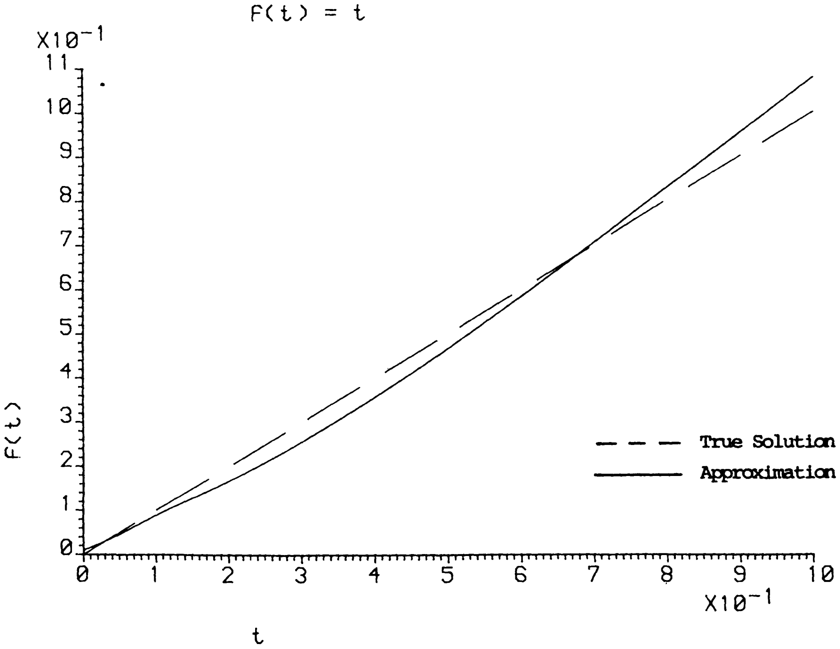


TABLE 4

i	approximation	true	abs(rel. error)
1	0.10132118D+00	0.10132118D+00	0.47806143D-08
2	0.25330305D-01	0.25330296D-01	0.37407818D-06
3	0.11257905D-01	0.11257909D-01	0.36911414D-06
4	0.63329359D-02	0.63325740D-02	0.57152623D-04
5	0.40523891D-02	0.40528473D-02	0.11306718D-03
6	0.28129488D-02	0.28144773D-02	0.54308482D-03
7	0.20604720D-02	0.20677793D-02	0.35338538D-02
8	0.15661765D-02	0.15831435D-02	0.10717267D-01
9	0.12234402D-02	0.12508788D-02	0.21935502D-01
10	0.77135429D-03	0.10132118D-02	0.23870383D+00
11	0.81764634D-03	0.83736515D-03	0.23548648D-01
12	0.42059704D-05	0.70361933D-03	0.99402238D+00
13	0.40759671D-05	0.59953363D-03	0.99320144D+00

Eigenvalues of the Kernel K(s,t)

TABLE 5.

which has true solution $f(t) = t(1-t)$. The kernel $K(s,t)$ has

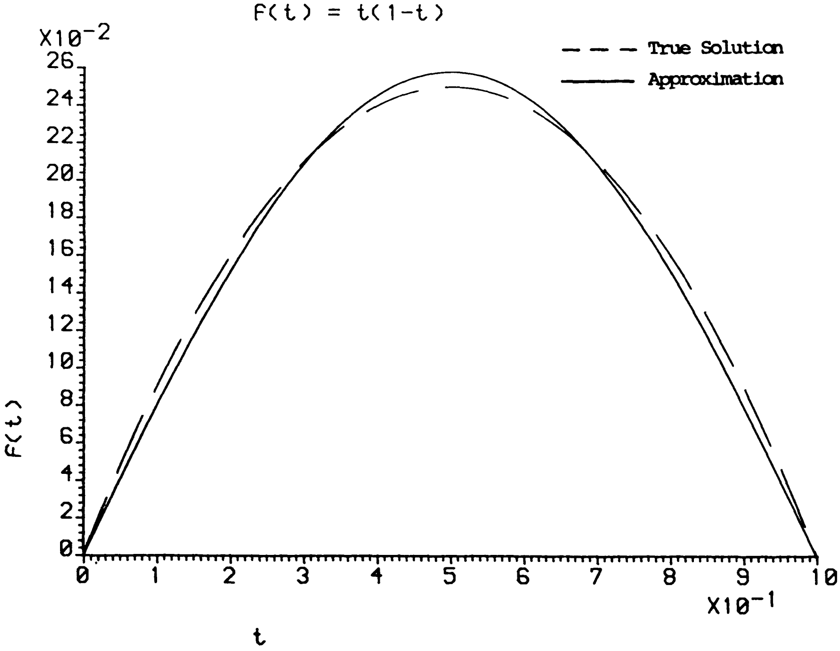


TABLE 6.

eigenvalues, $\lambda_i = (i\pi)^{-2}$, and corresponding eigenfunctions

$$\phi_i(s) = \sqrt{2}\,\sin(i\pi s), \qquad i = 1, 2, \dots\ .$$

In Table 5 we show the approximations to the first 11 eigenvalues, and Table 6 shows the approximation to $f(t)$ obtained using the corresponding eigenfunctions as the basis, with $g(s)$ measured exactly.

As with the previous methods the approximation was found to deteriorate with increasing levels of noise in the data, and the value of λ which minimizes $V(\lambda)$ was correspondingly difficult to determine accurately. This problem may be overcome by considering a higher degree of smoothing.

REFERENCES

1. C.W. Groetsch, *The theory of Tikhonov Regularisation for Fredholm Equations of the first kind,* Research Notes in Mathematics Vol 105, Pitman, Boston, 1984.

2. A.N. Tikhonov, *Solution of Incorrectly Formulated Problems and the Regularization Method,* Soviet Math. Dokl. **4** (1963) 1035-1038; *Regularization of Incorrectly Posed Problems,* Soviet Math. Dokl. **4** (1963), 1624-1627.

3. G. Ribeére, *Regularisation d'opératuers,* Rev. Francaise Informat. Recherche Opérationnelle **1** (1967) No. 5, 57-79.

4. C.T.H. Baker, L. Fox, D.F. Mayers, and K.W. Wright, *Numerical solution of Fredholm integral equations of the first kind.* Comput. J. **7** (1964), 141-148.

5. B.A. Lewis, *On the numerical solution of fredholm integral equations of the first kind,* J. Inst. Math. Appl. **16** (1975), 207-220.

6. G. Wahba, *Practical approximate solutions to linear operator equations when the data are noisy,* SIAM J. Numer. Anal. **14** (1977), 651-667.

7. D. Kimeldorf and G. Wahba, *Some results on Tchebycheffian spline functions,* J. Math. Anal. Applic. **33** (1971), 82-95.

8. F. O'Sullivan and G. Wahba, *A cross validated bayesian retrieval algorithm for non-linear remote sensing experiments,* J. Comput. Physics **59** (1985), 441-445.

Computational Mathematics Group, Royal Military College of Science,
Shrivenham, Swindon, Wiltshire, 5N6 8LA, England