SIAM J. SCI. STAT. COMPUT.
Vol. 4, No. 2, June 1983

0196-5204/83/0402-0013 $01.25/0

MULTIDIMENSIONAL ADDITIVE SPLINE APPROXIMATION*

JEROME H. FRIEDMAN,† ERIC GROSSE‡ AND WERNER STUETZLE§

Abstract. We describe an adaptive procedure that approximates a function of many variables by a sum of (univariate) spline functions s_m of selected linear combinations $a_m \cdot x$ of the coordinates

$$\phi(x) = \sum_{1 \leqq m \leqq M} s_m(a_m \cdot x).$$

The procedure is nonlinear in that not only the spline coefficients but also the linear combinations are optimized for the particular problem. The sample need not lie on a regular grid, and the approximation is affine invariant, smooth, and lends itself to graphical interpretation. Function values, derivatives, and integrals are inexpensive to evaluate.

Key words. multivariate approximation, surface fitting, projection pursuit

1. Introduction. Multidimensional surface approximation is recognized as an important problem for which several methodologies have been developed. The aim is to construct an approximation $\phi(x)$ to a p-dimensional surface $y = f(x)$ on the basis of (possibly noisy) observations $\{(y_i, x_i)\}_{1 \leqq i \leqq n}$. Most existing methods, such as tensor product splines, kernels, and thin plate splines (for a survey, see Schumaker [1976]), are linear in that

$$\phi(x) = \sum_{1 \leqq i \leqq n} w_i y_i,$$

where the weights $\{w_i\}$ depend only on x and $\{x_i\}_{1 \leqq i \leqq n}$, but not on $\{y_i\}_{1 \leqq i \leqq n}$. These methods have the advantage that they are straightforward to compute and their theory is tractable. In practice, however, they are limited because they cannot take advantage of special properties of the surface. Due to the inherent sparsity of high-dimensional sampling, procedures successful in high dimensions must be adaptive and thus nonlinear.

In this paper we describe an adaptive procedure that approximates $f(x)$ by a sum of (univariate) spline functions s_m of selected linear combinations $a_m \cdot x$ of the coordinates

$$\phi(x) = \sum_{1 \leqq m \leqq M} s_m(a_m \cdot x). \tag{1}$$

The procedure is nonlinear in that not only the spline coefficients but also the linear combinations are optimized for the particular problem.

2. The algorithm. The spline function s_m along $a_m \cdot x$ is represented as a sum of j_m B-splines (de Boor [1978]) of order q

$$s_m(a_m \cdot x) = \sum_{1 \leqq j \leqq j_m} \beta_{mj} B_{mj}(a_m \cdot x). \tag{2}$$

* Received by the editors May 8, 1980, and in revised form January 15, 1982. This work was supported by the Department of Energy under contracts DE-AC03-76SF00515 and DE-AT03-81-ER108 43, the Office of Naval Research under contract ONR N00014-81-K-0340, and the National Science Foundation under grant MCS 78-17697.

† Stanford Linear Accelerator Center, Stanford, California 94305.

‡ Computer Science Department, Stanford University, Stanford, California 94305. Present address: Bell Laboratories, Murray Hill, New Jersey 07974.

§ Stanford Linear Accelerator Center and Department of Statistics, Stanford University, Stanford, California 94305.

The approximation $\phi(x)$ (given by (1) and (2)) is specified by the directions $\{a_m\}_{1\leqq m\leqq M}$, the knot sequences along $a_m \cdot x$ for $1\leqq m\leqq M$, and the B-spline coefficients $\{\beta_{mj}\}_{1\leqq m\leqq M, 1\leqq j\leqq j_m}$. For particular $\{a_m\}$, the knots are placed heuristically and then the $\{\beta_{mj}\}$ are determined by (linear) least squares. The residual sum of squares from this fit is taken to be the inverse figure of merit for $\{a_m\}_{1\leqq m\leqq M}$.

Following Friedman and Stuetzle [1981], the approximation is constructed in a stepwise manner: given $\{a_m\}_{1\leqq m\leqq M-1}$, find a_M to optimize the figure of merit of $\{a_m\}_{1\leqq m\leqq M}$. Terminate when the figure of merit is below a user specified threshold.

3. Implementation. A difficult part of the algorithm is finding each direction a_m. We perform a numerical search using a Rosenbrock method (Rosenbrock [1966]). This method is easily modifiable to search over the unit sphere. We have found empirically that each iteration of the optimizer requires approximately $3.5p$ function evaluations, where p is the dimension of x. Two iterations are nearly always sufficient. As the search usually starts far from the solution and the solution does not have to be obtained with high precision, it does not seem likely that optimization procedures that estimate the Hessian would do better.

For high dimensionality, the computation is dominated by the evaluations of the object function. Since it is not crucial to find the precise optimum, considerable savings are achieved by substituting a similar, but much less expensive figure of merit during the search for a new direction. For this figure of merit not only the previously found directions but also the corresponding spline coefficients are held fixed. For a given direction, the residuals are modelled by (basically) a moving average smooth (see Friedman and Stuetzle [1981]). The characteristic bandwidth (the fraction of observations over which averaging takes place) is taken to be inversely proportional to the number of knots. The residual sum of squares from the smooth is the figure of merit used for the smooth. Solving the least squares problem for the original figure of merit requires

$$O\left[n\left(\sum_{1\leqq m\leqq M} j_m\right)^2\right]$$

operations, while the new figure of merit can be evaluated in roughly n operations using updating formulas for the moving average. The least squares problem has to be solved only once for each iteration to determine the new model after a_m has been found.

To solve the least squares problem, we form the normal equations and use a pseudo-inverse, since the design matrix might not have full rank. The singularity which arises from the inclusion of a constant term for each direction is remedied by simply dropping one column per direction from the design matrix. Higher order singularities caused, for example, by the linear terms for three co-planar directions, are not explicitly taken care of, but are handled by the pseudo-inverse.

Our knot placement procedure is motivated by the sequential nature of the algorithm. At each iteration, the knot positions are required for the least squares fit, after the new direction has been found. Our model at this point is the spline fit of the previous iteration, plus the moving average smooth along the newly found direction. The knot placement is based on the residuals $\{r_i\}$ from this model. Multidimensional structure in these residuals due to incompleteness of the model manifests itself as high local variability in the scatterplots of r_i against $a_m \cdot x_i$. In order to preserve the ability of fitting this structure in further iterations, it is important to avoid accounting for it by spurious fits along existing directions. For this reason we place fewer knots in regions of higher local variability. Since the residuals change, the knots are replaced along all directions at each iteration.

The knots along a direction a_m are placed as follows: the smooth described above is applied to $\{(r_i, a_m \cdot x_i)\}_{1 \leqq i \leqq n}$ and the local variability v_i at each point is taken to be the average squared residual from its local linear fit. The Winsorized local variabilities are defined by

$$w_i = \begin{cases} 2\bar{v} & \text{if } v_i > 2\bar{v}, \\ \frac{1}{2}\bar{v} & \text{if } v_i < \frac{1}{2}\bar{v}, \\ v_i & \text{otherwise} \end{cases}$$

(where $\bar{v} = (1/n) \sum_{1 \leqq i \leqq n} v_i$), and then are scaled so that $\sum_{1 \leqq i \leqq n} 1/w_i = 1$. The knots $\{t_i\}$ are placed to divide the line into intervals with equal content of $1/w_i$:

$$\text{for each } l, \quad \frac{1}{j_m - q + 1} = \sum_{a_m \cdot x_l \in [t_l, t_{l+1}]} \frac{1}{w_i}.$$

4. Procedure parameters. The operation of the procedure is controlled primarily by two parameters; these are the number of knots taken along each direction and the termination threshold. Both parameters can be adjusted using graphical output produced by the program. The adequacy of the number of knots and their placement can be judged by examination of the residuals from the final model plotted against each $a_m \cdot x$. A systematic pattern in any one of these plots indicates that either the number of knots is too small or that the knot placement algorithm did not perform well. Another indication that the number of knots might be insufficient is that the procedure chooses nearly the same direction twice, thereby effectively doubling the number of knots placed along that direction.

The value set for the termination threshold determines the number of terms making up the model. Various criteria can be used to decide whether a particular term should be included. In the case of noisy data, one can ask whether a term is significantly different from zero (given all previous terms), or whether the addition of the term reduces the predictive mean squared error of the model. Also, considerations outside the data having to do with the problem setting can influence such a decision. In order to judge statistical significance, it is necessary to know, by how much one would expect an additional term to increase the figure of merit if there were no structure in the residuals. This can be estimated with a permutation test. The residuals (from the previous terms) are randomly permuted among the observations, thereby guaranteeing no structure in the (permuted) data. MASA is applied to these residuals and the increase in figure of merit noted. This process can be repeated, obtaining a (null) distribution of the figure of merit. Either formal or informal hypothesis testing techniques can then be used to judge whether the nonpermuted figure of merit is significant.

The optimal number of terms with respect to prediction error can be estimated by cross validation. The observations are randomly divided into L (typically 5–10) subsamples. Each of the subsamples are in turn set aside and the model constructed from the remaining observations. Each observation is set aside exactly once. The mean squared prediction error averaged over the set aside observations is taken as an estimate of the model mean squared error. Such an estimate can be made for models with differing numbers of terms and that model minimizing the cross validated mean-squared error estimate is then selected. Both permutation tests and cross validation can be implemented in a small driver routine which calls MASA repeatedly.

5. Examples. In this section we present and discuss the results of applying the multidimensional spline approximation method (MASA) to four examples. (A

FORTRAN program implementing MASA is available from the authors.) The first three examples were suggested elsewhere for testing surface approximation procedures. The function in the fourth example was studied in connection with a problem in mathematical genetics.

The first example is taken from Friedman [1979]. In this example uniformly distributed random points $\{x_i | 1 \leqq i \leqq 200\}$ were generated in the six-dimensional hypercube $[0, 1]^6$. Associated with each point x_i was a surface value

$$y_i = 10 \sin(\pi x_i(1) x_i(2)) + 20[x_i(3) - 0.5]^2 + 10x_i(4) + 5x_i(5) + 0x_i(6) + \varepsilon_i,$$

where the $\{\varepsilon_i\}$ were independent identically distributed standard normal. The inverse figures of merit for the approximation with $M = 1, \cdots, 4$ terms were 6.71, 4.29, 1.87, 0.97. In three restarts, the figure of merit did not decrease below 0.86, so $M = 4$ was chosen. The four linear combinations and the corresponding spline functions are shown in Figs. 1a–1d. (The function value is plotted on the vertical axis, $a \cdot x$ on the horizontal axis. The "+" signs on the bottom of the graph indicate the knot positions. A "+" sign followed by a number indicates multiple knots. For completeness, the program parameters are also listed; see comments in the program source code for a detailed explanation.) The spline along the first linear combination (Fig. 1a) is seen to model the linear part of the surface. The second term in the approximation (Fig. 1b) models the additive quadratic dependence on $x(3)$. The final two terms (Figs. 1c, 1d) model the interaction between $x(1)$ and $x(2)$. The L_2 norm of the error $\|f - \phi\|_2$ was 0.57.

Although the full advantages of MASA compared to other procedures are realized in higher dimensional or noisy settings, we applied it to two bivariate examples used

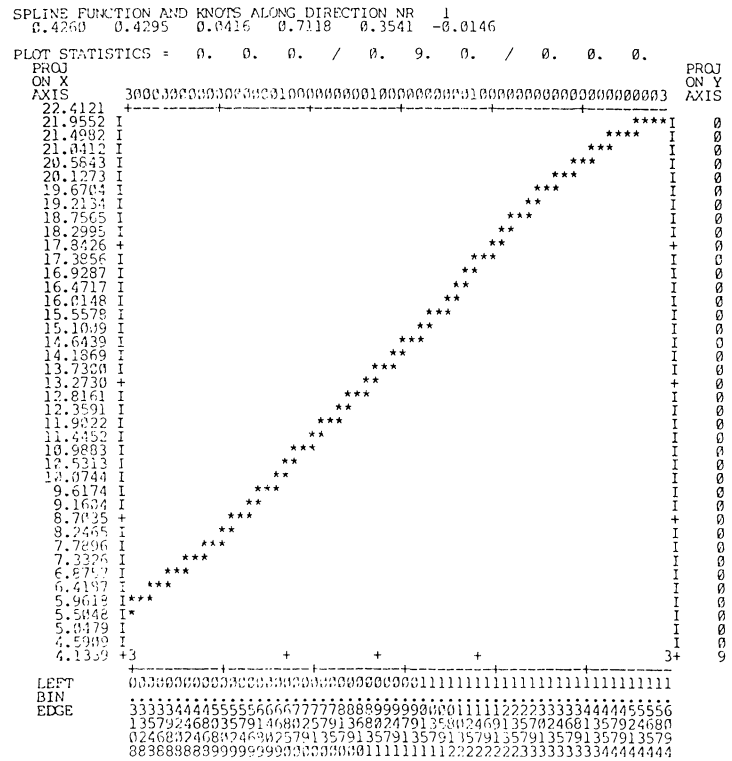

FIG. 1a

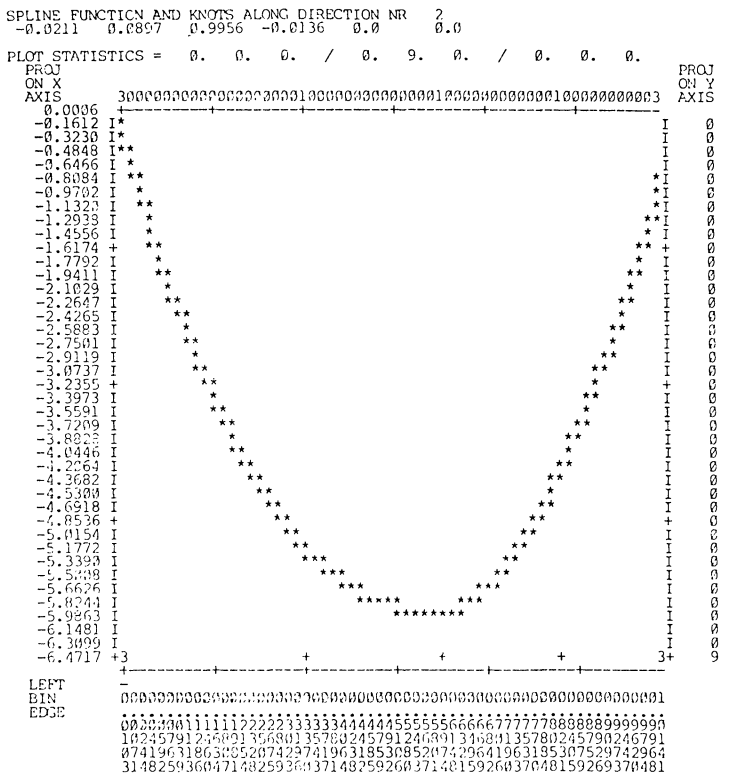
SPLINE FUNCTION AND KNOTS ALONG DIRECTION NR 2
-0.0211 0.0897 0.9956 -0.0136 0.0 0.0
PLOT STATISTICS = 0. 0. 0. / 0. 9. 0. / 0. 0. 0.
PROJ ON X AXIS
PROJ ON Y AXIS
LEFT
BIN
EDGE

FIG. 1b

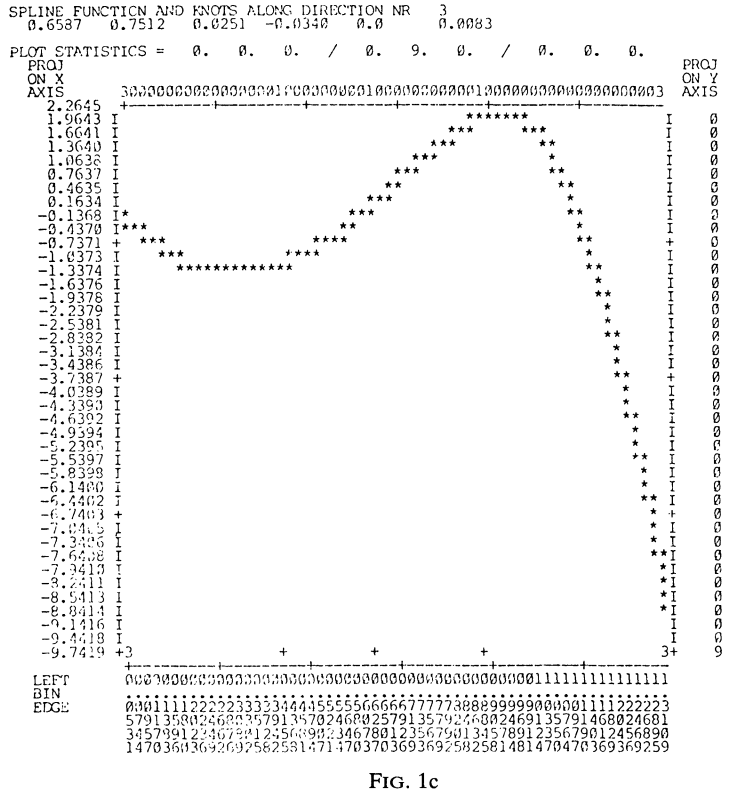
SPLINE FUNCTION AND KNOTS ALONG DIRECTION NR 3
0.6587 0.7512 0.0251 -0.0340 0.0 0.0083
PLOT STATISTICS = 0. 0. 0. / 0. 9. 0. / 0. 0. 0.
PROJ ON X AXIS
PROJ ON Y AXIS
LEFT
BIN
EDGE

FIG. 1c

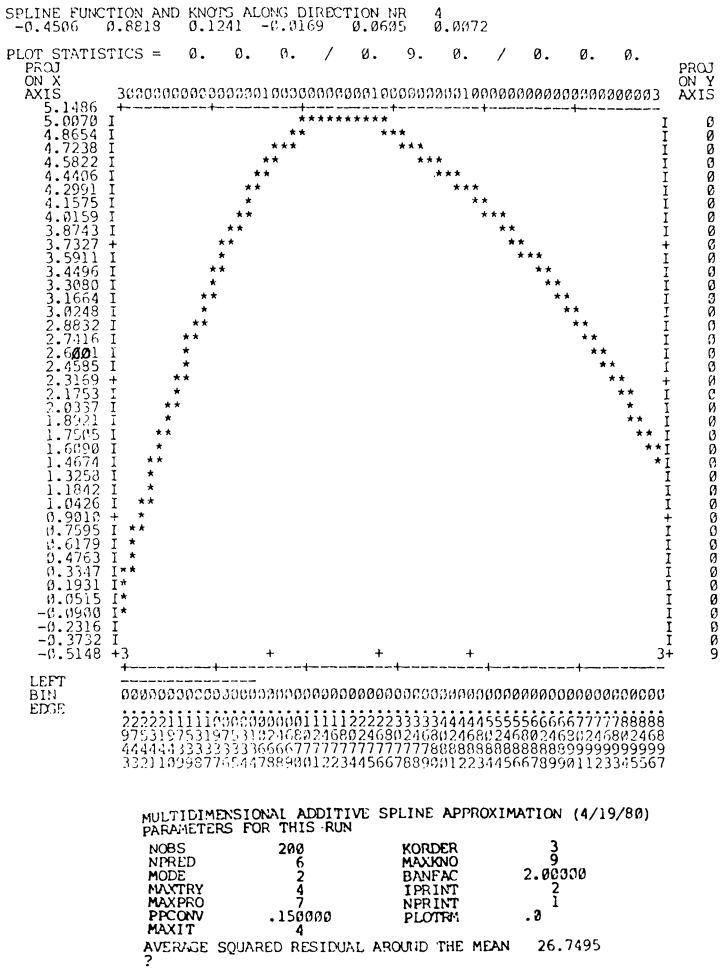

FIG. 1d

by Franke [1979] to compare a number of interpolatory surface approximation schemes. For both examples 100 uniformly distributed random points in the unit square $[0, 1]^2$ were generated. The function in Franke's first example is

$$f(x, y) = 0.75 \exp\left[-\frac{(9x-2)^2+(9y-2)^2}{4}\right] + 0.75 \exp\left[-\frac{(9x+1)^2}{49} - \frac{9y+1}{10}\right]$$

$$+ 0.5 \exp\left[-\frac{(9x-7)^2+(9y-3)^2}{4}\right] + 0.2 \exp\left[-(9x-4)^2-(9y-7)^2\right].$$

Considerations similar to those in the previous example led to an approximation with three terms. The linear combinations and corresponding spline functions are shown in Figs. 2a–2c.

The function in Franke's second example is

$$f(x, y) = \tfrac{1}{9}[\tanh(9y - 9x) + 1].$$

For this case the approximation used only one term, shown in Fig. 3.

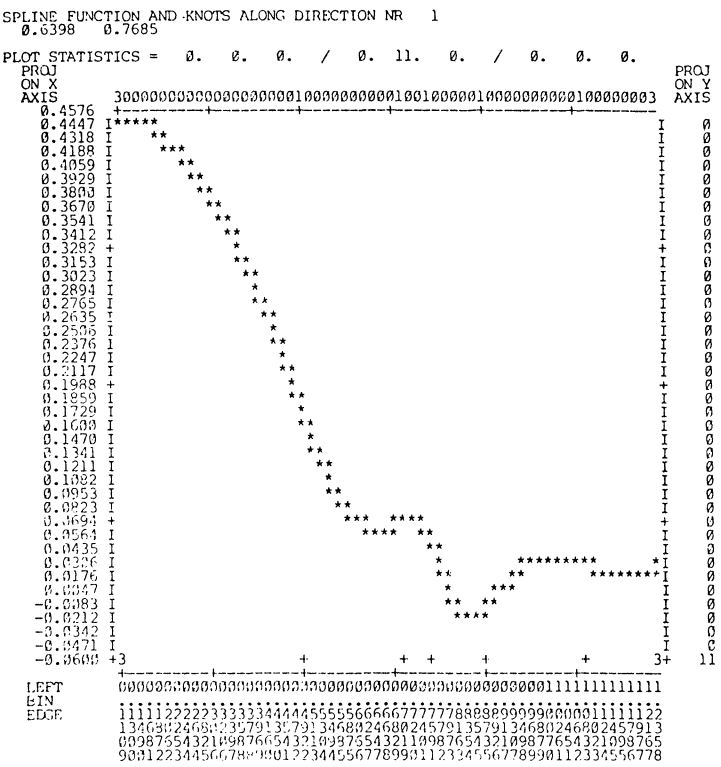
SPLINE FUNCTION AND KNOTS ALONG DIRECTION NR 1
0.6398 0.7685
PLOT STATISTICS = 0. 0. 0. / 0. 11. 0. / 0. 0. 0.
PROJ ON X AXIS
PROJ ON Y AXIS
0.4576
0.4447
0.4318
0.4188
0.4059
0.3929
0.3800
0.3670
0.3541
0.3412
0.3282
0.3153
0.3023
0.2894
0.2765
0.2635
0.2506
0.2376
0.2247
0.2117
0.1988
0.1859
0.1729
0.1600
0.1470
0.1341
0.1211
0.1082
0.0953
0.0823
0.0694
0.0564
0.0435
0.0306
0.0176
0.0047
-0.0083
-0.0212
-0.0342
-0.0471
-0.0600
LEFT BIN EDGE
RIGHT BIN EDGE

FIG. 2a

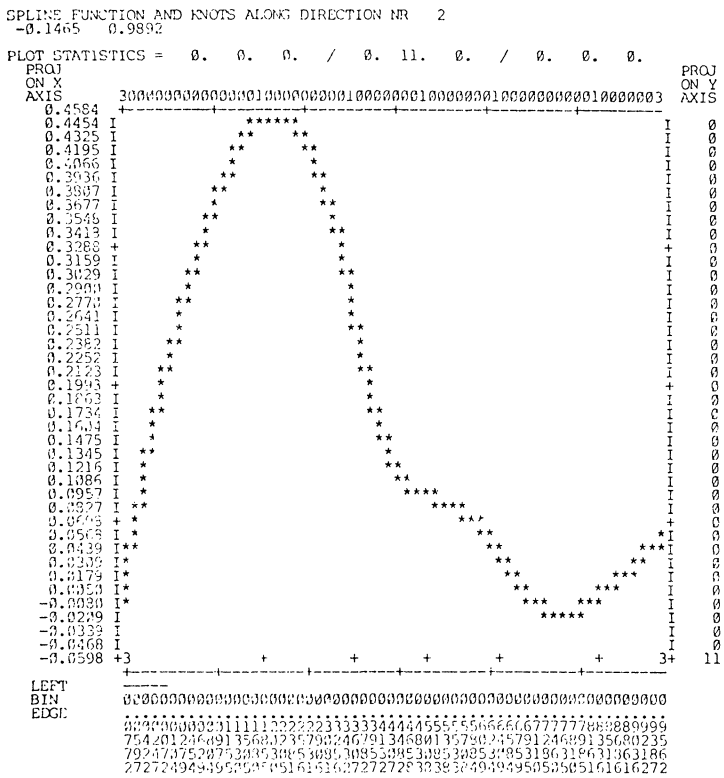
SPLINE FUNCTION AND KNOTS ALONG DIRECTION NR 2
-0.1465 0.9892
PLOT STATISTICS = 0. 0. 0. / 0. 11. 0. / 0. 0. 0.
PROJ ON X AXIS
PROJ ON Y AXIS
0.4584
0.4454
0.4325
0.4195
0.4066
0.3936
0.3807
0.3677
0.3548
0.3418
0.3288
0.3159
0.3029
0.2900
0.2770
0.2641
0.2511
0.2382
0.2252
0.2123
0.1993
0.1863
0.1734
0.1604
0.1475
0.1345
0.1216
0.1086
0.0957
0.0827
0.0698
0.0568
0.0439
0.0309
0.0179
0.0050
-0.0080
-0.0209
-0.0339
-0.0468
-0.0598
LEFT BIN EDGE
RIGHT BIN EDGE

FIG. 2b

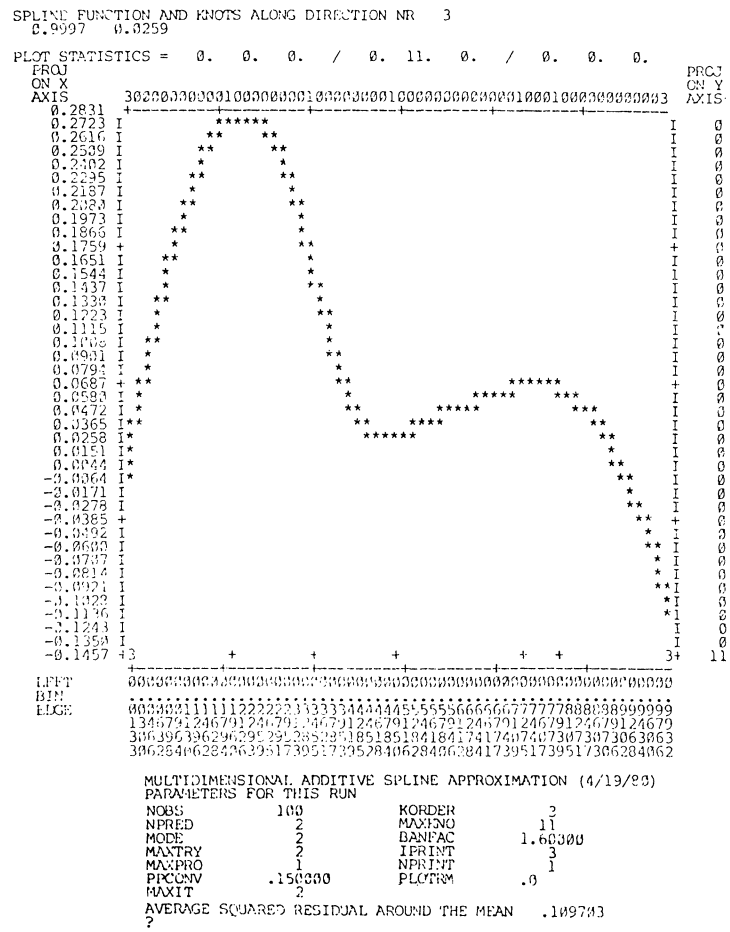

FIG. 2c

Since different random points were used in Franke's and our tests, precise comparisons are not possible. On the first example, MASA gave roughly an order of magnitude larger errors than the best methods in Franke's trials (global basis function methods) while on the second example, MASA gave an order of magnitude smaller errors than the best methods. These results are not surprising since the peak-shaped basis functions of the global basis methods are especially suited for representing the peaks of the first example, whereas the ridge-shaped basis functions of MASA are especially suited to the second example. Unfortunately, peak-shaped basis functions are not appropriate for moderate or higher dimensionality. The difficulty is that in order to achieve a smooth fit, the width of the basis peaks needs to be comparable to the distance between data points. For n uniformly distributed random points in a p-dimensional hypercube $[0, 1]^p$, the typical nearest neighbor distance is $(1/n)^{1/p}$. In particular for $n = 1000$ and $p = 10$, this distance is 0.5, and for $p = 20$ is 0.7. Thus variation of the surface over distances small compared to such large interpoint distances cannot be well approximated with these global basis functions methods.

Our final example is a 19-dimensional function encountered by Carmelli and Cavalli [1979]. An important question is the structure of this function near its

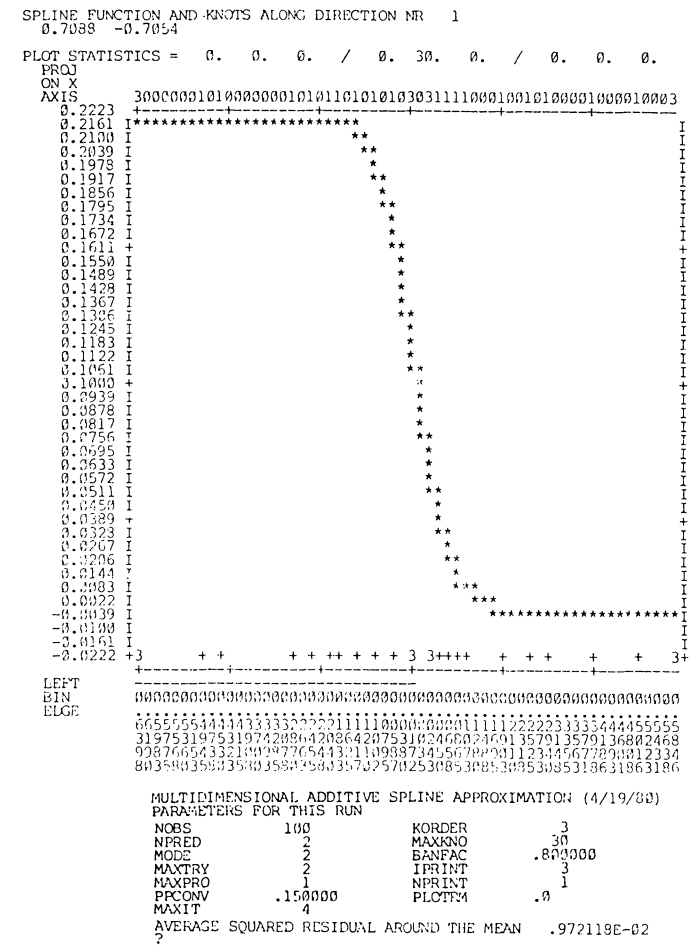

FIG. 3

minimum. We sampled the function at 200 points uniformly distributed in a small hypercube centered at the minimum found by numerical optimization and applied MASA. The inverse figure of merit for the best constant fit was 13.3. The inverse figure of merit for $M = 1$ was 0.78. In 30 restarts, the figure of merit did not decrease below 0.42. Figure 4 gives the linear combination and corresponding spline function. This shows that most of the structure in the likelihood function is revealed in this one projection. The structure certainly would not be easy to find by just looking at the definition of the function, and we know of no other approximation method that would yield this kind of information.

6. Discussion. MASA can be expected to work well to the extent that the surface can be approximated by a function of the form (1). Of course in the limit $M \to \infty$ all smooth surfaces can be represented by (1), but even for moderate M functions of this form constitute a rich class.

As seen in the previous section, an advantage of using essentially one-dimensional basis functions is the possibility of graphical interpretation. The entire model can be represented by graphing $s_m(a_m \cdot x)$ against $a_m \cdot x$ and by specifying $\{a_m\}_{1 \leqq m \leqq M}$ (perhaps

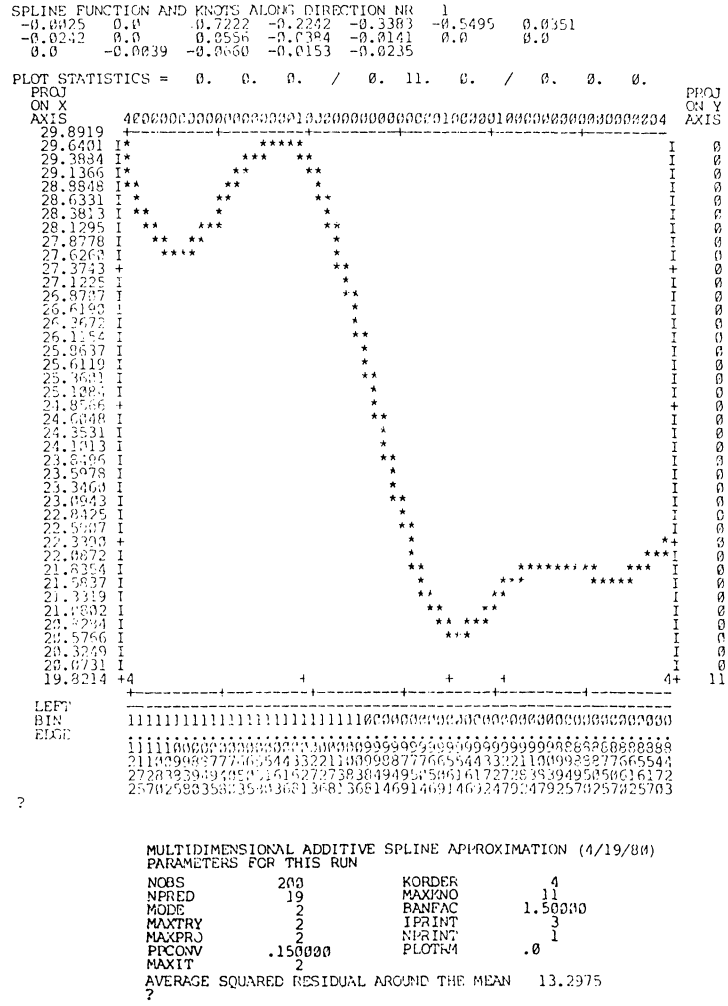

FIG. 4

graphically for $p = 2$ or 3). Additionally the graphical output is very helpful for setting the main procedure parameters, the number of knots along each direction and the termination threshold. Proper termination of the algorithm can be assured by monitoring at each iteration the plot of the residuals from the model of the previous iteration along the newly found direction.

The problem of sparse sampling in high dimensions is not encountered, since MASA is fitting one-dimensional projections of the entire sample. The sample need not lie on a regular grid, and the approximation is affine invariant and smooth. Function values, derivatives, and integrals are inexpensive to evaluate. In addition, since the approximation is locally quadratic for $q = 3$, optimization algorithms can be expected to converge rapidly. As only the directions, the knot positions and the B-spline coefficients have to be stored, MASA produces a very parsimonious description of the surface.

REFERENCES

CARL DE BOOR [1978], *A Practical Guide to Splines*, Springer-Verlag, New York.

DORIT CARMELLI AND L. L. CAVALLI-SFORZA [1979], *The genetic origin of the Jews: a multivariate approach*, Human Biology, 51, pp. 41–61.

WILLIAM S. CLEVELAND [1979], *Robust locally weighted regression and smoothing scatterplots*, J. Amer. Statist. Assoc., 74, pp. 829–836.

RICHARD FRANKE [1979], *A critical comparison of some methods for interpolation of scattered data*, Naval Postgraduate School report NPS-53-79-003.

JEROME H. FRIEDMAN AND WERNER STUETZLE [1981], *Projection pursuit regression*, J. Amer. Statist. Assoc., 76, pp. 817–823.

H. H. ROSENBROCK [1960], *An automatic method for finding the greatest or least value of a function*, Computer J., 3, pp. 175–184.

LARRY L. SCHUMAKER [1976], *Fitting surfaces to scattered data*, in Approximation Theory III, G. G. Lorentz, C. K. Chui and L. L. Schumaker, eds. Academic Press, New York, pp. 203–268.