

## Nonlinear Nonparametric Function Estimation:

$$y = f(x) + \varepsilon$$

any mapping  $f: \mathcal{R}^m \rightarrow \mathcal{R}^n$  where  $m \geq n$ ,

SMART, MARS and other algorithms

### 1 Projection Pursuit Regression:

SMART algorithm (Smooth Multiple Additive Regression Technique) (Friedman e.a),

It is based in the following expansion of  $f$ :

$$f(x) \approx \sum_{i=1}^M \beta_i f_i(\alpha_i' x)$$

where  $\beta_i \in \mathcal{R}^n$ ,  $\alpha_i \in \mathcal{R}^m$ ,  $f_i: \mathcal{R}^n \rightarrow \mathcal{R}^m$ .

Properties: (Diaconis e.a.)

- For many functions --e.g. polynomials-- the equality is also true,
- For others -- e.g.  $f(x_1, x_2) = e^{x_1 x_2}$  the sum needs an infinite number of terms in order to be exact.

The SMART estimator of  $f$  takes the form

$$\hat{f}(x) = \sum_{i=1}^M \hat{\beta}_i \hat{f}_i(\hat{\alpha}_i' x)$$

$\hat{\beta}_i \in \mathcal{R}^n$ ,  $\hat{\alpha}_i \in \mathcal{R}^m$ ,  $\hat{f}_i: \mathcal{R}^n \rightarrow \mathcal{R}^m$  are smooth functions.

Observe a set of data  $\{x_t, y_t, t=1, \dots, n\}$ ,

$\hat{f}(x)$  is obtained by the method of least squares: minimize

$$L_2 = \sum_{t=1}^n (y_t - \hat{f}(x_t))^2$$

over the three groups of parameters:  $\{\hat{\beta}_i\}_1^M$ ,  $\{\hat{\alpha}_i\}_1^M$ ,  $\{\hat{f}_i\}_1^M$

### SMART ALGORITHM:

Proceed sequentially:

For each  $i$  obtain  $\hat{\beta}_i$  that minimizes  $L_2$  given that all the other parameters are known. This is the linear regression estimator.

$\hat{\alpha}_i$  are obtained by a Gauss-Newton algorithm.

$\hat{f}_i$  are the expected values conditioned on the value of  $\hat{\alpha}_i' X_j$ .

The expected value can be calculated using a local average or a spline.

These three steps are repeated until the change in  $L_2$  is less than in two consecutive iterations.

Option: Add model selection

### Multivariate Additive Regression Splines(MARS)

We have seen that the main criticism of the SMART algorithm for fitting models is that is expensive computationally. Another criticism that has appeared in the literature is that it does not work very well if the function we are estimating is not very smooth.

In this section we will study a new statistical methodology for nonparametric function estimation known as Multivariate Additive Regression Splines(MARS) which was introduced by Friedman. As we will see here MARS does overcome the problems of the SMART algorithm.

The ideas of MARS come from the one dimensional case, where the method of splines is the most widely used method for function estimation. Consider the one dimensional model

$$y = f(x) + \text{ERROR}.$$

Smoothing Splines: The method of smoothing splines consists of dividing the domain of  $x$  in  $K+1$  regions by the ordered points  $t_1, \dots, t_k$ , and consider the basis functions

$$\{1, x, x^2, \{x-t_1\}_+^3, \dots, \{x-t_k\}_+^3\} = \{B_i(x)\}$$

where the  $a_j$ 's are fitted by least squares

The method of splines consists of dividing the domain of  $x$  in  $(K+1)$  regions by the ordered points  $t_1, \dots, t_k$ , and consider the basis of functions  $\{ \{x^j\}_1^q, \{x-t_j\}_1^K \}$  which will be denote by  $B_j(x)_1^{K+q}$ .

Then the spline estimator of  $f$  is

$$f(x) = \sum_{i=1}^{K+q} a_i B_i(x),$$

where the  $a_i$ 's are fitted by least squares.

The generalization of this method to dimensions greater than one is conceptually trivial, and it works in two dimensions, but for dimensions greater than two is computationally impractical in most situations.

In the remaining of this section we will describe the algorithm for recursive partitioning which is the basis for MARS,

we will describe the MARS algorithm, and we will show examples.

Recursive partitioning is a method of estimating  $f$ , which consists of fitting a step function by sequentially partitioning the domain of  $f$  into regions and assigning a value to each region. It was introduced as part of the methodology of classification and regression trees (CART). The estimator of  $f$  is

$$f(x) = \sum_{i=1}^m a_i B_i(x)$$

where the  $B_i(x)$  are indicator functions over sets that make a partition of the domain of  $f$ . The criteria that will be optimized is

$$L_2(\hat{f}) = \sum_{i=1}^N (y_i - \hat{f}(x_i))^2$$

The parameters  $\{a_i\}$  are estimated by least squares. The algorithm goes as follows

## MARS ALGORITHM

### START RECURSIVE PARTITIONING ALGORITHM

STEP 1: Set  $m=2$  and find the component of  $x$ , namely  $x^*$ , a fixed value  $z_l$  and the basis functions

$$B_1(x) = H(-(x^* - z_l))$$

$$B_2(x) = H+(x^* - z_l)$$

that minimize  $L_2(\hat{f})$ .  $H(x)$  is the heavy side function which takes value one for non-negative argument and zero otherwise.

STEP  $i$ : Set  $m=i+1$  and find the component of  $x$ , namely  $x^*$ ,  $z_i$ , and replace the basis function  $B_j$  by

$H+(x^* - z_i) B_j(x)$  and add the basis function  $B_{i+1}(x) = H-(y^* - z_i)$  that minimize  $L_2(\hat{f})$ .

The MARS algorithm is essentially the same as recursive partitioning but there are a few modifications.

(i) Replace the two-sided basis functions  $H-(x^* - z^*)$  by two-sided truncated power basis functions  $[+(x^* - z^*)]_+^q$  where the  $[x]$  gives  $x$  for non-negative values of  $x$  and zero otherwise.

(ii) Intermediate functions are not replaced.

(iii) Products are restricted to factors with different variables.

If we are fitting a map in  $k$  dimensions we apply the algorithm to each of the coordinates. This is problematic because MARS uses a large amount of memory for each fit. The most important advantages of MARS over SMART are first of all that it will give better results when the true map is not very smooth, and in addition the computing time used by MARS is much less than the one taken by SMART.

## References

D. Asimov: The Grand Tour: A tool for viewing multidimensional data, SIAM J. Sci. and Stat. Comp. **6**, 128-143 (1985).

A. Buja, D. Asimov and C. Hurley: "Methods for Subspace Interpolation in Dynamic Graphics," Bellcore Technical Memorandum TM-ARH-015639

L. Breiman, J.H. Friedman. Estimating Optimal Transformations for Multiple Regression and Correlation. *J. Amer. Statist. Assoc.* 80 (1985) p. 580-619.

J. Cabrera, D. Cook Projection Pursuit Indices based on Fractal Dimension. Proc. of the 24th Symp. on the Interface between Comput. Sci. and Statist., (Springer-Verlag, New York, 1991) .

D. Cook, A. Buja, J. Cabrera. Direction and Motion Control in the Grand Tour. Proc. of the 23rd Symp. on the Interface between Comput. Sci. and Statist., (Springer-Verlag, New York, 1991).

R.D. De Vaux, D.C. Psichogios, L.H. Ungar. A Comparison of Two Non-Parametric Estimation Schemes: MARS and Neural Networks. Technical Report SOR-92-01, Dept. Stat. and O.R. Princeton University(1992).

P. Diaconis, M. Shahshahani. On nonlinear functions of linear combinations. *SIAM J. Sci. Statist. Comput.* 5 (1984) p. 175--191.

J.H. Friedman, W. Stuetzle. Projection Pursuit Regression. *J. Amer. Statist. Assoc.* 76 (1981) p. 817-823.

J.H. Friedman. SMART: User's guide. Technical Report LCS01, Dept. Stat., Stanford University(1984).

J.H. Friedman. Multivariate Adaptive Regression Splines, *The Annals of Statistics*, 19:1 (1991) p. 1-141.