Statistica Sinica **5**(1995), 737-748

AN ADAPTIVE EXPANSION METHOD FOR REGRESSION

Michael LeBlanc

Fred Hutchinson Cancer Research Center

Abstract. This article describes a new non-parametric regression method that extends additive regression techniques to allow modeling of interactions among predictor variables. The proposed models consist of sums of smooth functions of one or more predictor variables. Each term involving more than one predictor is assumed to be a composition of bivariate functions of simpler terms in the model. The method is demonstrated on simulated and real data sets and predictions are compared to those from additive regression models and Friedman's (1991) multivariate adaptive regression spline (MARS) models.

Key words and phrases: Non-parametric regression, adaptive methods, smoothing.

1. Introduction

Consider a regression problem where the response y and p predictor variables $(x_1, \ldots, x_p)$ are related by $y = f(x_1, \ldots, x_p) + \epsilon$, where ϵ is a mean zero noise term. Assume we have data consisting of n independent observations of the response and the predictors $(y_i, x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$. There can be several possible goals for a regression analysis of the data. Interest may focus on identifying which predictors are associated with the response, studying estimates of the regression function f or predicting the response for new observations.

Additive models (eg. Hastie and Tibshirani (1990)) are a useful and increasingly popular class of models used to approximate f. While the class of additive models is considerably more flexible than traditional linear models, additive models may still not capture complicated response structure involving interactions with two or more predictors. Several techniques have been proposed for more general modeling and variable selection. Examples of regression procedures which include variable selection and interaction modeling are the projection pursuit regression technique (Friedman and Stuetzle (1981)), the classification and regression tree algorithm (CART) (Breiman, Friedman Olshen and Stone (1984)) and the multivariate adaptive regression splines technique (MARS) (Friedman (1991)). Other methods that allow for estimation of more general response structure include connectionist networks; see for instance, Hinton (1989).

The proposed prediction method extends additive modeling. The approximation for f is a sum of smooth function terms where a term involving a single predictor variable is a smooth univariate function and a term involving more than one predictor variable is a composition of bivariate functions of more simple terms in the model. Therefore, the new model includes the additive model as a special case.

An important motivation for the method is that familiar univariate and bivariate smoothers can be used as building blocks along with the backfitting algorithm for estimation, even for models with higher order interaction terms. The method also includes forward model selection and backward model deletion steps and a generalized cross-validation (GCV) score for selecting model complexity. The method will likely be most useful for low to moderately high dimensional problems ($p \leq 25$).

2. The Model

Consider an approximation to f

$$f^\star = \sum_1^m f_j(x_1, \ldots, x_p).$$

If each term, f_j, is a function of a single predictor the model is called additive. Other models can be expressed as expansions of low dimensional functions. For instance, more general models can be constructed with bivariate functions, $f_j(x_1, x_2)$, of the predictor variables. We allow a further generalization of the form of f_j, but still control the complexity of the allowable functions. Each of the terms, f_j, is defined as a composition of bivariate functions of simpler terms in the model and other predictors. An example of a model in this class is

$$f^\star = f_1(x_1)+f_2(x_2)+f_3(f_2(x_2), x_4)+f_4(f_1(x_1), x_3)+f_5(f_3(f_2(x_2), x_4), x_5). \quad (2.1)$$

Note, in the first two variable function, $f_3(f_2(x_2), x_4)$, the first argument, $f_2(x_2)$, is a univariate term in the model and in the higher order function, $f_5(f_3(f_2(x_2), x_4), x_5)$, the first argument, $f_3(f_2(x_2), x_4)$, is also a more simple term in the model.

The current implementation of the procedure restricts the arguments of bivariate functions to include one simpler term in the model and one predictor variable not yet in the model. Note, that every complex term enters as a composition; for instance a bivariate interaction must be of the form $g(h(x_1), x_2)$. However, this doesn't make interpretation more difficult, because the function can also be represented as $f(x_1, x_2)$ by using contour plots or profile plots with

the original predictors as the axes. Alternatively, the algorithm could be modified to only use the recursive form for interactions involving three or more variables.

It is important that the model approximation involves component functions of dimensions ($d \leq 2$), since this will allow the use of one and two dimensional smoothers as simple "modules" to estimate the terms in the model. Development of a flexible prediction method was the primary motivation for the form of the new model. However, the low dimensional expansion in terms of univariate and bivariate functions also allows an interesting model interpretation. The interpretation of a composition term is the "effect modification" of a lower order term in the model. For instance, $f_k(f_j, x_l)$ represents an effect modification of lower order term f_j in the model by variable x_l. Graphically, a term in the model involving k predictor variables can be explained by a sequence of $k-1$ bivariate function plots involving simpler terms in the model. Bivariate functions are usually the highest dimensional functions that are still easy to think about, represent and communicate. An example is given in Section 4.1.

In addition to generalizing additive models, the form of the model is also related to the form of the functions fitted in the MARS procedure which uses a sum of products of piecewise linear basis functions to model interactions. In a MARS model, a term can be created by multiplying a current basis function $B_j(x)$ by a piecewise linear function $(x - t_{kj})^+$

$$(x - t_{kj})^+ = \begin{cases} x - t_{kj}, & \text{for } x - t_{kj} \geq 0, \\ 0, & \text{for } x - t_{kj} < 0, \end{cases}$$

or $(t_{kj} - x)^+$, where the predictor variable to be included in the current basis function is not already an argument of $B_j(x)$. Therefore, the MARS model is a special case of the new model where the bivariate interactions are restricted to the form $f_k(f_j, x) = B_j(x) \times (x - t_{kj})^+$ and $f_{k+1}(f_j, x) = B_j(x) \times (t_{kj} - x)^+$.

While using a general bivariate interaction will cost more degrees of freedom than a simple product interaction built up in MARS, the degrees of freedom for the new method is controlled by adaptively selecting the spans for the smoothers. In addition, the added generality can cost fewer degrees of freedom, if the interaction is quite complicated and would require several MARS basis functions (and knot value optimizations) to be accurately approximated. While in most examples we have considered the procedure gives prediction errors close to that of the MARS procedure, the new method yields smaller prediction errors than MARS for an example with complicated interactions (given in Section 4.1) and a real data set with complicated spatial and temporal patterns of disease (described in Section 4.2). In addition, for those examples the new method gives substantially smaller prediction errors than additive models.

As with the additive model, the components of the new model are not unique. In the population case, the following identifiability constraints can be used to help interpret of the model terms. The univariate functions can be restricted as in additive models

$$E\{f_k(x_i)\} = 0$$

if a constant term is included in the model.

The higher order terms f_j can be restricted so that

$$E\{f_k(f_j, x_i)|x_i\} = E\{f_k(f_j, x_i)|f_j\} = 0. \tag{2.2}$$

To impose the constraint, one can first fit an additive model $g(f_j) + h(x_i)$ to f_k. Then, $f_k^\star(f_j, x_i) = \hat{f}_k(f_j, x_i) - \hat{g}(f_j) - \hat{h}(x_i)$ satisfies Equation (2.2). The important aspect of removing the additive components is that the residual function can be described as "interaction". This method of removal was proposed by Hastie and Tibshirani (1990, Section 9.5.3) for hierarchical models. For the data case, an approximation to the above removal method uses smoothers to fit an additive model to the estimated higher order term.

3. Estimation

The model is built using forward and backward stepwise techniques. For a given model, we want to estimate all the univariate and bivariate component functions f_j; for example, functions $f_1, \ldots, f_5$ in model (2.1). The backfitting algorithm is used for estimation, which allows us to fit the expansion described above by repeated use of univariate and bivariate smoothers.

3.1. Forward selection and backward deletion of model terms

Algorithm A – Forward model selection

1. Initialize the model basis set, $H^1 = \{1\}$ and $k = 1$. Let $\hat{y}_i^{(k)}, i = 1, \ldots, n$, be the fitted values based on H^k. For k=1, $\hat{y}_i^{(k)}$ is the average of the $y_i's$.
2. Add a term of the form $f_j(h_1, h_2)$ to the model where $h_1 \in H^k$ and h_2 is one of the predictor variables $(x_1, \ldots, x_p)$ that is not an argument of h_1. The term $f_j(1, h_2)$ denotes a univariate smooth function of h_2.
   For estimating $f_j(h_1, h_2)$, we apply the one/two dimensional loess smoother (Cleveland and Devlin (1988) and Cleveland, Grosse and Shyu (1992); available in the statistical package S) to the residual $y_i - \hat{y}_i^{(k)}$. GCV is applied to select the smoothing parameters and the best term (h_1, h_2), (see Section 3.3). The basis set is then augmented to $H^{k+1} = H^k \bigcup \{f_j(h_1, h_2)\}$. Note, if $h_1 = 1$ then a univariate smoother is used.

3. After each term is added to the model, re-estimate all terms by the backfitting algorithm (Algorithm C). (See Section 3.2.)
4. Repeat steps 2 and 3 until there are m_o terms in the full model M_0.

Algorithm B – Backwards deletion

1. Begin with full model M_0.
2. Define a terminal term to be any term that does not appear as an argument in a more complicated term. Remove the terminal term corresponding to the smallest reduction in GCV. Then refit the model using the backfitting algorithm.
3. Repeat 2 until the null model is reached.
4. Choose the model that minimizes the GCV in the backward deletion model sequence. The GCV score is discussed in Section 3.3.

3.2. Backfitting

The form of the backfitting algorithm is similar to that used for fitting the additive model. Denote a tentative model from Step 2 of Algorithm A or B by

$$y = \sum_{j=1}^{m} f_j(x_1, \ldots, x_p) + \epsilon.$$

Recall that by our construction, each f_j is either univariate or bivariate, taking the form of $f_j(h_1, h_2)$.

Algorithm C – Backfitting

1. Set $f_j^{(0)} = f_j$ and $q = 1$.
2. For each j, $j = 1, \ldots, m$ where m is the number of terms in the model, let $r_j^{(q)} = y - \sum_{k<j}^{p} f_k^{(q)} - \sum_{k>j}^{p} f_k^{(q-1)}$, the residual of the fit which includes all the simpler terms updated in the current iteration and the other terms updated in the previous iteration. Fit $r_j^{(q)}$ with a function $\tilde{f}_j^{(q)}$ of the form $f_j(h_1, h_2)$ where h_1 is one of the simpler updated terms in the model $(f_l^{(q)}, l < j)$ and h_2 is one of the predictors $(x_1, \ldots, x_p)$. Loess smoothing and GCV as described in Section 3.3 are applied here to find the best fit. Set $f_j^{(q)} = w\tilde{f}_j^{(q)} + (1 - w)f_j^{(q-1)}$ where w is a number between 0 and 1. Replace q by $q + 1$.
3. Repeat Step 2 until convergence.

Allowing the estimated smooth function arguments depending on simpler terms in the model adds considerable complexity to the problem of studying

properties of the algorithm. We do not, as yet, have general results regarding the convergence of the estimated component functions or regarding possible multiple local minima. However, it appears that finding the "best" function in the large class of functions described by such a model, is not critical for getting good predictions. For instance, the first two or three steps in the backfitting algorithm gave good predictions for the examples considered and the results were relatively close to those obtained by using a larger number of steps in the backfitting algorithm.

Applications have shown that an update factor or under-relaxation factor (e.g. see Buja, Hastie and Tibshirani (1989)) $w < 1$ is sometimes needed to achieve continued reduction in residual error with further iterations. For instance, a relatively small w is often needed for data with a low signal to noise ratio. Initially, we set $w = 1$ and monitor the residual sums of squares during backfitting to see if a smaller w is required.

During backfitting, constraints such as (2.2), on the higher order terms are ignored. The removal of additive components to facilitate interpretation of "interactions", can be done by fitting an additive model to each higher order fitted term after backfitting.

3.3. GCV score

The selection of terms to enter the model, span of the smoother, and overall model selection uses a GCV score. The GCV score corresponding to a model M is defined as

$$\mathrm{GCV}(M) = \sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2/(n(1 - c(M)/n)^2),$$

where $c(M)$ is an approximation to the degrees of freedom of the model. The GCV score should account for the increased variability due to the model complexity and variable selection process. The procedure uses an approximate degrees of freedom function $c(M)$ of the following form

$$c(M) = \sum_{j=1}^{m}(c_0 + c_j(d_j, p)),$$

where c_0 is a penalty associated with span selection for a given smoother (we typically let $c_0 = 1$). The approximation to the degrees of freedom used up by the smoothing and variable selection is denoted by $c_j(d_j, p)$ and depends on d_j, the approximate number of degrees of freedom corresponding to the outer-most smoother used in constructing term j (based on the trace of the smoother matrix)

and the number of predictors, p. A simple choice is $c_j(d_j, p) = d_j + 1$, the degrees of freedom for the smoother plus one for the variable selection. We have also tried another approximation to adjust for the adaptiveness of selecting the best predictor among p predictor variables. Details of the calculations are given in LeBlanc (1993).

The default spans of the smoothers used in the algorithms are $(1, .5, .2)$ for the univariate smoother and $(.9, .7, .3)$ for the bivariate smoother. These spans were chosen so that a relatively wide range of functions from "wiggly" to very smooth could be approximated. The spans are different for the bivariate and univariate smoothers for two reasons. First, the middle and small spans are larger for the bivariate smoother to control the effective degrees of freedom since a relatively small span on a bivariate smoother uses up many more degrees of freedom relative to a univariate smoother. Second, the largest span for the bivariate smoother is smaller than the univariate smoother because an approximately linear bivariate smooth function is not of interest since the same function can be obtained by a sum of two univariate functions.

4. Examples

In this section we report on the application of the proposed procedure to simulated and real data. We report the relative residual error,

$$\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 / \sum_{i=1}^{n}(y_i - \overline{y})^2,$$

and estimated relative prediction error (based on a test sample)

$$\sum_{i=1}^{n_T}(y_{Ti} - \hat{f}(x_{Ti}))^2 / \sum_{i=1}^{n_T}(y_{Ti} - \overline{y}_T)^2,$$

where y_{Ti} and x_{Ti} are the test sample observations, n_T is the number of test observations and $\overline{y}_T$ is the mean of the test sample responses.

4.1. Complex effect modification

Sometimes response variables need to be described by a complicated non-linear model of more than two dimensions. We consider the low dimensional expansion approximation of the following function with noise:

$$f = \sin(5\pi x_1)(5 + 8\exp(-3x_3)(2x_2 - 1)^2), \tag{4.1}$$

where the predictors x_1, x_2, x_3 were drawn independently from the standard uniform distribution and where response errors were generated from a Gaussian

distribution with standard deviation .25 so that the signal to noise ratio was approximately 4 to 1. The sample had 200 observations. The approximation obtained by the new method was of the form

$$\hat{f} = s_1(x_1) + s_2(s_1(x_1), x_2) + s_3(s_2(s_1(x_1), x_2), x_3) \tag{4.2}$$

with 37.5 effective degrees of freedom.

Based on one thousand test observations, the relative prediction error was estimated to be .086 for model (4.2) and it increased to .470 for an additive model. The estimated relative prediction error for the MARS methods was .171. A paired t-test statistic between MARS and the new method on the squared prediction errors for the test data observations indicates a significant difference ($t = 22.2, p < .0001$). The new procedure gave smaller prediction errors than the MARS procedure on some similar simulated problems where a complex main effect is modified by other variables. The reason for the improvement is due to the special form of the interaction for which MARS would require a large number of basis functions to obtain an accurate approximation.

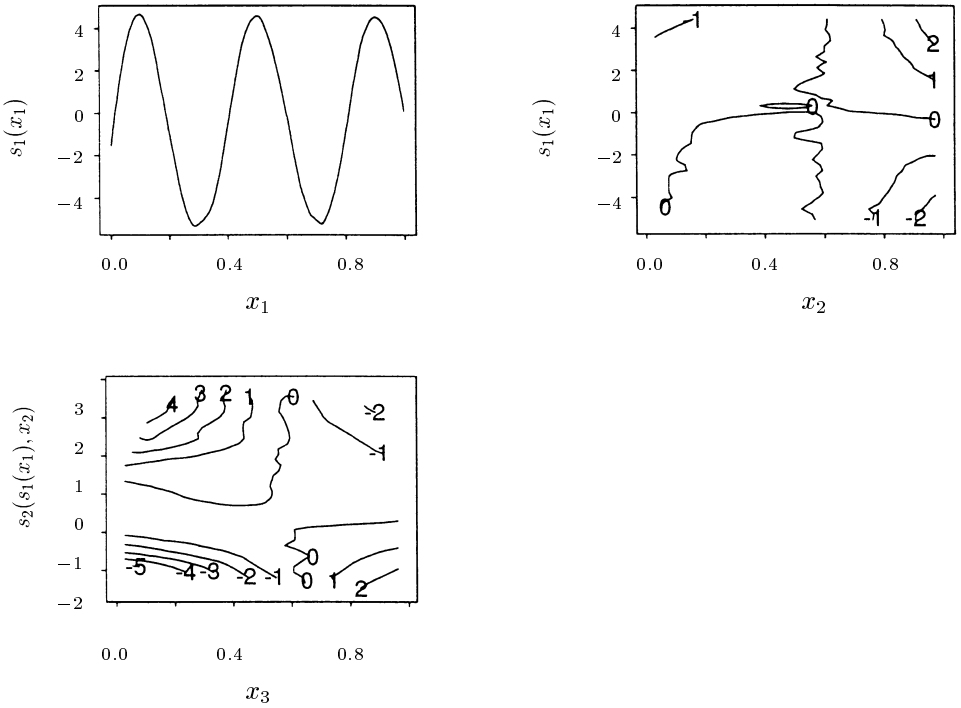

Figure 1. Complex effect modification

Figure 1 shows the approximation of the 3-dimensional surface by profile plots. For instance, the main effect term captures much of the complicated effect in x_1; the second term describes a modification of the main effect x_1 by x_2 to better approximate the function. For large values of x_2 and large negative or positive values values of $s_1(x_1)$ the sinusoid effect is increased in magnitude. A similar effect was missed by the estimation procedure for small x_2 due to the noise. The third panel of Figure 1 shows that for large values of x_3 the prediction rule is decreased in magnitude.

While examples of complicated effect modification are not common in some fields, they can occur in the modeling of complicated spatial or temporal relationships such as those present in the example given in Section 4.2.

4.2. Mumps data

This example considers a data set based on the monthly reported cases of mumps by state for 1953-1989 and is from the National Notifiable Disease Surveillance System directed by the Centers for Disease Control. The new method is used to model mumps incidence as a function of time (month and year) and location (the latitude and longitude of the approximate center of each state.) Several other authors have used the mumps data to investigate graphical and computational methods, for instance Chaudhuri, Huang, Loh and Yao (1994) and Burr and Gomatam (1991).

While there are many complexities in the data, we present a simple analysis with a goal of assessing the predictive performance of the new technique in a real data regression problem with complicated non-linearity. A 10% random sample ($n = 648$ observations) from the data for the years 1965-1981 and in the 48 contiguous states are used to develop a model.

The logarithm of the rate (number of cases divided by estimates of the population of the states in 1975) is used as the response. Chaudhuri, Huang, Loh and Yao (1994) also model the logarithm of rates. The latitude and longitude of the approximate center of states is from the `state.x77` data set in the S statistical language.

The model selected by the procedure is

$$\begin{aligned}\hat{f} =& s_1(\texttt{year}) + s_2(\texttt{mon}) + s_3(\texttt{lon}) + s_4(s_3(\texttt{lon}), \texttt{lat}) \\ &+ s_5(s_1(\texttt{year}), \texttt{lat}) + s_6(s_5(s_1(\texttt{year}), \texttt{lat}), \texttt{lon}). \qquad (4.3)\end{aligned}$$

The method identifies a striking seasonal effect in the mumps rates $s_2(\texttt{mon})$ which is shown in Figure 2. The rates were lowest in the late summer and early fall and peaked in early spring.

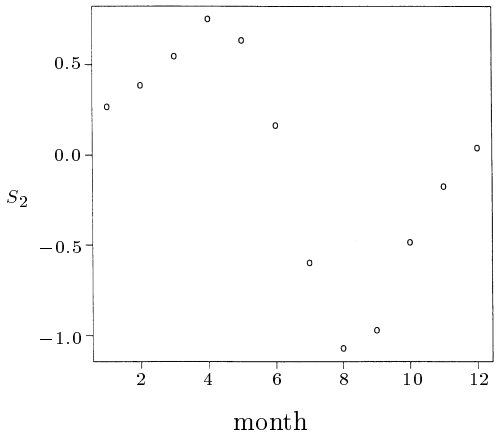

Figure 2. The seasonal component of the mumps incidence model

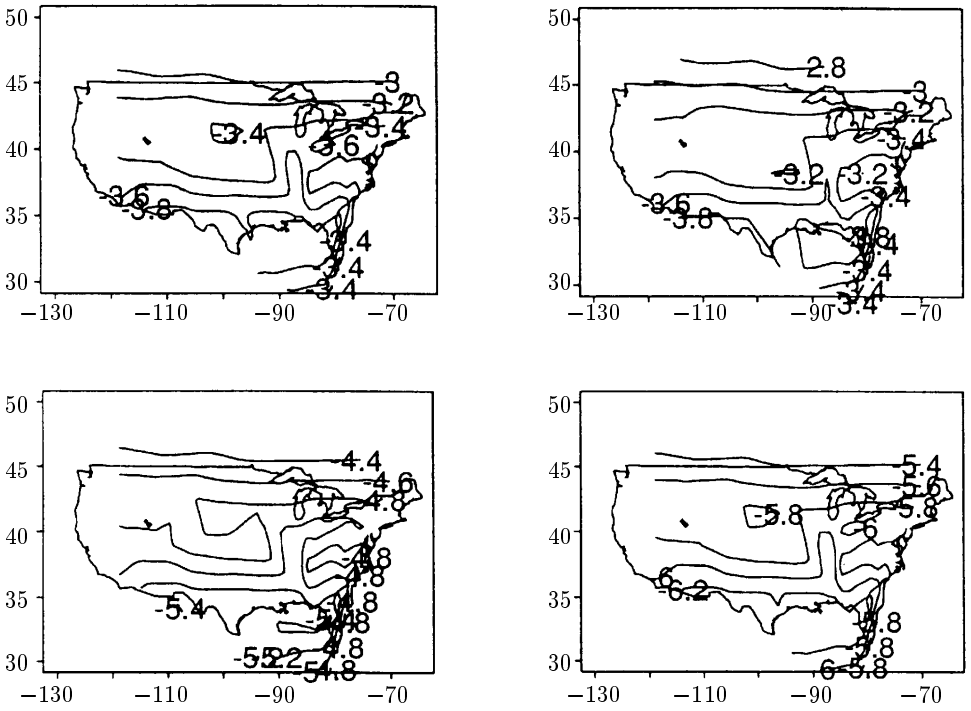

Figure 3. Spatial and temporal components in the fitted mumps model for years 1966, 1971, 1976 and 1981, for panels from the upper left to the lower right.

The main effect and interaction terms involving year, latitude and longitude are added together and presented as a series of contour plots in Figure 3. There was a strong decreasing trend over time which was due to the increasing use of mumps vaccine after 1967. The rates were also generally larger for more northern states. There were several states which tended to have relatively high rates in the Great Lakes region of the United States; this was not entirely identified by the procedure using the relatively small sample of approximately 10% of data in the time period.

A test sample of approximately 20% of the remaining observations ($n = 1218$) was used to estimate the prediction error. The relative estimated prediction error for the method was .436. An additive model was fitted to the data with span= .3 for the smoothers and yielded a substantially larger prediction error, .556, on the test sample. The MARS method was also applied to the same data set and gave relative prediction error on the test sample equal to .484. A paired t-test statistic on the squared prediction errors between MARS and the new method is 3.2 ($p < .001$), indicating a small but significant improvement compared to the MARS predictions for this example.

5. Discussion

The regression method developed in this paper can be a useful adaptive extension to additive modeling. For the examples, the new procedure gives prediction errors that are better than additive regression models and MARS models. In addition, the new method also leads to a general effect modification interpretation of higher order terms. The procedure could be modified to further improve predictions; for instance: GCV optimization could be done over a larger number of spans.

Acknowledgments

The author would like to thank Rob Tibshirani for helpful discussions and the associate editor and referees for many useful suggestions. This work was supported by the Natural Sciences and Engineering Research Council of Canada. When the author was a member of the Department of Preventive Medicine and Biostatistics, University of Toronto.

References

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). Classification and Regression Trees. Wadsworth, Belmont.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models (with discussion). Ann. Statist. **17**, 453-555.

Burr, D. and Gomatam, S. (1991). Analysis of the TB and Mumps data. Proc. ASA Section on Statistical Graphics.

Chaudhuri, P., Huang, M., Loh, W. and Yao, R. (1994). Piecewise-polynomial regression and regression trees. Statistica Sinica **4**, 143-167.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. J. Amer. Statist. Assoc. **83**, 596-610.

Cleveland, W. S., Grosse, E. and Shyu, W. M. (1992). Statistical Models in *S*, chapter Local Regression Modeling. Wadsworth International Group.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). Ann. Statist. **19**, 1-141.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. J. Amer. Statist. Assoc. **76**, 817-823.

Hastie, T. and Tibshirani, R. J. (1990). Generalized Additive Models. Chapman and Hall, London.

Hinton, G. E. (1989). Connectionist learning procedures. Artificial Intelligence **40**, 185-234.

LeBlanc, M. (1993). An adaptive expansion method for regression. Technical Report, Department of Statistics, University of Toronto.

Fred Hutchinson Cancer Research Center, 1124 Columbia St., Seattle, WA 98104, U.S.A.

(Received July 1993; accepted March 1995)