MASTER COPY KEEP THIS COPY FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

AD-A271 259

2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
July 93	Technical

Generalized Regression Trees: Function Estimation via Recursive Partitioning and Maximum Likelihood

5. FUNDING NUMBERS

DAAL03-91-G-0222

6. AUTHOR(S)

Probal Chaudhuri Wei-Yin Loh
Wen-Da-Lo Ching-Chi[illegible]ng

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Wisconsin
Madison, WI 53706

93-24868

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

ARO 28679.13-MA

11. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

DTIC ELECTE OCT 21 1993 S A D

13. ABSTRACT

A method that blends tree-structured nonparametric regression with classical maximum likelihood is used in a generalized regression setting. The function estimates constructed are piecewise polynomials and are produced together with decision trees containing useful information on the regressors. Fitting is carried out by applying maximum likelihood estimation to subsets of the data, where the subsets are selected via recursive partitioning and cross-validation pruning. Examples of Poisson and logistic regression trees are given to illustrate the method applied to count and binary response data. Large-sample properties of the estimates are derived under appropriate regularity conditions

14. SUBJECT TERMS	15. NUMBER OF PAGES
Generalized linear models, Anscombe residual, pseudo residual, Vapnik-Chervonenkis class, consistency	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 903

July 1993

GENERALIZED REGRESSION TREES: FUNCTION ESTIMATION VIA RECURSIVE PARTITIONING AND MAXIMUM LIKELIHOOD

by

Probal Chaudhuri
Wen-Da-Lo
Wei-Yin Loh
Ching-Ching Yang

Generalized Regression Trees: Function Estimation via Recursive Partitioning and Maximum Likelihood*

Probal Chaudhuri Wen-Da Lo Wei-Yin Loh
Ching-Ching Yang

Indian Statistical Institute, Calcutta, Chung Cheng Institute of Technology, Taiwan, University of Wisconsin, Madison, and Feng Chia University, Taiwan

Abstract

A method that blends tree-structured nonparametric regression with classical maximum likelihood is used in a generalized regression setting. The function estimates constructed are piecewise polynomials and are produced together with decision trees containing useful information on the regressors. Fitting is carried out by applying maximum likelihood estimation to subsets of the data, where the subsets are selected via recursive partitioning and cross-validation pruning. Examples of Poisson and logistic regression trees are given to illustrate the method applied to count and binary response data. Large-sample

*Chaudhuri's research was partially supported by a grant from the Indian Statistical Institute. Loh's research was partially supported by ARO grant DAAL03-91-G-0111.

properties of the estimates are derived under appropriate regularity conditions.

Key words and phrases: Generalized linear models, Anscombe residual, pseudo residual, Vapnik-Chervonenkis class, consistency.

1 Introduction: Motivation and main ideas

Consider a general regression set up in which a real-valued response Y is related to a real or a vector-valued regressor X through an appropriate probability model, which characterizes the nature of the dependence of Y on X. To be more specific, let us denote the conditional density or mass function of Y given $X = x$ as $f\{y|g(x)\}$, where the form of f is known but g is an unknown function, which happens to be the parameter of interest here. There are plenty of examples that arise in practice and fit into this structure. Some well-known cases, which have received extensive attention in the literature, are the logistic regression model (when the response Y is binary, and $g(x)$ is the "logit" of the conditional probability parameter given $X = x$), the Poisson regression model (when Y is a nonnegative integer-valued random variable with a Poisson distribution, and $g(x)$ is related to its unknown conditional mean given $X = x$), and more generally, models that are popularly called generalized linear models (GLM) (Nelder and Wedderburn 1972, McCullagh and Nelder 1989), where g is related to the link function. On the other hand, $g(x)$ may be the unknown location parameter associated with the conditional distribution of Y given $X = x$. In other words, Y may satisfy the equation $Y = g(X) + \epsilon$, where the conditional distribution of ϵ can be normal, Cauchy or exponential power (see, e.g., Box and Tiao 1973) with center at zero.

We are interested in the situation where no finite-dimensional parametric model is imposed on g, and it is assumed to be a smooth function with an appropriate degree of smoothness. Nonparametric estimation of the functional parameter g has been explored by Cox and O'Sullivan (1990), Gu (1990), Hastie and Tibshirani (1986, 1990), O'Sullivan, Yandell and Raynor (1986), Staniswalis (1989), Stone (1986, 1991a), and others, who considered various nonparametric smoothers when the conditional distribution of the response given the regressor is assumed to have a known shape (e.g., the conditional distribution may possess a GLM-type exponential structure).

In the case of the usual regression set up, where $Y = g(X) + \epsilon$ with $E(\epsilon|X) = 0$, several attempts have been made to estimate g by recursively partitioning the regressor space and then constructing a regression estimate in each partition using the method of least squares. Important developments along this direction are AID (Sonquist 1970, Sonquist, Baker and Morgan 1973), CART (Breiman, Friedman, Olshen and Stone 1984) and SUPPORT (Chaudhuri, Huang, Loh and Yao 1993). The purpose of this article is to explore recursive partitioning algorithms and related likelihood-based nonparametric function estimates in a generalized regression setting.

Two significant advantages enjoyed by recursive partitioning and tree-structured regression are:

- the decision tree as well as the intermediate and terminal nodes created by the partitioning algorithm may provide valuable information about the regressors, and

- the estimates constructed in each terminal node has a simple functional form. This permits their statistical properties to be studied and lends

insight into the nature of the relationship between the response and the regressors within a node.

Besides, the adaptive nature of a recursive partitioning algorithm allows varying degrees of smoothing over the regressor space so that the terminal nodes may have variable sizes in terms of numbers of observations contained in those nodes as well as the diameters of the sets in the regressor space to which they correspond. The main motivation behind such adaptive variable smoothing is to take care of heteroscedasticity as well as the possibility that the amount of smoothness in the functional parameter g may be different in different parts of the regressor space. This is an improvement over most of the earlier nonparametric estimation techniques in generalized regression, which concentrated either on adaptive but fixed smoothing (i.e., using a smoothing parameter whose value is constant over the entire regressor space) or on deterministic smoothing.

The general methodology explored in this paper consists of two fundamental steps.

1. Observations are recursively and adaptively divided into subsets so that the unknown function g can be satisfactorily approximated by a simple function (e.g., a constant, a linear function or a polynomial of suitable degree) in each subset.

2. The function g is estimated from the data in each terminal node by a polynomial using maximum likelihood. Estimates of the derivatives of g are given by the corresponding derivatives of the fitted polynomial.

The recursive partitioning algorithm used to create the terminal nodes and the nature of the function fitted will depend on the problem. In Sections 2

and 3, we give some algorithms and examples for illustration (see also Ciampi and Thiffault (1989)).

Adaptive recursive partitioning algorithms construct random subsets of the regressor space which form the terminal nodes. A serious technical barrier in studying the analytic properties of the likelihood-based function estimates is the randomness in these subsets. Our key tool in coping with this situation is a well-known combinatorial result in Vapnik and Chervonenkis (1971). In Section 4, we investigate the large sample statistical properties of the estimates that are constructed via recursive partitioning of the regressor space followed by maximum likelihood estimation of g by piecewise polynomials. We will consider a very general setting to get good theoretical insights into the performance of the estimates, and to derive some technical results under mild regularity conditions.

Friedman's (1991) MARS combines spline fitting with recursive partitioning to produce continuous function estimates. The complexity of the estimates makes interpretation difficult and theoretical analysis of their statistical properties extremely challenging. In the SUPPORT method of Chaudhuri et al. (1993), a weighted averaging technique is used to combine piecewise-polynomial fits into a smooth one. An identical technique can be used here to create a smooth estimate from a discontinuous piecewise-polynomial estimate without altering the asymptotic properties of the original estimate. Friedman (1991) gives some proposals for applying MARS to logistic regression problems, and Buja, Duffy, Hastie and Tibshirani (1991) and Stone (1991b) comment on possible extensions and modifications of MARS to GLM-type exponential response problems. The methodology presented and analyzed in this article has a clear edge over all these proposals

because of its simplicity and more direct approach. It is hoped that this will make it more appealing to users. It definitely helps in interpreting the estimates and in studying their statistical properties.

2 Algorithms for Poisson and logistic regression trees

Algorithms for fitting Poisson and logistic regression trees are briefly described in this section. Each algorithm has three main components, namely:

1. A method to select the variable and the splitting value to be used at a partition.
2. A method to determine the size of the tree.
3. A method to fit a model to each terminal node.

There are many reasonable solutions for each component, and several of them are described and implemented in FORTRAN 77 in Lo (1993) and Yang (1993). In the examples in this paper, two-sample tests for means and variances are used to find splitting variables (Huang 1989, Chaudhuri et al. 1993). CART's method of cost-complexity pruning (with cost defined as deviance) is used to determine the size of a tree. Finally, a loglinear model or a linear logistic regression model is fitted to each terminal node. We begin with Poisson regression.

2.1 Poisson regression

The following sequence of computations is performed at each node t.

1. A Poisson loglinear model is fitted to the data in t.

2. Let $m_i = EY_i$ and let $\hat{m}_i$ be its value estimated from the model. Also let y_i denote the observed value of Y_i. The adjusted Anscombe residual (Pierce and Schafer 1986)

$$r_i = \{y_i^{2/3} - (\hat{m}_i^{2/3} - (1/9)\hat{m}_i^{-1/3})\}/\{(2/3)\hat{m}_i^{1/6}\}$$

   is calculated for each y_i in t. (Yang, 1993, discusses the advantages of this residual over unadjusted Anscombe, Pearson, and deviance residuals.)

3. Observations with nonnegative r_i are classified as belonging to Group 1 and the others to Group 2.

4. Two-sample t-statistics to test for differences in means and variances between the two groups along each covariate axis are computed. (The latter test is Levene's, 1960, test; see Chaudhuri et al. (1993).) The rationale is that if the model fits adequately, the residuals should look like noise and there would be little difference between the means and variances of the two groups. Otherwise, one or more of the test statistics may be expected to be large. This method has proven to be effective for tree-structured classification (Loh and Vanichsetakul 1988) and regression with censored data (Ahn and Loh 1994). Its principal advantage over the exhaustive search strategies of AID and CART is computational speed.

5. The covariate used to split the node is the one that possesses the most significant t-statistic among all the tests.

July 25, 1993

6. The cut-point for the selected covariate is the average of the two group means along that covariate. Observations with covariate values less than or equal to the cut-point are channeled to the left subnode and the remainder to the right subnode

7. After an overly large tree is constructed, the nodes are pruned back following CART's pruning method with cost-complexity defined as residual deviance plus a constant times the number of terminal nodes of the tree. As in CART, 10-fold cross-validation is used to determine the constant and hence the amount of pruning.

8. The final tree is the one that has the smallest cross-validation estimate of deviance.

2.2 Logistic regression

Because of the 0-1 nature of the Y-variable in logistic regression applications, the definition of residuals in the preceding algorithm needs to be modified as follows. Otherwise, the algorithm is similar to that for Poisson regression.

1. The Y-values are first smoothed using a weighted average (similar to the LOWESS method of Cleveland (1979)) to give a preliminary estimate p_i^* of the probability $p_i = P(Y_i = 1)$. This estimate is called a "pseudo-observation."

2. A second estimate $\hat{p}_i$ of this probability from a logistic regression model fitted to the node is obtained.

3. The "pseudo-residual," $r_i^* = (p_i^* - \hat{p}_i)/\hat{\sigma}(p_i^*)$, is computed for each observation. Here $\hat{\sigma}(p_i^*)$ is an estimate of standard deviation proposed

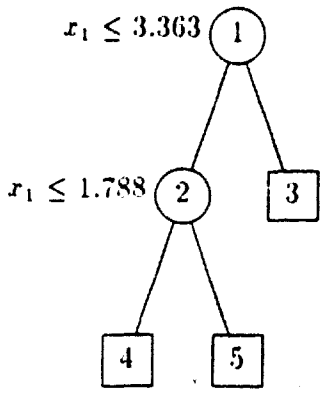

Figure 1: Pruned tree with 10-fold cross-validation for Poisson example. The true model is $\log(m) = 2\sin(x_1 - x_2) + x_2$. The loglinear models in the terminal nodes are given by $\log(m) = f(i)$, $i = 3, 4, 5$, where i denotes node number and $f(3) = 5.117 - 1.474x_1 + 2.286x_2$, $f(4) = -0.534 + 1.789x_1 - 0.407x_2$ and $f(5) = 1.146 + 0.222x_1 + 0.977x_2$.

by Fowlkes (1987), whose simulations suggest that the pseudo-residual is approximately standard normal and independent of the fitted value for large samples.

4. The pseudo-residual is used in place of the adjusted Anscombe residual in the algorithm for Poisson regression trees.

3 Numerical examples

Two examples are given in this section to illustrate the algorithms. In the first example, 100 independent (x_1, x_2) pairs are simulated, with x_1 and x_2 independent uniformly distributed random variables over the intervals $(0, 2\pi)$ and $(0, 2)$, respectively. For each pair, a Poisson response is generated

with mean m given by

$$\log m = 2\sin(x_1 - x_2) + x_2.$$

A plot of m versus the regressors is shown in Figure 2(a). Applying our Poisson tree algorithm with loglinear fits at each node and 10-fold cross-validation pruning gives a tree with three terminal nodes as shown in Figure 1. The corresponding piecewise-loglinear estimated surface is shown in Figure 2(b). The fit is remarkably good, even though it is made up of three separate pieces.

For the second example, we simulate 300 independent observation vectors (Y_i, X_{i1}, X_{i2}), $i = 1, \ldots, 300$, where X_{i1} and X_{i2} are uniformly and independently distributed on the square $(-1.5, 1.5) \times (-1.5, 1.5)$, and Y_i is Bernoulli with probability $p_i = P(Y_i = 1)$ given by

$$\log\{p_i/(1 - p_i)\} = x_{i1} + \sin(\pi x_{i2}).$$

A plot of p_i versus x_{i1} and x_{i2} is shown in Figure 3. Figure 4 shows a tree with six terminal nodes constructed by our logistic regression tree algorithm. The fitted functions at the terminal nodes are $\log\{p_i(1 - p_i)\} = f(i)$, where i denotes the node number and

$$\begin{aligned} f(4) &= 1.391 - 0.492x_1 + 1.477x_2, \\ f(6) &= -0.184 - 0.076x_1 - 0.002x_2, \\ f(8) &= 0.706 + 0.962x_1 - 3.198x_2, \\ f(9) &= -6.420 - 0.855x_1 + 4.998x_2, \end{aligned}$$

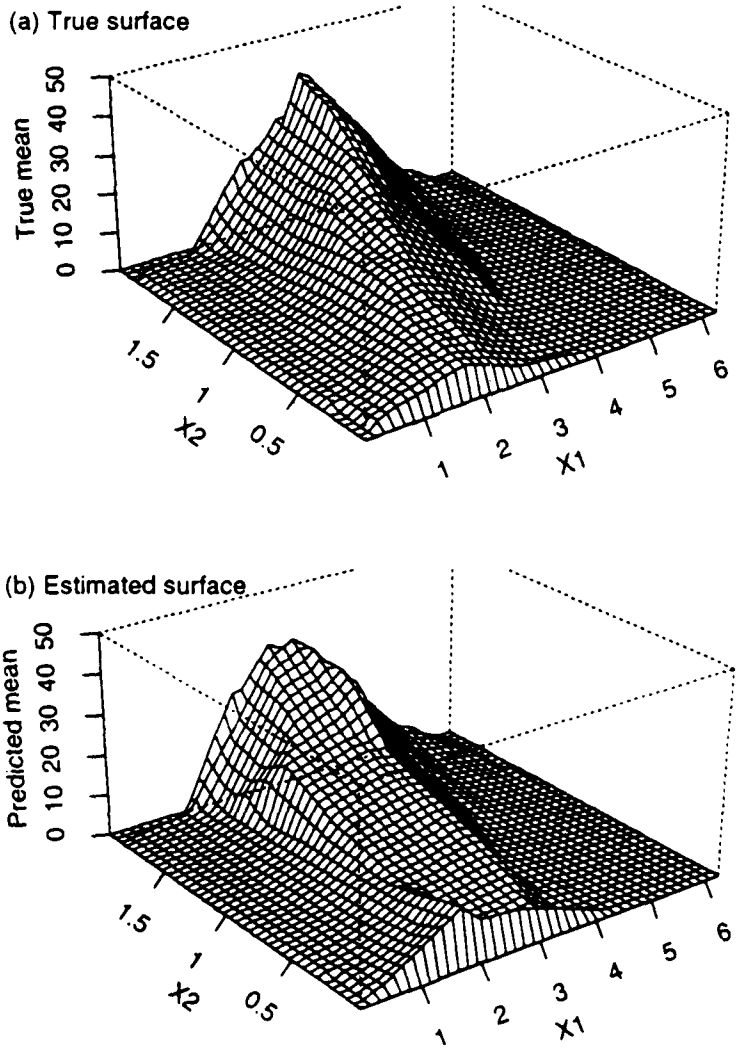

Figure 2: True and estimated surfaces for Poisson regression example. The estimated surface is composed of three discontinuous loglinear models.

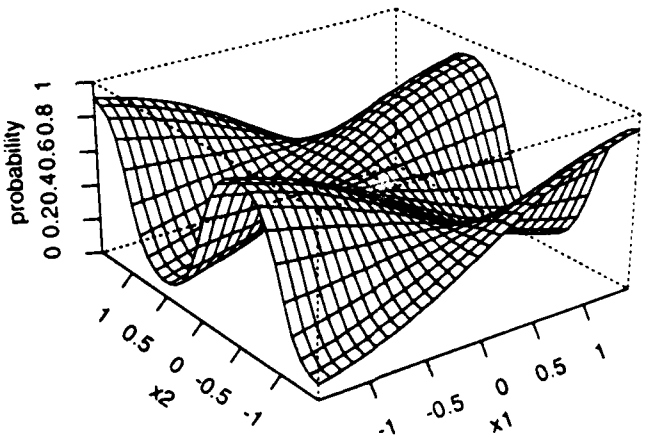

Figure 3: True function for logistic regression example.

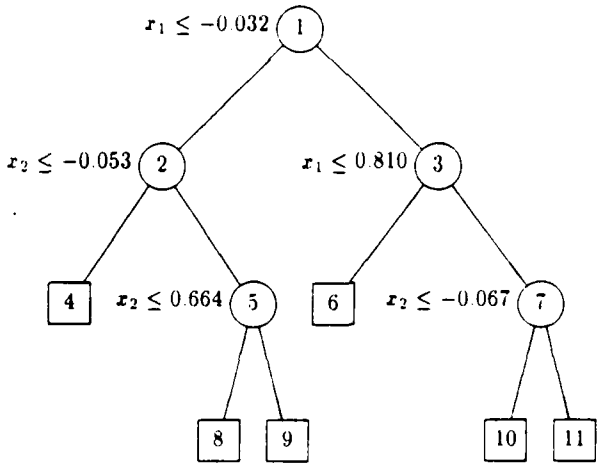

Figure 4: Pruned tree for logistic regression example.

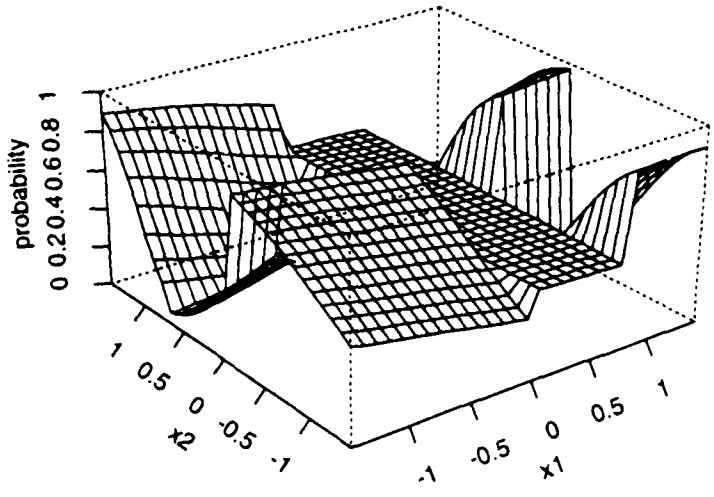

Figure 5: Unsmoothed estimate of the function for logistic regression example.

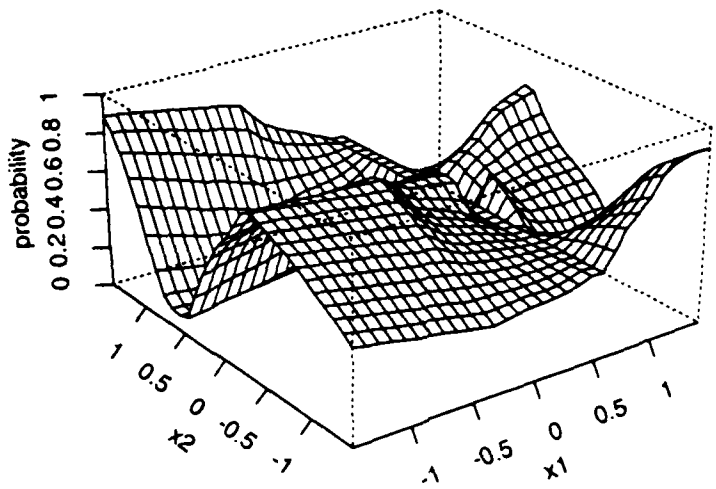

Figure 6: Smoothed estimate of the function for logistic regression example.

$$f(10) = -3.218 + 0.147x_1 - 3.552x_2,$$
$$f(11) = 1.279 + 1.399x_1 - 4.311x_2.$$

The unsmoothed and smoothed function estimates are plotted in Figures 5 and 6, respectively. The smoothing is achieved by weighted averaging using trapezoidal weights (see Lo (1993) for details).

4 Statistical properties of estimates: Some technical results

Assume that $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ are independent data points, where the response Y_i is real-valued and the regressor X_i is d-dimensional.

As before, let $f\{y_i|g(x_i)\}$ be the conditional pdf/pmf of Y_i given $X_i = x_i$. We wish to estimate the function g over a compact set $C \subset R^d$. Let T_n be a random partition of C (i.e., $C = \cup_{t \in T_n} t$), which is generated by some adaptive recursive partitioning algorithm applied to the data, and it is assumed to consist of polyhedrons having at most M (a fixed positive integer) faces. We will denote the diameter of a set $t \in T_n$ by $\delta(t)$ (i.e., $\delta(t) = \sup_{x,y \in t} |x - y|$), which will be assumed to be positive for each set $t \in T_n$. For $t \in T_n$, $\bar{X}_t$ will denote the average of the X_i's that belong to t. Also, assuming that the function g is m-th order differentiable ($m \geq 0$), let us write its Taylor expansion around $\bar{X}_t$ as

$$g(x) = \sum_{u \in U} (u!)^{-1} D^u g(\bar{X}_t)(x - \bar{X}_t)^u + r_t(x, \bar{X}_t) .$$

Here $U = \{u | u = (v_1, v_2, \ldots, v_d), [u] \leq m\}$, where $[u] = v_1 + v_2 + \ldots + v_d$ and the v_i's are nonnegative integers. For $u \in U$, D^u is the mixed partial differential operator with index u, $u! = \prod_{i=1}^d v_i!$, and for $x = (z_1, z_2, \ldots, z_d)$, $x^u = \prod_{i=1}^d z_i^{v_i}$ (with the convention that $0! = 1$ and $0^0 = 1$). We impose the following condition (cf. Condition (a) in Chaudhuri et al. (1993)) concerning the behavior of the remainder term in the above Taylor expansion.

Condition 1 $\max_{t \in T_n} \sup_{x \in t} \{\delta(t)\}^{-m} |r_t(x, \bar{X}_t)| \xrightarrow{P} 0$ *as* $n \to \infty$.

Observe that if g is continuously differentiable with derivatives up to order m on an open set in R^d that contains the compact set C and the diameters of the sets in T_n shrink (i.e., if $\sup_{t \in T_n} \delta(t) \xrightarrow{P} 0$) in probability as $n \to \infty$ (cf. Condition (12.9) in Breiman et al. (1984)), the above condition automatically holds. However, even if some of the sets in T_n do not shrink

as n grows. Condition 1 may still be true. In any case, Condition 1 implies that the function g can be uniformly well approximated by polynomials of degree smaller than or equal to m on each of the sets in T_n when n is large.

For $\Theta = (\theta_u)_{u \in U}$, let us define the polynomial $P(x, \Theta, X_t)$ in x as

$$P(x, \Theta, X_t) = \sum_{u \in U} \theta_u (u!)^{-1} \{\delta(t)\}^{-|u|} (x - X_t)^u .$$

Following the estimation procedure described in the previous sections, let $\hat{\Theta}_t$ be the estimate obtained by applying the maximum likelihood technique to the data points (Y_i, X_i) for which $X_i \in t$. In other words,

$$\hat{\Theta}_t = \arg\max_{\Theta} \prod_{X_i \in t} f\{Y_i | P(X_i, \Theta, X_t)\} .$$

We will now state a couple of conditions concerning the distribution of the X_i's in the regressor space. For $X_i \in t$, let Γ_i be the $s(U)$-dimensional column vector with components given by $(u!)^{-1}\{\delta(t)\}^{-|u|}(X_i - X_t)^u$, where $u \in U$. Here $s(U)$ is the size of the finite set U, which is defined earlier. Next, denote by D_t the $s(U) \times s(U)$ matrix defined as $\sum_{X_i \in t} \Gamma_i \Gamma_i^T$, where T indicates transpose.

Condition 2 *Let N_t = the number of X_i's that belong to t, and $N_n = \min_{t \in T_n} \{\delta(t)\}^{2m} N_t$. Then $N_n / \log n \xrightarrow{P} \infty$ as $n \to \infty$.*

Condition 3 *Let λ_t be the smallest eigenvalue of $N_t^{-1} D_t$ and let $\lambda_n = \min_{t \in T_n} \lambda_t$. Then λ_n remains bounded away from zero in probability as $n \to \infty$.*

Clearly, Condition 2 ensures that there will be sufficiently many observations in each of the sets in T_n (cf. Condition (12.8) in Breiman et al. (1984)

and Condition (b) in Chaudhuri et al. (1993)). Condition 3, on the other hand, guarantees that for large sample size, each of the matrices D_i's will be nonsingular and nicely behaved (cf. Condition (c) in Chaudhuri et al. (1993)) with a high probability. In a sense, it ensures regularity in the behavior of the Fisher information matrix associated with the finite-dimensional model fitted to the conditional distribution within each set in T_n. Note that we are fitting a polynomial of a fixed degree with a finite number of coefficients to the data points corresponding to any set in T_n.

Finally, we will state a Cramér-type regularity condition on the conditional distribution of the response given the regressor. This condition is absolutely crucial in establishing desirable asymptotic behavior of our estimates, which are constructed via maximum likelihood technique.

Condition 4 *Let us view the pdf/pmf $f(y|s)$ as a function of two variables so that s becomes a real-valued parameter varying in a bounded open interval J. Here J is such that as x varies over some open set containing C, $g(x)$ takes its values in J. The support of $f(y|s)$ for any given $s \in J$ is the same, and it does not depend on s. Also, $\log\{f(y|s)\}$ is three times continuously differentiable w.r.t. s for any given value of y, and let $A(y|s)$, $B(y|s)$ and $H(y|s)$ be the first, second and third derivatives respectively of $\log\{f(y|s)\}$ w.r.t. s. The random variable $A(Y|s)$ has zero mean, and the mean of $B(Y|s)$ is negative and stays away from zero as s varies in J. Here Y has pdf/pmf $f(y|s)$, and there exists a nonnegative function $K(y)$ which dominates each of $A(y|s)$, $B(y|s)$ and $H(y|s)$ for all values of $s \in J$ (i.e., $|A(y|s)| \leq K(y)$, $|B(y|s)| \leq K(y)$ and $|H(y|s)| \leq K(y)$). Further, if $M(w,s)$ denotes the moment generating function of $K(Y)$ defined as $M(w,s) = E[\exp\{wK(Y)\}]$ with Y having pdf/pmf $f(y|s)$, $M(w,s)$ re-*

mains bounded as w varies over an open interval around the origin and s varies over J.

It is appropriate to note here that Condition 4 is trivially satisfied when the response Y is binary in nature, or more generally, when its conditional distribution given the regressor is binomial, and s is the logit of the probability parameter such that the probability remains bounded away from 0 and 1. As a matter of fact, this condition will hold whenever the conditional distribution of the response belongs to a standard exponential family (e.g., binomial, Poisson, exponential, gamma, normal, etc.), and s is the natural parameter taking values in a bounded interval. If $f(y|s)$ happens to be a location model with s behaving like a location parameter varying over a bounded parameter space, Condition 4 remains true for several important cases like the Cauchy or an exponential power distribution (see e.g., Box and Tiao (1973)). In a sense, this condition can be viewed as an extension of Condition (12.12) in Breiman et al. (1984) and Condition (d) in Chaudhuri et al. (1993).

Theorem 1 *Suppose that Conditions 1 through 4 hold. Then there is a choice of the maximum likelihood estimate $\hat{\Theta}_t$ (possibly a local maximizer of the likelihood) for every $t \in T_n$ such that given any $u \in U$,*

$$\max_{t \in T_n} \sup_{x \in t} |D^u P(x, \hat{\Theta}_t, \tilde{X}_t) - D^u g(x)| \xrightarrow{P} 0 \quad \text{as } n \to \infty.$$

The above theorem guarantees that there exists a choice of the maximum likelihood estimate $\hat{\Theta}_t$ for each $t \in T_n$ so that the resulting piecewise polynomial estimates of the function g and its derivatives are all consistent. Now, it can very well happen that the estimate $\hat{\Theta}_t$ is only a local maximizer of the

likelihood instead of being a global maximizer. For instance, the likelihood based on the data points corresponding to a set in T_n may have multiple maxima. However, when the conditional distribution of the response given the regressor belongs to a standard exponential family, strict concavity of the loglikelihood guarantees uniqueness of the maximum likelihood estimate in large samples. In the special case when we fit a constant (i.e., a polynomial of degree zero) to the data points corresponding to each set in T_n using the maximum likelihood approach, Theorem 1 gives a useful generalization of the consistency result that holds for piecewise constant tree-structured regression estimates discussed in Breiman et al. (1984). The piecewise polynomial estimates of g and its derivatives are not continuous everywhere in the regressor space. Smooth estimates, which can be constructed by combining the polynomial pieces by means of smooth weighted averaging, will be consistent provided the weight functions are chosen properly. Theorem 2 in Chaudhuri et al. (1993) describes a way of constructing families of smooth weight functions that will give smooth and consistent estimates of g and its derivatives.

5 Appendix: The proofs

We begin by proving some preliminary results that will be used in the proof of Theorem 1. Unless stated otherwise, all vectors are assumed to be column vectors and a superscript T denotes transpose.

Lemma 1 *Under Conditions 1, 2 and 4, we have*

$$\max_{t \in T_n} N_t^{-1} \{\delta(t)\}^{-m} \left| \sum_{X_i \in t} [A\{Y_i | P(X_i, \Theta_t^*, \tilde{X}_t)\}] \Gamma_i \right| \xrightarrow{P} 0 \text{ as } n \to \infty .$$

Here Θ_t^* is the $s(U)$-dimensional vector with a typical entry $\{\delta(t)\}^{[u]} D^u g(\bar{X}_t)$ where $u \in U$. In other words, $P(x, \Theta_t^*, \bar{X}_t)$ is nothing but the Taylor polynomial of $g(x)$ expanded around $\bar{X}_t$.

Proof: First observe that a straight forward application of the mean value theorem of differential calculus yields the following

$$\begin{aligned} & N_t^{-1}\{\delta(t)\}^{-m} \sum_{X_i \in t} [A\{Y_i | P(X_i, \Theta_t^*, \bar{X}_t)\}] \Gamma_i \\ &= N_t^{-1}\{\delta(t)\}^{-m} \sum_{X_i \in t} [A\{Y_i | g(X_i)\}] \Gamma_i \\ &\quad - N_t^{-1}\{\delta(t)\}^{-m} \sum_{X_i \in t} \{r_t(X_i, \bar{X}_t) B(Y_i | Z_i)\} \Gamma_i \end{aligned} \tag{1}$$

where Z_i is a random variable that lies between $P(X_i, \Theta_t^*, \bar{X}_t)$ and $g(X_i)$. In view of Condition 4, the conditional mean of $A\{Y|g(X)\}$ given $X = x$ is zero, and if we denote its conditional moment generating function by $M_1(w|x)$, there exist constants $k_1 > 0$ and $\rho_1 > 0$ such that $M_1(w|x) \leq 2\exp(k_1 w^2/2)$ for all $x \in C$ and $0 \leq w \leq \rho_1$ (see the arguments at the beginning of Lemma 12.27 in Breiman et al. (1984)). At this point, pretend that t is a fixed non-random polyhedron in R^d, all the data points X_i's that fall in t form a collection of deterministic points in C, and the corresponding $A\{Y_i|g(X_i)\}$'s form a set of independent random variables such that the distribution of $A\{Y_i|g(X_i)\}$ is the same as the conditional distribution of it given X_i in the original problem. Note that Γ_i is an $s(U)$-dimensional vector with each component bounded in absolute value by 1. The arguments used in handling the "variance term" in the proof of Theorem 1 in Chaudhuri et al. (1993) imply that there exist constants $k_2 > 0$, $k_3 > 0$ and $\rho_2 > 0$ (which depend only on the compact set C, the integer $s(U)$ and the constants k_1, ρ_1) such

that

$$\begin{aligned}\Pr\left(\{\delta(t)\}^{-m}N_t^{-1}\left|\sum_{X_i\in t}[A\{Y_i|g(X_i)\}]\,\Gamma_i\right| > \rho\right) &\le k_2\exp[-k_3\{\delta(t)\}^{2m}N_t\rho^2] \\ &\le k_2\exp(-k_3N_n\rho^2),\end{aligned}$$

whenever $\rho \le \rho_2$. Observe that the first inequality above is a consequence of Lemma 12.26 in Breiman et al. (1984), which can be applied to each real-valued component of the $s(\Gamma)$-dimensional vector that appears here. Recall at this point that each set in T_n is a polyhedron in R^d having at most M faces. The fundamental combinatorial result of Vapnik and Chervonenkis (1971) (Dudley 1978, Section 7) now implies that there exists a collection $\mathcal{C}$ of subsets of the set $\{X_1, X_2, \ldots, X_n\}$ such that $\#(\mathcal{C}) \le (2n)^{M(d+2)}$, and for any polyhedron with at most M faces, there is a set $t^* \in \mathcal{C}$ with the property that $X_i \in t$ if and only if $X_i \in t^*$. Hence, even for a collection like T_n consisting of random polyhedrons generated by an adaptive recursive partitioning algorithm, we must have the following exponential bound for the conditional probability given the X_i's and T_n (i.e., after the sets in T_n are specified).

$$\begin{aligned}\Pr\left(\max_{t\in T_n}\{\delta(t)\}^{-m}N_t^{-1}\left|\sum_{X_i\in t}[A\{Y_i|g(X_i)\}]\,\Gamma_i\right| > \rho \,\middle|\, X_1, X_2, \ldots, X_n, T_n\right) \\ \le (2n)^{M(d+2)}k_2\exp(-k_3N_n\rho^2).\end{aligned}$$

It now follows from Condition 2 that

$$\max_{t\in T_n}\{\delta(t)\}^{-m}N_t^{-1}\left|\sum_{X_i\in t}[A\{Y_i|g(X_i)\}]\,\Gamma_i\right| \xrightarrow{P} 0 \text{ as } n \to \infty .$$

July 25, 1993

For the second term on the right of (1), we have using Conditions 1, 2 and 4

$$\begin{aligned}
&\max_{t\in T_n} N_t^{-1}\{\delta(t)\}^{-m}\left|\sum_{X_i\in t}\{r_t(X_i,\bar{X}_t)B(Y_i|Z_i)\}\,\Gamma_i\right| \\
&\quad\le \left[\max_{t\in T_n}\{\delta(t)\}^{-m}\sup_{x\in t}|r_t(x,\bar{X}_t)|\right]\left\{\max_{t\in T_n} N_t^{-1}\sum_{X_i\in t}K(Y_i)|\Gamma_i|\right\} \\
&\quad\xrightarrow{P} 0 \text{ as } n\to\infty.
\end{aligned}$$

Note that we are using the fact that $\max_{t\in T_n} N_t^{-1}\sum_{X_i\in t}K(Y_i)|\Gamma_i|$ remains bounded in probability as $n\to\infty$ in view of the boundedness of the vectors Γ_i's and Conditions 2 and 4. In fact, if $\mu(x)$ denotes the conditional mean of $K(Y)$ given $X = x$, arguments identical to those used in handling the first term on the right of (1) yield

$$\max_{t\in T_n} N_t^{-1}\left|\sum_{X_i\in t}\{K(Y_i)-\mu(X_i)\}|\Gamma_i|\right| \xrightarrow{P} 0 \text{ as } n\to\infty.$$

This completes the proof of the lemma.

Lemma 2 *Consider the $s(U)\times s(U)$ matrix*

$$-N_t^{-1}\sum_{X_i\in t}[B\{Y_i|P(X_i,\Theta_t^*,X_t)\}]\,\Gamma_i\Gamma_i^T$$

and let $\gamma(t)$ denote its smallest eigenvalue. Define $\gamma_n = \min_{t\in T_n}\gamma(t)$. Then, under Conditions 1 through 4, γ_n remains positive and bounded away from zero in probability as $n\to\infty$.

Proof: Once again, let us write using the mean value theorem of differential calculus

$$N_t^{-1} \sum_{X_i \in t} [B\{Y_i, P(X_i, \Theta_t^*, X_i)\}]' \Gamma_i \Gamma_i^T$$
$$= \sum_{X_i \in t} [B\{Y_i, g(X_i)\} - c(X_i)]' \Gamma_i \Gamma_i^T + N_t^{-1} \sum_{X_i \in t} c(X_i) \Gamma_i \Gamma_i^T$$
$$+ N_t^{-1} \sum_{X_i \in t} \{r_t(X_i, X_i) H(Y_i, V_i)\} \Gamma_i \Gamma_i^T, \qquad (2)$$

where $c(x)$ is the conditional mean of $B\{Y, g(X)\}$ given $X = x$, and V_i is a random variable that falls between $g(X_i)$ and $P(X_i, \Theta_t^*, X_i)$. Now it is obvious from Conditions 3 and 4 that if $\eta_n = \min_{t \in T_n} \eta(t)$, where $\eta(t)$ is the smallest eigenvalue of the matrix $-N_t^{-1} \sum_{X_i \in t} c(X_i) \Gamma_i \Gamma_i^T$, then η_n remains positive and bounded away from zero in probability as $n \to \infty$. On the other hand, the first term on the right of (2) can be handled in the same way as the first term on the right of (1) in the proof of Lemma 1 to yield

$$\max_{t \in T_n} N_t^{-1} \sum_{X_i \in t} [B\{Y_i, g(X_i)\} - c(X_i)] \Gamma_i \Gamma_i^T \xrightarrow{P} 0 \text{ as } n \to \infty$$

Note that the arguments, which exploit Conditions 2 and 4 and were applied to each component of the $s(C)$-dimensional vector appearing as the first term on the right of (1), can be easily modified for each entry of the $s(C) \times s(C)$ matrix here. Finally, using conditions 1, 2 and 4, and arguments that are virtually the same as those employed to treat the second term on the right of (1) in the proof of Lemma 1, we obtain the following result for the third

term on the right of (2):

$$\begin{aligned}
&\max_{t\in T_n} N_t^{-1}\left|\sum_{X_i\in t}\{r_t(X_i,\hat{X}_t)H(Y_i|V_i)\}\Gamma_i\Gamma_i^T\right| \\
&\quad\le \left\{\max_{t\in T_n}\sup_{x\in t}|r_t(x,\hat{X}_t)|\right\}\left\{\max_{t\in T_n} N_t^{-1}\sum_{X_i\in t}K(Y_i)|\Gamma_i\Gamma_i^T|\right\} \\
&\quad\xrightarrow{P} 0 \text{ as } n\to\infty.
\end{aligned}$$

This completes the proof of the lemma.

Proof of Theorem 1: First note that the assertion in the Theorem will follow if we can show that there exist choices for the maximum likelihood estimates $\hat{\Theta}_t$'s such that

$$\max_{t\in T_n}\{h(t)\}^{-m}\,\|\hat{\Theta}_t-\Theta_t^*\| \xrightarrow{P} 0 \text{ as } n\to\infty.$$

For $t\in T_n$, let $l_t(\Theta)$ denote the loglikelihood based on the observations (Y_i, X_i) such that $X_i\in t$. That is, $l_t(\Theta)=\sum_{X_i\in t}\log\{f(Y_i|V_i,X_i,\Theta,X_i)\}$. For $\rho>0$, define the event $E_t(\rho)$ as follows. $E_t(\rho)=\{l_t(\Theta)$ is a concave (it can be locally concave) function in a neighborhood of Θ_t^* with radius $\{h(t)\}^m\rho$, i.e., for Θ satisfying $\{h(t)\}^{-m}\|\Theta-\Theta_t^*\|\le\rho$, and it has a maximum (which can be a local maximum if $l_t(\Theta)$ has several maxima in the interior of this neighborhood$\}$. Note that the occurrence of this event implies that the maximum likelihood equation obtained by differentiating $l_t(\Theta)$ w.r.t. Θ will have a root $\hat{\Theta}_t$ such that $\{h(t)\}^{-m}\|\hat{\Theta}_t-\Theta_t^*\|<\rho$. Now, a Taylor expansion of $l_t(\Theta)$ around Θ_t^* yields

$$l_t(\Theta) = l_t(\Theta_t^*) + \sum_{X_i\in t}(\Theta-\Theta_t^*)^T\Gamma_i\dot{\ell}(Y_i,P,X_i,\Theta_t^*,X_i)+$$

$$+ (1/2) \sum_{X_i \in t} (\Theta - \Theta_t^*)^T \Gamma_i \Gamma_i^T B\{Y_i, P(X_i, \Theta_t^*, \bar{X}_t)\}(\Theta - \Theta_t^*)$$
$$+ (1/6) \sum_{X_i \in t} \{(\Theta - \Theta_t^*)^T \Gamma_i\}^3 H(Y_i|W_i), \tag{3}$$

where W_i is a random variable lying between $P(X_i, \Theta_t^*, \bar{X}_t)$ and $P(X_i, \Theta, \bar{X}_t)$. For the third term on the right of (3), recall that the Γ_i's are bounded vectors. Also, for Θ in a sufficiently small neighborhood of Θ_t^*, we have $\sum_{X_i \in t} |H(Y_i|W_i)| \leq \sum_{X_i \in t} K(Y_i)$ in view of Condition 4. It now follows from Lemmas 1 and 2 and some of the arguments used in their proofs that there exists $\rho_3 > 0$ such that whenever $\rho \leq \rho_3$, we must have

$$\Pr\left\{\bigcap_{t \in T_n} E_t(\rho)\right\} \to 1 \text{ as } n \to \infty.$$

The proof of the theorem is now complete.

References

Ahn, H. and Loh, W.-Y. (1994). Tree-structured proportional hazards regression modeling. *Biometrics*. To appear.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading.

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

Buja, A., Duffy, D., Hastie, T. and Tibshirani, R. (1991). Comment on "Multivariate adaptive regression splines", *Annals of Statistics* **19**: 93-99.

Chaudhuri, P., Huang, M.-C., Loh, W.-Y. and Yao, R. (1993). Piecewise-polynomial regression trees. *Statistica Sinica*. To appear.

Ciampi, A. and Thiffault, J. (1989). Pruning regression trees for censored survival data: The RECPAM approach. *Communications in Statistics, Part A* **18**: 3373–3388.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatter plots. *Journal of the American Statistical Association* **74**: 829–836.

Cox, D. D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Annals of Statistics* **18**: 1676–1695.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability* **6**: 899–929. Corr: **7**, 909–911.

Fowlkes, E. B. (1987). Some diagnostics for binary logistic regression via smoothing. *Biometrika* **74**: 503–515.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Annals of Statistics* **19**: 1–67.

Gu, C. (1990). Adaptive spline smoothing in non-Gaussian regression models. *Journal of the American Statistical Association* **85**: 801–807.

Hastie, T. J. and Tibshirani, R. J. (1986). Generalized additive models (with discussion). *Statistical Science* **1**: 297–310.

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

Huang, M.-C. (1989). *Piecewise-linear Tree-structured Regression*. PhD thesis, University of Wisconsin, Madison.

Levene, H. (1960). Robust tests for equality of variances. *in et al.* Olkin, I. (ed.), *Contributions to Probability and Statistics*, Stanford University Press, Stanford, pp. 278–292.

Lo, W.-D. (1993). *Logistic Regression Trees*. PhD thesis, University of Wisconsin, Madison.

Loh, W.-Y. and Vanichsetakul, N. (1988). Tree-structured classification via generalized discriminant analysis (with discussion). *Journal of the American Statistical Association* **83**: 715–728.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Chapman and Hall, London.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**: 370–384.

O'Sullivan, F., Yandell, B. S. and Raynor, W. J. J. (1986). Automatic smoothing of regression functions in generalized linear models, *Journal of the American Statistical Association* **81**: 96–103.

Pierce, D. A. and Schafer, D. W. (1986). Residuals in generalized linear models, *Journal of the American Statistical Association* **81**: 977–986.

Sonquist, J. N. (1970). Multivariate model building, *Technical report*, Institute for Social Research, University of Michigan.

Sonquist, J. N., Baker, E. L. and Morgan, J. A. (1973). Searching for structure, *Technical report*, Institute for Social Research, University of Michigan.

Staniswalis, J. G. (1989). The kernel estimate of a regression function in likelihood-based models, *Journal of the American Statistical Association* **84**: 276–283. Corr: **85**, 1182.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models, *Annals of Statistics* **14**: 590–606.

Stone, C. J. (1991a). Asymptotics for doubly flexible log-spline response models, *Annals of Statistics* **19**: 1832–1854.

Stone, C. J. (1991b). Comment on "Multivariate adaptive regression splines", *Annals of Statistics* **19**: 113–115.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities, *Theory of Probability and Its Applications* **16**: 264–280.

Yang, C.-C. (1993). *Tree-structured Poisson Regression*, PhD thesis, University of Wisconsin, Madison.