Lecture Notes in Mathematics

Edited by A. Dold and B. Eckmann

757

Smoothing Techniques for Curve Estimation

Proceedings, Heidelberg 1979

Edited by
Th. Gasser and M. Rosenblatt

Springer-Verlag
Berlin Heidelberg New York

Lecture Notes in Mathematics

Edited by A. Dold and B. Eckmann

757

Smoothing Techniques for Curve Estimation

Proceedings of a Workshop held in Heidelberg,
April 2 – 4, 1979

Edited by
Th. Gasser and M. Rosenblatt

Springer-Verlag
Berlin Heidelberg New York 1979

Editors

Th. Gasser
Dept. of Biostatistics
Central Institute of Mental Health
J 5, P.O. Box 5970
D-6800 Mannheim

M. Rosenblatt
Dept. of Mathematics
University of California
San Diego, La Jolla
California 92032
USA

AMS Subject Classifications (1970): 62G05, 62G20, 65D07, 65D10

ISBN 3-540-09706-6 Springer-Verlag Berlin Heidelberg New York
ISBN 0-387-09706-6 Springer-Verlag New York Heidelberg Berlin

Library of Congress Cataloging in Publication Data
Main entry under title:
Smoothing techniques for curve estimation.
(Lecture notes in mathematics; 757)
"The workshop . . . has taken place as part of the activities of the
Sonderforschungsbereich 123,
"Stochastic Mathematical Models."
Bibliography: p.
Includes index.
1. Estimation theory--Congresses. 2. Curve fitting--Congresses. I. Gasser, Theodor A., 1941-
II. Rosenblatt, Murray. III. Series: Lecture notes in mathematics (Berlin); 757.
QA3.L28 no. 757 [QA276.8] 510'.8s [519.5'4] 79-22814
ISBN 0-387-09706-6

Printed in Germany

Printing and binding: Beltz Offsetdruck, Hemsbach/Bergstr.
2141/3140-543210

P R E F A C E

The workshop 'Smoothing Techniques for Curve Estimation' has taken place as part of the activities of the Sonderforschungsbereich 123, "Stochastic Mathematical Models". The participants and the organizer agree that it was a lively and successful meeting. Our thanks go to the Deutsche Forschungsgemeinschaft for enabling this meeting as part of the visiting program of the Sonderforschungsbereich 123. Our hope is that the ties founded or strengthened during the meeting will continue to be fruitful.

Heidelberg, July 1979

TABLE OF CONTENTS

NONPARAMETRIC CURVE ESTIMATION:

Some Introductory Remarks

Th. Gasser
Zentralinstitut für Seelische Gesundheit
Abteilung Biostatistik
Postfach 5970
D-6800 Mannheim 1

M. Rosenblatt
University of California, San Diego
La Jolla, California 92032/USA

The workshop on smoothing techniques for curve estimation was organized because of the increasing theoretical and applied interest in such questions. Making a histogram of a data set is a time-honored tool in statistics as well as in other areas. The notion of representing and smoothing data in more flexible ways arises naturally. Given any way of representing a function, one can adapt such a representation to give a method of smoothing data. In this way one obtains attractive alternatives to a parametric analysis. There is interest in nonparametric curve estimation because parametric models require assumptions which are often unwarranted and not checked when entering a new field in the empirical sciences. The availability of computer equipment, especially with graphics terminals,allow one to avoid undue assumptions and "let the data speak for themselves".

We first make some historical remarks. One of the earliest papers to suggest using kernel estimates of density functions was that of Rosenblatt (1956). A few years later on further results on the large sample behavior of such estimates were obtained in Bartlett (1963) and Parzen (1962). Estimates making use of a Fourier representation were suggested in a paper of Censov (1962). These techniques have been used by Tarter and Raman (1972) in a biomedical context. Nearest neighbor estimates have been proposed for density estimation in Loftsgaarden and Quesenberry (1965), and for regression analysis in Stone (1977). In a recent paper of Mack and Rosenblatt (1979), the large sample behavior of nearest neighbor density estimates is determined. Spline methods have also been of considerable interest and were proposed as a basis for density estimation in Boneva, Kendall and Stefanov (1971). The large sample pro-

perties of cubic spline density estimates are discussed in the paper of Lii and Rosenblatt (1975). Both spline and kernel estimates are used to estimate the probability density of the derivative of turbulent velocity readings. This is of interest in a modified model of Kolmogorov where it is suggested that this probability density should be approximately log normal. This is not consistent with the analysis out in the tails (say beyond three sigma) of the distribution. Spline estimates have been used in an analysis arising in an archaeological context in Kendall (1974). The use of nonparametric (say kernel) regression estimates to deal with the important problem of calibrating radiocarbon and bristlecone pine dates has been proposed in Clark (1974). The use of kernel estimates for discriminant analysis in a multidimensional context has been examined by Van Ness and Simpson (1976) and evaluated favorably.

One of the important questions when dealing with such nonparametric estimates is that relating to the choice of bandwidth or the degree of smoothing. Papers of Rosenblatt (1971) and Bickel and Rosenblatt (1973) have considered global behavior and global measures of deviation of kernel density estimates. An elegant and powerful result of Komlos et al (1975) is very useful in obtaining such results. Silverman (1978, 1979) has used related ideas and suggested an interesting way of choosing a bandwidth based on them. A number of applications are discussed in Silverman (1979).

We should now like to mention another area in which nonparametric curve estimates have been useful, that of growth and development. The analysis of the Zürich longitudinal growth study (Stützle, 1977; Largo et al 1978, Stützle et al 1979) is a case in point. The analysis started with the classical problem of height growth, and in particular of the pubertal growth spurt. Polynomial models were recognized as unsuitable. The logistic and the Gompertz functions have been used and compared (Marubini et al, 1971), but do not fit well over the whole period of growth. This limits the span of interpretation seriously. Bock et al (1973) have proposed a double logistic model between 1 and 18 years of age (addition of two logistic functions, associated with pubertal and prepubertal growth respectively). The fit is not good, and there is an age-dependent bias; the model has also led to qualitatively dissatisfying 'facts': It suggests that the difference between boys and girls resides primarily in the prepubertal parameters, contrary to everyday experience, and that the pubertal component starts in early childhood. Preece and Baines

(1978) recently introduced a parametric family which gives a better fit, as measured by the residual sum of squares.

Smoothing procedures offer an alternative for obtaining a set of interpretative parameters. Tanner et al (1966) carried out smoothing by eye, a procedure which is time-consuming, not reproducible and biased. (Tanner et al, 1966) had a too accentuated pubertal spurt). The bias of spline or kernel smoothing depends in a simple way on the function to be estimated. In the Zürich growth study, cubic smoothing splines (following Reinsch, 1967) have been used (Largo et al, 1978). The choice of the smoothing parameter is critical and should ideally be determined from the data. A cross-validation procedure suggested by Wahba and Wold (1975) gave in general good results. A general feature of splines smoothing is the relatively high cost in computer time and/or core (particularly annoying with large data sets encountered in neurophysiology). This draws our attention to alternatives, as e.g. kernel estimates.

REFERENCES:

Bartlett, M.S. (1963): Statistical estimation of density functions. Sankhya Sec. A 25 245-254

Boneva, L.I., Kendall, D.G., and Stefanov, I. (1971): Spline transformations, J. Roy. Statist. Soc. B. 33, 1-70

Bickel, P.J. and Rosenblatt (1973): On some global measures of the deviations of density function estimates. Ann. Statist. 1, 1071-95

Bock, R.D., Wainer, H. Petersen, A., Thissen, D., Murray, J., Roche, A. (1973): A parametrization for individual human growth curves, Human Biology 45, 63-80

Censov, N.N. (1962): Evaluation of an unknown distribution density from observations. Soviet Math. 3, 1559-1562

Clark, R.M. (1974): A survey of statistical problems in archaeological dating. J. Multiv. Anal. 4, 308-326

Kendall, D.G. (1974): Hunting quanta. Phil. Trans. Roy. Soc. London, A 276, 231-266

Komlos, J., Major, P. and Tusnady, G. (1975): An approximation of partial sums of independent random variables. Zeit. für Wahr. 32, 111-131

Largo, R.H., Stützle, W., Gasser, T., Huber, P.J., Prader, A. (1978): A description of the adolescent growth spurt using smoothing spline functions. Annals of Human Biology, in print.

Lawton, W.H., Sylvestre, E.A., Maggio, M.S. (1972): Self modeling nonlinear regression. Technometrics 14, 513-532

Lii, K.S., Rosenblatt, M. (1975): Asymptotic behavior of a spline estimate of a density function. Comput. Math. Appl. 1, 223-235

Mack, Y.P. and Rosenblatt, M. (1979): Multivariate k-nearest neighbor density estimates. J. Multiv. Anal. 9, 1-15

Marubini, E., Resele, L.F., Barghini, G. (1971): A comparative fitting of gompertz and logistic functions to longitudinal height data during adolescense in girls. Human Biology 43, 237-252

Parzen, E. (1962): On the estimation of a probability density and mode. Ann. Math. Statist. 33, 1065-1076

Preece, M.A., Baines, M.J. (1978): A new family of mathematical models describing the human growth curve. Annals of Human Biology, 5, 1-24

Reinsch, Ch. (1967): Smoothing by spline functions. Num. Math. 10, 177-183

Rosenblatt, M. (1956): Remarks on some nonparametric estimates of a density function. Ann. Math. Statist. 27, 832-837

Rosenblatt, M. (1970): Density estimates and Markov sequences. In Nonparametric Techniques in Statistical Inference, M. Puri ed. 199-210

Rosenblatt, M. (1971): Curve estimates. Ann. Math. Statist. 42, 1815-1842

Silverman, B.W. (1978): Choosing a window width when estimating a density. Biometrika 65, 1-11

Silverman, B.W. (1979): Density estimation: are theoretical results useful in practice? - presented at a meeting in honor of W. Hoeffding

Stützle, W. (1977): Estimation and parametrization of growth curves. Thesis 6039 ETH Zürich

Stützle, W., Gasser, Th., Largo, R., Huber, P.J., Prader, A., Molinari, L. (1979): Shape-invariant modeling of human growth. Mimeographed manuscript, 1979

Tanner, J.M., Whitehouse, R.H., Takaishi, M. (1966): Standards from birth to maturity for height, weight, height velocity and weight velocity: British Children. Archives of Disease in Childhood 41, 451-471, 613-635

Tarter, M. and Raman, S. (1972): A systematic approach to graphical methods in biometry. Proceedings of 6th Berkeley Symposium vol. IV, 199-222

Van Ness, J.W. and Simpson, C. (1976): On the effects of dimension in discriminant analysis. Technometrics 18, 175-187

Wahba, G., Wold, S. (1975): A completely automatic French curve: Fitting spline functions by cross-validation. Communications in statistics 4, 1-17

A TREE-STRUCTURED APPROACH TO NONPARAMETRIC MULTIPLE REGRESSION

*Jerome H. Friedman**
Stanford Linear Accelerator Center
Stanford, California 94305/USA

Introduction

In the nonparametric regression problem, one is given a set of vector valued variables $\underline{X}$ (termed carriers) and with each an associated scalar quantity Y (termed the response). This set of carriers and associated responses $\{Y_i,\underline{X}_i\}$ $(1\leq i\leq N)$ is termed the training sample. In addition (usually at some later time), one is given another set of vector valued variables $\{\underline{Z}_j\}$ $(1\leq j\leq M)$ without corresponding responses and the problem is to estimate each corresponding response using the values of its carriers and the training sample. That is:

$$\hat{Y}(\underline{Z}_j) = \text{Rule}\ [\underline{Z}_j,\ \{Y_i,\underline{X}_i\}\ (1 \leq i \leq N)]\ (1 \leq j \leq M).$$

The rule for performing the estimation is usually referred to as the model or regression function.

In addition to this basic predictive role, there are usually other data analytic goals. One would like the model to reveal the nature of the dependence of the response on the respective carriers and lend itself to easy interpretation in a similar manner to the way parametric models often do via the fitted values of their parameters.

Binary Regression Tree

The nonparametric regression models discussed herein are based on binary trees. A binary tree is a rooted tree in which every node has either two sons (nonterminal nodes) or zero sons (terminal nodes). Figure 1 illustrates a simple binary tree.

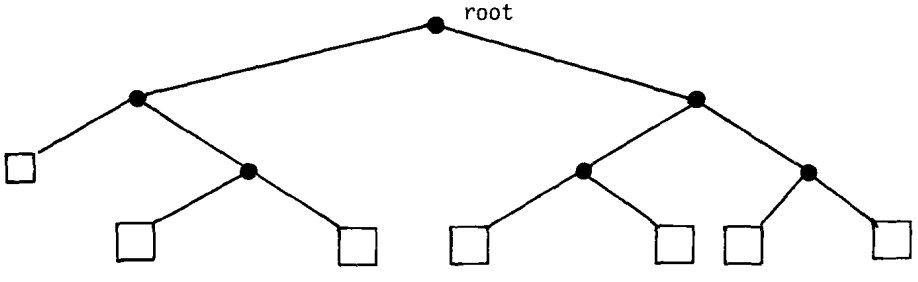

Figure 1

*This work is part of a joint research effort by Leo Breiman, Jerome Friedman, Lawrence Rafksy and Charles Stone. Work partially supported by the Department of Energy under contract number EY-76-C-03-0515.

For these models, each node t represents:

1) a subsample S_t of the training sample,
2) a subregion R_t of the carrier data space,
3) a linear model $L_t(\underline{X}) = \underline{A}_t \cdot \underline{X} + B_t$ to be applied to $\underline{X} \in R_t$.

(For the models discussed in this report, the subsample S_t, represented by node t, is just the set of training vectors that lie in its corresponding subregion R_t.)

In addition, each nonterminal node represents:

4) a partitioning or splitting of R_t into two disjoint subregions $R_{l(t)}$ and $R_{r(t)}$

$(R_{l(t)} \cup R_{r(t)} = R_t$ and $R_{l(t)} \cap R_{r(t)} = 0)$

and a corresponding partitioning of S_t into two disjoint subsets $S_{l(t)}$ and $S_{r(t)}$.

The binary regression tree is defined recursively: let t_o be the root node and

S_{t_o} = entire training sample

R_{t_o} = entire carrier data space

$L_{t_o}(\underline{X})$ = linear (least squares fit) of Y on $\underline{X}$ using S_{t_o}.

Let t be a nonterminal node with left and right sons l(t) and r(t) respectively. Then

$R_{l(t)}$ and $R_{r(t)}$ are the subregions defined by the partitioning of t,

$S_{l(t)}$ and $S_{r(t)}$ are the subsamples defined by the partitioning of t.

The linear models associated with the left and right sons are derived from the parent model by modifying the dependence on one of the carriers J_t:

$$L_{l(t)} = L_t + a_{l(t)} X(J_t) + b_{l(t)}$$
$$L_{r(t)} = L_t + a_{r(t)} X(J_t) + b_{r(t)}. \qquad (1)$$

To construct the model one then needs:

1) a training sample $\{Y_i, \underline{X}_i\}$ $(1 \le i \le N)$
[This allows the definition of the root node R_{t_o}, S_{t_o}, $L_{t_o}(\underline{X})$],

2) a splitting rule which consists of
   a) a prescription for partitioning R_t into $R_{l(t)}$ and $R_{r(t)}$ $(S_{l(t)}$ and $S_{r(t)})$,
   b) a prescription for updating the model (choosing values for J_t, $a_{l(t)}$, $a_{r(t)}$, $b_{l(t)}$, $b_{r(t)}$, to get $L_{l(t)}$ and $L_{r(t)}$ (thereby defining the two son nodes of t),

3) stopping (termination) rule for deciding when not to split a node, thereby making it a terminal node.

Splitting Rule

The situation at a node that is to be split is depicted in Figure 2.

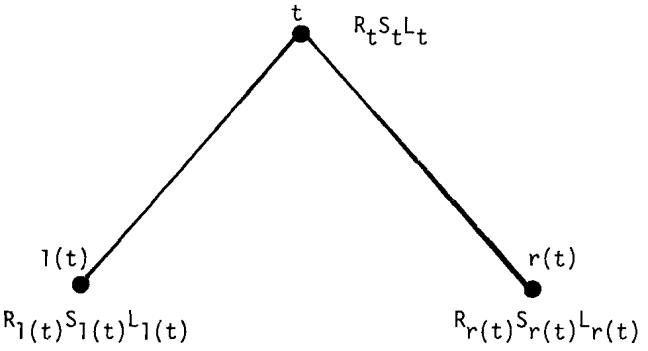

Figure 2

One has the subregion (subsample) and model associated with the parent [R_t (S_t) and L_t] and one would like to define the corresponding quantities for the two sons so as to best improve the fit of the model to the training sample. Let

$$\hat{Q}_t = \sum_{i \in S_t} [Y_i - L_t(\underline{X}_i)]^2 \tag{2}$$

be the empirical residual sum of squares associated with the parent and $\hat{Q}_{l(t)}$ and $\hat{Q}_{r(t)}$ be the corresponding quantities for the two sons. Then

$$\hat{I}_t = \hat{Q}_t - \hat{Q}_{l(t)} - \hat{Q}_{r(t)} \tag{3}$$

is an estimate of the improvement as a result of splitting node t. A reasonable goal is then to choose the partitioning so as to maximize $\hat{I}_t$ subject to possible limitations such as continuity and computability.

Since $L_{l(t)}$ and $L_{r(t)}$ are linear models (on $R_{l(t)}$ and $R_{r(t)}$) and $R_{l(t)} \cup R_{r(t)} = R_t$, one can think of $[L_{l(t)}, L_{r(t)}]$ as a piecewise-linear model on R_t. From (1)

$$\begin{aligned} L_{l(t)} - L_t &= a_{l(t)}\, X(J_t) + b_{l(t)} \\ L_{r(t)} - L_t &= a_{r(t)}\, X(J_t) + b_{r(t)} \end{aligned} \tag{4}$$

so that we want to choose the parameters on the RHS of (4) to best fit the residuals

$$r_i = Y_i - L_t(\underline{X}_i) \qquad (i \in S_i) \tag{5}$$

to the model associated with the parent node.

Consider the residuals (5) as a function of each of the carriers X(j) in turn. If $L_t(\underline{X})$ provides an adequate description of the dependence of the response on X(j), then there should be little structure in the values of the residuals when ordered on X(j). That is, a plot of r versus X(j) would resemble that of Figure 3a.

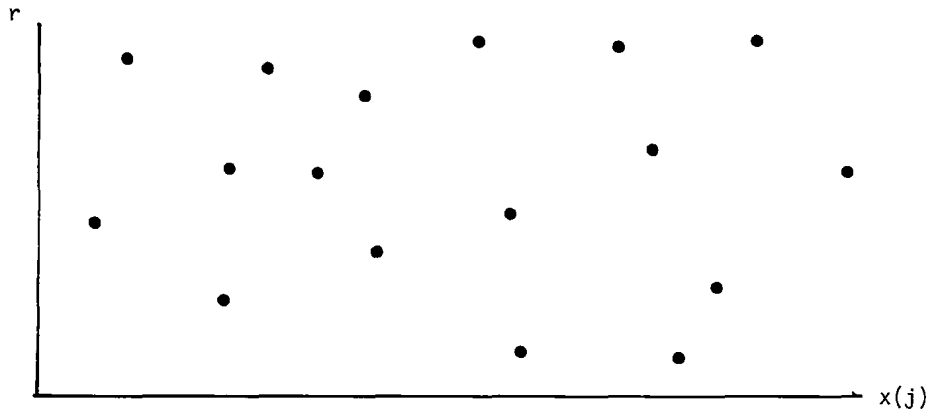

Figure 3a

On the other hand, considerable structure in the residuals (e.g., Figure 3b) would indicate that $L_t(\underline{X})$ does not provide an adequate description of the dependence of the response on X(j).

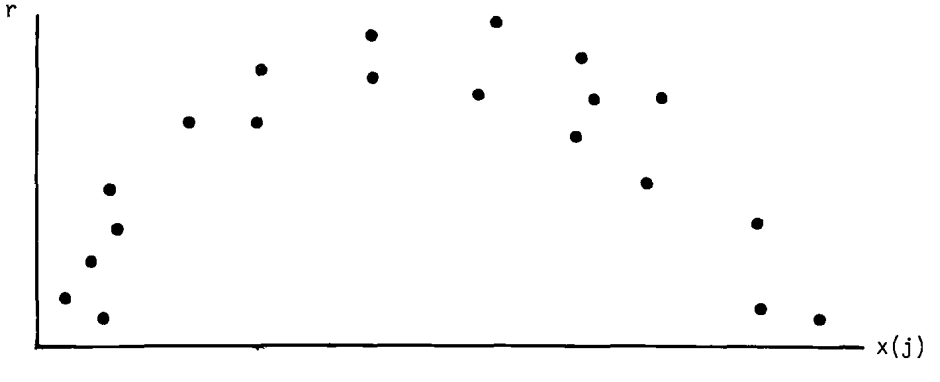

Figure 3b

The example of Figure 3b indicates a possible quadratic dependence of the residuals (and hence the response) on carrier X(j).

These observations motivate our splitting procedure. Each carrier X(j) is considered in turn. For each, a (univariate) continuous piecewise linear model is fit to the residuals from $L_t(\underline{X})$. That is, the model

$$\begin{aligned} r &= a_{lj}\,[X(j) - s_j] + b_j \qquad X(j) \le s_j \\ r &= a_{rj}\,[X(j) - s_j] + b_j \qquad X(j) > s_j \end{aligned} \tag{6}$$

is fit to $\{r_i,\ X_i(j)\}$ $(i \in S_t)$ by minimizing

$$\begin{aligned} Q_j = &\sum_{i=1}^{k} [r_i - a_{lj}\,(X_i(j) - s_j) - b_j]^2 \\ &+ \sum_{i=k+1}^{\#S_t} [r_i - a_{rj}\,(X_i(j) - s_j) - b_j]^2 \end{aligned} \tag{7}$$

with respect to j, a_{lj}, a_{rj}, b_j, and s_j. Here the $X_i(j)$ are ordered in ascending value and $X_k(j) \le s_j$ and $X_{k+1}(j) > s_j$. That is, the best (in the least squares sense) continuous piecewise linear fit (with s_j as the knot) is made to the residuals versus each carrier X(j) and the best fit (over the carriers) is chosen.

Let the optimum values found for j, a_{lj}, a_{rj}, b_j, s_j be represented by J, a_l, a_r, b and s respectively. These solution values are used to both define the partitioning and update the model:

$$\begin{aligned} &\text{For } \underline{X} \in R_t\text{:} \\ &\quad \text{If } X(J) \le s, \text{ then } \underline{X} \in R_{l(t)} \\ &\quad \text{If } X(j) > s, \text{ then } \underline{X} \in R_{r(t)} \\ &\quad L_{l(t)}\,(\underline{X}) = L_t(X) + a_l\,[X(J) - s] + b \\ &\quad L_{r(t)}\,(\underline{X}) = L_t(\underline{X}) + a_r\,[X(J) - s] + b\ . \end{aligned} \tag{8}$$

If the model associated with the parent node is

$$L_t\,(\underline{X}) = \sum_{j=1}^{p} A_t(j)\,X(j) + B_t$$

then from (8), the corresponding quantities for the son nodes are:

$$\begin{aligned} A_{l(t)}(j) &= A_{r(t)}(j) = A_t(j) \quad j \ne J \\ A_{l(t)}(J) &= A_t(J) + a_l \\ A_{r(t)}(J) &= A_t(J) + a_r \\ B_{l(t)} &= B_t - a_l\ s + b \\ B_{r(t)} &= B_t - a_r\ s + b. \end{aligned} \tag{9}$$

Thus, the models associated with the left and right sons differ from the parent and each other only in their dependence on carrier J, and the constant terms are adjusted for continuity at the split point s.

After the split is made and the model updated for the two son nodes, the above procedure is applied recursively to l(t) and r(t) and their sons and so on until the nodes meet a terminal condition. This stops the splitting making terminal nodes. Starting with the root, this recursive procedure then defines the entire regression tree.

Stopping (Termination) Rule

The recursive splitting described above cannot continue indefinitely. At some point, the cardinality of the subsample $\#(S_t)$ will be too small to reliably estimate the parameters for defining the splitting and updating the model. Thus, a sufficient condition for making a node terminal is that the size of its subsample is too small to continue splitting.

Using this condition as the sole one for termination, however, can cause serious overfitting. Basically, a split should not be made if it is not worthwhile. That is, it does not improve the model fit. The quantity $\hat{I}_t$ (3) is an estimate of the improvement in the fit as a result of splitting node t. This quantity is always positive, indicating that the empirical residual sum of squares will always improve as a result of choosing the optimum splitting. However, since the empirical residual sum of squares is an optimistically biased estimate of the true residual sum of squares from the model, a positive value for $\hat{I}_t$ does not guarantee a positive value for the true improvement I_t. A more reasonable criterion would be:

If $\hat{I}_t > k$ accept split at t and continue, otherwise make t a terminal node.

The quantity k is a parameter of the procedure, the interpretation of which is discussed below. Although lack of sufficient fit improvement (as estimated by $\hat{I}_t$) is a necessary condition for making t a terminal node, it is not sufficient. It is possible that a particular split, although not yielding much improvement itself, can make it possible for further splitting to make dramatic improvements. This would be the case, for example, if there were substantial interaction effects between pairs or sets of carriers. A sufficient condition for making a node terminal would be if its split and all further splits of its descendants yield insufficient empirical improvement. This is illustrated in Figure 4.

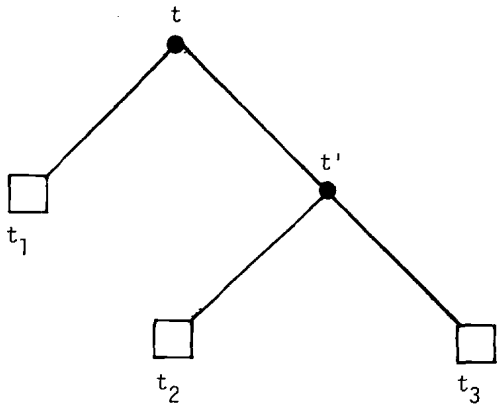

Figure 4

Here node t is split forming son nodes t_1 (which subsequently becomes terminal) and t'. Right son t' is further split forming nodes t_2 and t_3 which become terminal. The improvement associated with node t (and all further splits) is then defined to be

$$\hat{I}_t = \hat{Q}_t - \hat{Q}_{t_1} - \hat{Q}_{t_2} - \hat{Q}_{t_3} \tag{10}$$

That is the difference between the empirical residual sum of squares at node t and the sum of those associated with all terminal descendants of t. A reasonable condition for making t a terminal node is then

$$\text{If } \hat{I}_t \leq 2\ k \text{ make t terminal, otherwise accept split at t.} \tag{11}$$

The factor of two on the RHS of the inequality comes from the fact that two splits were required to form these three terminal nodes and this introduces even more optimistic bias than just one split. The condition (11) can be rewritten

$$\text{If } \hat{Q}_t + k \leq \hat{Q}_{t_1} + k + \hat{Q}_{t_2} + k + \hat{Q}_{t_3} + k \tag{12}$$

make t terminal,

Otherwise, accept split at t.

This suggests associating a cost C_t with each node t of the tree as follows:

$$\begin{aligned} &\text{If t is terminal} \quad C_t = \hat{Q}_t + k \\ &\text{If t is nonterminal} \quad C_t = \sum_{i \in t} C_{t_i} \end{aligned} \tag{13}$$

where the summation is over all terminal descendants of t. The decision to make a node terminal or not is then taken so as to minimize this cost. Note that if both sons of t [l(t) and r(t)] are terminated according to this prescription, then

$$\sum_{i \in t} C_{t_i} = C_{l(t)} + C_{r(t)} \,. \tag{14}$$

This suggests the following "bottom-up" recombination procedure for terminating the regression tree. First, the splitting procedure is applied as far as possible, terminating only for insufficient subsample cardinality. The nonterminal nodes of the resulting tree are then each considered in inverse order of depth. (The depth of a node is the number of nodes in the path from it to the root.) At each such node, the following termination rule is applied:

$$\underline{\text{If}}\ \hat{Q}_t + k \leq C_{l(t)} + C_{r(t)}$$

$$\underline{\text{then}}\ \text{make t terminal and } C_t = \hat{Q}_t + k \qquad (15)$$

$$\text{Otherwise accept split at t and } C_t = C_{l(t)} + C_{r(t)}.$$

This bottom-up recombination procedure insures that a node is made terminal only if its splitting and all possible further splitting yields insufficient improvement to the fit of the model, as determined by the improvement threshold parameter k.

This bottom-up recombination algorithm can be more easily understood intuitively by considering the following optimization problem. Let $\mathcal{J}$ be the set of all possible trees obtained by <u>arbitrarily</u> terminating the splitting procedure of the previous section. Let $T \in \mathcal{J}$ be one such tree and define its size $|T|$ to be the number of its terminal nodes. Let $\hat{Q}(T)$ be the empirical residual sum of squares associated with the regression model defined by T. The optimization problem is to choose that tree $T_k \in \mathcal{J}$, such that $\hat{Q}(T_k) + k|T|$ is minimum (breaking ties by minimizing $|T|$). The quantity k is a positive constant called the complexity parameter and T_k is said to be the optimally terminated tree for complexity parameter k. The complexity parameter is the analogue for this procedure to the smoothness parameter associated with smoothing splines or the bandwidth parameter associated with kernal estimates. Since $\hat{Q}(T_k)$ is monotone decreasing with increasing $|T_k|$, the value of k limits the size of the resulting optimally terminated tree T_k. Larger values of k result in smaller trees.

It can be shown (Breiman and Stone, 1977) that the bottom-up recombination procedure described above is an algorithm for solving this optimization problem where the complexity parameter k is just the improvement threshold parameter of that procedure. Thus, although motivated heuristically, that procedure is seen to have a natural interpretation in terms of generating optimally terminated trees T_k.

The complexity parameter k is the only parameter associated with this model. Ideally, its value should be chosen to minimize the true residual sum of squares $Q(T_k)$ associated with the model. This quantity is, of course, unavailable since only the training sample is provided. One could apply crossvalidation (e.g., see Breiman and Stone, 1977) or bootstrapping (Efron, 1977) techniques to obtain a less biased estimate of $Q(T_k)$ than $\hat{Q}(T_k)$. These estimates could be performed for various values of k and the best one chosen based on those estimates. However, this procedure is quite expensive computationally and not always reliable. Fortunately, a simple graphical procedure