Tree-Structured Methods for Longitudinal Data

Author(s): Mark Robert Segal

Source: *Journal of the American Statistical Association*, Jun., 1992, Vol. 87, No. 418 (Jun., 1992), pp. 407-418

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2290271

Tree-Structured Methods for Longitudinal Data

MARK ROBERT SEGAL*

The thrust of tree techniques is the extraction of meaningful subgroups characterized by common covariate values and homogeneous outcome. For longitudinal data, this homogeneity can pertain to the mean and/or to covariance structure. The regression tree methodology is extended to repeated measures and longitudinal data by modifying the split function so as to accommodate multiple responses. Several split functions are developed based either on deviations around subgroup mean vectors or on two sample statistics measuring subgroup separation. For the methods to be computationally feasible, it is necessary to devise updating algorithms for the split function. This has been done for some commonly used covariance specifications: independence, compound symmetry, and first-order autoregressive models. Data analytic issues, such as handling missing values and time-varying covariates and determining appropriate tree size are discussed. An illustrative example concerning immune function loss in a cohort of human immunodeficiency virus (HIV)-seropositive gay men also is presented.

KEY WORDS: Covariance structure; Human immune virus (HIV); Missing values; Multiple response; Regression tree.

The covariation induced by making several observations of some continuous response on the same unit, as occurs with repeated measures designs, cluster designs and longitudinal studies, poses data analytic problems. Analysis of such designs that ignore the covariance structure are known to produce incorrect variance estimates. In the regression setting, where some continuous and repeatedly measured response is related to covariates of interest, these problems often are redressed by straightforward generalization of conventional (single outcome) methods, treating the covariation as nuisance. Thus, ordinary least squares become generalized least squares, and, for example, the generalized linear model framework (McCullagh and Nelder 1983) has been extended to account for correlated outcomes by Liang and Zeger (1986).

In this article I extend the tree-structured regression paradigm (Breiman, Friedman, Olshen, and Stone 1984) to such multiple response settings and in particular to longitudinal data. Tree techniques have proved successful in extracting meaningful subgroups in other contexts (e.g., regression, classification, and survival). Hence, if one analytic goal is identifying strata with common covariate values and homogeneous *multiple* outcomes, then trees will be potentially useful.

To make things more concrete, I first present some discussion and analysis surrounding a celebrated growth-curve dataset taken from the recent statistical literature. Additional motivation derives from existing research questions posed in the study of immune function loss among human immunodeficiency virus (HIV)-seropositive patients. These questions are not adequately handled by standard techniques. An illustrative analysis for such a seropositive cohort is presented in Section 3.

The first data set is dental measurements of 11 girls and 16 boys; the analyses appear in Lee (1988) and Rao (1987). Although both papers are concerned primarily with prediction, they do raise the issue of heterogeneity of outcome across strata, in terms of both mean and covariance. In his discussion of the paper by Rao, Draper (1987) presents gender-specific plots of the raw growth curves (see Figures 1 and 2) and observes the following: ". . . the boys exhibit systematically different departures from global linearity than the girls do, and display nonmonotonicities over time at a rate about twice that of the girls. In addition, the two groups exhibit different variability around their basic growth curve shapes, suggesting that both the choice of growth curve families and the cross-validation estimates of prediction error should be stratified on sex" (Draper 1987, p. 457). In his rejoinder, Rao asserts: "Of course, if clusters could be identified as with the dental data on boys and girls, they should be treated separately" (Rao 1987, p. 468). This same dataset is also analyzed by Lee (1988), who employs a first-order autoregressive structure for *strata-specific covariances* and, on the basis of the differences between Figures 3 and 4, which display likelihoods for the autoregressive parameters, suggests strata-specific (i.e., sex-specific) analyses.

Two remarks are in order. First, Rao's qualifier "*could be identified*" [italics supplied by this author] (Rao 1987, p. 468) underscores one objective of this article: the identification of clusters or subgroups of growth curves. Such a determination was possible with the dental data because of the study's limited extent; there was only one binary covariate (gender) and a total sample size of only 27. This facilitated visual examination of all individual growth curves. While such graphical appraisal ought be a precursor to modeling, it will not always suffice for subgroup identification. If there are large samples and hence many curves, and/or if there is high variability, then assessing such plots will be problematic. Further, if there are multiple covariates, especially continuous or multilevel categorical, then predetermining which strata to examine becomes difficult. All these concerns apply to the HIV example. Morrison (1976, p. 153) acknowledges the complexity of attempting stratification purely by inspecting the growth curve profiles and asserts that existing methods for determining profile groupings have not been adequately developed. This situation does not seem to have changed since 1976.

Second, the heterogeneity between clusters can manifest itself in two ways: (a) on the mean structure, and (b) on the

* Mark Robert Segal is Assistant Professor, Division of Biostatistics, University of California, San Francisco, CA 94143-0560. This work was supported by NIH Grants R29-GM45543 and R01-AI29162. The author thanks Andrew Moss and Steve Shiboski, who provided data from the San Francisco General Hospital Cohort, and two referees, the associate editor, and Peter Bacchetti, who contributed helpful comments.

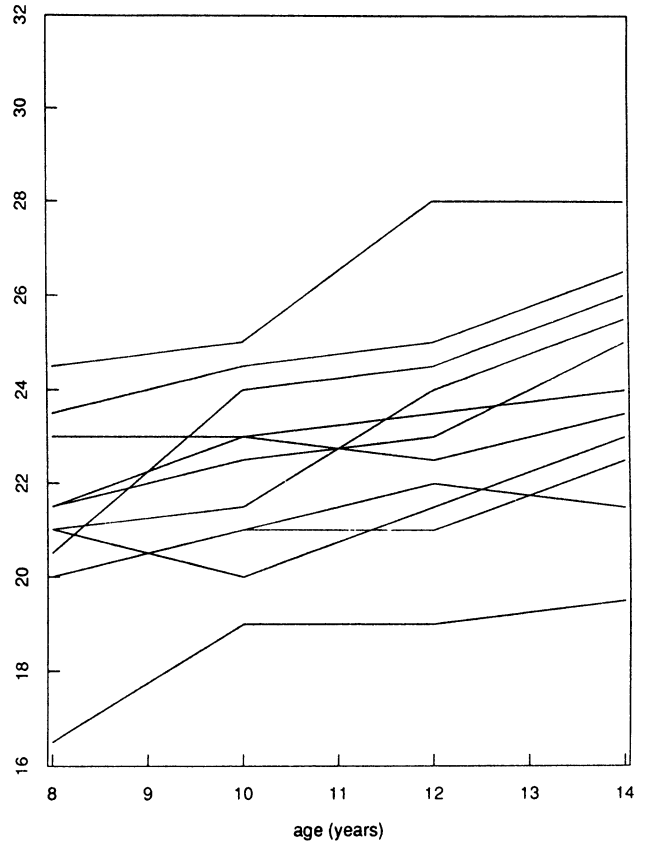

Figure 1. Superimposed Individual Growth Curves for the 11 Girls From the Dental Measurements Data Set.

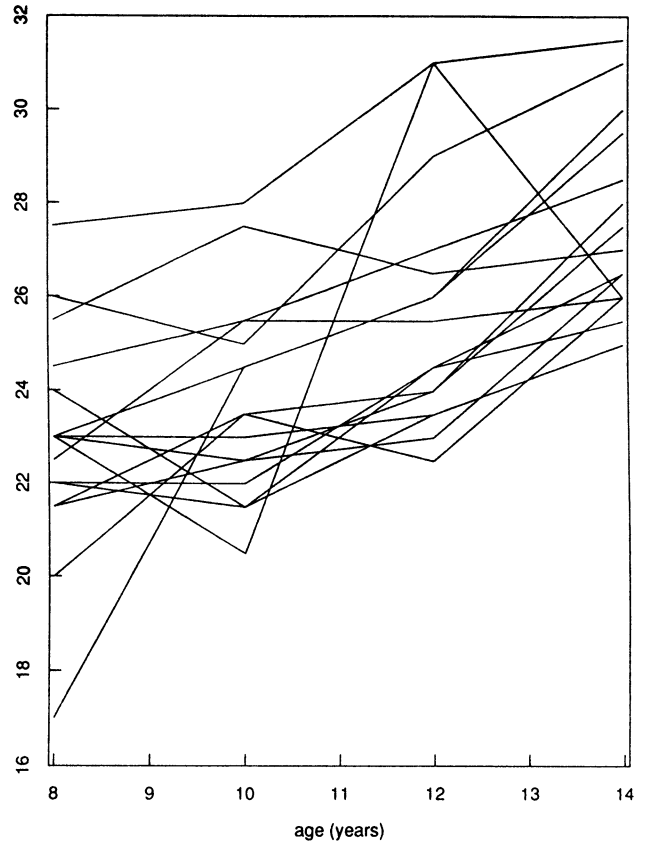

Figure 2. Superimposed Individual Growth Curves for the 16 boys From the Dental Measurements Data Set.

covariance structure. As an example of (a), the generally higher levels and greater slopes for the boys than for the girls suggest different parameter values or summaries for growth. With regard to (b), the seemingly greater within individual variability for the boys and the distinct likelihood functions and maximum likelihood estimators (MLE's) suggest that differing covariance parameters be used in the two strata. Several statistical papers have invoked the need for such strata-specific covariance structures, including Ware and Wu (1981, p. 433), Laird, Lange, and Stram (1987, Sec. 2.3); Schlucter (1988, Sec. 3.3); Lange and Laird (1989, Sec. 5.2); and DeGruttola, Lange, and Dafni (1991). Yet no methodology exists for determining if and when such strata-specific modeling should be performed.

The problem here is intermediary between classification and clustering; see Gnanadesikan (1977) for a similar example and discussion. It cannot be treated as a classification problem, because we do not know (in advance) what the classes are. Likewise, we are not contending with a pure clustering problem, because we have both response *and* covariate measurements. One proposal (C. R. Rao, personal communication, 1988) is to use some clustering algorithm on the multiple responses and then informally examine how covariates distribute within and between these clusters. The regression tree methodology affords a more algorithmic approach that produces interpretable subgroups; see Section 4.

A related area is that of *tracking,* a term used by epidemiologists to designate individuals whose growth curves conform to some normative population curve; see Ware and Wu (1981). The tree paradigm will determine whether there are subgroups of individuals (defined by common covariate values) who track differently. The extraction of such subgroups is commonly sought in risk and medical studies and is not facilitated by standard regression approaches, as is illustrated in Section 4 and in Segal and Bloch (1989).

Section 1 gives a brief overview of how regression trees are developed in the single-outcome setting and the modifications necessary to extend to multiple response. Section 2 deals with some important implementation issues: (a)

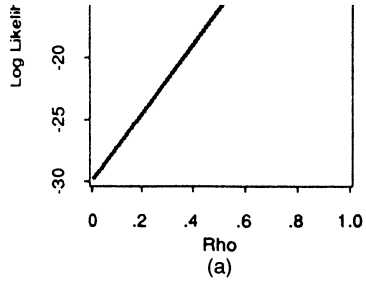

Figure 3. Autoregressive Likelihood for Girls.

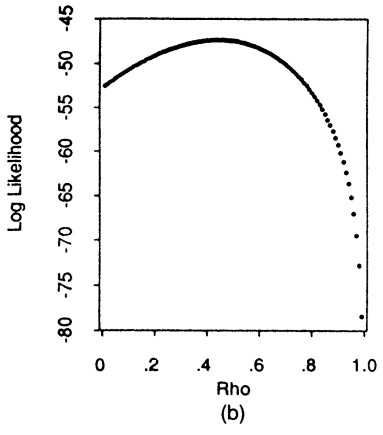

Figure 4. Autoregressive Likelihood for Boys.

specifying reasonable covariance families, (b) developing updating algorithms for the resultant split functions (c) handling missing values for both covariates and responses, (d) determining appropriate tree size, and (e) handling time-varying covariates. Section 3 features a tree-structured analysis of an HIV-seropositive cohort. Finally, Section 4 details some of the strengths and limitations of the regression tree approach.

1. REGRESSION TREE METHODOLOGY

1.1 Univariate Response

The definitive reference for tree techniques is "Classification and Regression Trees" by Breiman, Friedman, Olshen, and Stone (1984), hereinafter referred to as CART. Some familiarity with both tree techniques and the associated terminology is assumed. This section describes the required background so that the modifications needed to facilitate the use of tree methods with longitudinal data can be understood.

Attention is restricted to the familiar regression setting with p predictors $X_1, X_2, \ldots, X_p$ and continuous response Y. I assume that complete data are available on all subjects; CART Chapter 5 describes how to handle missing data. A regression tree is grown as follows. For each subgroup or *node;*

1. Examine every allowable split on each predictor variable.
2. Select and execute (create left and right daughter nodes) the *best* of these splits.

Steps 1 and 2 are then reapplied to each of the daughter nodes, and so on. The initial or *root* node comprises the entire sample. What constitutes an allowable split in Step 1 is defined in CART Chapter 2. Briefly, the covariates are examined univariately, for ordered covariates, an allowable split is into two subsamples (nodes) such that the covariate values in one node are all greater than those in the complementary node. The allowable splits therefore preserve ordering. For unordered categorical variables, any split into two disjoint subsets of the categories is permitted. "Best" in Step 2 is assessed in terms of the *split function* $\phi(s, g)$ that can be evaluated for any split s of any node g. Two such split functions are espoused in CART: least squares (LS) and least absolute deviations (LAD). The LS split function is made explicit here so that subsequent reformulations can be referenced.

Let g designate a node of the tree. That is, g contains a subsample $\{(x_i', y_i)\}$, where $x_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$. Let N_g be the total number of cases in g and let $\bar{y}(g) = (1/N_g) \times \sum_{i \in g} y_i$ be the response average for node g. Then, the within node sum-of-squares is given by $SS(g) = \sum_{i \in g} (y_i - \bar{y}(g))^2$. Now suppose a split s partitions g into left and right daughter nodes g_L and g_R. The LS split function is $\phi(s, g) = SS(g) - SS(g_L) - SS(g_R)$, and the best split s^* of g is the split such that $\phi(s^*, g) = \max_{s \in \Omega} \phi(s, g)$, where Ω is the set of all allowable splits s of g.

A LS regression tree is constructed by recursively splitting nodes so as to maximize the above ϕ function. The function is such that we create smaller and smaller nodes of progressively increased homogeneity on account of the nonnegativity of ϕ: $\phi \geq 0$ since $SS(g) \geq SS(g_L) + SS(g_R)\ \forall s$. This nonnegativity also holds for least absolute deviations and is an essential property of a split function.

1.2 Multiple Response

We now consider the situation in which, besides the vector of covariates, each individual has a $T \times 1$ vector of responses $y_i' = (y_{i1}, y_{i2}, \ldots, y_{iT})$. Initially, we assume that neither the response nor covariate vector has any missing components and that the data are equally spaced. Departures from these two assumptions are discussed in Sections 2.3 and 2.6, respectively. Define $V(\theta, g)$ to be the model covariance matrix of the responses for node g depending on unknown parameters θ. Allowing $\hat{\theta}$ to be the $T(T+1)/2$ sample covariances s_{jk} enables us to proceed without making any assumptions on the covariance structure. However, both efficiency and interpretation gains can be made by restricting the dimension of θ. Instability resulting from overparameterizing covariance matrices is well known; see Diggle (1988). Therefore, we will always use low dimensional θ. For notational clarity, we will suppress the dependence of V on g and θ when there is no ambiguity. Let $\mu(g)$ be the $T \times 1$ vector of response means for individuals within a given node g.

The split function ϕ must be modified to handle multiple response data. There are two sorts of split functions: those focusing on the mean structure with the covariance as nuisance and those in which the covariance structure is of primary interest.

1.2.1 Mean Structure. An immediate generalization of the least squares split function for the single outcome case given above is obtained by replacing $SS(g)$ with

$$SS(g) = \sum_{i \in g} (y_i - \mu(g))' V(\theta, g)^{-1} (y_i - \mu(g)). \quad (1)$$

Then, the split function ϕ_m for evaluating a split s of g into g_L and g_R is as before:

$$\phi_m(s, g) = SS(g) - SS(g_L) - SS(g_R). \quad (2)$$

As written, this function allows for a different covariance matrix for each of g, g_L, and g_R because the parameter estimates $\hat{\theta}$, $\hat{\theta}_L$, $\hat{\theta}_R$ can differ. It may seem desirable to "update"

the covariance parameters in this fashion, thereby allowing for a more locally determined covariance matrix. But in this formulation, ϕ_m could be negative. Hence, maximizing ϕ_m would not necessarily be improving homogeneity. Thus, we impose that for each candidate split the covariance parameters be determined from the parent node g so that

$$V(\theta, g) = V(\theta_L, g_L) = V(\theta_R, g_R), \tag{3}$$

and only—as part of the splitting process—update the mean function. This ensures that $\phi_m \geq 0$. Additionally, the computation is reduced. Of course, having determined and implemented the best split, the resultant daughter nodes become the new parent nodes and the covariance parameters are reestimated for each.

As noted in Segal (1988), any two-sample statistic provides a split function that optimizes between node separation rather than within node homogeneity. A possibility here is Hotelling's T^2:

$$T^2 = \frac{N_L N_R}{(N_L + N_R)} (\mu(g_L) - \mu(g_R))' S^{-1} (\mu(g_L) - \mu(g_R)), \tag{4}$$

where N_L and N_R are the sample sizes for the left and right daughter nodes, respectively, and S is the pooled sample covariance matrix. However, ϕ_m and T^2 differ only in the covariance structure used; the same split will maximize both if we take $\theta = s_{jk}$ i.e. $V = S$. The utility of using two-sample statistics in the single-outcome setting derives from the use of two-sample linear *rank* statistics. Such rank-based statistics are not available in higher dimensions.

1.2.2 Covariance Structure. As illustrated in Figure 2, heterogeneity in longitudinal data also can affect covariances. Recent papers dealing with the generalized linear model framework with univariate outcome (e.g., Nelder and Pregibon 1987, Sec. 3.3) have modeled variance heteroscedasticity as a function of covariates. Here, we seek to elicit *covariance* heteroscedasticity as a function of covariates, again using the regression tree paradigm. Before prescribing a split function to measure covariance heterogeneity, it is necessary to account for the mean structure. Diggle (1988) advocates fitting flexible mean functions to avoid inducing spurious correlation in the residuals. One such flexible approach is multivariate adaptive regression splines (Friedman 1991). The split functions for detecting covariance heterogeneity are applied to residuals after accounting for the mean structure.

Analogous to within node measures of loss, we consider functions that assess how closely the sample covariance matrix conforms to the hypothesized covariance matrix. Here, conformity is measured via a matrix norm:

$$\begin{aligned}\phi_c(s, g) = \log(\|S(g) - V(\theta, g)\|) \\ - \log(\|S(g_L) - V(\theta_L, g_L)\|) \\ - \log(\|S(g_R) - V(\theta_R, g_R)\|).\end{aligned} \tag{5}$$

The preceding form is motivated by analogy with the likelihood ratio test for equality of covariance matrices; see below. For the dental data example, we have $g_L \equiv$ girls, $g_R \equiv$ boys, $V \equiv$ first-order autoregressive model, and $\hat{\theta}_L = .9$, as distinct from $\hat{\theta}_R = .5$. The matrix norm $\|\cdot\|$ can be selected in accordance with what constitutes a meaningful distance measure for the problem at hand. However, such appropriateness often is unclear. A common choice is squared Euclidean norm, which affords simple updating algorithms for several basic choices for V; see Section 2.2. An alternate loss function is presented in Loh (1991).

Akin to the usage of two-sample statistics as split functions, we can rework a likelihood ratio test for the equality of covariance matrices into a split function. The value of the likelihood ratio test for equality of two covariance matrices, maximized over all candidate splits, will select for the division that most separates in terms of covariance structure. Here, we assume two multivariate normal populations with zero mean vectors and covariance matrices given by $V(\theta_L, g_L)$ and $V(\theta_R, g_R)$. For the case where θ is unrestricted, let the unbiased sample covariance matrices be $S(g_L)$ and $S(g_R)$. Then, under the null hypothesis $H_0: V(g_L) = V(g_R)$, the pooled estimate of common covariance is $S = (N_L + N_R)^{-1}(N_L S(g_L) + N_R S(g_R))$. The likelihood ratio split function is

$$\begin{aligned}\phi_{LR}(s, g) = (N_L + N_R)\log|S| - N_L \log|S(g_L)| \\ - N_R \log|S(g_R)|.\end{aligned} \tag{6}$$

Maximizing ϕ_{LR} is equivalent to minimizing $N_L \log|S(g_L)| + N_R \log|S(g_R)|$, because the first term is constant for all splits s. Correction factors exist for improving the approximation to an asymptotic χ^2 distribution. These should be incorporated, not for testing purposes but because if the split functions are not comparable (due to differing degrees of departure from χ^2), then maximizing the corresponding criterion function is not meaningful. When θ is restricted to either first-order autoregressive or compound symmetry models, simplified forms for the determinants result. However, obtaining updating formula for maximum likelihood estimates of the parameters is now complicated, especially for the autoregressive case. Some updating and estimation strategies are given in Section 2.2.

2. IMPLEMENTATION ISSUES

2.1 Choosing V

Choosing V appropriately is an important and difficult problem. The issues here parallel those for metric selection in multivariate classification problems; see Gnanadesikan (1977, chap. 4). To the extent that the role of regression trees is exploratory, there is no need to lock into any one specification for either the split function ϕ or the covariance structure V. Indeed, trying a variety of specifications helps elucidate the sensitivity to particular model assumptions. Nevertheless, some guidelines and justifications are warranted.

The modeling of, or adjustment for, covariance heterogeneity via a *single* process (e.g., autoregressive) with varying (strata-specific) parameters is seemingly reasonable. The covariation of the responses is determined by some process

intrinsic to those responses. Proposing that this covariation varies across subgroups in conjunction with varying parameters of the process is plausible.

For most applications, there will not be any mechanistic basis for specifying what is an appropriate process. Inasmuch as first-order autoregressions (AR1) and compound symmetry (CS) models represent extremes in terms of autocorrelation and are simple single-parameter processes, they constitute natural candidates as choices for V; see Lee (1988). Note that the AR1 autocorrelation function decays geometrically, $\rho(t) = \gamma^t$, whereas the CS autocorrelation function is constant, $\rho(t) = \rho$. Another natural candidate is independence, with identity correlation matrix and no correlation parameters.

The need for parsimony with respect to correlation parameters already has been argued; see Section 1.2. A number of two-parameter families have been proposed that interpolate between the above extremes and afford better descriptions of the empiric autocorrelation function of the responses for many biological measurements. These include autocorrelation functions given by $\rho(t) = \rho + (1 - \rho)\gamma^t$ (Chi and Reinsel 1989; Lange, Little, and Taylor 1989) and $\rho(t) = \gamma^{t^\nu}$ (Diggle 1988; Muñoz, et al. 1990), as well as the more familiar ARMA(1, 1) and AR2 models. Algorithms for estimation and updating for these two-parameter covariance specifications are being developed.

In summary, choosing V ideally should be based on external (mechanistic) considerations. In the absence of such information, selection should not be limited to a single choice. Rather, a number of simple covariance models whose autocorrelation functions approximate the empiric autocorrelation function of the multiple responses should be tried.

2.2 Updating Algorithms

Tree-structured methods trade assumptions for computing. The process of determining what constitutes the best split for a given node requires an exhaustive search among the numerous allowable splits. If the split function is not easily evaluated, then the amount of computation involved can make for prohibitively long run times. What has facilitated computational feasibility for the univariate LS and LAD split functions is the availability of updating formulas. These allow the value of $\phi(s^{\text{new}}, g)$ to be simply computed from $\phi(s^{\text{old}}, g)$ and $y' = (y_1, y_2, \ldots, y_T)$, where s^{new} and s^{old} differ by having one observation (x', y') "switch" from g_R^{old} to g_L^{new}.

Updating formulas are obtained by substituting for V in terms of the model parameters, expanding the quadratic form, and dropping terms that are constant over the differing allowable splits. This can be attempted for any covariance specification. The extent of the simplification will depend on the complexity of V^{-1}.

For example, consider an AR1 process, which has autocorrelation function $\rho(t) = \gamma^t$ and covariance matrix

$$V = (v_{jk}) = \frac{\sigma^2}{(1 - \gamma^2)}\gamma^{|j-k|}. \tag{7}$$

We can ignore the scale parameter σ^2 in maximizing ϕ_m, because it factors out. The $T \times T$ matrix V^{-1} is tri-diagonal, with sub- and super-diagonals equal to $-\gamma$. The main diagonal has entries $1 + \gamma^2$ apart from the $(1, 1)$ and (T, T) elements which are equal to 1.

Equivalent to maximizing ϕ_m as given in (2) for AR1 processes is minimizing

$$\phi_m^*(s, g) = N_L\left[\sum_{j=1}^{T}(\mu_j^L)^2 - 2\gamma\sum_{j=1}^{T-1}\mu_j^L\mu_{j+1}^L + \gamma^2\sum_{j=2}^{T-1}(\mu_j^L)^2\right] + N_R\left[\sum_{j=1}^{T}(\mu_j^R)^2 - 2\gamma\sum_{j=1}^{T-1}\mu_j^R\mu_{j+1}^R + \gamma^2\sum_{j=2}^{T-1}(\mu_j^R)^2\right], \tag{8}$$

where $\mu_j^L = \mu_j(g_L)$ and similarly for μ_j^R. This expression is easily updated, because the quantities N_L, N_R, μ_j^L, and μ_j^R all change in a straightforward fashion ($O(1)$ operations) as an observation (x', y') switches from g_R to g_L. Thus, for example, $N_L \to N_L + 1$ and $\mu_j^L \to (N_L + 1)^{-1}(N_L\mu_j^L + y_j)$. The parameter γ is estimated by regression of y_{ij} on $y_{i,j-1}$ for all observations in the parent node g; see Little and Rubin (1987, Sec. 8.6.2) and Section 2.3.2 in this article. In accordance with (3), γ is not updated as part of the splitting process.

An alternate specification for V is that of compound symmetry. This often is appropriate for repeated measures or clustered data designs in which a common correlation exists between all observations on a given individual or cluster: $V = (v_{jk}) = \rho\ \forall j \neq k$. The expression to be minimized for CS is

$$\begin{aligned}\phi_m^*(s, g) = N_L\Bigg[&\{1 - (\rho/(1 + (T - 1)\rho))\} \\ &\times \sum_{j=1}^{T}(\mu_j^L)^2 - (\rho/(1 + (T - 1)\rho)) \\ &\times \sum_{j\neq k}\mu_j^L\mu_k^L\Bigg] \\ + N_R\Bigg[&\{1 - (\rho/(1 + (T - 1)\rho))\} \\ &\times \sum_{j=1}^{T}(\mu_j^R)^2 - (\rho/(1 + (T - 1)\rho)) \\ &\times \sum_{j\neq k}\mu_j^R\mu_k^R\Bigg].\end{aligned} \tag{9}$$

As for the AR1 case, the scale parameter σ^2 factors out and the expression involves the same readily updated quantities. The parameter ρ, which can be interpreted as an intraclass correlation, can be estimated by maximum likelihood assuming multivariate normality or by method of moments. When each individual contributes the same number of observations or each cluster is of the same size, the methods produce identical estimates:

$$\hat{\rho} = \frac{1}{s^2}\frac{1}{T(T-1)}\sum_{j\neq k}s_{jk},$$

where s_{jk} is the sample correlation between the jth and kth measures and s^2 is the unbiased sample estimate of σ^2. In

the event of differing numbers of observations on individuals, the likelihood must be maximized numerically, and it is computationally simpler to use the method of moments estimator. Donner and Koval (1980) note that the moment estimator can perform poorly when the number of observations per individual are disparate and the true correlation is .3 or greater.

As required, on setting $\gamma = 0$ in (8) and $\rho = 0$ in (9) we obtain identical expressions that correspond to the independence case $V = I$.

Updating formulas also can be obtained for split functions based on covariance heterogeneity. For example, maximizing ϕ_c in (5) with squared Euclidean matrix norm and compound symmetry covariance structure is equivalent to minimizing:

$$\phi_c^*(s, g) = \log\left(\sum_{j\neq k} s_{jk}^2(g_L) - T(T-1)\hat{\rho}^2(g_L)\right) + \log\left(\sum_{j\neq k} s_{jk}^2(g_R) - T(T-1)\hat{\rho}^2(g_R)\right).$$

Here, $s_{jk}^2(\cdot)$ is the unbiased sample covariance between the jth and kth measures evaluated only over the argument node. Similarly, $\hat{\rho}(\cdot)$ is estimated within each node. The expression is easily updated by virtue of $O(1)$ updating formulas for the constituent covariances:

$$C^{\text{new}} = C^{\text{old}} + \frac{(\sum_{i\in g_L} y_{ij} - N_L y_i)(\sum_{i\in g_L} y_{ik} - N_L y_k)}{N_L(N_L+1)}$$

$$s_{jk}^2(g_L^{\text{new}}) = C^{\text{new}}/(N_L - 1).$$

Here C^{old} is initially zero and $g_L^{\text{new}} = g_L \cup (x', y')$. Similar downdating formulas can be derived for $s_{jk}^2(g_R^{\text{new}})$. Further, because $\hat{\rho}$ is just an average of the $s_{jk}(\cdot)$'s, no additional up(down)dating is required to compute $\hat{\rho}$.

To obtain updating formulas for AR1 and CS covariance models for the split function ϕ_{LR}, as given by (6), assume that the scale parameter is constant: $\sigma^2 = \sigma^2(g_L) = \sigma^2(g_R)$, so that it again factors out. Not assuming this would not appreciably affect the ease of updating, because all determinants would just have an extra factor $[\sigma^2(\cdot)]^T$.

For an AR1 model with autoregressive parameter γ and correlation matrix $V(\gamma)$, the value of the determinant is $|V(\gamma)| = (1-\gamma^2)^{T-1}$; so maximizing the split function ϕ_{LR} is equivalent to minimizing

$$\phi_{LR}^*(s, g) = N_L(T-1)\log(1-\gamma^2(g_L)) + N_R(T-1)\log(1-\gamma^2(g_R)).$$

This is feasible, because we can obtain node-specific estimates for $\gamma(\cdot)$ by least squares which, in turn, admit simple updating.

For a CS model with correlation parameter ρ and correlation matrix $V(\rho)$, the value of the determinant is $|V(\rho)| = (1-\rho)^{T-1}(1+(T-1)\rho)$; so maximizing the split function ϕ_{LR} is equivalent to minimizing

$$\phi_{LR}^*(s, g) = N_L[(T-1)\log(1-\rho(g_L)) + \log(1+(T-1)\rho(g_L))] + N_R[(T-1)\log(1-\rho(g_R)) + \log(1+(T-1)\rho(g_R))].$$

In this instance, we obtain node-specific estimates for $\rho(\cdot)$ using the covariance updating methods as outlined.

2.3 Missing Values

Missing data are commonplace in epidemiologic longitudinal studies. Studies spanning several years are fraught with attrition, and to obtain valid inferences attention must be paid to the mechanisms generating the missing data; see Laird (1988). Discarding cases with missing components can result in appreciable information loss. The following subsections illustrate how the tree methodology handles both missing covariates and missing response.

2.3.1 Missing Covariates. The manner in which tree methods for longitudinal data cope with missing covariate values coincides with that used for conventional (univariate outcome) regression trees and is detailed in CART Section 5.3. The main idea is that of a *surrogate* splitting variable. Designate the best split of a node g into g_L^* and g_R^* by s^*. Let s_m be a competing split (into g_L' and g_R') of g, but on a different covariate than the best split. The split $\tilde{s}_{m_1}$ is defined as the (first) surrogate split if $\tilde{s}_{m_1}$ maximizes $|g_L^* \cap g_L'| + |g_R^* \cap g_R'|$ over all competing splits s_m. Here, $|\cdot|$ denotes the size of the argument set. Thus, $\tilde{s}_{m_1}$ is the split that best reproduces the optimal split s^* but on a different covariate.

The missing value algorithm then works as follows. If there are missing values for certain cases on certain covariates, $x_m : m \in M \subseteq \{1, 2, \ldots, p\}$, determine the best split s_m^*, on x_m, using all cases containing a value of x_m. Then, s^* is selected as the best of the s_m^*'s *and* the s_n^*'s, where $x_n : n \in M^c$ are the covariates without any missing components for all cases. Now, in assigning a case to g_L^* or g_R^* that is missing the covariate on which s^* is based, use $\tilde{s}_{m_1}$, the first surrogate. That is, the case gets assigned to g_L^* if $\tilde{s}_{m_1}$ would have assigned it to g_L', and conversely for g_R^*. If the case is also missing x_{m_1}, then allocate according to $\tilde{s}_{m_2}$, the second surrogate, and so on. Note that the treatment of missing covariates does depend on the missing data mechanism.

2.3.2 Missing Response. Handling missing components of an individual's response vector depends on (a) the covariance structure V used to model associations between responses, (b) the pattern of missingness, and (c) the missing data mechanism. I present a method for accommodating missing responses that makes particular assumptions about each of these aspects and some comments about generalizing to other situations. Assume (a) an AR1 covariance structure, (b) a consecutive sequence of missing values, and (c) a missing completely at random (MCAR) mechanism (Laird 1988). To use the split function as given by (8), we need estimates for γ and the missing responses.

Under the above assumptions, such estimates can be obtained by generalizing the EM algorithm developed by Little

and Rubin (1987, Sec. 8.3). They formulate the AR1 model as

$$(y_t|y_1, \ldots, y_{t-1}, \alpha, \gamma, \sigma^2) \sim N(\alpha + \gamma y_{t-1}, \sigma^2), \quad t = 2, \ldots, T.$$

The complete-data log-likelihood for individual i, ignoring the contribution of the marginal distribution of y_{i1}, is equivalent to the log-likelihood for the simple linear regression of y_{it} on $y_{i,t-1}$. Thus, we have simple closed-form expressions for the MLE's of the parameters $\theta = (\alpha, \gamma, \sigma^2)$ as functions of the complete-data sufficient statistics:

$$\hat{\alpha} = (S_1 - \hat{\gamma}S_2)/NT,$$

$$\hat{\gamma} = (S_5 - (NT)^{-1}S_1S_2)/(S_4 - (NT)^{-1}S_2^2)$$

$$\hat{\sigma}^2 = (S_3 - (NT)^{-1}S_1^2 - \hat{\gamma}^2(S_4 - (NT)^{-1}S_2^2))/NT, \quad (10)$$

where the complete-data sufficient statistics $S = (S_1, S_2, S_3, S_4, S_5)$ are

$$S_1 = \sum y_{it}, \qquad S_2 = \sum y_{i,t-1}, \qquad S_3 = \sum y_{it}^2,$$
$$S_4 = \sum y_{i,t-1}^2, \qquad S_5 = \sum y_{it}Y_{i,t-1},$$

with all sums being double over individuals ($i = 1$ to N) and time ($t = 2$ to T). MLE's for θ are obtained using the EM algorithm as follows: Let $\theta^{(n)} = (\alpha^{(n)}, \gamma^{(n)}, \sigma^{2(n)})$ be the estimate of θ at iteration n. The M step computes $\theta^{(n+1)}$ from (10) using complete-data sufficient statistics S replaced by estimates $S^{(n)}$ from the E step. The E step computes $S^{(n)}$ as

$$S_1^{(n)} = \sum y_{it}^{(n)}, \qquad S_2^{(n)} = \sum y_{i,t-1}^{(n)}$$
$$S_3^{(n)} = \sum \{(y_{it}^{(n)})^2 + k_{it}^{(n)}\}$$
$$S_4^{(n)} = \sum \{(y_{i,t-1}^{(n)})^2 + k_{t-1,t-1}^{(n)}\}$$
$$S_5^{(n)} = \sum \{y_{it}^{(n)}y_{i,t-1}^{(n)} + k_{t,t-1}^{(n)}\},$$

where

$$y_{it}^{(n)} = y_{it}, \quad \text{if } y_{it} \text{ is present,}$$
$$= E(y_{it}|y_{\text{obs}}, \theta^{(n)}), \quad \text{if } y_{it} \text{ is missing.}$$
$$k_{tu}^{(n)} = 0, \quad \text{if } y_{it} \text{ or } y_{iu} \text{ is present,}$$
$$= \text{Cov}(y_{it}, y_{iu}|y_{\text{obs}}, \theta^{(n)}), \quad \text{if } y_{it} \text{ and } y_{iu} \text{ are missing.}$$

Here, y_{obs} designates all nonmissing data.

The E step computations (i.e., the determination of $E(y_{it}|y_{\text{obs}}, \theta^{(n)})$ and $\text{cov}(y_{it}, y_{iu}|y_{\text{obs}}, \theta^{(n)})$) can be achieved through the use of standard sweep operations (see Little and Rubin 1987, Sec. 6.5) on the covariance matrix of the observations. For the AR1 model under consideration, this matrix is $T \times T$ with (j, k) element as given in (7). Because this matrix can be large, Little and Rubin (1987) exploit stationarity properties of the AR1 model to simplify the E step. If $y_{i,\text{mis}} = (y_{i,j+1}, y_{i,j+2}, \ldots, y_{i,j+k-1})$ is a sequence of missing values bounded by present observations y_{ij} and $y_{i,j+k}$, then:

(†) $y_{i,\text{mis}}$ is independent of the other missing values given y_{obs} and θ; and

(‡) the distribution of y_{mis} given y_{obs} and θ depends on y_{obs} only through the bounding observations.

This distribution is multivariate normal, with covariance matrix and means that are weighted averages of $\mu = \alpha/(1 - \gamma)$, y_{ij}, and $y_{i,j+k}$. The weights and covariance matrix depend only on the number of missing values in the sequence, $k - 1$, and not on the position in time j.

Little and Rubin demonstrate the operations involved and the resultant expressions for a missing sequence of length 1. For arbitrary k, we have:

$$E(y_{i,j+r}|y_{\text{obs}}, \theta) = \mu + \frac{\gamma^r(1 - \gamma^{2(k-r)})}{1 - \gamma^{2k}}(y_{ij} - \mu) + \frac{\gamma^{k-r}(1 - \gamma^{2r})}{1 - \gamma^{2k}}(y_{i,j+k} - \mu)$$

$$\text{cov}(y_{i,j+r}, y_{i,j+h}|y_{\text{obs}}, \theta) = \frac{\sigma^2}{1 - \gamma^2} \times \frac{\gamma^{h-r}(1 - \gamma^{2r})(1 - \gamma^{2(k-h)})}{1 - \gamma^{2k}},$$

where $1 \le r \le h \le k - 1$. Substituting $\theta = \theta^{(n)}$ gives expressions for $y_{i,j+r}^{(n)}$ and $k_{j+r,j+h}^{(n)}$ for the E step above. These are then used in the M step to generate updated parameter estimates $\theta^{(n+1)}$ as described. Starting parameter values $\theta^{(0)}$ are obtained from OLS regression on the complete cases. This procedure furnishes the necessary estimates to evaluate the split function (8), namely $\hat{\gamma}$ and $\hat{y}_{i,\text{mis}}$. However, a considerable computational burden is involved because strictly speaking, the EM algorithm must be invoked before attempting to split each node. Running EM just at the outset, before any splitting, effectively imputes $\hat{y}_{i,\text{mis}}$ under an assumption that $\mu(g)$ and $V(\theta, g)$ are constant for all subgroups g, which is contrary to the assumption of distinct subgroups and the goals of the analysis.

These computational difficulties arise in the most favorable setting. Different assumptions on the covariance structure and mechanisms and patterns of missingness make for even more complicated algorithms. For example, to ensure that properties (†) and (‡) hold for AR2 models, we need two nonmissing bounding observations both leading and trailing a sequence of missing observations. These properties are essential to simplifying E step computations. This requirement dramatically reduces the number of missing data patterns that can be handled with a resultant loss in efficiency with which the correlation parameters are estimated. Further, the above scheme for imputing $\hat{y}_{i,\text{mis}}$ does not account for imputation variance, which would entail appreciably more computation. Finally, the MCAR assumption is likely to be violated in practice. More probable is nonignorable nonresponse, which mandates modeling the mechanism creating the missing data, an often uncertain exercise. On account of these problems, methods for handling missing responses have not progressed beyond the above AR1 case.

2.4 Determining Tree Size

A crucial aspect of the regression tree methodology is determining what constitutes a reasonable size for the tree; that is, how many splits should be implemented. The manner in which this issue is resolved receives considerable attention

in CART, Chapter 3. There, the shortcomings of using stopping rules based on either node sizes becoming too small or the improvement as measured by ϕ being insufficient are argued. Essentially, the problems derive from smallness or insufficiency having to be gauged relative to preset thresholds. Misspecification of these thresholds can result in overfitting or underfitting. These difficulties are redressed by a pruning algorithm: (a) grow a very large tree initially so as to capture all potentially important splits; (b) collapse this back up—using complexity cost as defined below—creating an hierarchical sequence of trees; and (c) select an optimal tree from this sequence using cross-validation. Discussion of (c) is deferred to CART; here, we focus on (b) and other issues particular to the multiple-response setting.

Let $R(g)$ be the cost of a node g. For example, in splitting based on ϕ_m as given by (2), we can take $R(g)$ to coincide with $SS(g)$ as given in (1). Now, define $R(G)$ as the cost of a given tree G: $R(G) = \sum_{g\in\tilde{G}} R(g)$, where $\tilde{G}$ is the collection of terminal nodes of G. Further, define the *complexity* of G as $|\tilde{G}|$, the number of terminal nodes of G. Then, the *cost-complexity* of a tree G is $R_\alpha(G) = R(G) + \alpha|\tilde{G}|$, $\alpha \geq 0$.

Initially, a large tree $G_{\max}$ is grown using the goodness-of-split function ϕ of choice. The size of this tree is not critical provided it is large enough. To this end, a user-defined parameter $N_{\min}$ is specified such that every terminal node of $G_{\max}$ satisfies $N_g \leq N_{\min}$; see CART Section 3.2. We use $N_{\min} = 5$ as a default setting. For each value of α, we find the subtree $G(\alpha)$ of $G_{\max}$ that minimizes $R_\alpha(G)$. If α is small, the penalty for large $|\tilde{G}|$ also will be small and hence $G(\alpha)$ will be large. As $\alpha \uparrow$, $|\tilde{G}(\alpha)| \downarrow$. Finally, for α sufficiently large, $|\tilde{G}(\alpha)| = 1$ and the minimal cost-complexity tree is the root node (the entire original sample), because any splitting will increase the cost-complexity. Determining the values of α that correspond to a change in $G(\alpha)$, and what that change is, is described in CART Section 3.3. The estimation problem of selecting, by cross-validation, a best tree from the sequence so defined is described in CART Section 3.4. We note, as in Rice and Silverman (1991), that with such dependent data, cross-validation provides approximately unbiased estimates of $R(g)$ by leaving out groups of individuals rather than groups of observations.

The only modification needed to accommodate multiple response is the imposition of a more stringent selection criterion. For the single-response variable situation, the CART monograph advocates selecting as an optimal tree the smallest tree with a cross-validated error rate within one standard error of the tree possessing the minimal error rate. The rationale behind this rule was that the cross-validated error rates themselves are subject to variability, so choosing the simplest tree with an accuracy comparable to the tree with the smallest error is reasonable. When contending with longitudinal data or multiple response, there is reason to be even more conservative. This derives from seeking to limit the number of covariance parameters fitted. Instead of making some ad hoc recommendation as to what multiplier should be attached to the standard error to achieve this greater simplicity, it is proposed that the plot of cross-validated standard errors for $\tilde{G}(\alpha)$ versus $|\tilde{G}(\alpha)|$ be examined. A characteristic feature of such plots is a flat valley surrounding the minimum. Smaller trees in this region are candidates for further scrutiny. Details on this approach are provided in CART Section 3.4.3 and in Segal (1988).

Some caveats concerning the usage of cross-validation error estimates in tree selection are warranted. First, it is problematic to compare such estimates across models using different covariance structures. This is due to differing constant terms being ignored in the computations; see Rice and Silverman (1991). Second, even for a particular covariance structure the sequence of cross-validated error estimates can be used only as a guide, because the strict nesting that pertained in the univariate outcome case no longer holds due to the reestimation of the covariance parameters. If we imposed (3) throughout—that is, did not update covariance parameters after splitting—this concern would vanish.

Because of these concerns, additional assessments of tree stability are required. Thus, tree selection is based also on sensitivity analyses whereby a variety of different split functions and covariance structures are used and the stable features of the resultant trees retained. Both approaches are used in the subsequent example.

2.5 Time-Varying Covariates

One strength of longitudinal studies is that they facilitate the assessment of how change in the response variable relates to change in covariates. It is therefore important to be able to use such time-varying covariates. Covariates are used in tree-structured methods to determine what constitutes an allowable split. At present, no convincing technique for defining splits on time-varying covariates has been devised. For ordered covariates, the difficulty lies in formulating interpretable splits that preserve ordering with respect to both time and the variable itself. The only strategy that has been implemented to date collapses these two components into one by replacing the time-varying covariate with low-order polynomial approximations. In particular, linear summaries have been used. Each time-varying covariate is regressed, within individuals, against time. The resultant slopes and intercepts then constitute covariates that are submitted to the regression tree algorithm. Such an approach is only reasonable to the extent that the linear regression adequately captures the time-varying covariate. Splits performed on the resultant "intercept covariate" can be interpreted as defining subgroups based on the level of the time-varying covariate, whereas those performed on the "slope covariate" define subgroups based on linear change in the time-varying covariate. Derived covariates capturing higher-order curvature and/or variability also could be used.

A referee suggested another possibility for handling time-varying covariates. The accommodation does not involve using time-varying covariates as part of the splitting process, but rather modifying the split criteria. Presently, in (1), the sum of squared deviations is about the mean vector $\mu(g)$. However, estimating the mean at each time within a node g can be thought of as fitting a flexible function of time. Hence, the methodology is already handling one time-varying covariate, namely time. Now, instead of just estimating a multivariate mean within each node, it is possible to fit a function $f(g)$, within node g, of one or more time-varying

covariates. Then, $f(g)$ replaces $\mu(g)$ both in evaluating (1) and summarizing the "growth curve" for subgroup g (c.f. Figure 6). Such a procedure allows for irregularly spaced observations (see Sec. 2.6) and multiple time-varying covariates. The increased computation involved in fitting $f(\cdot)$ can be contained by prescribing simple parametric forms for $f(\cdot)$. Nevertheless, with the emergence of very efficient algorithms for computing nonparametric smooths (e.g., splines), it is possible to retain the fully nonparametric treatment.

2.6 Unequally Spaced Data

Until now, the assumption has been that the observations are spaced equally in time. This is not too restrictive, because longitudinal studies are designed this way. Furthermore, when unequally spaced data results from missing observations occurring under an equally spaced design, the missing value strategies of Section 2.3 can be used. When the observations are highly irregular, we are obliged to "borrow strength" through some sort of smoothing (parametric or nonparametric), as just discussed. The principal difficulty in dealing with highly irregularly spaced data lies in estimating the covariance parameters. Of course, for independence or CS covariance models, the irregularity is not problematic, because in these models the correlation does not vary with time. The AR1 model has a continuous time analog: $\rho(t) = \exp(-\gamma t)$, which in principle allows for irregularly spaced data. However, simple methods of parameter estimation (moments, least squares) will be very unstable; see Diggle (1988, Sec. 4). Attempts at improving stability by averaging over comparable lags essentially revert to the equi-spaced observation setting. Work is in progress to investigate estimation with such unequally spaced data. Additionally, new updating methods will need to be devised since the inverse covariance matrices are more complex.

3. EXAMPLE: HIV IN SAN FRANCISCO

3.1 Background

The natural history of acquired immunodeficiency syndrome (AIDS) still is not fully understood. Following infection with HIV, subjects progress to a constellation of clinical symptoms defining AIDS and then to death. These symptoms result from deterioration of the immune system. It is of considerable clinical and epidemiological importance to understand the nature of this immune function decay, which will facilitate improvement in the timing and evaluation of therapies and help project the course of the epidemic. The latency period for AIDS (time from infection to symptoms) is long, with the median latency $\simeq$ 10 years (Muñoz et al. 1989; Bacchetti and Moss 1989), and variable. To try and explain this variability in terms of immune function decay, markers of immune function such as CD4 T lymphocytes and β_2 microglobulin are regularly measured on longitudinal cohorts of HIV-seropositive and seroconverting individuals. It has been hypothesized that the variable latencies are due to the existence of subgroups experiencing distinct patterns of immune function loss. Such subgroups include (a) subjects experiencing rapid CD4 decline after infection followed by gradual decay, (b) subjects having an initial CD4 elevation following infection with subsequent linear decay, (c) subjects possessing linear decay from time of infection to development of AIDS, and (d) subjects possessing linear decay from time of infection until just before development of AIDS when there is accelerated loss. The purpose of the subsequent analysis is to determine whether subgroups exist and, if so, whether these subgroups can be characterized in terms of covariates.

3.2 San Francisco General Hospital Cohort

The composition and recruitment of the San Francisco General Hospital Cohort were described in Moss, et al. (1988). Briefly, recruitment of 462 homosexual men from sexually transmitted disease clinics, sexual partners of patients with AIDS, and the San Francisco community occurred from 1983–1984. The study design called for these individuals to be seen annually, at which time sociodemographic, medical history, and behavioral information was obtained via questionnaire, sera was procured, and immune function markers were measured. Of the 462 men in the cohort, 288 were HIV-seropositive at entry. The analysis here is restricted to 95 of these men who provided data on β_2 microglobulin for their first five annual visits. Increased values of β_2 microglobulin indicate worsened prognosis. Unfortunately, there were insufficient seroconverters (15) for this group to be analyzed; only the seroconverters can provide an approximate infection date, and using data from seropositives can produce biased estimates, as indicated by Brookmeyer and Gail (1987). Thus, the analysis must be viewed as mainly illustrative. The results could be meaningful if all subjects had been infected at roughly the same time. The baseline distribution of β_2 microglobulin does not refute such an assumption.

The covariates submitted to the tree-structured regression procedure included age; education; race; number of past episodes of syphilis, gonorrhea, genital herpes, and hepatitis B; number of male sex partners in the preceding year; history of blood transfusion; and smoking and alcohol consumption. All the covariate information was extracted from the subject's baseline data. Thus, all the covariates are time-independent. A subject's response vector consisted of the five consecutive annual β_2 microglobulin determinations. β_2 microglobulin was analyzed because it had proved prognostically important for time to AIDS in this cohort (Moss et al. 1988). Like CD4 counts, however, β_2 microglobulin is a very noisy measure with high within-individual variability.

3.3 Results

Empiric correlations of β_2 microglobulin between all pairs of visits were computed so as to suggest appropriate choices for V. The results indicated that the autocorrelation function was intermediary between an AR1 and CS model. Therefore, both possibilities were explored, as was the independence correlation structure. The split function used was ϕ_m as given by (2). The regression tree that resulted from assuming either independence or CS is displayed in Figure 5. The tree that emerges under an AR1 assumption features identical co-

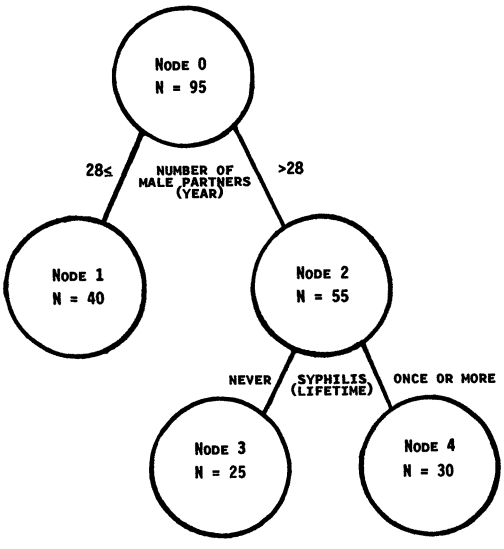

Figure 5. β_2 Microglobulin Regression Tree.

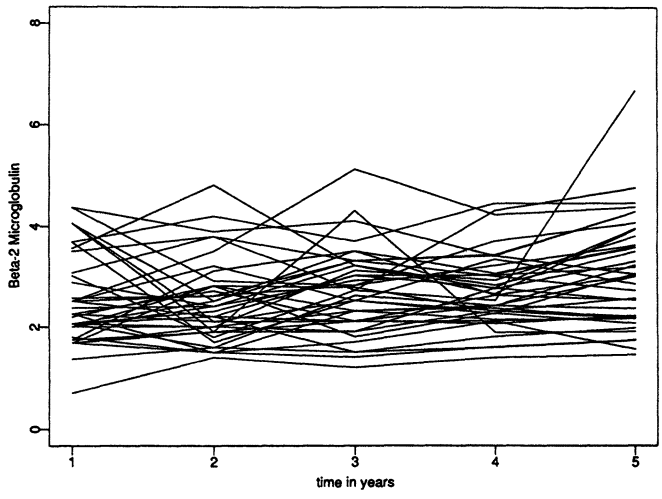

Figure 7. Node 1: Number of Partners < 28.

variates; however, the first split on number of partners occurs at a somewhat lower split point (15 instead of 28). Thus, the displayed tree structure is reasonably stable. Figure 6 depicts the mean response profiles (over time) for each of the subgroups in Figure 5; Figures 7–10 display subgroup-specific individual response vectors.

The following conclusions can be drawn. The first split on the covariate number of male partners in the preceding year at a cut-point of 28 represents the best subdivision of the sample as gauged by ϕ_m. The less sexually active subjects (node 1) have uniformly better (lower) mean levels of β_2 microglobulin than their more sexually active counterparts (node 2). Further, the rate at which the sexually active individuals suffer immune function deterioration is greater on average than that of the less sexually active subjects. The sexually active subjects are then subdivided on the basis of past syphilis. Those subjects with no past syphilis (node 3) have levels and rates of immune function loss that are almost identical to the less sexually active subgroup (node 1). Conversely, those subjects with one or more past episodes of syphilis (node 4) experience the fastest rate of immune function loss.

It is important to be aware of possible proxy behavior for the variable number of partners. More sexually active individuals were likely to have been infected earlier. Hence this variable might be acting as a proxy for time since seroconversion, and the differing β_2 microglobulin levels for differing numbers of partners may simply reflect differing infection times. Resolving this issue requires data on a large sample of seroconverters.

Tree-structured regression methods as developed here for longitudinal studies are flexible in that no parametric form is prescribed for the growth curves under study. To examine the extent to which this flexibility mattered in the current analysis, I summarized each individual's five vector of response measures by a simple linear regression against time. The slopes from these regressions then constituted a single-outcome measure that captured rate of change in β_2 microglobulin. This derived outcome variable, along with the original covariates, were then analyzed by conventional regres-

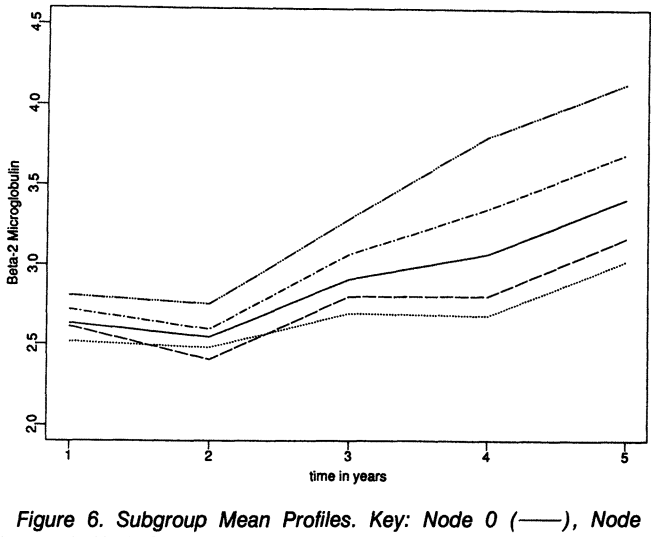

Figure 6. Subgroup Mean Profiles. Key: Node 0 (——), Node 1 (· · · · ·), Node 2 (-·-·-), Node 3 (--), Node 4 (-· · ·-).

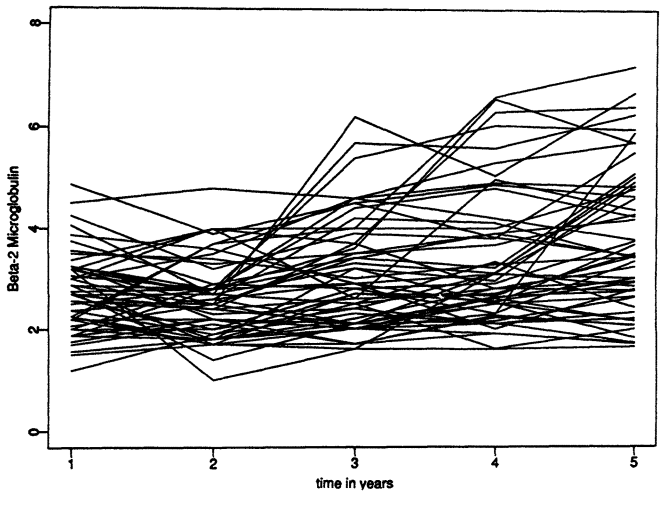

Figure 8. Node 2: Number of Partners > 28.

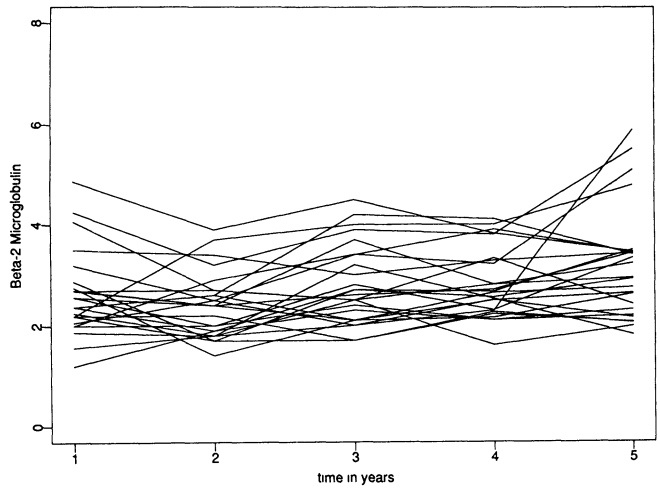

Figure 9. Node 3: Number of Partners > 28, Never Syphilis.

sion tree methods. The prominent split variables were history of genital herpes and age at recruitment. These are not strongly correlated with number of partners or syphilis. Furthermore, the composition of the terminal nodes was different for the two analyses. Had a similar tree structure emerged, it would be possible to infer that the growth curve information was adequately summarized by a slope. However, the differences suggest that the nonlinear changes in β_2 microglobulin are important.

Another remark concerning the versatility of regression tree methods is warranted. Several of the continuous covariates used had highly skewed distributions. For example, the covariate number of male partners in the preceding year has a long right tail. However, there is no need to symmetrize such covariates, because the tree structure is invariant under monotone transformations of the covariates. Thus, if a particular best split occurs on covariate X at value x, then reconstruction of the tree using covariate $W = g(X)$ instead of X will yield an identical tree structure with a split on W at value $w = g(x)$ for monotone $g(\cdot)$.

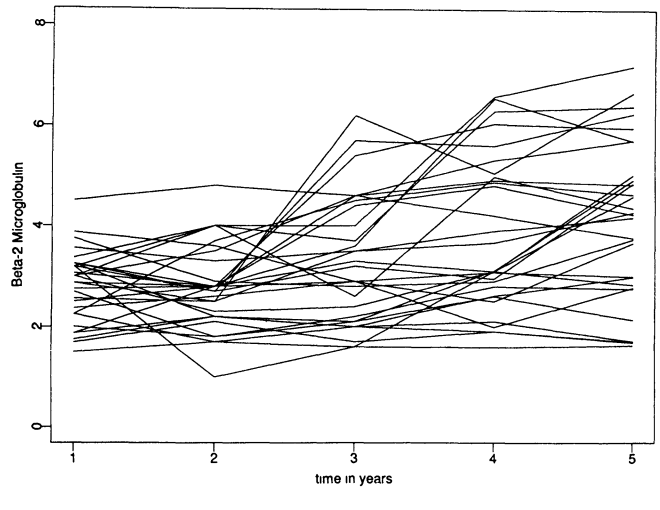

Figure 10. Node 4: Number of Partners > 28, Ever Syphilis.

4. DISCUSSION

It is purposeful to make explicit the limitations of the regression tree approach to identifying subgroups in longitudinal data settings. These limitations result from the imposition of assumptions and their potential violation. First, the collection of splits that determine the putative subgroups is restrictive for continuous or ordered covariates; only binary splits perpendicular to the covariate axes are allowed. This restriction is mandated by computational considerations. Although it can be argued that a ternary (or higher order n-ary) split can be captured by two $(n - 1)$ binary splits, the fact that the algorithms are single-step optimizations (no look ahead is entertained) makes such arguments tenuous. In any event, only a limited set of subgroups is examined. Thus, any claims at having—or not having—identified subgroups can only be stated relative to this set. A counterpoint is that the standard set of splits examined (as defined in CART Chapter 2 and Sec. 2.1) is (a) a priori plausible, (b) extensive, and (c) readily interpretable. This class of splits is advocated by Breiman and Friedman (1988), whereas Loh and Vanichsetakul (1988) prefer splits based on linear combinations of covariates and allow for multiway splits.

It is pertinent to contrast this with unconstrained grouping of the responses (by some clustering algorithm) with subsequent examination of how covariates distribute within and between subgroups, as suggested by Rao (C. R. Rao, personal communication, 1988). The principal difficulty here will be in attaching "labels" or interpreting the resulting subgroups. Examining how the covariates distribute with respect to the subgroups may reveal "clean" divisions (e.g., into boys and girls), but equally they may defy interpretation. The problems facing standard regression techniques in identifying subgroups can be illustrated in the standard multiple linear regression setting with a single-response variable. Subgroups can be constructed by first computing predicted values $\hat{y}_i = \hat{\alpha} + \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$ and then ordering these predicted values and dividing on the basis of percentiles; for example, for five groups, the divisions would be at the quintiles. Again, the drawback to such an approach is that the subgroups so formed may defy description in that individuals with very different covariates could be placed in the same subgroup. Attempting to overcome this by grouping on the percentiles of the covariates is also problematic. Without loss of generality, suppose all coefficients are positive: $\beta_j > 0 \; \forall j$. Then, to obtain a subgroup with "high" response values, we may posit that it contain all individuals whose covariate values are, say, in the upper quintile for each covariate. The problem is that there may be no individuals possessing such a covariate profile. The flip side to all this is that tree-structured methods, although adept at identifying subgroups, do not succinctly describe covariate effects.

The second assumption is of some parametric form for the covariance function. This has all the attendant possibilities for misspecification. This imposition may appear all the more heavy-handed because the same form is assumed to hold over all subgroups, with provision only for the parameter(s) to change. These issues have been discussed in Section 2.1.

Another imposition is the use of a particular split function, a problem arising whenever a loss measure is prescribed. Thus, as in other areas, split functions should be chosen on grounds of appropriateness for the problem at hand, and the impact of particular choices should be explored. The split functions proposed here afford natural and computationally feasible options for the analysis of longitudinal and repeated measures data. Additional split functions can be entertained; for example, robust analogs based on L^1 norms could be considered. The difficulties here will be primarily computational.

The difficulties facing tree-structured methods for longitudinal data in handling the important practical problems of missing responses and time-varying covariates have been made evident. Their resolution is the subject of future research.

A final matter that has not been discussed is inference. Tree techniques have been used for exploratory and descriptive purposes. As exploratory tools, the examination of a large number of possible subgroups, without the imposition of parametric assumptions, facilitates the extraction of unanticipated structure. However, inference—the probability assessment of individual splits or collections of splits—has yet to be undertaken. Such assessments mandate placing distributional assumptions on the responses. Suppose, for example, we assume the multiple responses are multivariate normal and we use Hotelling's T^2 as given in (4) as a split function. In appraising the significance of a chosen split, it would be incorrect to reference the corresponding F distribution, because the chosen split represents the *maximum* of a number of F statistics. The exact distribution is intractable. In a related setting, Miller and Siegmund (1982) have obtained the asymptotic distribution for maximally selected χ^2 statistics. A similar endeavor is needed in the context of tree-structured regression along with accompanying small sample simulation studies. Until such developments have been realized, the role of regression trees will remain exploratory. Yet in fulfilling this role in the longitudinal data setting, tree-structured methods ought to prove valuable.

[*Received September 1990. Revised March 1991.*]

REFERENCES

Bacchetti, P., and Moss, A. R. (1989), "Incubation Period of AIDS in San Francisco," *Nature,* 338, 251–253.

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

Breiman, L., and Friedman, J. H. (1988), "Discussion of 'Tree-Structured Classification via Generalized Discriminant Analysis,' " by W.-Y. Loh and N. Vanichsetakul, *Journal of the American Statistical Association,* 83, 725–727.

Brookmeyer, R., and Gail, M. (1987), "Biases in Prevalent Cohorts," *Biometrics,* 43, 739–749.

Chi, E. M., and Reinsel, G. C. (1989), "Models for Longitudinal Data With Random Effects and AR1 Errors," *Journal of the American Statistical Association,* 84, 452–459.

DeGruttola, V., Lange, N., and Dafni, U. (1990), "Modeling the Progression of HIV Infection," *Journal of the American Statistical Association,* 86, 569–577.

Diggle, P. (1988), "An Approach to the Analysis of Repeated Measures," *Biometrics,* 44, 959–971.

Donner, A., and Koval, J. J. (1980), "The Estimation of Intraclass Correlation in the Analysis of Family Data," *Biometrics,* 36, 19–25.

Draper, D. (1987), "Discussion of 'Prediction of Future Observations in Growth Curve Models,' " by C. R. Rao, *Statistical Science,* 2, 434–471.

Friedman, J. H. (1991), "Multiple Adaptive Regression Splines," *The Annals of Statistics,* 19, 1–67.

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations,* New York: John Wiley.

Laird, N. M. (1988), "Missing Data in Longitudinal Studies," *Statistics in Medicine,* 7, 305–315.

Laird, N. M., Lange, N., and Stram, D. O. (1987), "Maximum Likelihood Computations with Repeated Measures: Application of the EM Algorithm," *Journal of the American Statistical Association,* 82, 97–105.

Lange, N., and Laird, N. M. (1989), "The Effect of Covariance Structure on Variance Estimation in Balanced Growth Curve Models With Random Parameters," *Journal of the American Statistical Association,* 84, 241–247.

Lange, K., Little, R. J. A., and Taylor, J. (1989), "Robust Statistical Modeling Using the t Distribution," *Journal of the American Statistical Association,* 84, 881–896.

Lee, J. C. (1988), "Prediction and Estimation of Growth Curves With Special Covariance Structures," *Journal of the American Statistical Association,* 83, 432–440.

Liang, K.-Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika,* 73, 13–22.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis with Missing Data,* New York: John Wiley.

Loh, W.-L. (1991), "Estimating Covariance Matrices," *The Annals of Statistics,* 19, 283–296.

Loh, W.-Y., and Vanichsetakul, N. (1988), "Tree-Structured Classification via Generalized Discriminant Analysis," *Journal of the American Statistical Association,* 83, 715–724.

Miller, R. G., Jr., and Siegmund, D. (1982), "Maximally Selected Chi Square Statistics," *Biometrics,* 38, 1011–1016.

McCullagh, P., and Nelder, J. (1983), *Generalized Linear Models,* New York: Chapman and Hall.

Morrison, D. F. (1976), *Multivariate Statistical Methods,* New York: McGraw-Hill.

Moss, A. R., Bacchetti, P., Osmond, D., Krampf, W., Chaisson, R. E., Stites, D., Wilber, J., Allain, J.-P., and Carlson, J. (1988), "Seropositivity for HIV and the Development of AIDS: Three-Year Follow-Up of the San Francisco General Hospital Cohort," *British Medical Journal,* 296, 745–750.

Muñoz, A., Wang, M.-C., Bass, S., Taylor, J. M. G., Kingsley, L. A., Chmiel, J. S., and Polk, B. F. (1989), "AIDS Free Time After HIV Seroconversion in Homosexual Men," *American Journal of Epidemiology,* 130, 530–539.

Muñoz, A., Carey, V., Schouten, J. P., Segal, M. R., and Rosner, B. (1990), "A Parametric Family of Correlation Structures for the Analysis of Longitudinal Data," Technical Report, Department of Biostatistics, Johns Hopkins School of Public Health.

Nelder, J. A., and Pregibon, D. (1987), "An Extended Quasi-Likelihood Function," *Biometrika,* 74, 221–232.

Rao, C. R. (1987), "Prediction of Future Observations in Growth Curve Models," *Statistical Science,* 2, 434–471.

Rice, J. A., and Silverman, B. W. (1991), "Estimating the Mean and Covariance Structure Nonparametrically When the Data Are Curves," *Journal of the Royal Statistical Society,* Ser. B, 53, 233–243.

Schlucter, M. D. (1988), "Analysis of Incomplete Multivariate Data Using Linear Models With Structured Covariance Matrices," *Statistics in Medicine,* 7, 317–324.

Segal, M. R. (1988), "Regression Trees for Censored Data," *Biometrics,* 44, 35–47.

Segal, M. R., and Bloch, D. A. (1989), "A Comparison of Proportional Hazards and Regression Trees for Censored Data," *Statistics in Medicine,* 8, 539–550.

Ware, J. H., and Wu, M. C. (1981), "Tracking: Prediction of Future Values From Serial Measurements," *Biometrics,* 37, 427–437.