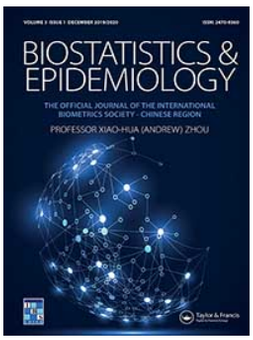

Taylor & Francis
Taylor & Francis Group

Biostatistics & Epidemiology

ISSN: 2470-9360 (Print) 2470-9379 (Online) Journal homepage: https://www.tandfonline.com/loi/tbep20

Regression trees for longitudinal data with baseline covariates

Madan Gopal Kundu & Jaroslaw Harezlak

To cite this article: Madan Gopal Kundu & Jaroslaw Harezlak (2019) Regression trees for longitudinal data with baseline covariates, Biostatistics & Epidemiology, 3:1, 1-22, DOI: 10.1080/24709360.2018.1557797

To link to this article: https://doi.org/10.1080/24709360.2018.1557797

BIOSTATISTICS & EPIDEMIOLOGY
2019, VOL. 3, NO. 1, 1–22
https://doi.org/10.1080/24709360.2018.1557797

Taylor & Francis
Taylor & Francis Group

Regression trees for longitudinal data with baseline covariates

Madan Gopal Kundu [a] and Jaroslaw Harezlak [b]

[a]AbbVie, North Chicago, IL, USA; [b]Department of Epidemiology and Biostatistics, Indiana University School of Public Health, Bloomington, IN, USA

ABSTRACT

Longitudinal changes in a population of interest are often heterogeneous and influenced by a combination of baseline factors. In such cases, classical linear mixed effects models [Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982;38:963–974.] for the mean structure provide poor fit to the data. We propose regression tree methodology for the longitudinal data identifying and characterizing homogeneous subgroups. Currently available regression tree construction methods are either limited to a repeated measures scenario or combine the heterogeneity among subgroups with the random inter-subject variability. We propose a longitudinal classification and regression tree (LongCART) algorithm under conditional inference framework that overcomes these limitations utilizing a two-step approach. The LongCART first selects the partitioning variable via a parameter instability test and then finds the optimal split for the selected partitioning variable. Thus, at each node, the decision of further splitting is type I error controlled, guarding against variable selection bias, over-fitting and spurious splitting. We obtained asymptotic results for the proposed instability test and examined its finite sample behavior through simulation studies. Comparative performance of LongCART algorithm was evaluated empirically via simulation studies. Finally, we applied LongCART to study the longitudinal changes in choline levels among HIV-positive patients.

ARTICLE HISTORY
Received 13 April 2017
Accepted 28 October 2018

KEYWORDS
LongCART; regression tree; instability test; longitudinal data; mixed models; score process; brownian bridge

1. Introduction

In longitudinal studies, repeated measurements of the outcome variable are often collected at irregular and possibly subject-specific time points. Parametric regression methods for analyzing such data have been developed by Laird and Ware [1] and Liang and Zeger [2] among others, and have been summarized by Diggle [3]. If the population under consideration is diverse and there exist several distinct subgroups within it, the true parameter values for the longitudinal mixed effects model may vary between these subgroups. In such cases, the traditional mixed effects models, for example, linear mixed effects model, which assumes a common parametric form for the mean structure may not be appropriate. For

CONTACT Jaroslaw Harezlak harezlak@iu.edu Indiana University School of Public Health, Bloomington, IN 47405, USA

example, Raudenbush [4] used a longitudinal depression study to argue that it is incorrect to assume that all the individuals in a given population will be experiencing either increasing or decreasing levels of depression. As another example, in clinical research, often the influence of biomarkers, for instance, pharmacogenetic biomarkers, on patients' response to a treatment are evaluated. Thus a diverse population may be a reality in both observational and experimental studies. In such instances, an assumption of a common parametric form for mean structure will mask important subgroup differences and will lead to erroneous conclusions. In our work, we are interested in the identification of meaningful and interpretable subgroups with differential longitudinal trajectories. We have proposed a regression tree construction technique for the longitudinal data, LongCART algorithm, using baseline characteristics as partitioning variables. The LongCART algorithm provides an improvement over the existing methods in one or more of the following aspects: (1) the test for the decision about further splitting at each node is type I error controlled via formal hypothesis testing and hence offers guard against variable selection bias, over-fitting and spurious splitting, (2) it is applicable when the measurements are taken at the subject-specific time points, (3) it does not merge the group differences with the random individual difference (captured by random effect components) and (4) it reduces computational time.

When the longitudinal profile in a population depends on the baseline covariates, the most common strategy is to include these covariates and their interactions with the time-varying factor in the model. However, this strategy has some inherent drawbacks including over-fitting (due to the inclusion of many interaction terms which are not required), estimation bias (due to possible misspecification of the functional form) and inability to capture nonlinear effects. Because of these drawbacks, a better strategy is to identify longitudinally homogeneous subgroups, possibly characterized by baseline covariates. One of the popular techniques to find homogeneous subgroups is latent class modeling (LCM) [5]. An alternative approach is to construct a regression tree with longitudinal data [6]. Advantages of regression tree technique over LCM are: (1) it characterizes the subgroups in terms of partitioning variables and (2) the number of the subgroups does not need to be known a-priori. In general, the thrust of any tree technique 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 covariance structure [6]. In our work, we focus on finding homogeneous groups for the mean structure.

Throughout this article, we refer to the regression tree with longitudinal data as 'longitudinal tree .' Figure 1 displays a toy example of a longitudinal tree. This longitudinal tree represents a heterogeneous population with three distinct subgroups in terms of their longitudinal profiles. These subgroups can be characterized by *gender* and *age*. Here, *gender* and *age* are baseline attributes. In each of the three subgroups, the longitudinal trajectory depends on the covariates $\mathbf{w} = [w_1, \ldots, w_q]^\top$, but these subgroups are heterogeneous in terms of the true coefficients ($\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ for subgroups 1, 2 and 3, respectively) associated with their longitudinal profiles. Consider the following form of a linear longitudinal mixed effects model

$$y_{it} = \beta_0^x + \beta_1^x t + \mathbf{w}_{it}^\top \boldsymbol{\beta}^x + \mathbf{z}_{it}^\top \mathbf{b}_i + \epsilon_{it}, \tag{1}$$

where i is the subject index and y,t and $\mathbf{w}$ denote the outcome variable, time and the vector of measurements of scalar covariates $w_1, \ldots, w_q$, respectively. Let $X_1^{G_1}, \ldots, X_S^{G_S}$ include all potential baseline attributes (with possible cut-off points $G_1, \ldots, G_S$, respectively) that

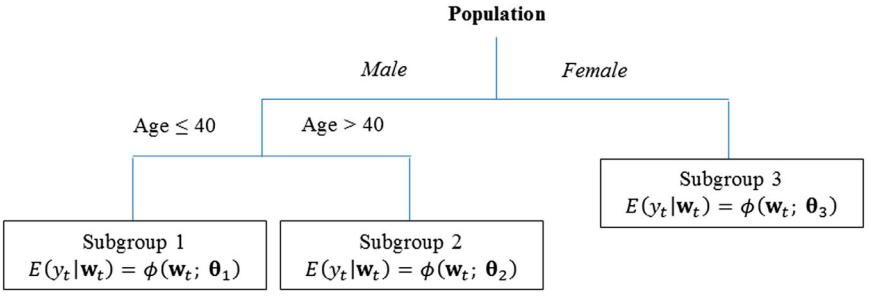

Figure 1. Sample longitudinal tree. The population consists of three subgroups and they differ in their longitudinal profiles. The model forms are same, but they are different in terms of coefficients – $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ for subgroups 1, 2 and 3, respectively. These subgroups are defined by the partitioning variables gender and age.

might influence the longitudinal trajectory in (1). The superscript x is added to the coefficients β_0, β_1 and $\boldsymbol{\beta}$ to reflect their possible dependence on these baseline attributes. Let $\boldsymbol{\theta}^x = (\beta_0^x, \beta_1^x, \boldsymbol{\beta}^x)^\top$. With such a model, 'homogeneity' refers to the situation when the true value of $\boldsymbol{\theta}^x$ remains the same for all the individuals in the entire population, *i.e.* $\boldsymbol{\theta}^x = \boldsymbol{\theta}$. When the longitudinal changes in the population of interest are *heterogeneous* there exist distinct subgroups differing in terms of the coefficients' true values, *i.e.* $\boldsymbol{\theta}^x \neq \boldsymbol{\theta}$. To model the influence of $\{X_1^{G_1}, \ldots, X_S^{G_S}\}$ on the longitudinal trajectory of y non-parametrically, we have used these baseline attributes as the partitioning variables for the construction of a longitudinal tree.

In constructing a longitudinal tree through binary partitioning, one way to choose a partition is via maximizing improvement in a goodness-of-fit criterion. For example, Abdolell [7] chose deviance as a goodness-of-fit criterion. They evaluated deviance at each split of a given partitioning variable and selected the partition with a maximum reduction in deviance for the binary splitting. In general, any exhaustive search framework without any formal test of statistical hypothesis like this is prone to over-fitting, variable selection bias even with the presence of pruning mechanism (see e.g. [8]) and also it is prone to spurious findings (see e.g. [9]). Furthermore, such methods are computationally expensive as these procedures require calculation of the goodness-of-fit criterion at each possible cut-off points over all available partitioning variables. For example, with S partitioning variables: $X_1^{G_1}, \ldots, X_S^{G_S}$, with cut-off points $G_1, \ldots, G_S$, respectively, the total number of the goodness-of-fit criterion calculations is $\sum_{s=1}^{S} (G_s - 1)$.

To avoid the problems associated with the exhaustive search strategies, we propose the LongCART algorithm for construction of a regression tree under the conditional inference framework of regression tree construction suggested by Hothorn et al. [8]. In this framework, in step 1, we first identify whether any partitioning variable is associated with the heterogeneity of response trajectory through formal statistical testing via a global 'test for parameter instability.' Parameter instability test is carried out for each partitioning variable separately with an adjustment for testing multiplicity. If one or more partitioning variables are found to be significantly associated with the heterogeneity of the response trajectory, the partitioning variable with the minimum p-value is selected as a splitting variable. Once the splitting variable is chosen, in step 2, the cut-off point with the maximum improvement

in goodness-of-fit criterion is used for binary splitting. If no partitioning variable is found to be significant in step 1, we stop the recursion. The key idea here is that we are combining the multiple testing procedures (step 1) with model selection (step 2) in order to control the type I error while taking the decision on splitting at each node. Such a step minimizes the selection bias in choosing the partitioning variable compared to the exhaustive search-based procedures where the partitioning variables with many unique values tend to have an advantage over the partitioning variables with fewer unique values [10–13].

The idea of *parameter instability test* was originally proposed in the time-series literature to test for structural change (see e.g. [14–16]). The purpose of *parameter instability test* in the context of regression tree is to detect any evidence of heterogeneity of model parameters across all of its cut-off points in a partitioning variable as has been used in previous tree construction algorithms (see e.g. [8,17,18]). The advantage of parameter instability test is that, for each partitioning variable, the test statistic has to be obtained only once under the homogeneity [17]. Various test statistics likelihood based score process has been proposed for *parameter instability tests*. For example, Zeileis et al. [17] used *supLM* test of Andrews[19] and obtained approximate p-values according to Hansen [20]. On the other hand, Hothorn et al. [8] proposed general form of test statistic and employed permutation based test strategy to obtain p-value . In our work, for *parameter instability test* with continuous partitioning variables, we considered the test statistic of Hjort and Koning [16] which converges to the supremum of Brownian Bridge process under null hypothesis of homogeneity. The distribution function of supremum of Brownian Bridge process is well established and can be expressed in finite closed form for any given accuracy level leading to relatively easier calculation of p-values . The advantage of this approach is that it is more principled than approximate permutation based test and p-values can be obtained relatively more easily compared to *supLM* test of Andrews. Unlike the aforementioned works, we have derived the asymptotic properties of the instability test for the continuous partitioning variables and explored its size and power through an extensive simulation study. For categorical partitioning variables with a small number of cut-off points, the test is derived in a straightforward way by employing asymptotic normality of the score functions.

Among the tree based methods, classification and regression tree (CART) methods [21] is the most popular one. Zeileis et al. [17] have extended the concept of CART methodology in the context of fitting cross-sectional generalized linear models (GLM). Binary partitioning for longitudinal data has been proposed first by Segal [6]. Segal's approach along with the other two regression tree construction methods proposed by De'Ath [22] and Larsen and Speckman [23] are restricted to longitudinal data with a regular structure, that is all the subjects have an equal number of repeated observations at the fixed time points [24]. Zhang [25] proposed multivariate adaptive splines to analyze longitudinal data, which can be used to generate regression trees for longitudinal data. Abdolell [7] used deviance as a goodness-of-fit criterion for binary partitioning. They controlled the level of Type I error via permutation test taking into account testing multiplicity. Sela and Simonoff [26] as well as Galimberti and Montanari [27] merged the subgroup differences with the random individual differences. Sela and Simonoff [26] constructed the *RE-EM* tree through an iterative two-step process. In the first step, they obtained the random effects' estimates and in the second step, they constructed the regression tree ignoring the longitudinal structure according to CART algorithm implemented in `rpart` package in `R`. They repeated these two steps until the estimates of the random effect converged in the

first step. Later Fu and Simonoff [28] proposed to construct *unbiased RE-EM* tree replacing CART algorithm by conditional inference tree [8]. On the other hand, Fokkema Pet al. [29] proposed to construct (G)LMM tree replacing CART algorithm by GLM tree algorithm of Zeileis et al. [17] in step 1. GUIDE [10,13] and MELT [30] algorithms also construct regression trees in two steps similar to the conditional inference framework [8]; however, these two algorithms employ chi-square test based on the residuals' direction only for the selection of partitioning variables in step 1 and, unlike *permutation based tests*, does not use the information from the full joint distribution. Furthermore, these two algorithms have been primarily developed for the fixed time-point scenario and the extension to random subject-specific time-point scenario have been proposed via an ad-hoc adjustment whereas random subject-specific time points fit naturally in the likelihood based score setting of LongCART algorithm.

The remainder of this paper is organized as follows. In Section 2 the longitudinal mixed effects models of interest are summarized. Tests for parameter instability for continuous and categorical partitioning variable cases are discussed separately in Section 3. Algorithm for constructing longitudinal regression trees along with the measures of improvement and a pruning technique are discussed in Section 4. Results from the simulation studies examining the performance of the instability test are provided in Section 5.1. Simulation results comparing LongCART algorithm with othe existing tree construction algorithms and linear mixed effects models are reported in Section 5.2. An application of LongCART is illustrated on the brain metabolite data collected from chronically HIV-infected patients in Section 6. The R-code for LongCART algorithm and the simulation code used in this article are available through webpage http://il-balds.org/software/.

2. Notation and statistical model

Let $\{y_{it}, \mathbf{w}_{it}\}$ be a set of measurements recorded on the ith subject ($i = 1, \ldots, N$) at time $t = (t_1, \ldots, t_{n_i})$, where y is a continuous scalar outcome; and $\mathbf{w}$ is the vector of measurements on scalar covariates $w_1, \ldots, w_q$. We assume that these covariates are linearly associated with y. In addition, for each individual, we observe a vector of attributes $(X_{1i}^{G_1}, \ldots, X_{Si}^{G_S})$ measured at baseline. We assume that $X_1^{G_1}, \ldots, X_S^{G_S}$ includes all potential baseline attributes that can influence the longitudinal trajectory of y and its association with covariates $w_1, \ldots, w_q$. Further, we do not assume the strict functional form of these baseline attributes' influence. We use the variables $X_1^{G_1}, \ldots, X_S^{G_S}$ as the candidate partitioning variables to construct a longitudinal regression tree to discover meaningful and interpretable subgroups with differential changes in y characterized by the $X_1^{G_1}, \ldots, X_S^{G_S}$.

When the longitudinal profile is homogeneous in the entire population, we can fit the following traditional linear mixed effects model for all N individuals [1]

$$y_{it} = \beta_0 + \beta_1 t + \mathbf{w}_{it}^\top \boldsymbol{\beta} + \mathbf{z}_{it}^\top \mathbf{b}_i + \epsilon_{it}, \tag{2}$$

where $\epsilon_{it} \sim N(0, \sigma^2)$ and $\mathbf{b}_i$ is the vector of random effects pertaining to subject i and distributed as $N(0, \sigma_b^2 \mathbf{D})$. By 'homogeneity' we mean that the true value of $\boldsymbol{\theta}^\top = (\beta_0, \beta_1, \boldsymbol{\beta}^\top)$ remains the same for all the individuals in the population. In fact, (2) is the simplified version of model in (1) under homogeneity.

We follow the common assumptions made in longitudinal modeling that $\mathbf{z}_{it}$ is a subset of $[\mathbf{w}_{it}^\top \quad t]^\top$; ϵ_{it} and $\mathbf{b}_i$ are independent; ϵ_{it} and $\epsilon_{i't'}$ are independent whenever $i \neq i'$ or $t \neq t'$ or both, and $\mathbf{b}_i$ and $\mathbf{b}_{i'}$ are independent if $i \neq i'$. Here, $\mathbf{w}_{it}^\top \beta$ is the fixed effect term and $\mathbf{z}_{it}^\top \mathbf{b}_i$ is the standard random effects term. For the i^{th} subject, we rewrite the Equation (2) as follows

$$\mathbf{y}_i = \mathbf{w}_i \boldsymbol{\theta} + \mathbf{z}_i \mathbf{b}_i + \epsilon_i, \tag{3}$$

where $\mathbf{y}_i^\top = (y_{i1}, \ldots, y_{in_i})$, $\mathbf{w}_i$ is the design matrix consisting of the intercept, time (t) and covariates ($\mathbf{w}$). n_i is the number of observations obtained from the i^{th} individual. The score function for estimating $\boldsymbol{\theta}$ under (3) is (see e.g. [31])

$$\mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta}) = \frac{d}{d\boldsymbol{\theta}} l(\mathbf{y}_i, \boldsymbol{\theta}) = \frac{1}{\sigma^2} \mathbf{w}_i^\top \mathbf{V}_i^{-1} (\mathbf{y}_i - \mathbf{w}_i \boldsymbol{\theta}),$$

where $\mathbf{V}_i = \mathbf{I} + \frac{\sigma_b^2}{\sigma^2} \mathbf{z}_i \mathbf{D} \mathbf{z}_i^\top$ and $\mathbf{e}_i = \mathbf{y}_i - \mathbf{w}_i \boldsymbol{\theta}$. Further, its variance is

$$\mathrm{Var}\left[\mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta})\right] = \mathbf{J}(\boldsymbol{\theta}) = \frac{1}{N} \mathbf{H}(\boldsymbol{\theta}),$$

where,

$$\mathbf{H}(\boldsymbol{\theta}) = -E\left[\frac{d}{d\boldsymbol{\theta}} \mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta})\right] = \frac{1}{\sigma^2} \mathbf{w}_i^\top \mathbf{V}_i^{-1} \mathbf{w}_i.$$

Maximum likelihood (ML) estimate of $\boldsymbol{\theta}$ obtained using all the observation from N subjects is valid only if the entire population under consideration is homogeneous. If the entire population is not homogeneous in terms of $\boldsymbol{\theta}$ then the likelihood estimate obtained considering all the subjects together are misleading; the extent and direction of ambiguity in the estimate will depend on the nature and proportion of heterogeneity in the sampled individuals. Therefore, under the assumption that $X_1^{G_1}, \ldots, X_S^{G_S}$ are the only attributes that influences the longitudinal profiles of y, it is important to decide first whether the true value of $\boldsymbol{\theta}$ remains the same for all the subgroups defined by $X_1^{G_1}, \ldots, X_S^{G_S}$ or not. In the next section, we describe statistical tests to assess whether the true value of $\boldsymbol{\theta}$ remains the same across all the values of a given partitioning variable.

3. Test for parameter instability

The purpose of *parameter instability test* is to test whether the true value of $\boldsymbol{\theta}$ remains the same across all distinct values of baseline attributes (i.e. partitioning variables). Let $X^G \in \{X_1^{G_1}, \ldots, X_S^{G_S}\}$ be any partitioning variable with G ordered cut-off points: $c_{(1)} < \ldots < c_{(G)}$ and $\boldsymbol{\theta}_{(g)}$ be the true value of $\boldsymbol{\theta}$ when $X^G = c_{(g)}$. Assume that there are m_g subject with $X^G = c_{(g)}$. We denote the cumulative number of subjects with $X^G \leq c_{(g)}$ by M_g. That is, $M_g = \sum_{j=1}^{g} m_j$ and $M_G = \sum_{j=1}^{G} m_j = N$. We want to conduct an omnibus test,

$$H_0 : \boldsymbol{\theta}_{(g)} = \boldsymbol{\theta}_0 \text{ vs. } H_1 : \boldsymbol{\theta}_{(g)} \neq \boldsymbol{\theta}_0.$$

Here, H_0 indicates the situation when parameter $\boldsymbol{\theta}$ remains constant (that is, *homogeneity*) and H_1 corresponds to the situation of parameter instability (that is, *heterogeneity*).

The parameter instability tests utilize the following properties of score function under H_0:

- A1: $E_{H_0}[\mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta}_0)] = 0$;
- A2: $\text{Var}_{H_0}[\mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta}_0)] = \mathbf{J}(\boldsymbol{\theta}_0) = \mathbf{J}$;
- A3: $\mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}})|_{H_0} \rightarrow^d N[0, \hat{\mathbf{J}}]$,

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ and $\hat{\mathbf{J}} = \mathbf{J}(\hat{\boldsymbol{\theta}})$. We discuss the instability test separately depending on whether the partitioning variable X^G is categorical or continuous.

3.1. Instability test with a categorical partitioning variable

It is straightforward to obtain a test for parameter instability using the properties A1–A3 when the partitioning variable, X^G, is categorical with a small number of categories (that is, $G \ll N$). Since the score functions $\mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}})$ are independent, we have under H_0, the following quantity

$$\chi^2_{cat} = \sum_{g=1}^{G} \left[\sum_{i=1}^{N} I(X_i^G = c_{(g)}) \mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}}) \right]^\top \left[m_g \hat{\mathbf{J}} \right]^{-1} \left[\sum_{i=1}^{N} I(X_i^G = c_{(g)}) \mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}}) \right]$$

is asymptotically distributed as χ^2 with $(G-1)p$ degrees of freedom where p is the dimension of $\boldsymbol{\theta}$. Here, $I(\cdot)$ is the indicator function. The reduction in p degrees of freedom is due to the estimation of p dimensional $\boldsymbol{\theta}$ from the data.

3.2. Instability test with continuous partitioning variable

For a continuous partitioning variable, number of cut-off point are usually high as almost all unique values except one of the extreme values may represent potential cut-off points. Our proposed instability test for continuous partitioning variable is based on score process. We begin by defining the following *score process*

$$\mathbf{W}_N(t, \boldsymbol{\theta}_0) = N^{-1/2} \sum_{i=1}^{M_g} \mathbf{u}(\mathbf{y}_i, \boldsymbol{\theta}_0) \qquad t \in [t_g, t_{g+1}),$$

where $t_g = M_g/N$. Under H_0, using multivariate version of Donsker's theorem and Cramér-Wold theorem (see e.g. [32]), it can be shown that

$$\mathbf{W}_N(t, \boldsymbol{\theta}_0) \rightarrow_d \mathbf{Z}(t),$$

where $\mathbf{Z}(t)$ is the zero-mean Gaussian process with $\text{cov}[\mathbf{Z}(t), \mathbf{Z}(s)] = \min(t, s)\mathbf{J}(\boldsymbol{\theta}_0)$. Since $\boldsymbol{\theta}_0$ is unknown in practice, we replace $\boldsymbol{\theta}_0$ by $\hat{\boldsymbol{\theta}}$ in score process

$$\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}) = N^{-1/2} \sum_{i=1}^{M_g} \mathbf{u}(\mathbf{y}_i, \hat{\boldsymbol{\theta}}).$$

It has been shown that the above estimated score process converges to Brownian Bridge process [16]. We present this result as following theorem and the proof of the theorem is outlined in Appendix.

Theorem 3.1: *Let's define the standardized estimated score process as*

$$\mathbf{M}_N(t, \hat{\boldsymbol{\theta}}) = \hat{\mathbf{J}}^{-1/2}\mathbf{W}_N(t, \hat{\boldsymbol{\theta}}).$$

Then under H_0,

$$\mathbf{M}_N(t, \hat{\boldsymbol{\theta}}) \rightarrow_d \mathbf{W}^0(t),$$

where $\mathbf{W}^0(t) = (W_1^0(t), \ldots, W_p^0(t))$ *is a vector with p independent standard Brownian Bridges as component processes.*

Since the limiting distribution is the vector of independent Brownian Bridge process, individual components of $\mathbf{M}_N(t, \hat{\boldsymbol{\theta}})$ is distributed as a standard Brownian Bridge, $W^0(t)$. That is,

$$M_N(t, \hat{\theta}_k) \rightarrow_d W^0(t) \quad k^{th} \quad (k = 1, \ldots, p).$$

The above weak convergence continues to hold for any 'reasonable' functionals (including supremum) of $M_N(t, \hat{\theta}_k)$ (see e.g. pp 509, Theorem 3.1 in [33]). Therefore,

$$D_k \equiv \max_{0 \le t \le 1} |M_N(t, \hat{\theta}_k)| = \max_{1 \le j \le N-1} |M_N(t, \hat{\theta}_k)| \rightarrow_d \max_{0 \le t \le 1} |W_k^0(t)| \equiv D. \tag{4}$$

D has known distribution function [32]

$$F_D(x) = 1 + 2\sum_{l=1}^{\infty} (-1)^l \exp(-2\, l^2 x^2).$$

Although this expression involves an infinite series, this series converges very rapidly. Usually, a few terms suffice for very high accuracy. This result can be used to formulate a test for instability of parameters at α level of significance as follows: (1) Calculate the value of the process D_k for each parameter $k = 1, \ldots, p$ and obtain the raw p-values. (2) Adjust the p-values according to a chosen multiple testing procedure. (3) Reject H_0 if the adjusted p-value for any of the processes, D_k, is less than α.

3.3. Instability test for multiple partitioning variables

In practice, we expect to have multiple partitioning variables. Let there be S partitioning variables: $\{X_1^{G_1}, \ldots, X_S^{G_S}\}$. We perform the p-value from instability test for each partitioning variable separately and adjust the p-values to control type-I error rate. Let the adjusted p-values be $p_1, \ldots, p_S$, respectively and $p_{min} = \min\{p_1, \ldots, p_S\}$. Then the partitioning variable with the smallest p-value (p_{min}) will be chosen as a partitioning variable for splitting if p_{min} is smaller than the nominal significance level. For further discussion please see Section 4.

3.4. Power under the alternative hypothesis

We consider the following form of Pitman's local alternatives [34] in the vicinity of H_0

$$\boldsymbol{\theta}_{(g)} = \boldsymbol{\theta}_0 + \boldsymbol{\delta} \circ \mathbf{h}\left(\frac{c_{(g)}}{c_{(G)}}\right)\frac{1}{\sqrt{N}} + O\left(\frac{1}{N}\right), \tag{5}$$

where $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)^\top$ is the vector containing degrees of departure from the null hypothesis and $\mathbf{h} = (h_1, \ldots, h_p)^\top$ is the vector containing magnitudes of departure. The operation $\circ$ denotes the point-wise multiplication, i.e.

$$\boldsymbol{\delta} \circ \mathbf{h}\left(\frac{c_{(g)}}{c_{(G)}}\right) = \left[\delta_1 h_1\left(\frac{c_{(g)}}{c_{(G)}}\right), \ldots, \delta_p h_p\left(\frac{c_{(g)}}{c_{(G)}}\right)\right]^\top$$

.

Theorem 3.2: *Under* (5), *the limiting distribution for the* χ^2_{cat} *is a non-central chi-square distribution*

$$\chi^2_{cat} \longrightarrow_d \chi'^2\left[(G-1)p, \sum_{g=1}^{G}\lambda_g^2\right],$$

where $\lambda_g = \mathbf{J} \cdot m_g \mathbf{h}(c_{(g)}/c_{(G)}) \cdot 1/\sqrt{N}$

Theorem 3.3: *Under* (5), *the limiting distribution for the canonical monitoring process is as follows*

$$\mathbf{M}_N(t, \hat{\boldsymbol{\theta}}) \longrightarrow_d \mathbf{J}^{1/2} \cdot t_g \cdot \boldsymbol{\delta} \circ (\bar{\mathbf{h}}_g - \bar{\mathbf{h}}) + \mathbf{W}^0(t) \qquad t \in [t_g, t_{g+1}),$$

where, $\bar{\mathbf{h}}_g = (1/M_g)\sum_{j=1}^{g} m_j \mathbf{h}(c_{(j)}/c_{(G)})$ *and* $\bar{\mathbf{h}} = \bar{\mathbf{h}}_G$

Proofs of these theorems are provided in the Appendix.

4. Longitudinal regression tree

We describe the proposed LongCART algorithm in Section 4.1 emphasizing the use of the parameter instability test. We provide a modified Akaike Information Criterion (AIC_T) in Section 4.2 to be used in model comparisons. Tree pruning is described in Section 4.3.

4.1. LongCART algorithm

When more than one partitioning variable is found to be significant at level α based on the parameter instability test, the LongCART selects the partitioning variable with the smallest p-value to split the node. Similar p-value methods have been used in other tree algorithms [8,13,17]). The advantage of p-value approach is that it offers unbiased partitioning variable selection when the partitioning variables are measured at different scales [8]. We propose the following algorithm to construct a regression tree for longitudinal data.

Step 1. Obtain the instability test's p-value for each partitioning variable separately. If there are multiple partitioning variables, adjust the α level that the p-values are compared to.

Step 2. Stop if no partitioning variable is significant at level α. Otherwise, choose the partitioning variable with the smallest p-value and proceed to Step 3.

Step 3. Consider all cut-off points of the chosen partitioning variable. At each cut-off point, calculate the improvement in the goodness-of-fit criterion (e.g. AIC) due to splitting.

Step 4. Choose the cut-off value that provides the maximum improvement in goodness-of-fit criterion and use this cut-off for binary splitting.

Step 5. Follow the Steps 1–4 for each non-terminal node.

4.2. Improvement

A measure of improvement due to regression tree can be provided in terms of likelihood function based criterion. For example, Akaike Information Criterion (AIC) for a tree T can be obtained as

$$\mathrm{AIC}_T = 2\sum_{k=1}^{|T|} l_k - 2\cdot|T|\cdot p,$$

where $|T|$ denotes the number of terminal nodes in T, l_k is the log-likelihood in kth terminal node and p is the number of estimated parameters in each node. If we denote the AIC obtained from the traditional linear mixed effects model without including partitioning variables as covariates at root node (that is, common parametric form for mean structure for the entire population) by AIC_0, the improvement due to regression tree can be measured as

$$\text{Improvement } (T) = \mathrm{AIC}_T - \mathrm{AIC}_0.$$

Since the overall model fitted to all the data is nested within the regression tree based model, a likelihood ratio test or test for deviance can be constructed as well to evaluate the overall significance of a given regression tree.

4.3. Pruning

The improvement in regression tree comes at a cost of adding complexity to the model. If we can summarize the complexity of a tree by the number of terminal nodes, the cost-adjusted AIC of a regression tree T can be defined as follows

$$\mathrm{AIC}_T(\gamma) = \mathrm{AIC}_T - \gamma(|T| - 1), \quad \gamma > 0,$$

where γ be the *cost* for each terminal node. The tree T offers improvement in terms of cost-adjusted AIC as long as $AIC_T(\gamma) > AIC_0$ where AIC_0 is AIC obtained over all data points at root node (i.e. without any tree structure). This is the case when $\gamma < (AIC_T - AIC_0)/(|T| - 1) \equiv \gamma_T^{max}$. In other words, the tree T stands beneficial as long as cost per each terminal node does not exceed γ_T^{max}. With this measure, one can choose the tree $\tilde{T}$ which offers maximum cost-adjusted AIC, as follows:

$$\tilde{T} : \gamma_{\tilde{T}}^{max} \geq \gamma_T^{max}, \quad \forall\ T.$$

5. Simulation study

We have explored the performance of instability test for continuous partitioning variables and the performance of the proposed LongCART algorithm as a whole through simulation studies.

5.1. Performance of instability test with continuous partitioning variable

Let X^G be a continuous partitioning variable with ordered cut-off points as $c_{(1)} \leq \ldots \leq c_{(G)}$. We first investigated the size of the test and then evaluated the power.

5.1.1. Size of the test

In order to examine the size of the test, we have considered a longitudinal model with single mean parameter. The observations for N subjects at $t = 0,1,2,3$ were generated from the following model

$$X^G = c_{(g)} : \; y_{it} = \beta_0 + b_i + \epsilon_{it} \tag{6}$$

with $\beta_0 = 2, b_i \sim N(0, 0.5^2)$ and $\epsilon_{it} \sim N(0, 0.2^2)$. The observations for X^G were generated from uniform(0,300). This simulation study was carried out for different sample sizes (N). For each N,10,000 Monte-Carlo samples were generated and in each sample, parameter instability test considering X^G as a partitioning variable was carried out as described in Section 3.2.

The observed percentiles of test statistic, D_k, and the size of the instability test are summarized in Table 1. In addition, the critical value D_α for test statistic at α level of significance (based on standard brownian bridge process; see Equation 4) were also provided. We can make following observations: (1) the size of the test does not exceed the nominal level, (2) the size of the test approaches to the desired significance level α with the increase in the sample size N and (3) the test is under-sized for smaller sample sizes. The severe problem with the size of the test for smaller sample size can be explained as follows. Calculation of test statistic, D_k, involves σ^2 and $\mathbf{V}_i$. However, in practice, the true values of σ^2 and $\mathbf{V}_i$ are unknown and we replace them by their estimates. A consistent estimator (e.g. ML- or REML-based) approaches the true value with an increasing sample size. However, the estimates might be biased for smaller sample sizes. To be precise, for smaller sample size, σ^2 and $\mathbf{V}_i$ may remain underestimated and this leads to smaller value of D_k which in turn results in a smaller size of the test. However, bias in the estimation of σ^2 and $\mathbf{V}_i$ fades away with the increase in N and this increases the size of the test. We observe this trend in Table 1 as the size of test approaches the nominal level of type I error with the increase in sample size. However, the size of test remains smaller than nominal level even for the reasonably large N. The reduced size has been also reported in other tests based on the Brownian Bridge process. For example, Kolmogorov–Smirnov test for normality (which also uses the Brownian Bridge as limiting distribution) is conservative [35–37]. As N exceeds 500, the size of the test is close to the nominal level of significance. As a remedy for smaller sample sizes, one might consider using a liberal α level or small sample distribution for D_k obtained through simulation.

Table 1. Observed size of proposed parameter instability test for continuous partitioning variable based on 10,000 simulations as discussed in Section 5.1.1.

(a) Observed size of test (%)					
			N		
α(%)	50	100	200	500	1000
1.25	0.54	0.56	0.89	1.02	0.95
1.67	0.75	0.85	1.10	1.33	1.29
2.50	1.20	1.46	1.77	2.04	1.94
5.00	2.78	3.35	3.48	4.07	4.19
10.00	5.66	7.14	7.19	8.37	8.53
20.00	13.05	14.73	15.83	16.97	17.14

(b) Observed $(1-\alpha)100^{th}$ percentile of test statistic (D_k)						
				N		
α(%)	D_α	50	100	200	500	1000
1.25	**1.5930**	1.4760	1.4938	1.5366	1.5643	1.5504
1.67	**1.5472**	1.4447	1.4532	1.4891	1.5147	1.4986
2.50	**1.4802**	1.3722	1.3998	1.4180	1.4392	1.4412
5.00	**1.3581**	1.2497	1.2924	1.2934	1.3154	1.3287
10.00	**1.2238**	1.1236	1.1585	1.1629	1.1901	1.1857
20.00	**1.0728**	0.9859	1.0045	1.0194	1.0350	1.0373

Note: Data were generated with constant intercept as per Equation (6). α and D_α indicate the nominal level of type I error and the critical value for test statistic D_k (based on standard brownian bridge process; see Equation 4), respectively. The simulation results are summarized for (a) Observed size of test (to be compared with α) and (b) observed $(1-\alpha)100^{th}$ percentile of D_k (to be compared with D_α). The proposed parameter instability test seems to be conservative; however, the size of the test approaches to nominal level with the increase in *N*.

5.1.2. Power

For this simulation, we considered constant intercept (β_0), but the slope (β_1) was dependent on continuous variable X_G. The data were generated for N subjects at $t = 0,1,2,3$ from the following model

$$\begin{aligned} X^G = c_{(g)} : \; y_{it} &= \beta_{0(g)} + \beta_{1(g)} t + b_i + \epsilon_{it}, \\ \beta_{0(g)} = \beta_0 \quad \beta_{1(g)} &= \beta_1 + \delta \cdot \frac{c_{(g)}}{c_{(G)}} \end{aligned} \tag{7}$$

We set $\beta_0 = 1$ and $\beta_1 = 2$. b_i, ϵ_{it} and X^G were generated similarly as before in Section 5.1.1. The parameter δ is indicator of degree of heterogeneity in β_1. The parameter β_1 is not homogeneous unless $\delta = 0$. Positive (negative) value of δ indicates increase (decrease) in β_1 with increase in X_G.

We have two parameters in the mean structure: β_0 and β_1, therefore, we can construct two instability tests – one for β_0 and another for β_1 (see Section 3.2). The p-values were adjusted according to the Hochberg's step-up procedure [38] to control the overall type I error rate at 5% level. We chose Hochberg's step-up procedure because it is relatively less conservative than the Bonferroni procedure [39]. However, in principle, any multiple comparison procedure can be applied here.

The observed power based on 10,000 simulation are displayed in Table 2. The Table 2 represents the observed power of parameter instability test associated with β_0 (i.e. $H_0 : \beta_0$ is homogeneous), β_1 (i.e. $H_0 : \beta_1$ is homogeneous) and overall instability test (H_0 : both β_0 and β_1 are homogeneous). As the absolute value of δ deviates from zero, the power to

Table 2. Observed power (%) of parameter instability test with continuous partitioning variable obtained based on 10,000 simulations as described in Section 5.1.2.

		Observed power (%)					
	Parameter instability				δ		
N	test	0	.25(−.25)	.50(−.50)	.75(−.75)	1.0(−1.0)	1.2(−1.2)
	for β_0	1.4	1.4(1.4)	1.6(1.6)	1.9(1.9)	2.3(2.3)	2.4(2.3)
50	for β_1	1.6	4.4(4.3)	16.9(16.6)	41.9(42.0)	70.2(70.6)	86.9(87.0)
	overall	2.9	5.6(5.5)	17.9(17.6)	42.6(42.5)	70.5(70.8)	87.0(87.1)
	for β_0	1.5	1.6(1.6)	2.0(2.1)	2.5(2.6)	3.0(3.0)	3.2(3.2)
100	for β_1	1.7	5.2(5.3)	18.7(19.7)	44.4(46.0)	72.9(73.9)	88.9(89.0)
	overall	3.1	6.6(6.7)	19.8(20.8)	45.0(46.6)	73.1(74.2)	89.0(89.1)
	for β_0	1.8	1.9(1.8)	2.2(2.2)	2.7(2.7)	3.3(3.3)	3.5(3.4)
200	for β_1	1.9	5.6(5.3)	20.7(19.8)	47.5(46.8)	75.7(75.2)	90.1(89.8)
	overall	3.6	7.4(6.8)	21.9(21.0)	48.2(47.4)	76.0(75.4)	90.6(89.9)
	for β_0	2.1	2.1(2.2)	2.7(2.5)	3.2(3.2)	3.6(3.7)	3.9(4.0)
500	for β_1	1.8	6.1(6.0)	21.4(20.1)	48.1(48.2)	76.6(76.6)	91.1(91.1)
	overall	3.7	7.8(7.8)	22.8(22.2)	48.8(49.1)	77.0(77.0)	91.3(91.2)

Note: Data were generated with constant intercept (β_0) and slope (β_1) dependent on continuous variable X_G (see Equation (7)). δ indicates the degree of departure for β_1 from homogeneity. $\delta = 0$ indicates β_1 does not depend on X_G. Positive (negative) values of δ indicate increase (decrease) in β_1 with increase in X_G. The table represents the observed power of parameter instability test associated with β_0 (i.e. $H_0 : \beta_0$ is homogeneous), β_1 (i.e. $H_0 : \beta_1$ is homogeneous) and overall instability test (H_0 : both β_0 and β_1 are homogeneous).

reject the homogeneity of β_1 and the power for overall parameter instability test increase. The power of test is close to 80% and approaching the 90% mark when $|\delta| > 1$.

Note that the sign of δ does not influence the power of the test. Further, the size of the test is very much in agreement with the first simulation study. As observed previously, the test is mildly conservative in the current simulation scenario as the observed level size of the test (see the power corresponding to $\delta = 0$ in Table 2) is consistently slightly below the nominal level of $\alpha = 0.05$.

5.2. Performance of regression tree for longitudinal data

In this simulation, we have evaluated the performance of LongCART algorithm compared to existing tree algorithms and linear mixed effects models when the population under consideration is truly heterogeneous. Following existing tree construction algorithms are considered: MVPART algorithm [22] (using `mvpart()` in `mvpart` package [40]), RE-EM tree method [26] (using `REEMtree()` in `REEMtree` package), unbiased RE-EM tree [28] (using `REEMctree()` available at http://people.stern.nyu.edu/jsimonof/unbiasedREEM/) and GLMM tree algorithm [29] (using `lmertree()` in `glmrtree` package).

We have simulated observations for $N = 300$ subjects and these subjects come from one of the four different subgroups defined by baseline characteristics X_1, X_2 (continuous) and X_3 with group sizes $f_1 = 70, f_2 = 50, f_3 = 50$ and $f_4 = 130$, respectively. Description of these subgroups is displayed in the form of a tree structure in Figure 2. In rth subgroup ($r = 1, \ldots, 4$), the values for continuous response variable y were generated at $t = 0,1,2,3$ according to following model:

$$y_{it} = \beta_{0r} + \beta_{1r} t + b_i + \epsilon_{it}; \quad i = 1, \ldots, f_r \tag{8}$$

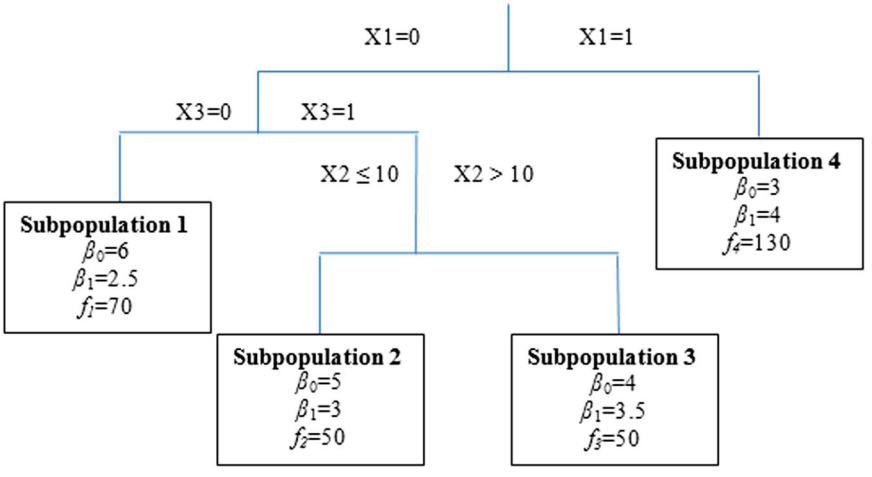

Figure 2. True tree structure for the simulation described in section 5.2. In *rth* subgroup, f_r observations were generated according to Equation (8) with specified β_0 and β_1.

where $b_i \sim N(0, 4)$ and $\epsilon_{it} \sim N(0, 1)$. As displayed in Figure 2, the true values of β_1 were set at 2.5,3.0,3.5 and 4.0 and for β_0, the true values were set at 6,5,4 and 3, for the four subgroups, respectively. The values for X_1 were set to 0 in subgroups 1–3 and to 1 in subgroup 4. The observations for X_2 were generated from Uniform(5, 15), Uniform(5, 10), Uniform(10, 15) and Uniform(5, 15), for subgroups 1–4, respectively. Baseline covariate X_3 takes value 0 for subgroup 1, and 1 for subgroups 2 and 3. For subgroup 4, observations for X_3 were generated from Bernoulli (0.5). In addition, we have also generated observations for additional two baseline covariates, X_4 and X_5, from Bernoulli(0.5) and Uniform(0,15), respectively, for entire population. Regression trees were constructed using X_1, X_2, X_3, X_4 and X_5 as partitioning variables. Further, LongCART was fitted with the following specifications: (1) the overall significance level of instability test was set at 5%, (2) minimum node size for further split was set at 40 and (3) minimum terminal node size was set at 20.

The simulation results comparing LongCART with the other existing algorithms are summarized in Table 3 and Figure 3 based on 1000 simulations. Algorithms such as MVPART, RE-EM tree and unbiased RE-EM tree generate regression trees with estimated mean at each time point for each terminal node, but they do not provide estimates of the

Table 3. Comparison of LongCART algorithm with the other tree fitting algorithms as described in Section 5.2.

		Number(%) of extracted subgroups			
	Median nodes	3	4*	5	MAPE
MVPART[22]	2	183 (18.3%)	54(5.4%)	20(2.0%)	0.338
RE-EM Tree [26]	9	0	0	0	0.235
unbiased RE-EM Tree[28]	12	0	0	0	0.192
GLMM tree[29]	3	167(16.7%)	155(15.5%)	102(10.2%)	0.400
LongCART	4	14(1.4%)	927(92.7%)	58(5.8%)	0.206

Notes: Simulation results are based on 1000 simulated datasets; MAPE: Mean absolute prediction error. *True number of node was 4.

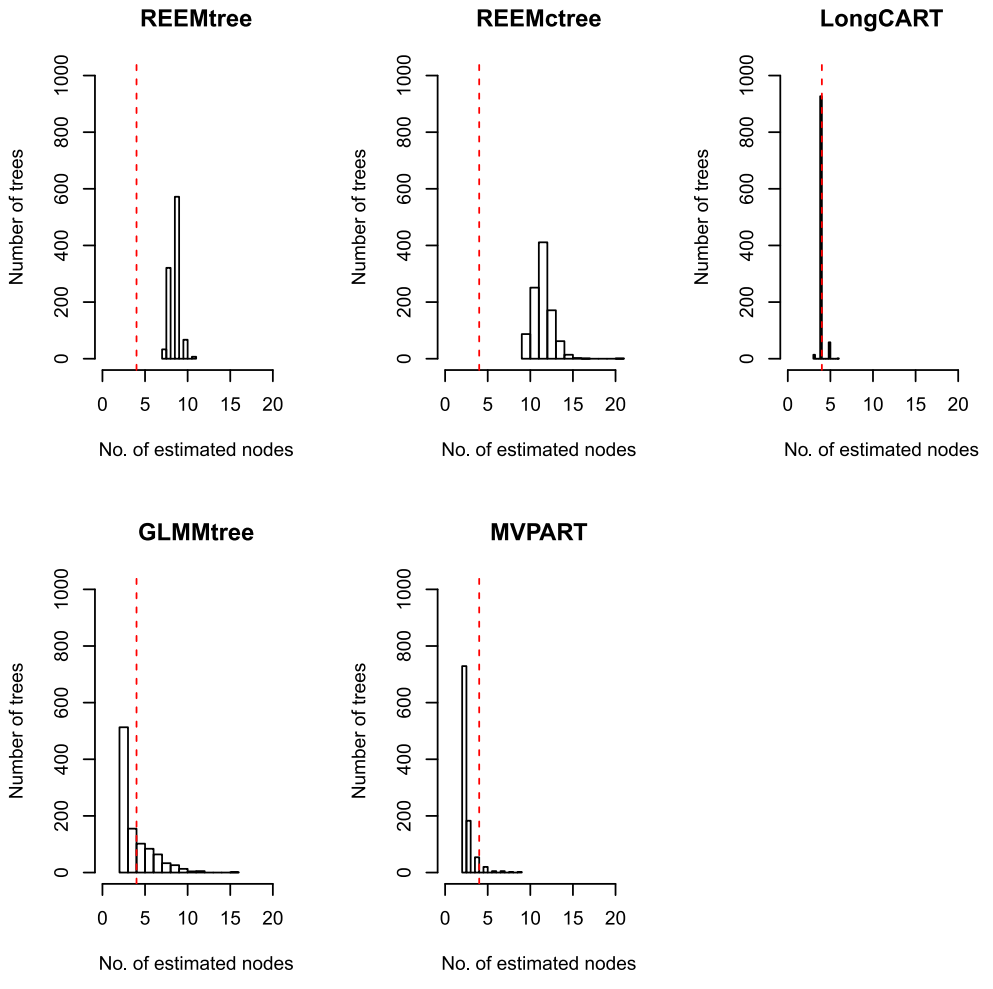

Figure 3. Number of tree nodes estimated by LongCART algorithm and other tree fitting algorithms as described in Section 5.2. The dotted line indicates the true number of nodes equal to 4.

regression coefficients and hence, comparisons were made using the mean absolute prediction error over the fixed timepoints ($t = 0,1,2,3$). In terms of number of nodes extracted by different tree algorithms, LongCART algorithm performs the best compared to the other methods. The LongCART algorithm extracted exactly the assumed four subgroups in 92.7% of the cases. Five subgroups were extracted in 5.8% of the cases and there were only 1.4% instances when three subgroups were extracted. In general, MVPART and GLMM tree algorithms underestimated the number of subgroups and consequently they had larger mean absolute prediction errors. On the other hand, RE-EM tree and unbiased RE-EM tree algorithms overestimated the numbers of subgroups with the medians equal to 9 and 12, respectively, leading to spurious splitting.

For the comparison with the standard linear mixed effects models, we considered seven linear mixed models (Model 1–Model 7). These models along with summary simulations results are presented in Table 4. To study the comparative performance of LongCART algorithm, we calculated the mean absolute deviation (MAD) in β_0 and β_1 in rth subgroup

Table 4. Comparison of LongCART algorithm with linear mixed effect models (Models 1–7) as described in section 5.2.

	Predictors
Model 1	t
Model 2	t, X_1, X_2, X_3
Model 3	$t, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3$
Model 4	$t, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3, X_1X_2X_3$
Model 5	$t, X_1, X_2, X_3, tX_1, tX_2, tX_3$
Model 6	$t, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3, tX_1, tX_2, tX_3, tX_1X_2, tX_1X_3, tX_2X_3$
Model 7	$t, X_1, X_2, X_3, X_1X_2, X_1X_3, X_2X_3, X_1X_2X_3, tX_1, tX_2, tX_3, tX_1X_2, tX_1X_3, tX_2X_3, tX_1X_2X_3$

		Subgroup 1		Subgroup 2		Subgroup 3		Subgroup 4	
	p	ϕ_0	ϕ_1	ϕ_0	ϕ_1	ϕ_0	ϕ_1	ϕ_0	ϕ_1
LongCART	8*	0.236	0.303	0.293	0.153	0.056	0.090	0.072	0.031
GLMM tree	6	0.892	0.185	1.113	0.369	0.443	0.060	0.553	0.049
Model 1	2	1.456	0.631	0.332	0.905	0.902	0.402	0.098	0.598
Model 2	5	1.358	0.660	0.281	0.906	0.902	0.402	0.098	0.598
Model 3	8	0.424	0.215	0.648	0.435	0.198	0.068	0.319	0.208
Model 4	9	0.274	0.321	0.324	0.290	0.071	0.121	0.118	0.072
Model 5	8	1.358	0.645	0.283	0.908	0.902	0.402	0.098	0.598
Model 6	14	0.284	1.195	2.104	1.488	0.059	1.667	2.919	2.367
Model 7	16	0.284	0.320	0.323	6.861	1.985	1.290	2.235	2.181

Note: $\phi_0 =$ Average MAD$(\hat{\beta}_0)$, $\phi_1 =$ Average MAD$(\hat{\beta}_1)$, p: No.of parameters in mean structure.

for each simulation as defined below

$$\text{MAD}(\hat{\beta}_{0r}) = \frac{1}{f_r}\sum_{j\in S_r}|\beta_{0r} - \hat{\beta}_{0j}|,$$

$$\text{MAD}(\hat{\beta}_{1r}) = \frac{1}{f_r}\sum_{j\in S_r}|\beta_{1r} - \hat{\beta}_{1j}|,$$

where β_{0r} and β_{1r} are the true values of β_0 and β_1 in the rth subgroup and $\hat{\beta}_{0j}$ and $\hat{\beta}_{1j}$ are the corresponding estimates for the jth individual applying longitudinal tree and then fitting mixed model in each subgroup. S_r is the set of indices for all individuals in the rth subgroup while f_r denotes its size.

The application of the LongCART algorithm shows comparatively larger improvements in the estimation of the coefficients in all four subgroups. Both the MAD$(\hat{\beta}_0)$ and MAD$(\hat{\beta}_1)$ were considerably smaller in LongCART compared to the Models 1–7. The improvement in the estimation of coefficients in regression tree was attributed to its ability to extract homogeneous subgroups and then fitting mixed model separately within each group. Models 1–7 includes either additive (Models 1–2) or an interaction (Models 3–7) effects; yet these models failed to capture the complexity of heterogeneous population, especially, in presence of continuous partitioning variable. Model 5 including the interaction terms with t and the partitioning variables is probably the most commonly used model in practice. However, the application of the LongCART algorithm offers a considerable improvement in the estimation compared to Model 5. Models 5–6 provide some improvement over regression tree in some of the subgroups. However, these improvements are comparatively rare and largely influenced by the fact how the subgroups are defined.

Apart from providing an improvement in estimation, the LongCART algorithm also identifies the meaningful subgroups defined by the partitioning variables which would remain unidentified otherwise.

6. Application

We applied the LongCART algorithm to study the changes in the concentration of a brain metabolite *choline* in gray matter among HIV patients enrolled in the HIV Neuroimaging Consortium (HIVNC) study [41]. Concentrations of choline were obtained via magnetic resonance spectroscopy (MRS). Choline is considered to be a marker of brain inflammation. It has been found in previous studies that the concentrations of choline were elevated in all three brain regions among HIV patients [42]. We considered a total of $\sum_{i=1}^{N} n_i = 780$ observations from $N = 239$ subjects. All the observations were within 3 years from baseline. The number of observations per subject ranged from 2 to 6 with median equal to 3. We estimated the overall significant decrease of 0.077 arbitrary unit (AU) per year (p-value = 0.003) in choline concentration suggesting overall beneficial effect of the antiretroviral therapy.

For the construction of regression tree we used baseline measurements of several clinical and demographic variables including sex, race, education, age, current CD4 count, nadir CD4 count, duration of HIV infection, duration of antiretroviral (ARV) treatment, duration of highly active antiretroviral therapy (HAART), plasma HIV RNA count, antiretroviral CNS penetration-effectiveness (CPE) score and AIDS dementia complex (ADC) stage as partitioning variables. In each node, we consider fitting the following model separately

$$y_{it} = \beta_0 + \beta_1 t + b_i + \epsilon_{it}, \tag{9}$$

where y_{it} indicates the measurement of the concentration of choline from the ith individual at time t (in years) and b_i is the subject-specific intercept. It was assumed that b_i and ϵ_{it} are independently and normally distributed with mean equal to zero. LongCART algorithm was applied with the following specifications: (1) the significance level for individual instability test was set to 5%, (2) the minimum node size for further split was set to 50 and (3) the minimum terminal node size was set to 25.

Figure 4 displays the estimated longitudinal regression tree with the estimates of β_0 and β_1 for each terminal node or subgroup and the plot of estimated linear trajectories within each subgroup. Duration of ARV treatment (p-value = 0.004) and HAART (p-value = 0.004) seem to influence the change in concentration of choline over time. Improvement in deviance due to an application of LongCART algorithm was 519 (log-likelihoods were −1427 vs. −1687; with 4 degrees of freedom). ARV treatment for over 7.5 years not only helped to reduce baseline concentration of choline but also resulted in a significant decrease of 0.094 per year (p-value = 0.015). A higher baseline value of choline concentration was observed among those who received ARV treatment for at most 7.5 years; however, a longer period of HAART therapy in them led to significant decrease of 0.196 per year (p-value = 0.041) in concentration over time. We did not observe any decrease among those who received ARV treatment for less than 7.5 years and HAART therapy for 2.64 years.

In summary, both the longer duration of ARV treatment and HAART resulted in the reduction of choline concentration. However, the rate of reduction is almost doubled

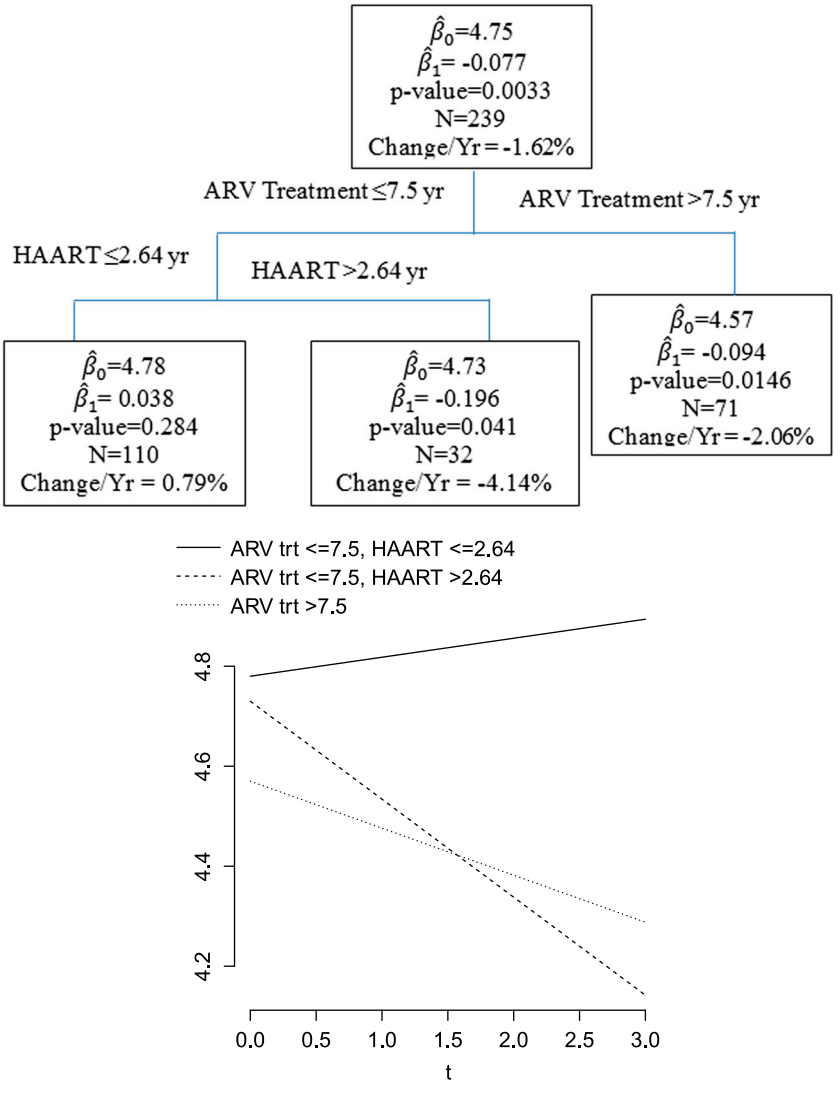

Figure 4. *Top panel.* Longitudinal regression tree obtained via LongCART algorithm for longitudinal change in *choline* concentration as discussed in Section 6. The *p*-value in each node corresponds to the estimate of the slope β_1. *Bottom panel.* Estimated linear trajectory for longitudinal change within each subgroup obtained via fitting mixed effect model of the form Equation (9). This regression tree suggests the duration of ARV treatment and HAART are significant determinants for the longitudinal change of choline.

(4.14% vs. 2.06%) when patients were on HAART compared to only ARV treatment (see Figure 4). This suggests that both ARV treatment and HAART are effective in controlling brain inflammation via reducing choline concentration. Finally, all these interpretable subgroups along with a significant improvement in overall model fit suggests underlying heterogeneity in the population in terms of longitudinal change in choline concentration. Thus considering a traditional linear mixed effects model for the entire population is not defensible.

7. Discussion

The longitudinal profile in a population may be influenced by several baseline characteristics. This may be true both in observational studies and clinical trials. The most common strategy to incorporate the effect of baseline variables in a traditional linear mixed effects model is to include these baseline characteristics and their interactions with the time-varying variables as covariates in the model. However, this approach has its own limitations as discussed in Section 1. Longitudinal trees, i.e. regression trees for longitudinal data, are extremely useful to identify the heterogeneity in longitudinal trajectories in a nonparametric way. We have proposed LongCART algorithm for the construction of longitudinal tree under conditional inference framework proposed by Hothorn et al. [8]. The LongCART algorithm identifies the splitting variable via formal hypothesis testing controlling type I error at each node; hence it offers protection against variable selection bias, over-fitting and spurious splitting. Additionally, LongCART algorithm substantially reduces the computation time as it first chooses the partitioning variable and then evaluates the goodness-of-fit criterion at all cut-off points of the selected partitioning variable only. Furthermore, the statistical tests implemented in the LongCART algorithm are based on the score process. Therefore, we can extend the scope of LongCART algorithm to other applications as long as we can obtain (or approximate) an expression for the score function and the Hessian matrix in a tractable form including the survival data with censoring, generalized linear mixed effects model (GLMM) and multiple response variable settings.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by National Institute of Mental Health [R01MH108467].

Notes on contributors

Madan Gopal Kundu is a Manager in the Data and Statistical Sciences (DSS) at AbbVie in Chicago, IL, USA.

Jaroslaw Harezlak is a Professor at the Department of Epidemiology and Biostatistics, Indiana University School of Public Health, Bloomington, IN, USA.

ORCID

Madan Gopal Kundu http://orcid.org/0000-0001-6616-5762
Jaroslaw Harezlak http://orcid.org/0000-0002-3070-7686

References

[1] Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982;38: 963–974.
[2] Liang KY, Zeger S. Longitudinal data analysis using generalized linear models. Biometrika. 1986;73(1):13–22.

[3] Diggle P, Heagerty P, Liang K, et al. Analysis of longitudinal data. Oxford: Oxford University Press; 2002.
[4] Raudenbush SW. Comparing personal trajectories and drawing causal inferences from longitudinal data. Annu Rev Psychol. 2001;52(1):501–525.
[5] Muthén B, Shedden K. Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics. 1999;55(2):463–469.
[6] Segal MR. Tree-structured methods for longitudinal data. J Am Stat Assoc. 1992;87(418):407–418.
[7] Abdolell M, LeBlanc M, Stephens D, et al. Binary partitioning for continuous longitudinal data: categorizing a prognostic variable. Stat Med. 2002;21(22):3395–3409.
[8] Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat. 2006;15(3):651–674.
[9] Negassa A, Ciampi A, Abrahamowicz M, et al. Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria. Stat Comput. 2005;15(3):231–239.
[10] Loh W. Regression trees with unbiased variable selection and interaction detection. Stat Sin. 2002;12(2):361–386.
[11] Shih Y-S. A note on split selection bias in classification trees. Comput Stat Data Anal. 2004;45(3):457–466.
[12] Strobl C, Boulesteix A-L, Augustin T. Unbiased split selection for classification trees based on the Gini index. Comput Stat Data Anal. 2007;52(1):483–501.
[13] Loh W-Y, Zheng W, others. Regression trees for longitudinal and multiresponse data. Ann Appl Stat. 2013;7(1):495–522.
[14] Brown RL, Durbin J, Evans JM. Techniques for testing the constancy of regression relationships over time. J Roy Stat Soc Ser B. 1975;Jan 1:149–192.
[15] Nyblom J. Testing for the constancy of parameters over time. J Am Stat Assoc. 1989;84(405):223–230.
[16] Hjort NL, Koning A. Tests for constancy of model parameters over time. J Nonparametr Stat. 2010;14(1-2):113–132.
[17] Zeileis A, Hothorn T, Hornik K. Model-based recursive partitioning. J Comput Graph Stat. 2008;17(2):492–514.
[18] Zeileis A, Hothorn T, Hornik K. party with the mob: model-based Recursive Partitioning in R. *R package vignette, version 1.0-19* 2010; Available at https://cran.r-project.org/web/packages/party/vignettes/MOB.pdf.
[19] Andrews DWK. Tests for parameter instability and structural change with unknown change point. Econometrica. 1993;61821–856.
[20] Hansen B. Approximate asymptotic p values for structuraS-change tests. J Bus Econ Stat. 1997;15(1):60–67.
[21] Breiman L, Friedman J, Stone C, et al. Classification and regression trees. Boca Raton (FL): Chapman & Hall/CRC; 1984.
[22] De'Ath G. Multivariate regression trees: a new technique for modeling species-environment relationships. Ecology. 2002;83(4):1105–1117.
[23] Larsen DR, Speckman PL. Multivariate regression trees for analysis of abundance data. Biometrics. 2004;60(2):543–549.
[24] Zhang H, Singer B. Recursive partitioning in the health sciences. New York (NY): Springer Verlag; 1999.
[25] Zhang H. Multivariate adaptive splines for analysis of longitudinal data. J Comput Graph Stat. 1997;87(418):74–91.
[26] Sela RJ, Simonoff JS. RE-EM trees: a data mining approach for longitudinal and clustered data. Mach Learn. 2012;86(2):169–207.
[27] Galimberti G, Montanari A. Regression trees for longitudinal data with time-dependent covariates. Classi Cluster Data Anal. 2002;Jan 1:391–398.
[28] Fu W, Jeffrey JS. Unbiased regression trees for longitudinal and clustered data. Comput Stat Data Anal. 2015;88:53–74.

[29] Fokkema M, Smits N, Zeileis A, Hothorn T, Kelderman H. Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees (No. 2015-10). Working Papers in Economics and Statistics. University of Innsbruck, Innsbruck, Austria. 2015.
[30] Eo SH, Cho H. Tree-structured mixed-effects regression modeling for longitudinal data. J Comput Graph Stat. 2014;23(3):740–760.
[31] Demidenko E. Mixed models: theory and applications. Hoboken (NJ): Wiley-Interscience; 2004.
[32] Billingsley P. Convergence of probability measures. New York (NY): Wiley-Interscience; 2009.
[33] Csõrgõ M. A glimpse of the impact of pál erd# ous on probability and statistics. Can J Stat. 2002;30(4):493–556.
[34] Pitman E. Notes on non-parametric statistical inference. New York (NY): Columbia University; 1949.
[35] Lilliefors HW. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J Am Stat Assoc. 1967;62(318):399–402.
[36] Massey Jr F. The Kolmogorov-Smirnov test for goodness of fit. J Am Stat Assoc. 1951;46(253):68–78.
[37] Birnbaum ZW. Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. J Am Stat Assoc. 1952;47(259):425–441.
[38] Hochberg Y. A sharper Bonferroni procedure for multiple tests of significance. Biometrika. 1988;75(4):800–802.
[39] Hochberg Y, Tamhane A. Multiple comparison procedures. New York: John Wiley & Sons; 2009.
[40] De'Ath G mvpart. Multivariate partitioning *R package version 0.1-6*. 2013.
[41] Gongvatana A, Harezlak J, Buchthal S, et al. others. Progressive cerebral injury in the setting of chronic HIV infection and antiretroviral therapy. J Neurovirol. 2013;19(3):209–218.
[42] Chang L, Ernst T, Witt M, et al. Relationships among brain metabolites, cognitive function, and viral loads in antiretroviral-naıve HIV patients. Neuroimage. 2002;17(3):1638–1648.

Appendix: Proofs

A.1 Proof of Theorem 3.1

Proof. Under H_0, by applying Taylor series expansion

$$\mathbf{W}_N(t,\hat{\boldsymbol{\theta}}) \doteq \mathbf{W}_N(t,\boldsymbol{\theta}_0) - t\,\mathbf{W}_N(1,\boldsymbol{\theta}_0),$$

where $A_n \doteq B_n$ means that $A_n - B_n$ tends to zero in probability. In the case of linear mixed effects models, this relationship is exact as the second derivative of the score function is equal to 0. That is, $\mathbf{W}_N(t,\hat{\boldsymbol{\theta}}) = \mathbf{W}_N(t,\boldsymbol{\theta}_0) - t\,\mathbf{W}_N(1,\boldsymbol{\theta}_0)$. Consequently,

$$\mathbf{W}_N(t,\hat{\boldsymbol{\theta}}) \to_d \mathbf{Z}(t) - t\cdot\mathbf{Z}(1) \equiv \mathbf{Z}^0(t).$$

The limit process $\mathbf{Z}^0(t)$ is a p-dimensional mean zero Brownian Bridge process with covariance function $\mathrm{cov}[\mathbf{Z}^0(t),\mathbf{Z}^0(s)] = s(1-t)\mathbf{J}(\boldsymbol{\theta}_0)$ for $s<t$. Therefore, under H_0

$$\mathbf{M}_N(t,\hat{\boldsymbol{\theta}}) = \hat{\mathbf{J}}^{-1/2}\mathbf{W}_N(t,\hat{\boldsymbol{\theta}}) \to_d \mathbf{W}^0(t),$$

where $\mathbf{W}^0(t) = (W_1^0(t),\ldots,W_p^0(t))$ is a vector with p independent standard Brownian Bridges as component processes.

A.2 Proof of Theorem 3.2

proof. Using Taylor series expansion we can write

$$f(\mathbf{y},\boldsymbol{\theta}_{(g)}) \doteq f(\mathbf{y},\boldsymbol{\theta}_0)\left\{1+\mathbf{u}(\mathbf{y},\boldsymbol{\theta}_0)^\top\boldsymbol{\delta}\circ\mathbf{h}\Big(\frac{c_{(g)}}{c_{(G)}}\Big)\frac{1}{\sqrt{N}}\right\}$$

Consequently,

$$E_{\boldsymbol{\theta}_g}[\mathbf{u}(\mathbf{y},\boldsymbol{\theta}_0)] = \int u(\mathbf{y},\boldsymbol{\theta}_0) f(\mathbf{y},\boldsymbol{\theta}_{(g)}) dy = E_{\boldsymbol{\theta}_0}[\mathbf{u}(\mathbf{y},\boldsymbol{\theta}_0)] + \mathbf{J}\cdot\boldsymbol{\delta}\circ\mathbf{h}\Big(\frac{c_{(g)}}{c_{(G)}}\Big)\frac{1}{\sqrt{N}}$$
$$= \mathbf{J}\cdot\boldsymbol{\delta}\circ\mathbf{h}\Big(\frac{c_{(g)}}{c_{(G)}}\Big)\frac{1}{\sqrt{N}}. \tag{A1}$$

It can be shown that

$$\mathrm{cov}_{H_1}[\mathbf{W}_N(t,\boldsymbol{\theta}_0)] = \mathrm{cov}_{H_0}[\mathbf{W}_N(t,\boldsymbol{\theta}_0)] + O\left(\frac{1}{N}\right) \doteq \mathbf{J}. \tag{A2}$$

Proof of Theorem 3.2 follows from the definition of non-central chi-square distribution.

A.3 Proof of Theorem 3.3

Proof. Using (A1) and (A2),

$$E_{H_1}[\mathbf{W}_N(t,\boldsymbol{\theta}_0)] = \mathbf{J}\frac{1}{N}\sum_{i=1}^{M_g}\delta\circ\mathbf{h}\Big(\frac{c_{(g)}}{c_{(G)}}\Big) = \mathbf{J}\cdot t_g\cdot\boldsymbol{\delta}\circ\bar{\mathbf{h}}_g \qquad t\in[t_g,t_{g+1}).$$

This time using the FCLT along with Cramer-Wold device we can show that

$$\mathbf{W}_N(t,\boldsymbol{\theta}_0) \longrightarrow_d \mathbf{J}\cdot t_g\cdot\delta\circ\bar{\mathbf{h}}_g + \mathbf{Z}(t) \qquad t\in[t_g,t_{g+1}).$$

Therefore, for $t\in[t_g,t_{g+1})$,

$$\mathbf{W}_N(t,\hat{\boldsymbol{\theta}}) = \mathbf{W}_N(t,\boldsymbol{\theta}_0) - t_g\,\mathbf{W}_N(1,\boldsymbol{\theta}_0) + o_p(1) \longrightarrow_d \mathbf{J}\cdot t_g\cdot\delta\circ(\bar{\mathbf{h}}_g-\bar{\mathbf{h}}) + \{\mathbf{Z}(t) - t\cdot\mathbf{Z}(1)\}.$$

Thus under H_1,

$$\mathbf{M}_N(t,\hat{\boldsymbol{\theta}}) = \hat{\mathbf{h}}^{-1/2}\mathbf{W}_N(t,\hat{\boldsymbol{\theta}}) \longrightarrow_d \mathbf{J}^{1/2}\cdot t_g\cdot\boldsymbol{\delta}\circ(\bar{\mathbf{h}}_g-\bar{\mathbf{h}}) + \mathbf{W}^0(t) \qquad t\in[t_g,t_{g+1}).$$