Tree-structured Regression Modeling for Correlated Data

Soo-Heang EO

Department of Statistics

Korea University

Oct 6, 2014

Table of Contents

Background

Tree-structured Methods

Model-based Trees

Main references

Proposal

References

Correlated Data Analysis

Many kind of data, either observational or from designed experiments have a clustered structure:

- Students in a school
- Patients at a clinic
- Workers in a department
- Repeated measurements on an individual

Observations from the same cluster are possibly correlated while observations from distinct clusters are independent.

Motivating Example 1 - Longitudinal Study

- **Wage data**: Hourly wage of high-school dropouts
- 888 male high-school dropouts (246 Black, 204 Hispanic, 438 White)
- Response is *hourly wage* (in 1990 dollars)
- Predictor variables are:
  - `exper`: duration of the work experience (0.01 ~ 12.70 yrs)
  - `hgc`: highest grade completed (6 ~ 12)
  - `race`: individual's race, namely, White, Black, and Hispanic
- The aim of the study is to determine what predictors, especially **trends over time**, have significant effects on the wage.
- Data from the National Longitudinal Survey of Youth

Motivating Example 1 - Longitudinal Study

- Questions in the analysis of longitudinal data
  - How does the response change over time?
  - Can we predict the differences in these changes?
- Two popular approaches
  - **Parametric**: Fit a mixed model (also called individual growth model, random coefficient model, multilevel model, and hierarchical linear model) and deduce the effect of predictor variables from the regression coefficients (Fitzmaurice, Laird, and Ware, 2011)
  - **Non-parametric**: Cluster the subject trajectories, then test each predictor variable for its effect on the clusters (Genolini and Falissard, 2010)

Motivating Example 1 - Longitudinal Study

Linear mixed model (Singer and Wilett, 2003)

$$\begin{aligned}\log(wage) &= \beta_0 + \beta_1 hgc + \beta_2 exper + \beta_3 black + \beta_4 hisp \\ &\quad + \beta_5 exper \times black + \beta_6 exper \times hisp \\ &\quad + b_0 + b_1 exper + \epsilon,\end{aligned}$$

Assumptions & limitations:

- $b_0 \sim \mathcal{N}(0, \sigma_0^2)$ and $b_1 \sim \mathcal{N}(0, \sigma_1^2)$; all independent
- Log transformation of `wage` to address skewness, linearize individual wage trajectories, and overcome range restriction
- Predictions of `wage` requires exponentiation of fitted values of $\log$(wage): OLS fit on log-dollar scale not best for dollar scale

Motivating Example 1 - Longitudinal Study

"Analyses not shown here suggest that we cannot distinguish statistically between the trajectories of Hispanic and White dropbouts."

by **Singer and Wilett (2003, p. 149)**

	Value	Std.Error	DF	t-value	p-value
(Intercept)	1.382	0.059	5511	23.43	0.000
hgc	0.038	0.006	884	5.94	0.000
exper	0.047	0.003	5511	14.57	0.000
black	0.006	0.025	884	0.25	0.804
hisp	-0.028	0.027	884	-1.03	0.302
exper×black	-0.015	0.006	5511	-2.65	0.008
exper×hisp	0.009	0.006	5511	1.51	0.131

Motivating Example 1 - Longitudinal Study

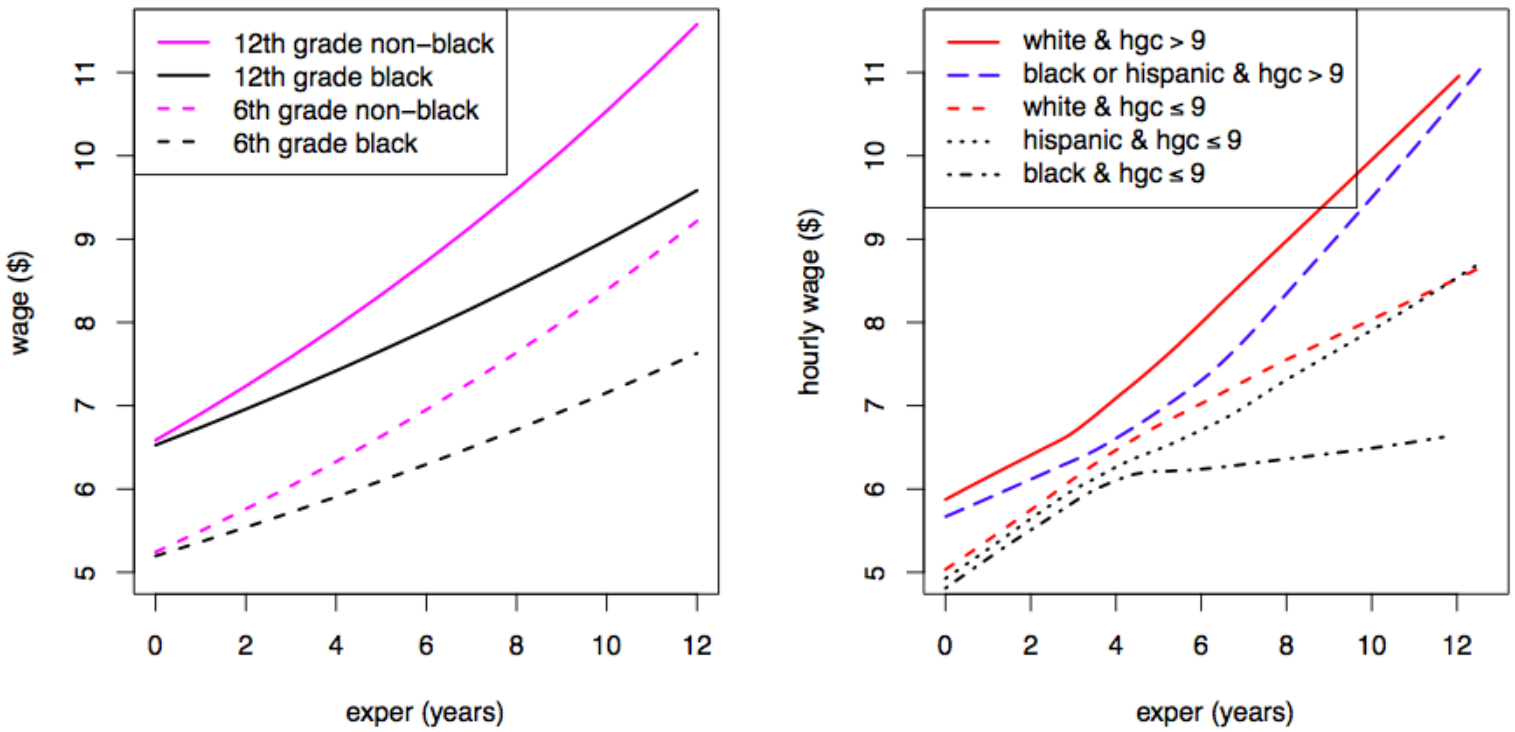
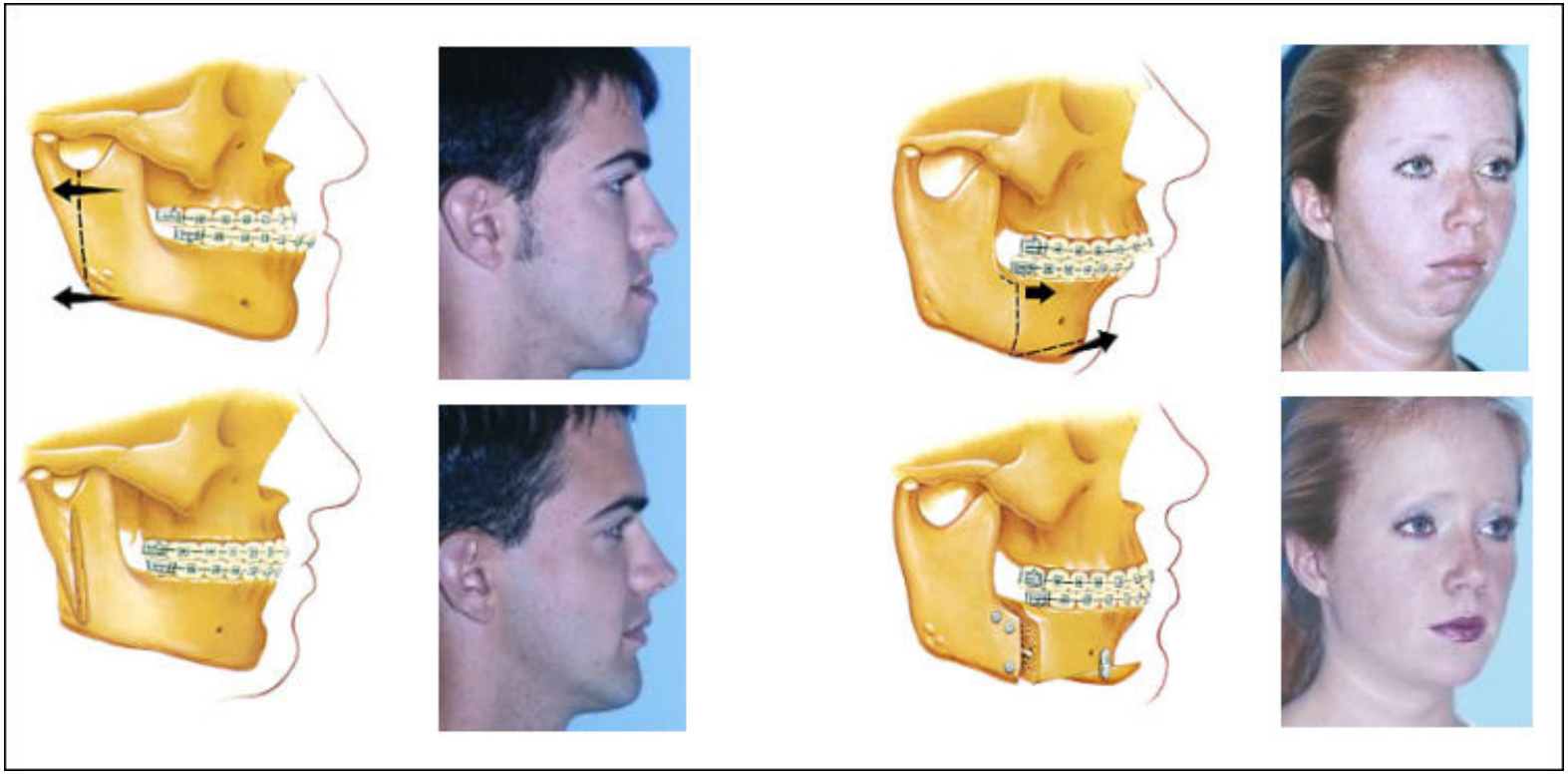


Figure: LME fits

Motivating Example 2 - Orthodontics Study

Cosmetic surgery to correct conditions of the jaw and face ...(from *Wikipedia*).



The goal is to seek functional rehabilitation and facial esthetic. For more details, see `http://www.augersmiles.com`.

Motivating Example 2 - Orthodontics Study

Orthodontics study belongs to landmark-based geometric morphometrics, also known as **landmark data** (Dryden and Mardia, 1992).

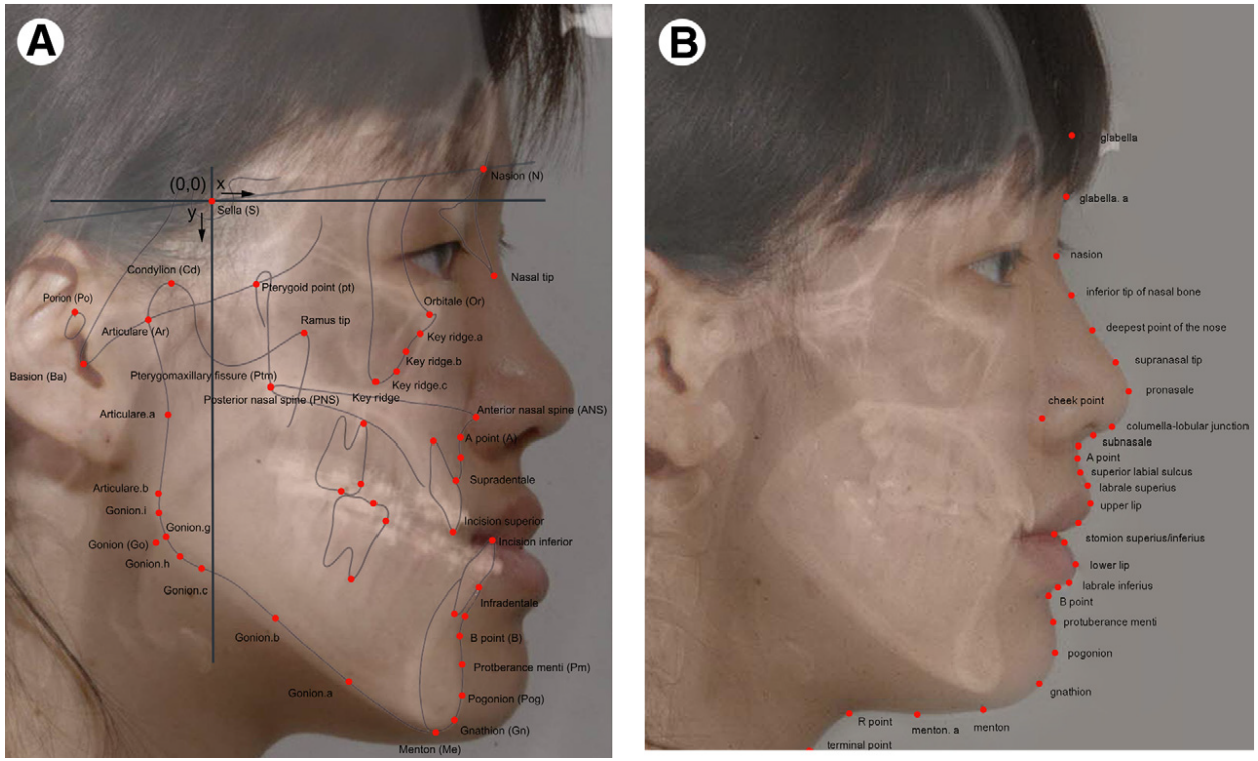


Figure: Skeletal landmarks (left) and skin landmarks (right)

Motivating Example 2 - Orthodontics Study

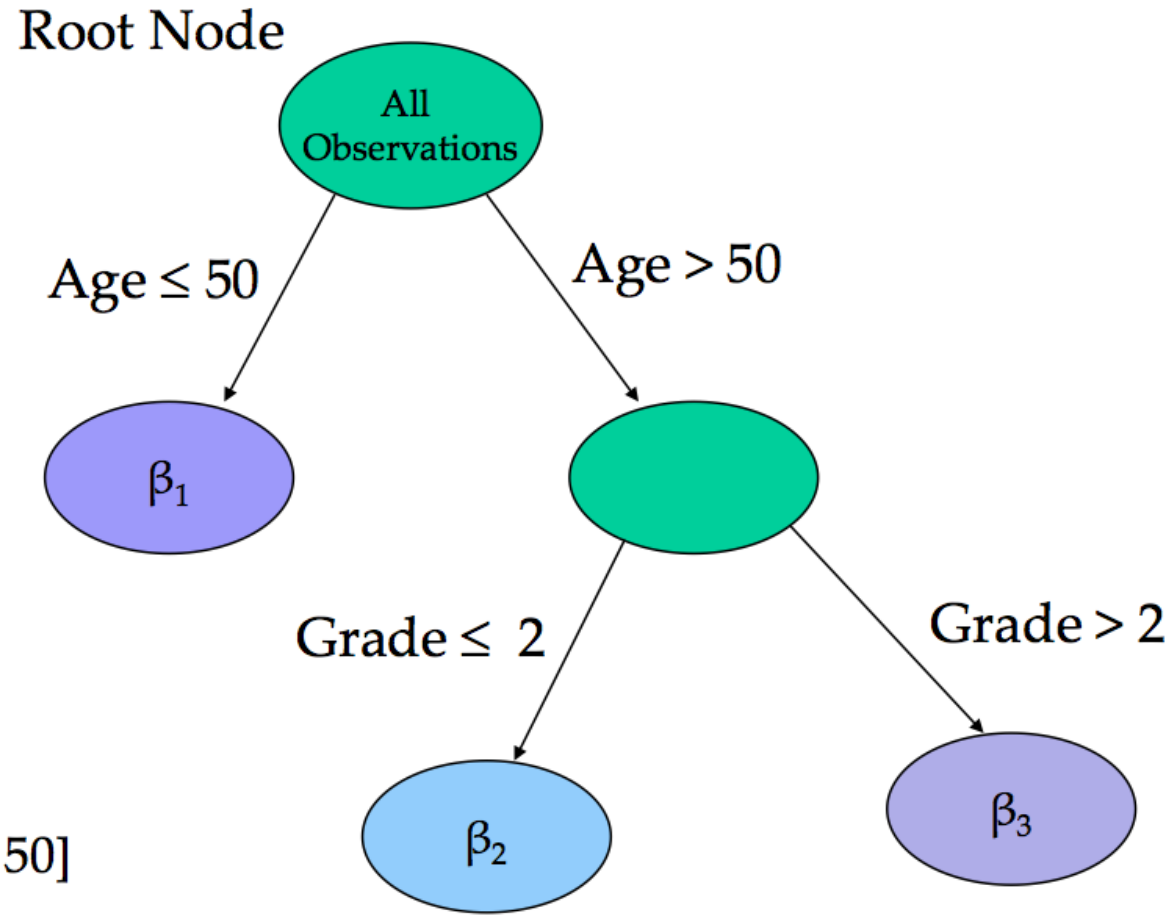
- The data consists of **318** patients collected from Seoul National University Dental Hospital (Suh *et al.*, 2012; Lee *et al.*, 2014; Suh *et al.*, 2014).
- *Our goal is* **to predict multivariate correlated response (Y)** *using* **landmark information** *(X) and* **other treatments** *(Z).*

Type	Variable	Description
Y	$Y_1, \ldots, Y_{64}$	skin landmarks after surgery
X	$X_1, \ldots, X_{64}$	skin landmarks before surgery
	$X_{65}, \ldots, X_{238}$	movements of skeletal landmarks
Z	`age`	age at surgery
	`gender`	gender
	`time`	time after surgery
	`mx`	type of surgery (1:bi-maxillar, 0:not)
	`genio`	type of genioplasity (1:genioplasity, 0:not)
	`class`	type of cranfacial class (2:class2 3:class3)
	`asym`	amount of mandibular asymmetry

Motivating Example 2 - Orthodontics Study

- Questions in the analysis of landmark data
  - How is the response influenced by one another?
  - Can we predict the difference between before and after surgery?
- Previous approaches
  - Ingervall *et al.* (1995) - **correlation analysis**.
  - Chew *et al.* (2008) - **OLS regression**.
  - Suh *et al.* (2012) - **PLS regression**.
    - Harder to interpret
    - Not reveal the relationship between variables

Classification and Regression Trees (CART)



$\beta_1 = E[T \mid Age \leq 50]$

$\beta_2 = E[T \mid Age > 50 \ \& \ Grade \leq 2]$

$\beta_3 = E[T \mid Age > 50 \ \& \ Grade > 2]$

Breiman, et al. (1984)

Classification and Regression Trees (CART)

1. Fit a constant, the node mean $\bar{y}$ at each node
2. Use sum of squared deviations $\sum_i (y_i - \bar{y})^2$ as node impurity
3. Choose the split that maximizes the decrease in node impurity
4. Use the sample $\bar{y}$ in node t as predicted value.
5. Prune tree using test sample or cross-validation
6. Use surrogate spilts to deal with missing values

Classification and Regression Trees (CART)

A *naive* approach is to evaluate the reduction of impurity, $\Delta(\cdot)$, over all possible splits, and select the split set with the greatest reduction of impurity,

$$\underset{A_{|Z}}{\operatorname{argmax}}\,\Delta(t) \quad = \quad R(t) - [R(t_L) + R(t_R)], \tag{1}$$
$$\forall A_{|Z} \in \{a_{1|Z}, a_{2|Z}, \ldots, a_{c|Z}\}, \forall Z \in \{Z_1, Z_2, \ldots, Z_r\},$$

where

- (t, t_L, t_R): a node t and its left and right child node
- Z: a candidate split variable
- $A_{|Z}$: a candidate split set given Z

We call this approach an **exhaustive search (ES)** algorithm (Breiman *et al.*, 1984).

Residual Analysis Tree (GUIDE)

- Proposed by Loh (2002, Sinica; 2009, AOAS) based on Loh and Shih (1996, Sinica) and Kim and Loh (2001, JASA)
- Define a residual from fitting a node model at node t as

$$r_i = y_i - \hat{y}_i, \quad i \in t. \tag{2}$$

- If the fitted model is correct, *the residuals should be randomly distributed* over each split variable. The randomness means that the fitted model is sufficient to explain the data so that no further split is needed.
- If the fitted model is not correct, the residuals have a systematic pattern against Z. Based on residuals, we can select a split variable to segment the whole data space into $\mathcal{B}$ subspaces (by default, $\mathcal{B} = 2$).

$$r_i^s = \text{sgn}(r_i) := \begin{cases} -1, & \text{if } r_i \leq 0, \\ +1, & \text{if } r_i > 0. \end{cases} \tag{3}$$

Residual Analysis Tree (GUIDE)

- Select a partitioning variable by performing main- and interaction- effects tests for **the randomness of residuals** against Z.
- Obtain a test statistic by a main-effects test.

  (a) Divide cases into three or four groups for ordered Z or number of categories for categorical Z.

  (b) Construct a table with signs of residuals as rows and groups as columns.

  (c) Compute the Wilson-Hilferty χ^2_1 approximation statistic, $W_M(Z)$.
- Do an interaction-effects test $W_I(Z_i, Z_j)$ if $\max_i W_M(X_i) \leq \chi^2_{1,\alpha}$

  where $\alpha = 0.05/K$, $\beta = 0.1/(K(K-1))$ and

  K means the number of non-constant split variables

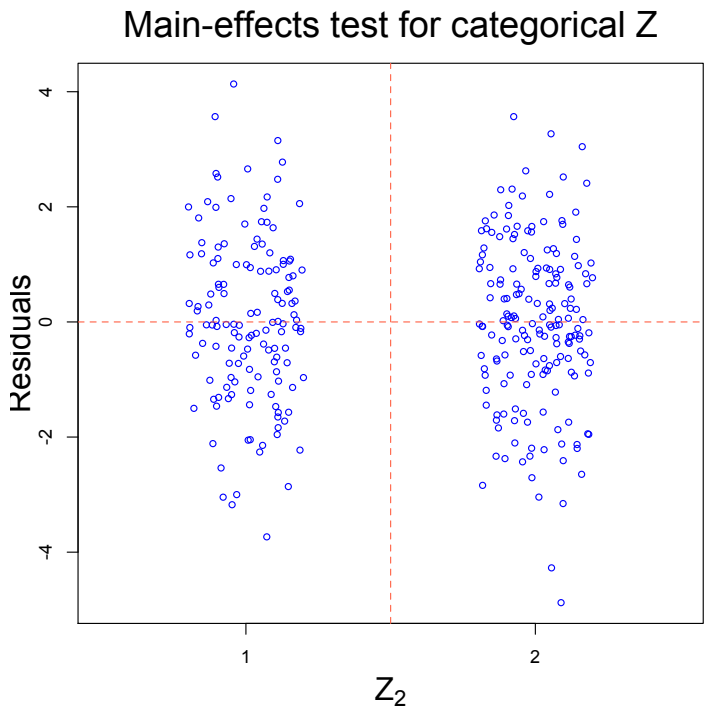
  (a) Find $W_I(Z_i, Z_j)$ for each pair of predictor variables

  (b) $\max_i W_I(Z_i, Z_j) > \chi^2_{1,\beta}$, select the pair with largest $W_I(Z_i, Z_j)$.

  (c) Otherwise, select variable with largest $W_M(Z_i)$ via a Two-level search.

Residual Analysis Tree (GUIDE)

Main-effects test for selecting a split variable Z



Residuals	G_1	G_2	G_3
+	49	88	17
–	39	82	43
$W_M(Z) = 9.42$, p-value $= 0.002$			

Residuals	G_1	G_2
+	49	105
–	42	122
$W_M(Z) = 1.21$, p-value $= 0.271$		

Conditional Inference Tree (CTREE)

- Proposed by Hothorn, Hornik, and Zeilies (2006, JCGS)
- **Use conditional permutation tests to select split variables**
  - Requires computation of p-values, either by exact calculation, Monte Carlo simulation, or asymptotic approximation
- Use stopping rules controlled by Bonferroni adjustments without pruning
- Surrogate splits are used to deal with missing values
- Permutation tests (with subnode label as response variable) are used to find the surrogate variables.

Conditional Inference Tree (CTREE)

Selection of split set or points

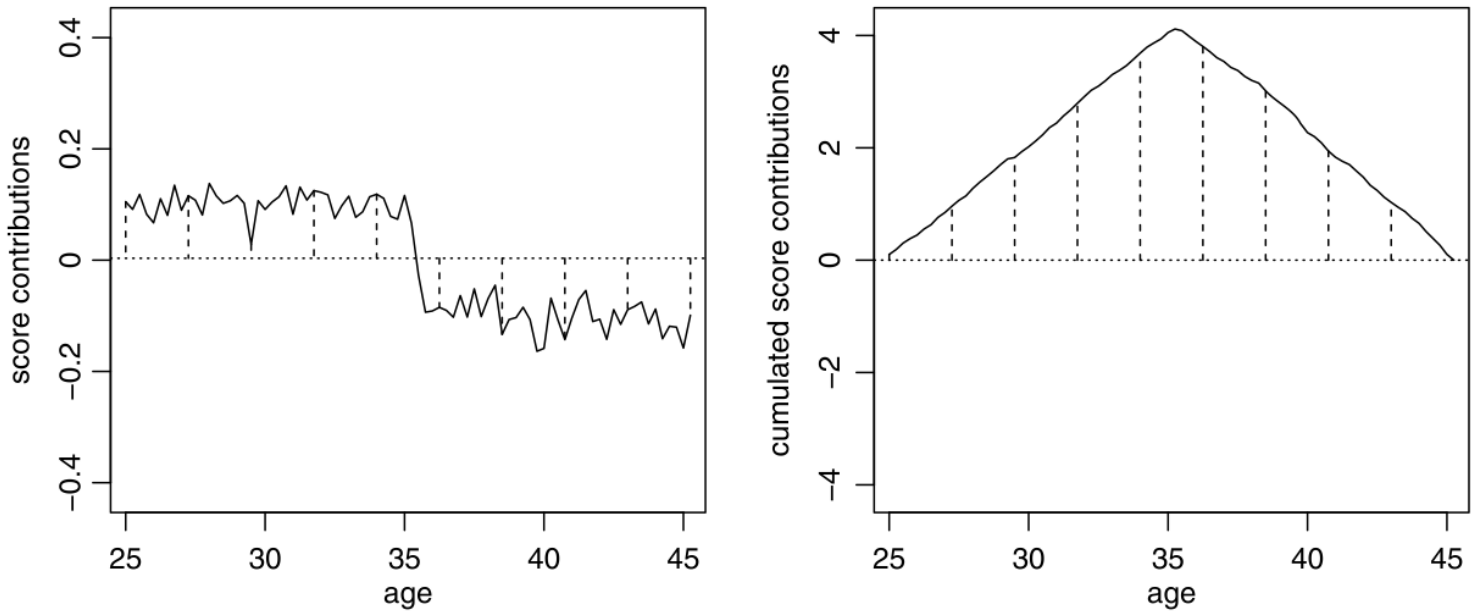


FIGURE 2.
Structural change in the variable age (artificial data for illustration purposes). In the *left plot*, the individual score contributions are ordered with respect to the variable. The *dashed lines* indicate deviations from the overall mean zero, which are positive before the structural change and negative afterward. In the *right plot*, the positive and negative deviations are cumulated and the structural change is now noticeable from the peak in the cumulative sum process.

Figure: Concept of an instability test (Strobl et al., 2013)

Model-based CART

Alexander and Grimshaw (1996, JCGS) proposed **treed regression algorithm** that combines the merits of CART and OLS.

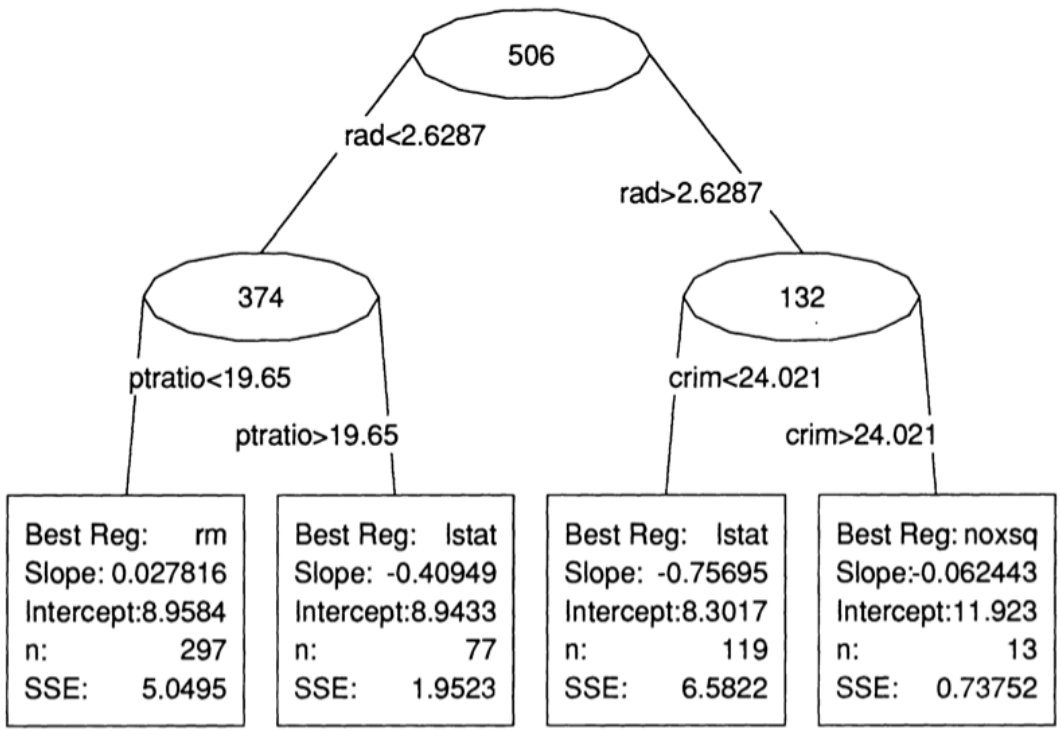


Figure 9. Monotone treed regression model of level two (four terminal nodes) of the log of the median value of owner-occupied homes using 13 explanatory variables measured on Boston area census tracts in 1970 from an example in Belsley, Kuh, and Welsch (1980).

Previous works on model-based regression trees

- Alexander and Grimshaw (1996, JCGS): CART with simple regression
- Chan and Loh (2003, JCGS): GUIDE with logistic regression
- **Zeilies et al. (2008, JCGS) and Rusch et al. (2013): CTREE with GLM**
- **Sela and Simonoff (2012) and Hajjem et al. (2011): CART with LME**
- Brandmaier *et al.* (2013): CART with SEMs
- **Loh and Zheng (2013, AOAS) and Eo and Cho (2014, JCGS): GUIDE with LME**
- Simonoff (2013): CART with linear multilevel model
- Strobl *et al.* (2013): CTREE with Rasch model
- Rusch *et al.* (2013b, AOAS): CTREE with logistic regression
- Rusch *et al.* (2013c, AOAS): CTREE with latent dirichlet allocation
- Tutz and Berger (2014): CTREE with GLM
- ...

Main References

1. Segal, M. R. (1992). Tree-structured methods for longitudinal data. *Journal of the American Statistical Association*, **87**, 407-418.
2. Sela, R. J., and Simonoff, J. S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine Learning*, **86**, 169-207.
3. Eriksson, L., Trygg, J., and Wold, S. (2009). PLS trees: a top-down clustering approach. *Journal of Chemometrics*, **23**, 569-580.
4. Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, **17**, 492-514.
5. Loh, W.-Y., and Zheng, W. (2013). Regression trees for longitudinal and multi response data. *The Annals of Applied Statistics*, **7**, 495-522.

Summary

- Segal (1992) extended CART by using local means in a node impurity.

  Difficult to solve when q is large

- Hajjem et al. (2011) and Sela and Simonoff (2012) extended CART with LME to treat **time-varying covariates**.

  Not consider trends over time & variable selection bias

- Eriksson et al. (2009) extended CART with PLS.

  Vague impurity function and split variables

- Zeileis et al. (2008) and Rusch et al. (2013a) extended CTREE with GLM.

  Not done with LME yet; no interaction effect test

- Loh and Zheng (2013) extended GUIDE with splines.

  Difficult to solve when q is large & no consideration of random effects

Longitudinal CART (Segal, 1992)

		CART algorithm	SEGAL algorithm
Model		$y_i = \beta_0 + \varepsilon_{ij}$	$y_{ij} = \beta_{0j} + \varepsilon_{ij}$
Split	Split Mtd.	Binary Split	
	Split Ft.	$\phi_m(s, g) = SS(g) - SS(g_L) - SS(g_R)$.	
	R(t)	$SS(g) = \Sigma(y_i - \bar{y}(g))^2$	$SS(g) = \sum_{i \in g} (y_i - \mu(g))' V(\theta, g)^{-1} (y_i - \mu(g))$.
	Cov Str.		$V(\theta, g) = V(\theta_L, g_L) = V(\theta_R, g_R)$
Stopping		Cost-Complexity Pruning	
Missing		Surrogate rule	

Figure: Comparison between CART and Longitudinal CART (Eo, 2009)

RE-EM tree (Sela and Simonoff, 2012)

Hajjem et al. (2011, SPL) and Sela and Simonoff (2012) independently proposed **random effect with EM algorithm tree model**.

$$y_{it} = f(X_i) + Z_{it} b_i + \epsilon_{it},$$

$$b_i \sim N(0, D), \quad \epsilon_{it} \sim N(0, \sigma^2 I),$$

$$i = 1, 2, \ldots, n.$$

- If the random effects, b_i, were known, the model implies that we could fit a regression tree $y_{it} - Z_{it} b_i$ to estimate f.
- If the fixed effects, f, were known and can be represented as a linear function, then we could estimate the random effects using a traditional mixed effects linear model with $f(X_i)$.

RE-EM tree (Sela and Simonoff, 2012)

For a known cluster: Prediction $= \hat{f}(x_{ij}) + Z_i \hat{b}_i$.

For a new cluster: Prediction $= \hat{f}(x_{ij})$.

Step 0. Set $r = 0$. Let $\hat{b}_{i(0)} = 0$, $\hat{\sigma}^2_{(0)} = 1$, and $\hat{D}_{(0)} = I_q$.

Step 1. Set $r = r + 1$. Update $y^*_{i(r)}$, $\hat{f}(X_i)_{(r)}$, and $\hat{b}_{i(r)}$

i) $y^*_{i(r)} = y_i - Z_i \hat{b}_{i(r-1)}$, $i = 1, \ldots, n$,

ii) Let $\hat{f}(X_i)_{(r)}$ be estimated from a tree (or forest) algorithm with $y^*_{i(r)}$ as responses and X_i as covariates,

iii) $\hat{b}_{i(r)} = \hat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} \left(y_i - \hat{f}(X_i)_{(r)} \right)$, $i = 1, 2, \ldots, n$,

where $\hat{V}_{i(r-1)} = Z_i \hat{D}_{(r-1)} Z_i^T + \hat{\sigma}^2_{r-1} I_{n_i}$, $i = 1, 2, \ldots, n$.

Step 2. Update $\hat{\sigma}^2_{(r)}$, and $\hat{D}_{(r)}$ using

$$\hat{\sigma}^2_{(r)} = N^{-1} \sum_{i=1}^{n} \left\{ \hat{\epsilon}^T_{i(r)} \hat{\epsilon}_{i(r)} + \hat{\sigma}^2_{(r-1)} [n_i - \hat{\sigma}^2_{(r-1)} tr(\hat{V}_{i(r-1)})] \right\}$$

$$\hat{D}_{(r)} = n^{-1} \sum_{i=1}^{n} \left\{ \hat{b}_{i(r)} \hat{b}^T_{i(r)} + [\hat{D}_{(r-1)} - \hat{D}_{(r-1)} Z_i^T \hat{V}^{-1}_{i(r-1)} Z_i \hat{D}_{(r-1)}] \right\},$$

Figure: Main algorithm of RE-EM tree

PLS tree (Eriksson, 2009)

- **Use the first score vector of PLS as the basis for the selection of splitting.**
- Project data onto the first PC *recursively*.
- Allow the responses to influence the clustering for the analysis of rank deficient data with many variables, not ordinary dataset.
- No detailed comparison between CART and PLS trees.
- This algorithm comes from a clustering point of view, not decision tree.

Model-based CTREE (Zeileis et al., 2008)

1. **Fit a parametric model in each node, e.g., based on maximum likelihood (Zeileis et al, 2008, JCGS) or GLM (Rusch et al., 2013).**
2. Assess whether parameter estimates are stable with respect to each split variable, using Bonferroni-adjusted p-values (instability tests).
3. If minimum p-value is sufficiently small, select the most unstable variable and split the node into two. Otherwise stop.

Model-based CTREE (Zeileis et al., 2008)

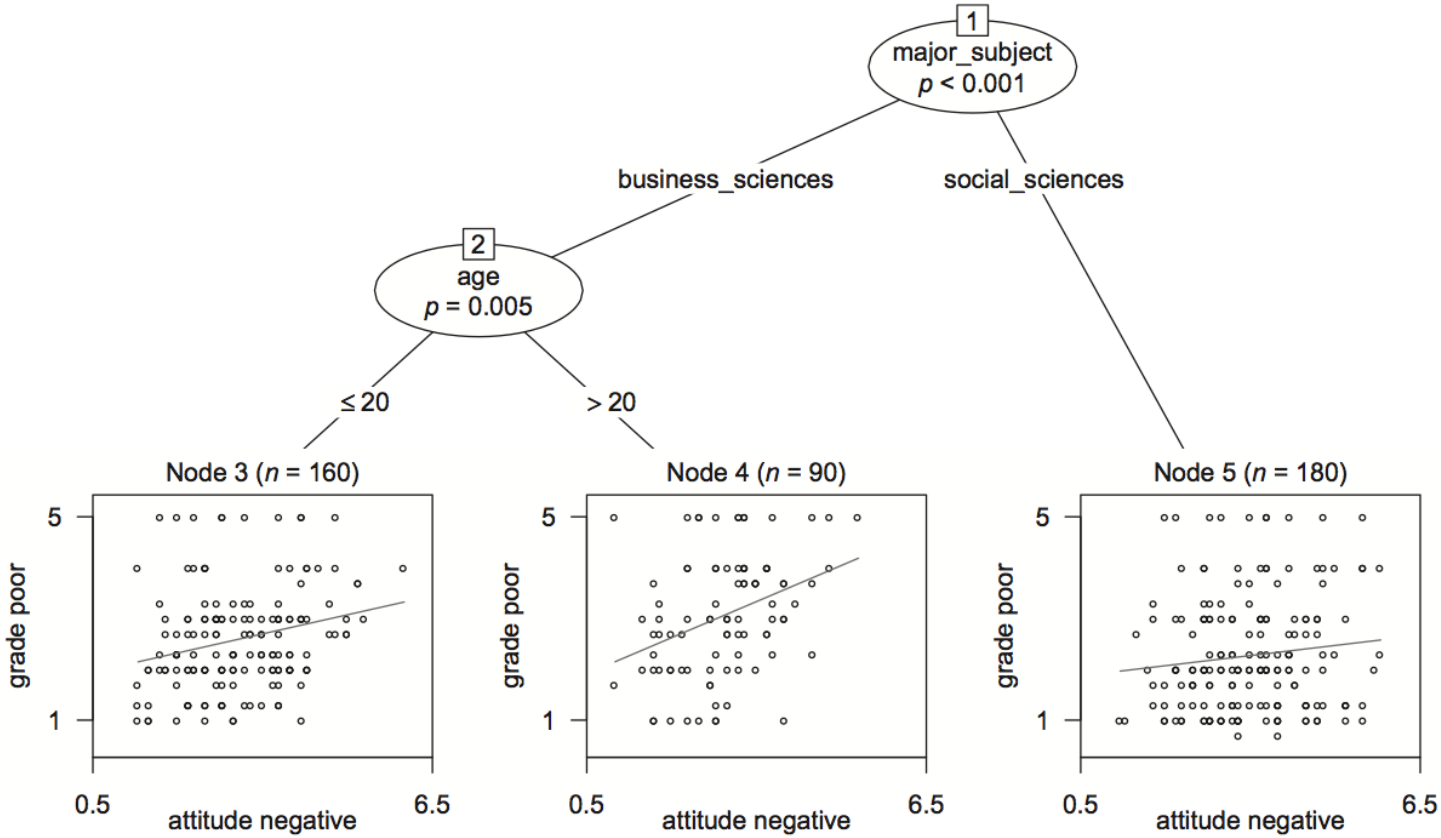


Figure 9. Model-based partition for the attitude toward statistics data. The model of interest relates the final statistics grade to the Cognitive Competence score obtained in the first week. (For the interpretation, note that the Cognitive Competence score was recoded such that high values correspond to a negative attitude, i.e., agreement with items such as "I will find it difficult to understand statistical concepts," and numerically high grades indicate poor performance.)

Figure: Example of model-based CTREE (Strobl et al., 2009)

Longitudinal GUIDE (Loh and Zheng, 2013)

1. Treat each data point as a curve (trajectory).
2. **Fit a mean curve (lowess or smoothing spline) to data in the node.**
3. Group trajectories into classes according to shapes relative to mean curve.
4. For each X variable, find p-value of χ^2 test of class vs. X.
5. Select X with smallest p-value to split node.
6. For each split point, fit a mean curve to each child node.
7. Select the split that minimizes sum of squared deviations or trajectories from mean curves in two child nodes.
8. Stop splitting when sample size in node is too small.
9. Prune the tree using cross-validation.

Longitudinal GUIDE (Loh and Zheng, 2013)

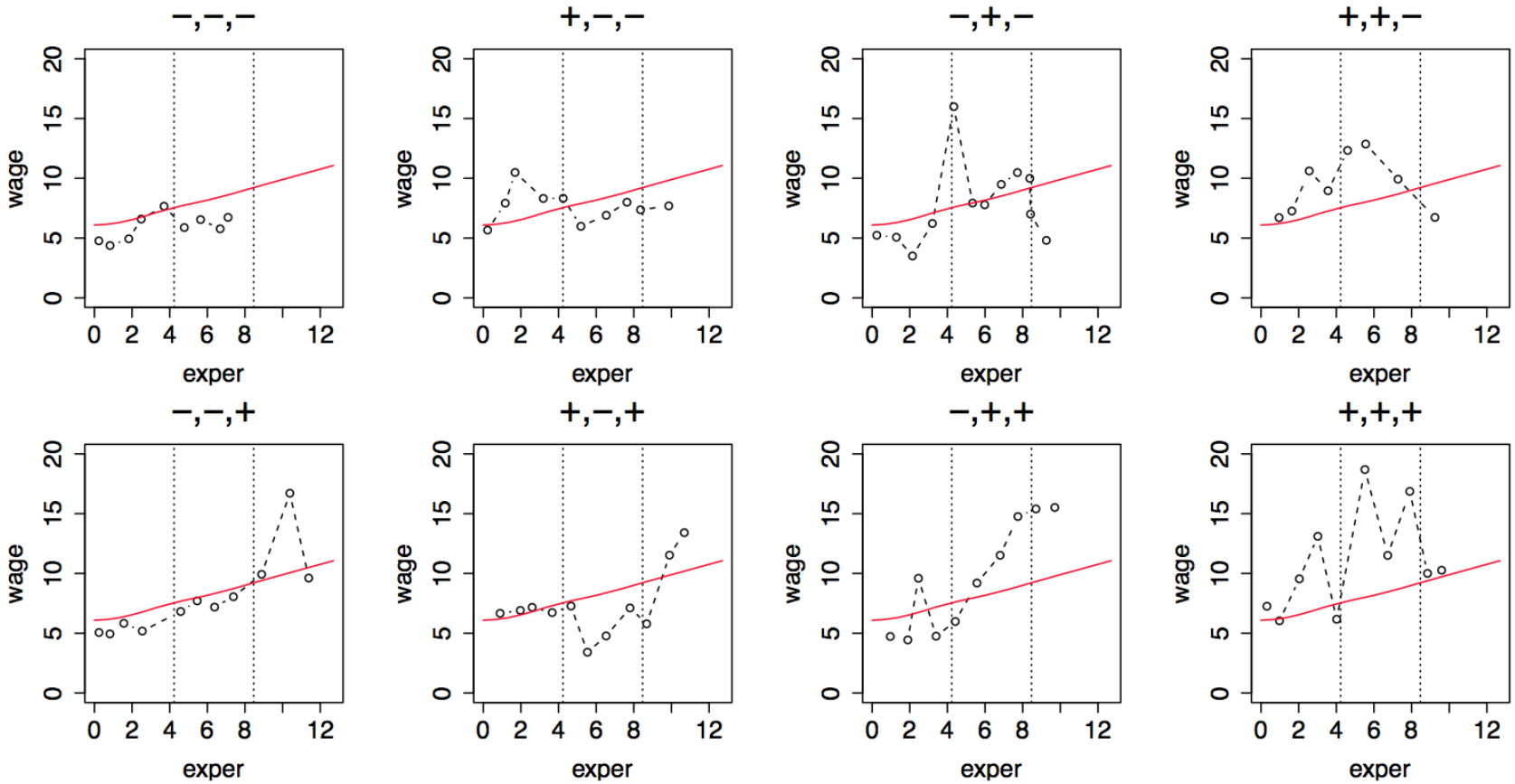


Figure: Some individual trajectories

Proposal 1 - Mixed-Effect Longitudinal Tree

- Mixed-effects model with a **random intercept** and a **fixed time effect** as a basis model at each node:

$$y_{ij} = \alpha_i + \beta(time)_{ij} + \epsilon_{ij}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, q_n, \tag{4}$$

where

  - $\alpha_i \sim N(\alpha_0, \sigma_a^2)$, β is a fixed unknown parameter,
  - $\epsilon \sim N(0, \Lambda)$, where Λ is a certain covariance matrix.

- Let β_{ti} be a slope of subject i and β_t be the common slope of all the subjects at node t with an impurity function as

$$R(t) = \sum_{i \in t} (\hat{\beta}_{ti} - \hat{\beta}_t)^2. \tag{5}$$

Proposal 2 - PLS TREE

Decompose $\mathbf{X}_t$ into orthogonal scores $\mathbf{T}_t$ and loadings $\mathbf{P}_t$

$$\mathbf{X}_t = \mathbf{T}_t \mathbf{P}_t, \tag{6}$$

regressing $\mathbf{y}_t$ not on $\mathbf{X}_t$ itself but on the first a columns of $\mathbf{T}_t$. Then,

$$\hat{\boldsymbol{\beta}}_t^{PCR} = \mathbf{P}_t(\mathbf{T}_t^\top \mathbf{T}_t)^{-1}\mathbf{T}_t^\top \mathbf{y}_t. \tag{7}$$

Use PLS to find projection directions $\mathbf{w}_k$

$$\max_{\mathbf{w}_k} cov(\mathbf{X}_t^\top \mathbf{w}_k, \mathbf{y}_t) \tag{8}$$

s.t. $\mathbf{w}_k^\top \mathbf{w}_k = 1$ and $\mathbf{w}_k^\top \mathbf{X}_t \mathbf{X}_t^\top \mathbf{w}_{k'} = 0$ for $1 \leq k' < k$.

$$\hat{\boldsymbol{\beta}}_t^{PLS} = \mathbf{W}(\mathbf{W}^\top \mathbf{X}_t^\top \mathbf{X}_t \mathbf{W})^{-1}\mathbf{W}^\top \mathbf{X}_1^\top \mathbf{y}_t. \tag{9}$$

Proposal 3 - Unified Tree Structure

- We propose new algorithm of split point selection by using constant spline model.
- We propose new stopping rule called **M-step stopping rule** for the determination of right tree size.

  1. Let t be any node and t_L and t_R be its left and right child node.

  $$\delta(t) = R(t) - [R(t_L) + R(t_R)].$$

  We split node t if $\delta(t) > \epsilon$ and go to Step 2 if $\delta(t) \leq \epsilon$.

  2. if $\delta(t_L) > \epsilon (\delta(t_R) > \epsilon)$, then go to Step 1, letting t_L (or t_R) as node t.
  3. Declare node t as a terminal node if $\delta(t_L) \leq \epsilon$ and $\delta(t_R) \leq \epsilon$.

  For $M > 2$, declare node t as a terminal node when there is no improvement as all M-step nodes.

Motivating Example 2 - Orthodontics Study

Orthodontics study belongs to landmark-based geometric morphometrics, also known as **landmark data** (Dryden and Mardia, 1992).

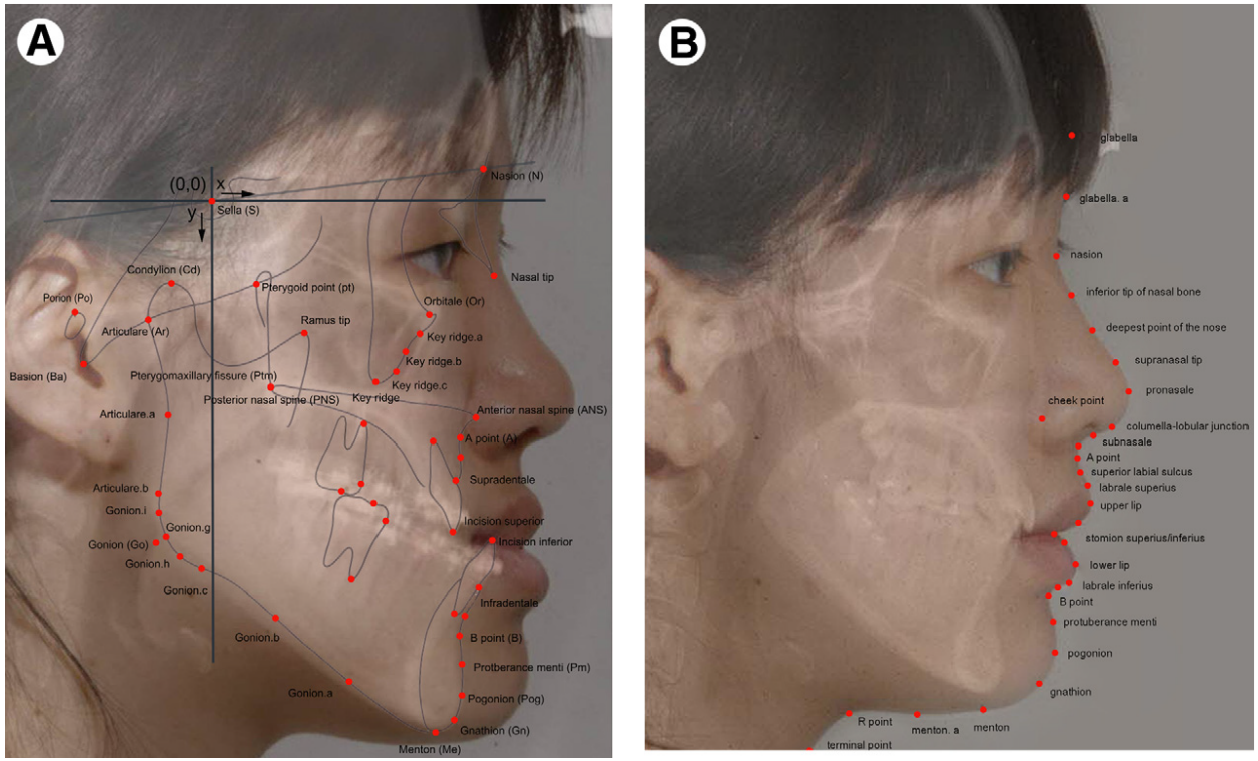


Figure: Skeletal landmarks (left) and skin landmarks (right)

References

- Alexander, W. P., and Grimshaw, S. D. (1996) Treed regression, *JCGS*, **5**, 156-175.
- Boulesteix, A-L., and Strimmer, K. (2006). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data, *Brief Bioinform*, **8**, 32-44.
- Eo, S-H and Cho, H (2014). **Tree-structured mixed-effects regression modeling for longitudinal data**,*JCGS*, .
- Erriksson, L., Trygg, J., and Wold, S. (2009). **PLS trees, a top-down clustering approach**, *J.Chem*, **23**, 569-580.
- Kettaneh, N., Berglund A., and Wold, S. (2005). PCA and PLS with very large data sets, *CSDA*, **48**, 69-85.
- Loh, W-Y. (2002). Regression trees with unbiased variable selection and interaction detection, *Stat Sinicai*, **12**, 361-386.
- Loh, W-Y. (2009).**Improving the precision of classification trees**, *AOAS*, **3**, 1710-1737.
- Loh, W-Y., and Zheng, W (2013). Regression trees for longitudinal and multi-response data, *AOAS*, **7**, 495-522.
- Suh, H-Y. *et al.* (2012). A more accurate method of predicting soft-tissue changes after mandibular setback surgery, *JOMS*, **70**, 553-562.
- Zeileis, A., Hothorn, T., and Hornik, K. (2008). Model-based recursive partitioning, *JCGS*, **17**, 492-514.