Recursive Partitioning for Longitudinal Markers Based on a U-Statistic

Shannon Stock*, Victor DeGruttola, Chengcheng Hu

Department of Biostatistics
Harvard School of Public Health

August 20, 2009

Outline

Outline

Outline

1. Clinical motivation
2. Recursive partitioning
    - Scoring function
    - Split function
    - Tree construction
3. Accommodating missing data
4. Simulation results
5. Application: ACTG Protocol 398
6. Discussion

Outline

1. Clinical motivation
2. Recursive partitioning
   - Scoring function
   - Split function
   - Tree construction
3. Accommodating missing data
4. Simulation results
5. Application: ACTG Protocol 398
6. Discussion

Outline

1. Clinical motivation
2. Recursive partitioning
   - Scoring function
   - Split function
   - Tree construction
3. Accommodating missing data
4. Simulation results
5. Application: ACTG Protocol 398
6. Discussion

Outline

1. Clinical motivation
2. Recursive partitioning
    - Scoring function
    - Split function
    - Tree construction
3. Accommodating missing data
4. Simulation results
5. Application: ACTG Protocol 398
6. Discussion

Clinical motivation

- Highly Active Anti-Retroviral Therapy (HAART) has improved length and quality of life of people infected with HIV.
- Interest lies in finding genotypic mutations associated with reduced virlogic response.
- CART: Used for a single time point.
- *Segal et al*. (1992) extended the CART framework to accommodate longitudinal outcomes, but this method requires parametric models.

Clinical motivation

- Highly Active Anti-Retroviral Therapy (HAART) has improved length and quality of life of people infected with HIV.
- Interest lies in finding genotypic mutations associated with reduced virlogic response.
- CART: Used for a single time point.
- *Segal et al*. (1992) extended the CART framework to accommodate longitudinal outcomes, but this method requires parametric models.

Clinical motivation

- Highly Active Anti-Retroviral Therapy (HAART) has improved length and quality of life of people infected with HIV.
- Interest lies in finding genotypic mutations associated with reduced virlogic response.
- CART: Used for a single time point.
- *Segal et al*. (1992) extended the CART framework to accommodate longitudinal outcomes, but this method requires parametric models.

Clinical motivation

- Highly Active Anti-Retroviral Therapy (HAART) has improved length and quality of life of people infected with HIV.
- Interest lies in finding genotypic mutations associated with reduced virlogic response.
- CART: Used for a single time point.
- *Segal et al.* (1992) extended the CART framework to accommodate longitudinal outcomes, but this method requires parametric models.

Clinical motivation

- We propose a recursive-partitioning approach for continuous longitudinal outcomes which uses the kernel of a U-statistic as the splitting criterion.
- Accommodates arbitrarily large set of covariates.
- Avoids need for parametric modeling of the relationship between observed virologic responses and covariates.
- Accommodates MAR longitudinal measurements.

Clinical motivation

- We propose a recursive-partitioning approach for continuous longitudinal outcomes which uses the kernel of a U-statistic as the splitting criterion.
- Accommodates arbitrarily large set of covariates.
- Avoids need for parametric modeling of the relationship between observed virologic responses and covariates.
- Accommodates MAR longitudinal measurements.

Clinical motivation

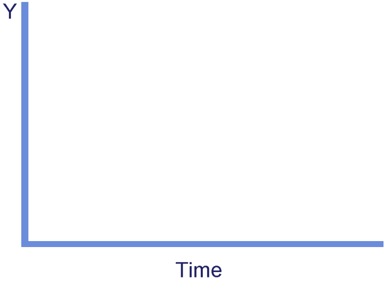
- We propose a recursive-partitioning approach for continuous longitudinal outcomes which uses the kernel of a U-statistic as the splitting criterion.
- Accommodates arbitrarily large set of covariates.
- Avoids need for parametric modeling of the relationship between observed virologic responses and covariates.
- Accommodates MAR longitudinal measurements.

Clinical motivation

- We propose a recursive-partitioning approach for continuous longitudinal outcomes which uses the kernel of a U-statistic as the splitting criterion.
- Accommodates arbitrarily large set of covariates.
- Avoids need for parametric modeling of the relationship between observed virologic responses and covariates.
- Accommodates MAR longitudinal measurements.

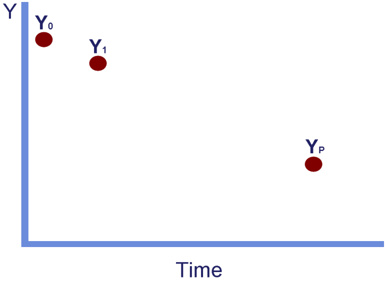
Recursive partitioning method:Notation

- Assume N enrolled subjects
- $P+1$ outcome measurements: $(Y_{i0}, Y_{i1}, ..., Y_{iP})$
- Observed at: $(T_0, T_1, ..., T_P)$
- R binary baseline covariates: $(X_{i1}, X_{i2}, ..., X_{iR})$



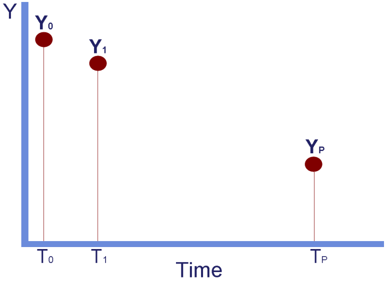
Notation

- Assume N enrolled subjects
- $P+1$ outcome measurements: $(Y_{i0}, Y_{i1}, \ldots, Y_{iP})$
- Observed at: $(T_0, T_1, \ldots, T_P)$
- R binary baseline covariates: $(X_{i1}, X_{i2}, \ldots, X_{iR})$



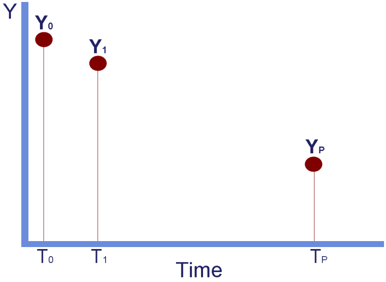
Notation

- Assume N enrolled subjects
- $P + 1$ outcome measurements: $(Y_{i0}, Y_{i1}, ..., Y_{iP})$
- Observed at: $(T_0, T_1, ..., T_P)$
- R binary baseline covariates: $(X_{i1}, X_{i2}, ..., X_{iR})$



Notation

- Assume N enrolled subjects
- $P + 1$ outcome measurements: $(Y_{i0}, Y_{i1}, ..., Y_{iP})$
- Observed at: $(T_0, T_1, ..., T_P)$
- R binary baseline covariates: $(X_{i1}, X_{i2}, ..., X_{iR})$



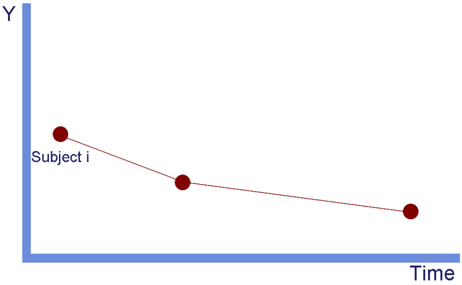
Scoring Function

- Motivated by *May and DeGruttola* (2007), compare subject i's k^{th} observation with subject j's l^{th} observation:

$$D\{(Y_{ik}, t_{ik}), (Y_{jl}, t_{jl})\} = \begin{cases} 1 & \text{if } Y_{ik} < Y_{jl} \text{ and } t_{ik} \leq t_{jl}, \\ -1 & \text{if } Y_{ik} > Y_{jl} \text{ and } t_{ik} \geq t_{jl}, \\ 0 & \text{otherwise} \end{cases}$$

Scoring Function

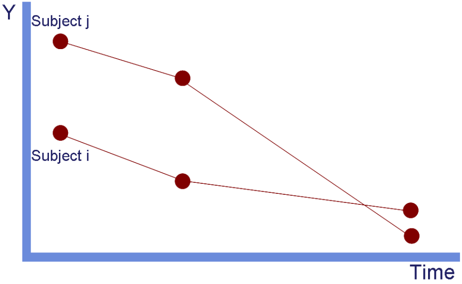
- Motivated by *May and DeGruttola* (2007), compare subject i's k^{th} observation with subject j's l^{th} observation:



$$D\{(Y_{ik}, t_{ik}), (Y_{jl}, t_{jl})\} = \begin{cases} 1 & \text{if } Y_{ik} < Y_{jl} \text{ and } t_{ik} \leq t_{jl}, \\ -1 & \text{if } Y_{ik} > Y_{jl} \text{ and } t_{ik} \geq t_{jl}, \\ 0 & \text{otherwise} \end{cases}$$

Scoring Function

- Motivated by *May and DeGruttola* (2007), compare subject i's k^{th} observation with subject j's l^{th} observation:

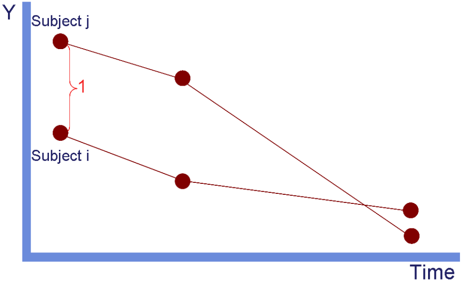


$$D\{(Y_{ik}, t_{ik}), (Y_{jl}, t_{jl})\} = \begin{cases} 1 & \text{if } Y_{ik} < Y_{jl} \text{ and } t_{ik} \leq t_{jl}, \\ -1 & \text{if } Y_{ik} > Y_{jl} \text{ and } t_{ik} \geq t_{jl}, \\ 0 & \text{otherwise} \end{cases}$$

Scoring Function

Outline
Clinical motivation
Recursive partitioning method
Missing Data
Simulation Results
Application: ACTG 398
Discussion

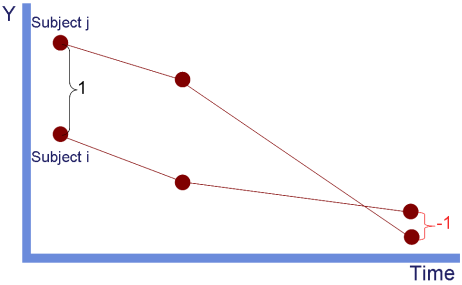
- Motivated by *May and DeGruttola* (2007), compare subject i's k^{th} observation with subject j's l^{th} observation:



$$D\{(Y_{ik}, t_{ik}), (Y_{jl}, t_{jl})\} = \begin{cases} 1 & \text{if } Y_{ik} < Y_{jl} \text{ and } t_{ik} \leq t_{jl}, \\ -1 & \text{if } Y_{ik} > Y_{jl} \text{ and } t_{ik} \geq t_{jl}, \\ 0 & \text{otherwise} \end{cases}$$

Scoring Function

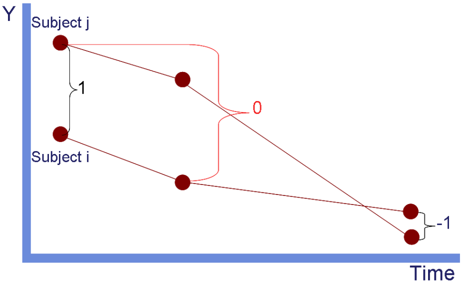
- Motivated by *May and DeGruttola* (2007), compare subject i's k^{th} observation with subject j's l^{th} observation:



$$D\{(Y_{ik}, t_{ik}), (Y_{jl}, t_{jl})\} = \begin{cases} 1 & \text{if } Y_{ik} < Y_{jl} \text{ and } t_{ik} \leq t_{jl}, \\ -1 & \text{if } Y_{ik} > Y_{jl} \text{ and } t_{ik} \geq t_{jl}, \\ 0 & \text{otherwise} \end{cases}$$

Scoring Function

- Motivated by *May and DeGruttola* (2007), compare subject i's k^{th} observation with subject j's l^{th} observation:

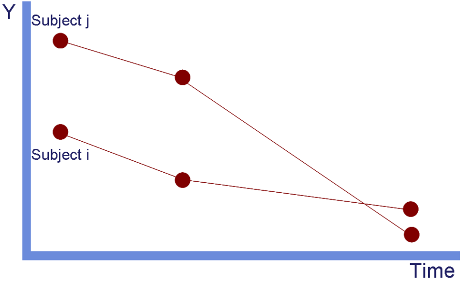


$$D\{(Y_{ik}, t_{ik}), (Y_{jl}, t_{jl})\} = \begin{cases} 1 & \text{if } Y_{ik} < Y_{jl} \text{ and } t_{ik} \leq t_{jl}, \\ -1 & \text{if } Y_{ik} > Y_{jl} \text{ and } t_{ik} \geq t_{jl}, \\ 0 & \text{otherwise} \end{cases}$$

Scoring Function

- Summary of the comparisons between subjects i and j :

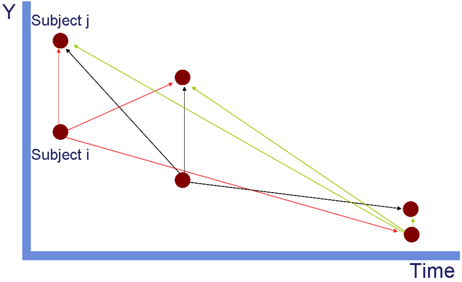
$$D(i,j) = \sum_{k=1}^{P} \sum_{l=1}^{P} D\{(Y_{ik}, t_{ik}), (Y_{jl}, t_{jl})\}$$



Scoring Function

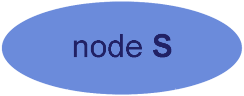
- Summary of the comparisons between subjects i and j:

$$D(i,j) = \sum_{k=1}^{P} \sum_{l=1}^{P} D\{(Y_{ik}, t_{ik}), (Y_{jl}, t_{jl})\}$$



Split Function

- For node S and covariate X_r:



- Goodness-of-split measure:

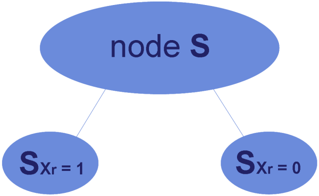
$$G(r,s) = \Big| \sum_{i \in S_{X_r=1}} \sum_{j \in S_{X_r=0}} D(i,j) \Big|$$

- Split based on:

$$G(r^*,s) = \max_{1 \le r \le R} G(r,s)$$

Split Function

- For node S and covariate X_r:



- Goodness-of-split measure:

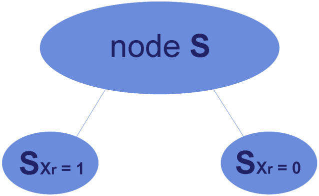
$$G(r,s) = \Big|\sum_{i \in S_{X_r=1}} \sum_{j \in S_{X_r=0}} D(i,j)\Big|$$

- Split based on:

$$G(r^*,s) = \max_{1 \le r \le R} G(r,s)$$

Split Function

- For node S and covariate X_r:



- Goodness-of-split measure:

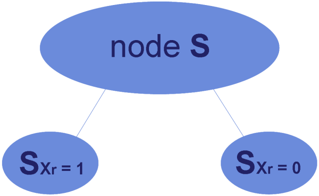
$$G(r,s) = \left| \sum_{i \in S_{X_r=1}} \sum_{j \in S_{X_r=0}} D(i,j) \right|$$

- Split based on:

$$G(r^*,s) = \max_{1 \le r \le R} G(r,s)$$

Split Function

- For node S and covariate X_r:



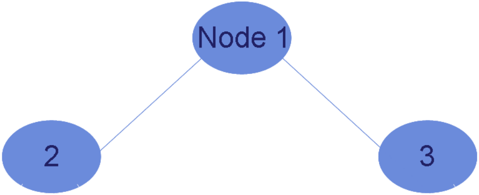
- Goodness-of-split measure:

$$G(r,s) = \left| \sum_{i \in S_{X_r=1}} \sum_{j \in S_{X_r=0}} D(i,j) \right|$$

- Split based on:

$$G(r^*,s) = \max_{1 \leq r \leq R} G(r,s)$$

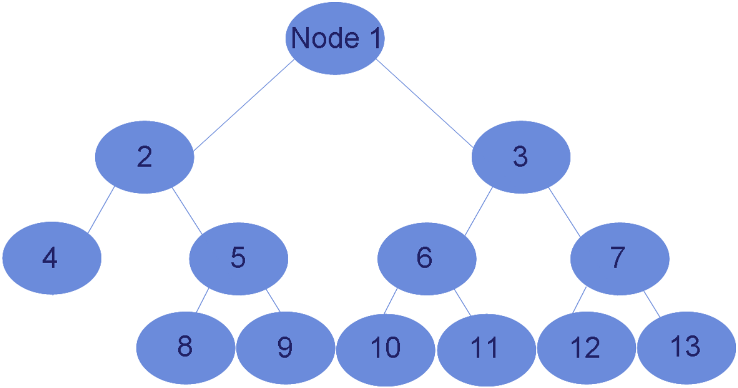
Tree Construction



- Split the root node at the covariate X_r based on:

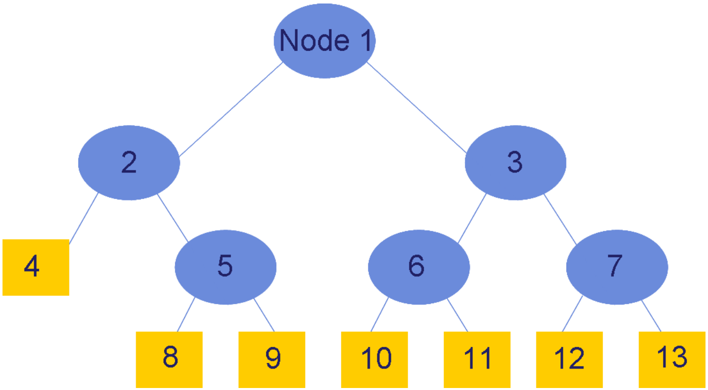
$$G(r^*, rootnode) = \max_{1 \le r \le R} G(r, rootnode)$$

Tree Construction



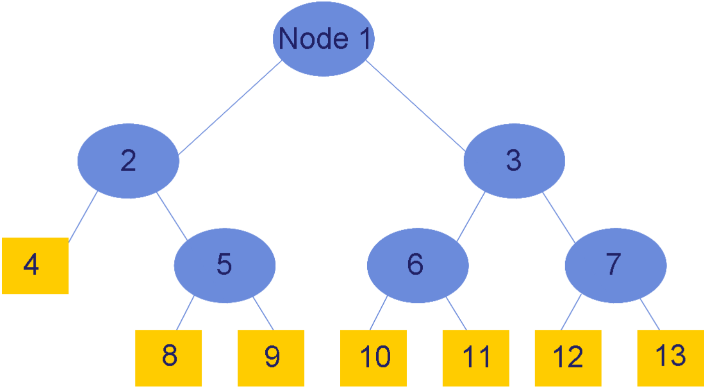
- Recursively split each resulting daughter node.

Tree Construction



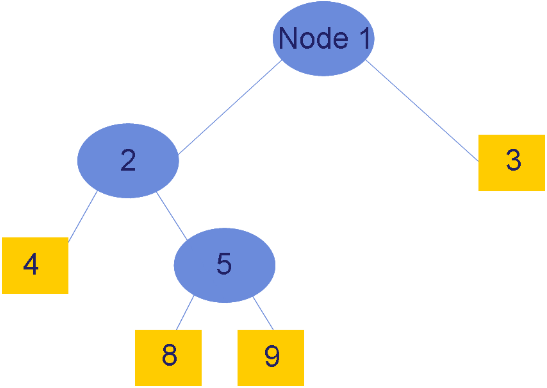
- STOP:
  - Node is homogenous w.r.t the outcome
  - A minimum node size is reached
  - The tree is saturated

Tree Construction



- Denote this "full tree" T_0
- Prune T_0 to an appropriate size

Tree Construction



- Denote this "full tree" T_0
- Prune T_0 to an appropriate size

Tree Construction

- An appropriately sized subtree can be found by adapting the bootstrap method of *Fan et al*.
- 1. Draw B boostrap samples L_b $b = 1, 2, ..., B$ from the complete data set L
- 2. Construct an ordered sequence of pruned subtrees for each of the B bootstrap samples
- 3. Compute the bootstrap estimator of $G_\alpha(T)$ using:

$$G^{(b)}(T_m) = G_L(L) - \frac{1}{b}\sum\{G_{L_b}(L_b) - G_L(L_b\} \qquad (1)$$

Tree Construction

- An appropriately sized subtree can be found by adapting the bootstrap method of *Fan et al*.
- 1. Draw B boostrap samples L_b $b = 1, 2, ..., B$ from the complete data set L
- 2. Construct an ordered sequence of pruned subtrees for each of the B bootstrap samples
- 3. Compute the bootstrap estimator of $G_\alpha(T)$ using:

$$G^{(b)}(T_m) = G_L(L) - \frac{1}{b} \sum \{ G_{L_b}(L_b) - G_L(L_b\} \quad (1)$$

Tree Construction

- An appropriately sized subtree can be found by adapting the bootstrap method of *Fan et al*.
- 1. Draw B boostrap samples L_b $b = 1, 2, ..., B$ from the complete data set L
- 2. Construct an ordered sequence of pruned subtrees for each of the B bootstrap samples
- 3. Compute the bootstrap estimator of $G_\alpha(T)$ using:

$$G^{(b)}(T_m) = G_L(L) - \frac{1}{b}\sum\{G_{L_b}(L_b) - G_L(L_b\} \quad (1)$$

Tree Construction

- An appropriately sized subtree can be found by adapting the bootstrap method of *Fan et al*.
- 1. Draw B boostrap samples L_b $b = 1, 2, ..., B$ from the complete data set L
- 2. Construct an ordered sequence of pruned subtrees for each of the B bootstrap samples
- 3. Compute the bootstrap estimator of $G_\alpha(T)$ using:

$$G^{(b)}(T_m) = G_L(L) - \frac{1}{b}\sum\{G_{L_b}(L_b) - G_L(L_b\} \quad (1)$$

Accommodating Missing Data

- When observations are MAR for subject i or j, the $E(D(i,j)) \neq 0$.
- Replace $D(i,j)$ with the alternative estimator:

$$\tilde{D}(i,j) = \sum_{k=1}^{ni} \sum_{l=1}^{nj} \frac{D(i,j)^{kl}}{\pi_i(t_k)\pi_j(t_l)}$$

- $\pi_i(t_k) =$ The probability of observing subject i's k^{th} outcome measurement at time t_k.
- $D(i,j)^{kl} =$ The value of the score comparing subject i's K^{th} observation with subject j's l^{th} observation.
- Estimate the $\pi_i(t_k)$ using:
  - Logistic regression
  - Recursive partitioning

Accommodating Missing Data

- When observations are MAR for subject i or j, the $E(D(i,j)) \neq 0$.
- Replace $D(i,j)$ with the alternative estimator:

$$\tilde{D}(i,j) = \sum_{k=1}^{ni} \sum_{l=1}^{nj} \frac{D(i,j)^{kl}}{\pi_i(t_k)\pi_j(t_l)}$$

- $\pi_i(t_k) =$ The probability of observing subject i's k^{th} outcome measurement at time t_k.
- $D(i,j)^{kl} =$ The value of the score comparing subject i's K^{th} observation with subject j's l^{th} observation.
- Estimate the $\pi_i(t_k)$ using:
    - Logistic regression
    - Recursive partitioning

Accommodating Missing Data

- When observations are MAR for subject i or j, the $E(D(i,j)) \neq 0$.
- Replace $D(i,j)$ with the alternative estimator:

$$\tilde{D}(i,j) = \sum_{k=1}^{ni} \sum_{l=1}^{nj} \frac{D(i,j)^{kl}}{\pi_i(t_k)\pi_j(t_l)}$$

- $\pi_i(t_k) =$ The probability of observing subject i's k^{th} outcome measurement at time t_k.
- $D(i,j)^{kl} =$ The value of the score comparing subject i's K^{th} observation with subject j's l^{th} observation.
- Estimate the $\pi_i(t_k)$ using:
  - Logistic regression
  - Recursive partitioning

Accommodating Missing Data

- When observations are MAR for subject i or j, the $E(D(i,j)) \neq 0$.
- Replace $D(i,j)$ with the alternative estimator:

$$\tilde{D}(i,j) = \sum_{k=1}^{ni} \sum_{l=1}^{nj} \frac{D(i,j)^{kl}}{\pi_i(t_k)\pi_j(t_l)}$$

- $\pi_i(t_k) =$ The probability of observing subject i's k^{th} outcome measurement at time t_k.
- $D(i,j)^{kl} =$ The value of the score comparing subject i's K^{th} observation with subject j's l^{th} observation.
- Estimate the $\pi_i(t_k)$ using:
  - Logistic regression
  - Recursive partitioning

Accommodating Missing Data

- When observations are MAR for subject i or j, the $E(D(i,j)) \neq 0$.
- Replace $D(i,j)$ with the alternative estimator:

$$\tilde{D}(i,j) = \sum_{k=1}^{ni} \sum_{l=1}^{nj} \frac{D(i,j)^{kl}}{\pi_i(t_k)\pi_j(t_l)}$$

- $\pi_i(t_k) =$ The probability of observing subject i's k^{th} outcome measurement at time t_k.
- $D(i,j)^{kl} =$ The value of the score comparing subject i's K^{th} observation with subject j's l^{th} observation.
- Estimate the $\pi_i(t_k)$ using:
    - Logistic regression
    - Recursive partitioning

Accommodating Missing Data

- When observations are MAR for subject i or j, the $E(D(i,j)) \neq 0$.
- Replace $D(i,j)$ with the alternative estimator:

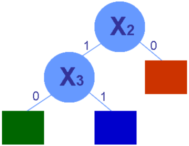
$$\tilde{D}(i,j) = \sum_{k=1}^{ni} \sum_{l=1}^{nj} \frac{D(i,j)^{kl}}{\pi_i(t_k)\pi_j(t_l)}$$

- $\pi_i(t_k) =$ The probability of observing subject i's k^{th} outcome measurement at time t_k.
- $D(i,j)^{kl} =$ The value of the score comparing subject i's K^{th} observation with subject j's l^{th} observation.
- Estimate the $\pi_i(t_k)$ using:
  - Logistic regression
  - Recursive partitioning

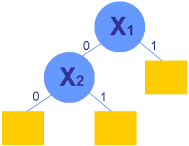
Simulation Results

- 1,000 simulated data sets
- $n = 1,000$ subjects, $P = 3$ outcome measurements, and $R = 5$ binary covariates

- Outcome trajectories:



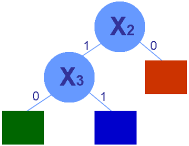
- Missing data:



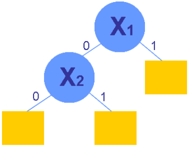
Simulation Results

- 1,000 simulated data sets
- $n = 1,000$ subjects, $P = 3$ outcome measurements, and $R = 5$ binary covariates

- Outcome trajectories:



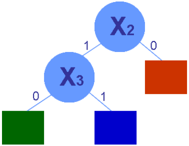
- Missing data:



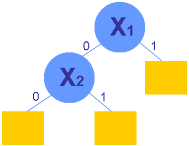
Simulation Results

- 1,000 simulated data sets
- $n = 1,000$ subjects, $P = 3$ outcome measurements, and $R = 5$ binary covariates

- Outcome trajectories:

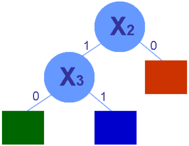


- Missing data:

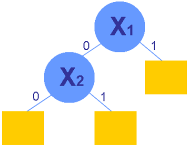


- 1,000 simulated data sets
- $n = 1,000$ subjects, $P = 3$ outcome measurements, and $R = 5$ binary covariates

- Outcome trajectories:

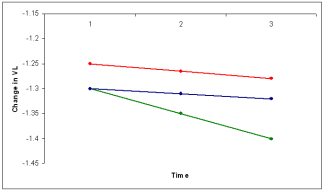


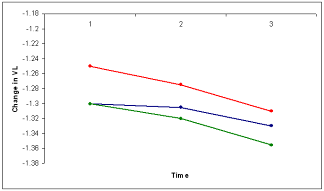
- Missing data:



Simulation Results

- Simulated outcome trajectories:





Simulation Results

Set	N	Cor.	Mis.	Trajectory	All	F,W,eW	F	F Incorrect	Other
1	1000	0.40	35%	non linear	0.783	0.121	0.055	0.034	0.007
2	1000	0.75	35%	non linear	0.264	0.122	0.138	0.451	0.025
3	1000	0.40	35%	linear	1	0	0	0	0
4	1000	0.75	35%	linear	0.879	0.095	0.014	0.002	0.010
5	1000	0.40	15%	non linear	0.725	0.190	0.031	0.049	0.005
6	1000	0.75	15%	non linear	0.216	0.199	0.110	0.456	0.019
7	1000	0.40	15%	linear	1	0	0	0	0
8	1000	0.75	15%	linear	0.910	0.080	0.006	0.004	0

Full = Complete data tree
W = IPW corrected tree (true weights)
eW = IPW corrected tree (estimated weights)

Application: ACTG 398 data

- A randomized trial comparing single- versus dual-PI regimens in combination with other drugs among treatment-experienced patients.
- $n = 440$ subjects with viral load measurements obtained at $t = 0, 14, 27, 56$ ~~weeks~~.
- Interest lies in determining the baseline patterns of drug resistance mutations associated with a reduced viral response.

Application: ACTG 398 data

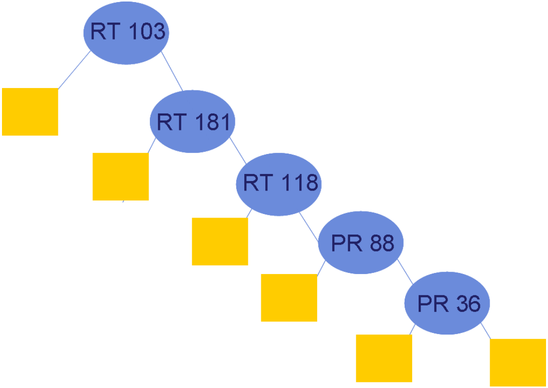
- A randomized trial comparing single- versus dual-PI regimens in combination with other drugs among treatment-experienced patients.
- $n = 440$ subjects with viral load measurements obtained at $t = 0, 14, 27, 56$ weeks.
- Interest lies in determining the baseline patterns of drug resistance mutations associated with a reduced viral response.

Application: ACTG 398 data

- A randomized trial comparing single- versus dual-PI regimens in combination with other drugs among treatment-experienced patients.
- $n = 440$ subjects with viral load measurements obtained at $t = 0, 14, 27, 56$ weeks.
- Interest lies in determining the baseline patterns of drug resistance mutations associated with a reduced viral response.

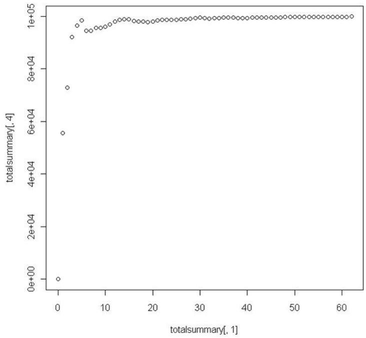
Application: ACTG 398 data

- IPW tree:



Application: ACTG Protocol 398

- Bootstrap selection of proper sized subtree:



Discussion

- CART is used to identify resistance mutations associated with virologic response at a single time point.
- Our recursive partitioning method accommodates a large set of covariates and longitudinal responses.
- Avoids parametric assumptions regarding the relationship between the observed responses and covariates.
- Accommodates MAR longitudinal measurements.
- Extensions:
  - Allow variable times of measurement
  - Accommodate time varying covariates
  - Allow multivariate longitudinal outcomes

Discussion

- CART is used to identify resistance mutations associated with virologic response at a single time point.
- Our recursive partitioning method accommodates a large set of covariates and longitudinal responses.
- Avoids parametric assumptions regarding the relationship between the observed responses and covariates.
- Accommodates MAR longitudinal measurements.
- Extensions:
  - Allow variable times of measurement
  - Accommodate time varying covariates
  - Allow multivariate longitudinal outcomes

Discussion

- CART is used to identify resistance mutations associated with virologic response at a single time point.
- Our recursive partitioning method accommodates a large set of covariates and longitudinal responses.
- Avoids parametric assumptions regarding the relationship between the observed responses and covariates.
- Accommodates MAR longitudinal measurements.
- Extensions:
  - Allow variable times of measurement
  - Accommodate time varying covariates
  - Allow multivariate longitudinal outcomes

Discussion

- CART is used to identify resistance mutations associated with virologic response at a single time point.
- Our recursive partitioning method accommodates a large set of covariates and longitudinal responses.
- Avoids parametric assumptions regarding the relationship between the observed responses and covariates.
- Accommodates MAR longitudinal measurements.
- Extensions:
  - Allow variable times of measurement
  - Accommodate time varying covariates
  - Allow multivariate longitudinal outcomes

Discussion

- CART is used to identify resistance mutations associated with virologic response at a single time point.
- Our recursive partitioning method accommodates a large set of covariates and longitudinal responses.
- Avoids parametric assumptions regarding the relationship between the observed responses and covariates.
- Accommodates MAR longitudinal measurements.
- Extensions:
  - Allow variable times of measurement
  - Accommodate time varying covariates
  - Allow multivariate longitudinal outcomes

Acknowledgements

- Xihong Lin
- Eric Tchetgen
- The ACTG 398 team