Robust and efficient estimation in the presence of a randomly censored covariate

Brian Richardson

November 12, 2024

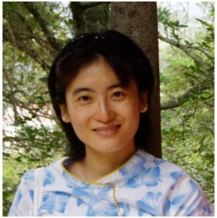
UNC GILLINGS SCHOOL OF GLOBAL PUBLIC HEALTH
Biostatistics

Acknowledgements

Seong-Ho Lee, PhD

University of Seoul

Yanyuan Ma, PhD



Pennsylvania State University

Karen Marder, MD

Columbia University Medical Center

Tanya Garcia, PhD

UNC Chapel Hill

This research was supported by the National Institutes of Neurological Disorders and Stroke (grant R01NS131225) and of Environmental Health Sciences (grant T32ES007018)

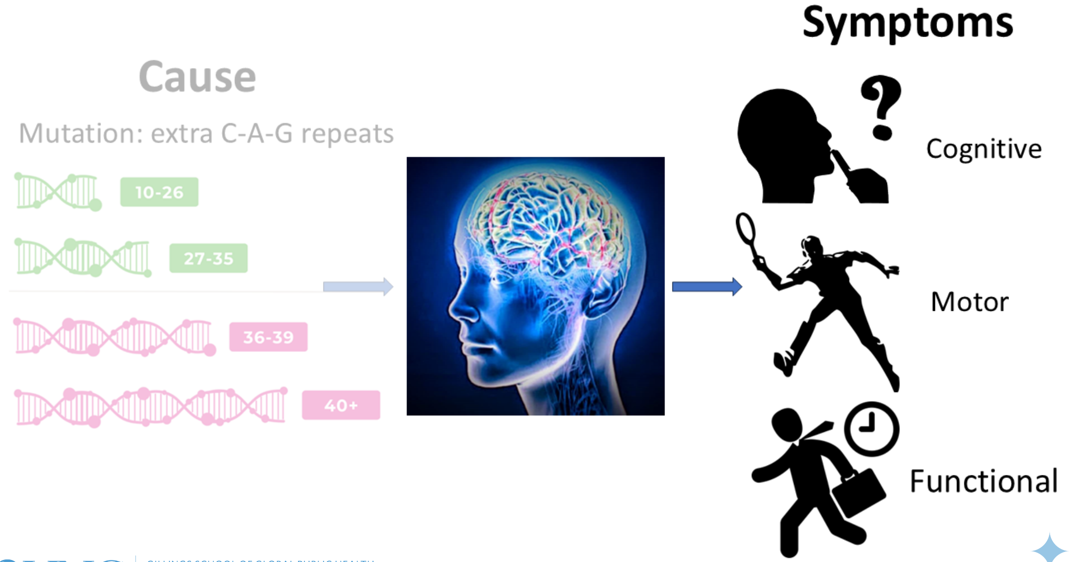
Huntington's Disease

Huntington's Disease

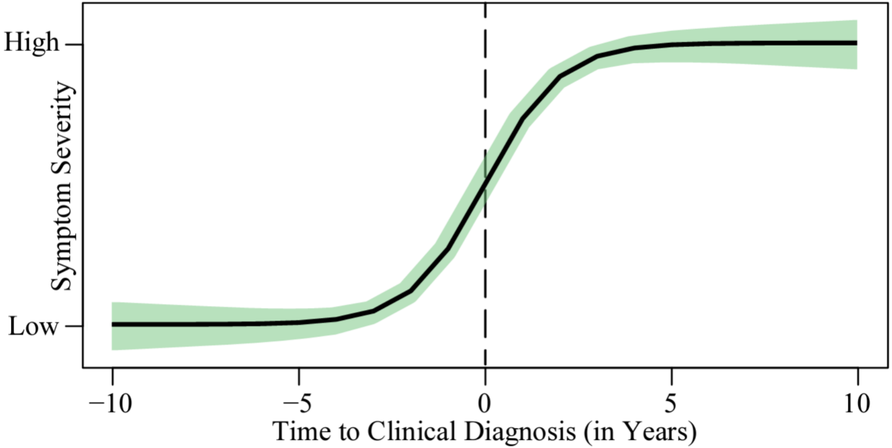
Cause

Mutation: extra C-A-G repeats

Huntington's Disease



Huntington's Disease



(Lotspeich et. al., 2024)

Censoring in Huntington's Disease Studies

- **model**: $\mathrm{E}(Y|X) = m(X, \beta)$

Censoring in Huntington's Disease Studies

- **model**: $\mathrm{E}(Y|X) = m(X, \beta)$
- **of interest**: parameter β characterizing $Y|X$

Censoring in Huntington's Disease Studies

- **model**: $\mathrm{E}(Y|X) = m(X, \beta)$
- **of interest**: parameter β characterizing $Y|X$
- **problem**: X censored by C

Censoring in Huntington's Disease Studies

- **model**: $\mathrm{E}(Y|X) = m(X, \beta)$
- **of interest**: parameter β characterizing $Y|X$
- **problem**: X censored by C
  - observe $W = \min(X, C)$

Censoring in Huntington's Disease Studies

- **model**: $\mathrm{E}(Y|X) = m(X, \beta)$
- **of interest**: parameter β characterizing $Y|X$
- **problem**: X censored by C
  - observe $W = \min(X, C)$
  - observe $\Delta = I(X \leq C)$

Censoring in Huntington's Disease Studies

- **model**: $\mathrm{E}(Y|X) = m(X, \beta)$
- **of interest**: parameter β characterizing $Y|X$
- **problem**: X censored by C
  - observe $W = \min(X, C)$
  - observe $\Delta = I(X \leq C)$
- **nuisance**: distributions $(f_X, f_C) = \eta$

Censoring in Huntington's Disease Studies

- **model**: $\mathrm{E}(Y|X) = m(X, \boldsymbol{\beta})$
- **of interest**: parameter $\boldsymbol{\beta}$ characterizing $Y|X$
- **problem**: X censored by C
  - observe $W = \min(X, C)$
  - observe $\Delta = I(X \leq C)$
- **nuisance**: distributions $(f_X, f_C) = \boldsymbol{\eta}$

"Humans are precious"

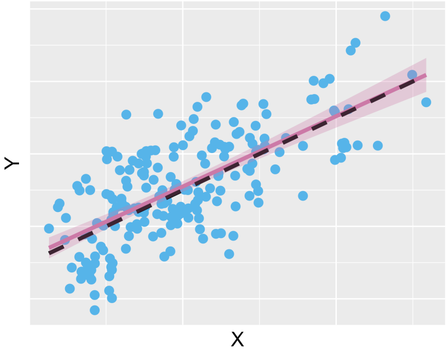
Censoring in Huntington's Disease Studies

- **model**: $\mathrm{E}(Y|X) = m(X, \beta)$
- **of interest**: parameter β characterizing $Y|X$
- **problem**: X censored by C
  - observe $W = \min(X, C)$
  - observe $\Delta = I(X \leq C)$
- **nuisance**: distributions $(f_X, f_C) = \eta$

"Humans are precious"

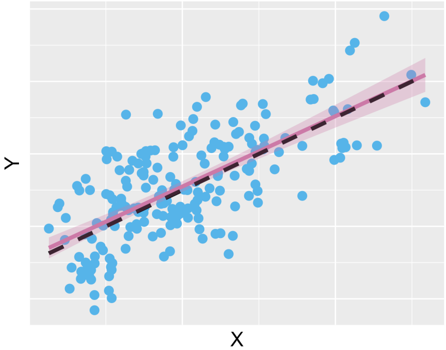
Want methods that are **robust** and **efficient**

Censored Covariates: a Simple Example



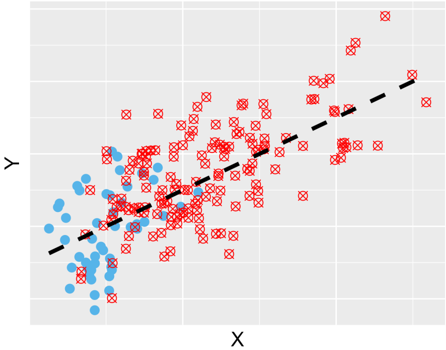
- Regression model: $\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$

Censored Covariates: a Simple Example



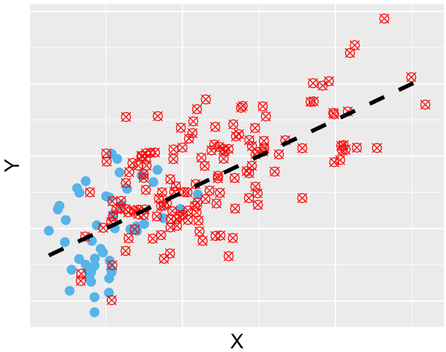
- Regression model: $\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$
- Estimate $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ with least squares/maximum likelihood
- Solve estimating equation $\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)(1, X_i)^T = \mathbf{0}$

Censored Covariates: a Simple Example



Problem: X is censored by a **random censoring time** C

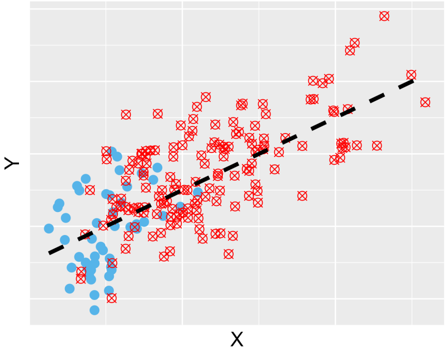
Censored Covariates: a Simple Example



Problem: X is censored by a **random censoring time** C

- $W = \min(X, C)$
- $\Delta = I(X \leq C)$

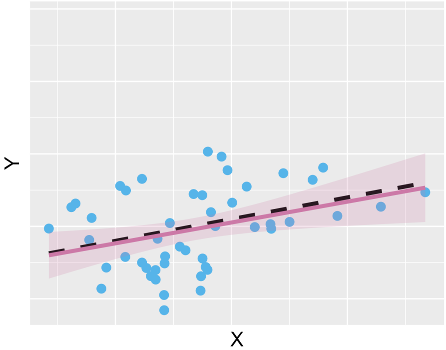
Censored Covariates: a Simple Example



Problem: X is censored by a **random censoring time** C

- $W = \min(X, C)$
- $\Delta = I(X \leq C)$
- assume: $C \perp\!\!\!\perp (X, Y)$

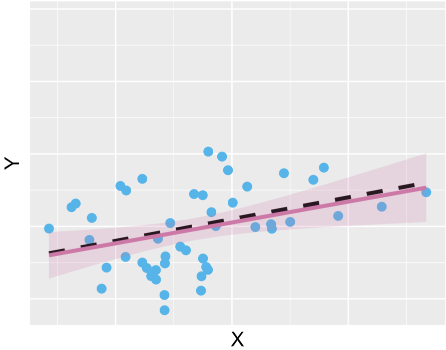
Complete Case Analysis



Only use *uncensored* observations

Solve estimating equation
$\sum_{i=1}^{n} \Delta_i (Y_i - \beta_0 - \beta_1 W_i)(1, W_i)^T = \mathbf{0}$

Complete Case Analysis

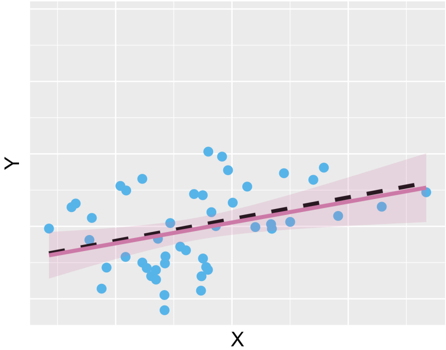


Only use *uncensored* observations

Solve estimating equation
$\sum_{i=1}^{n} \Delta_i (Y_i - \beta_0 - \beta_1 W_i)(1, W_i)^T = \mathbf{0}$

✓ Consistent

Complete Case Analysis



Only use *uncensored* observations

Solve estimating equation
$\sum_{i=1}^{n} \Delta_i (Y_i - \beta_0 - \beta_1 W_i)(1, W_i)^T = \mathbf{0}$

- ✓ Consistent
- ✗ Inefficient

Maximum Likelihood Estimation (MLE)

$$f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boxed{\boldsymbol{\alpha}}) \propto \underbrace{\{f_{Y|X}(y, w, \boldsymbol{\beta})\}^{\delta}}_{\text{uncensored}} \underbrace{\left\{\int_{w}^{\infty} f_{Y|X}(y, x, \boldsymbol{\beta}) \boxed{f_X(x, \boldsymbol{\alpha})} dx\right\}^{1-\delta}}_{\text{censored}}$$

Maximum Likelihood Estimation (MLE)

$$f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boxed{\boldsymbol{\alpha}}) \propto \underbrace{\{f_{Y|X}(y, w, \boldsymbol{\beta})\}^{\delta}}_{\text{uncensored}} \underbrace{\left\{\int_{w}^{\infty} f_{Y|X}(y, x, \boldsymbol{\beta}) \boxed{f_X(x, \boldsymbol{\alpha})} dx\right\}^{1-\delta}}_{\text{censored}}$$

$$\mathbf{S}_{\mathrm{ML}}(y, w, \delta, \boldsymbol{\beta}) \equiv \frac{\partial}{\partial \boldsymbol{\beta}} \log f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boldsymbol{\alpha}), \qquad \sum_{i=1}^{n} \mathbf{S}_{\mathrm{ML}}(Y_i, W_i, \Delta_i, \boldsymbol{\beta}) = \mathbf{0}$$

Maximum Likelihood Estimation (MLE)

$$f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boxed{\boldsymbol{\alpha}}) \propto \underbrace{\{f_{Y|X}(y, w, \boldsymbol{\beta})\}^{\delta}}_{\text{uncensored}} \underbrace{\left\{\int_{w}^{\infty} f_{Y|X}(y, x, \boldsymbol{\beta}) \boxed{f_X(x, \boldsymbol{\alpha})} dx\right\}^{1-\delta}}_{\text{censored}}$$

$$\mathbf{S}_{\mathrm{ML}}(y, w, \delta, \boldsymbol{\beta}) \equiv \frac{\partial}{\partial \boldsymbol{\beta}} \log f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boldsymbol{\alpha}), \qquad \sum_{i=1}^{n} \mathbf{S}_{\mathrm{ML}}(Y_i, W_i, \Delta_i, \boldsymbol{\beta}) = \mathbf{0}$$

✓ consistent

✓ fully efficient

Maximum Likelihood Estimation (MLE)

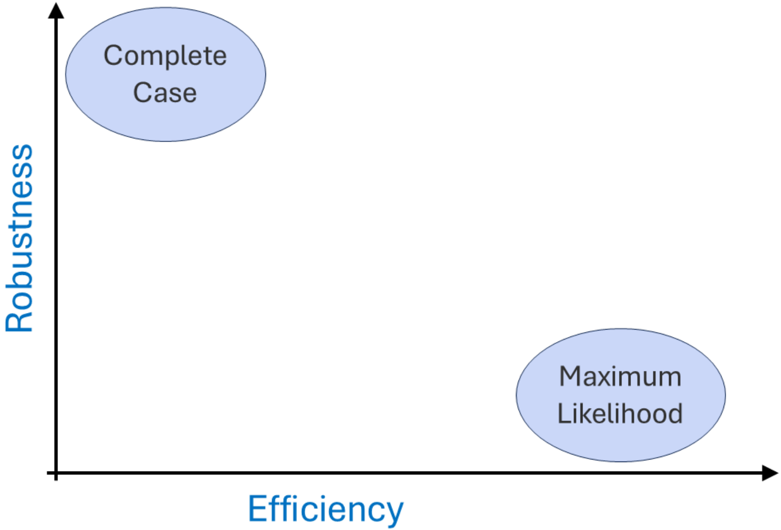
$$f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boxed{\boldsymbol{\alpha}}) \propto \underbrace{\{f_{Y|X}(y, w, \boldsymbol{\beta})\}^{\delta}}_{\text{uncensored}} \underbrace{\left\{\int_{w}^{\infty} f_{Y|X}(y, x, \boldsymbol{\beta}) \boxed{f_X(x, \boldsymbol{\alpha})} dx\right\}^{1-\delta}}_{\text{censored}}$$

$$\mathbf{S}_{\mathrm{ML}}(y, w, \delta, \boldsymbol{\beta}) \equiv \frac{\partial}{\partial \boldsymbol{\beta}} \log f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boldsymbol{\alpha}), \qquad \sum_{i=1}^{n} \mathbf{S}_{\mathrm{ML}}(Y_i, W_i, \Delta_i, \boldsymbol{\beta}) = \mathbf{0}$$

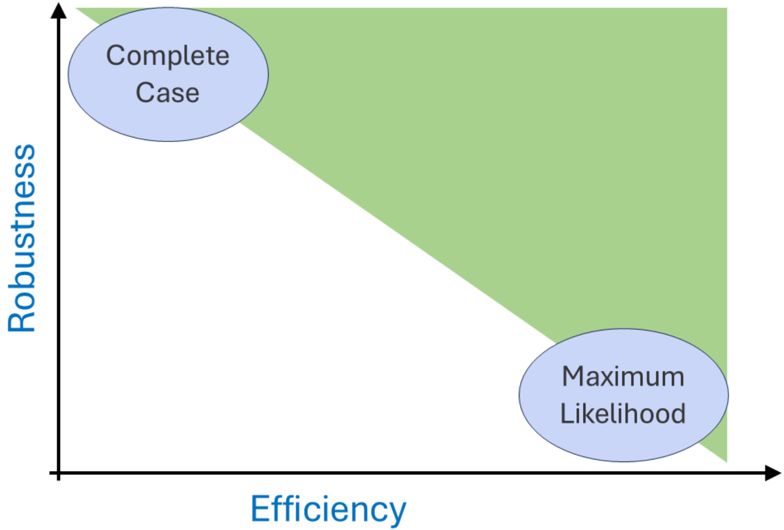
- ✓ consistent
- ✓ fully efficient

- ✗ inconsistent when model for f_X is incorrect

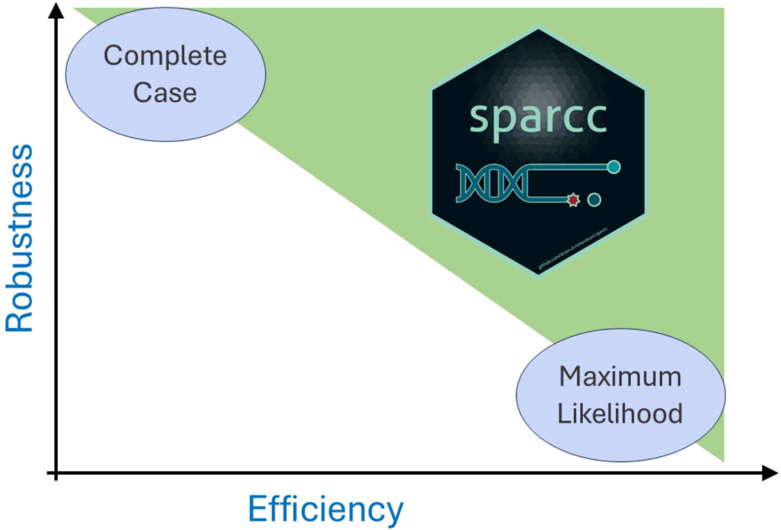
Existing Opportunity



Existing Opportunity



Existing Opportunity



The Semiparametric Recipe

- **model**: $\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$

The Semiparametric Recipe

- **model**: $\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$
- **of interest**: parameter $\boldsymbol{\beta}$ characterizing $Y|X$

The Semiparametric Recipe

- **model**: $\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$
- **of interest**: parameter $\boldsymbol{\beta}$ characterizing $Y|X$
- **nuisance**: distributions $(f_X, f_C) = \boldsymbol{\eta}$

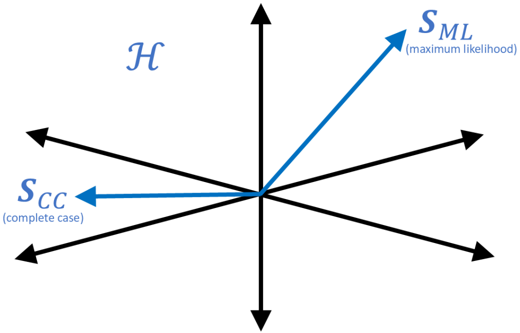
The Semiparametric Recipe

- **model**: $\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$
- **of interest**: parameter $\boldsymbol{\beta}$ characterizing $Y|X$
- **nuisance**: distributions $(f_X, f_C) = \boldsymbol{\eta}$
- **semiparametric**: avoid parametric assumptions for $\boldsymbol{\eta}$

The Semiparametric Recipe

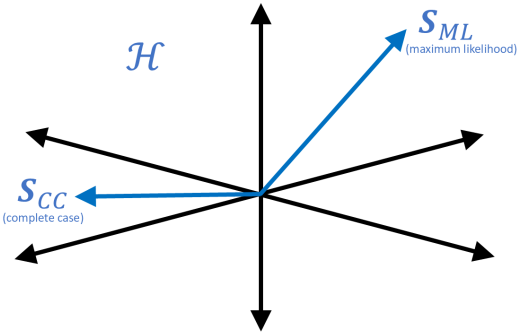
- **model**: $\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$
- **of interest**: parameter $\boldsymbol{\beta}$ characterizing $Y|X$
- **nuisance**: distributions $(f_X, f_C) = \boldsymbol{\eta}$
- **semiparametric**: avoid parametric assumptions for $\boldsymbol{\eta}$
- **goal:** derive the SPARCC estimator

The Semiparametric Recipe



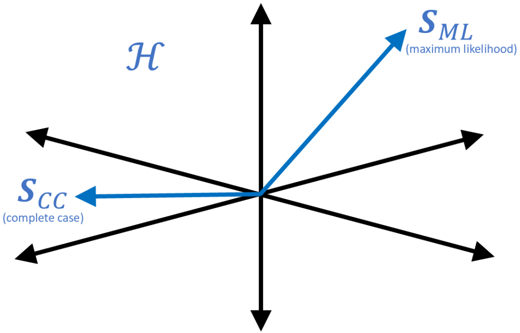
- **Hilbert space** of estimating functions

The Semiparametric Recipe



- **Hilbert space** of estimating functions
- **covariance inner product**
  $\langle \boldsymbol{h}, \boldsymbol{g} \rangle \equiv \mathrm{E}(\boldsymbol{h}^T \boldsymbol{g})$

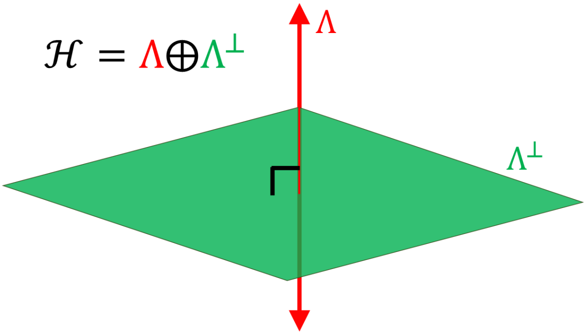
The Semiparametric Recipe



- **Hilbert space** of estimating functions
- **covariance inner product**
  $\langle \boldsymbol{h}, \boldsymbol{g} \rangle \equiv \mathrm{E}(\boldsymbol{h}^T \boldsymbol{g})$
- **orthogonal** $\Leftrightarrow$ **uncorrelated**

$$\boldsymbol{h} \perp \boldsymbol{g} \iff \langle \boldsymbol{h}, \boldsymbol{g} \rangle = 0$$

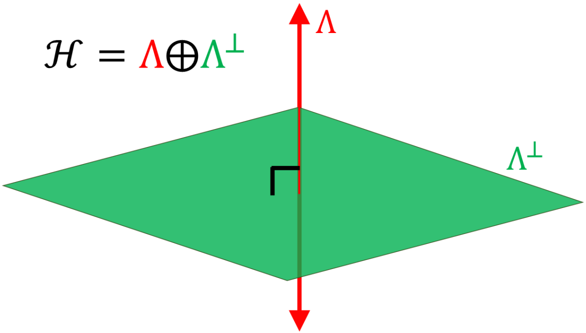
The Semiparametric Recipe



- construct Λ using **nuisance scores**

$$\partial \log f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boldsymbol{\eta}) / \partial \boldsymbol{\eta}$$

The Semiparametric Recipe

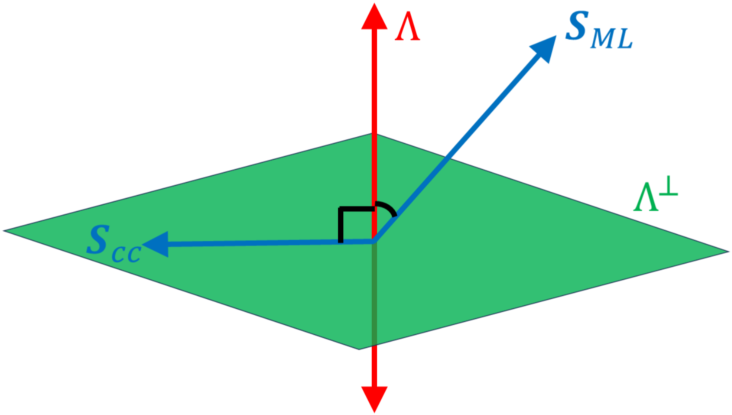


- construct Λ using **nuisance scores**

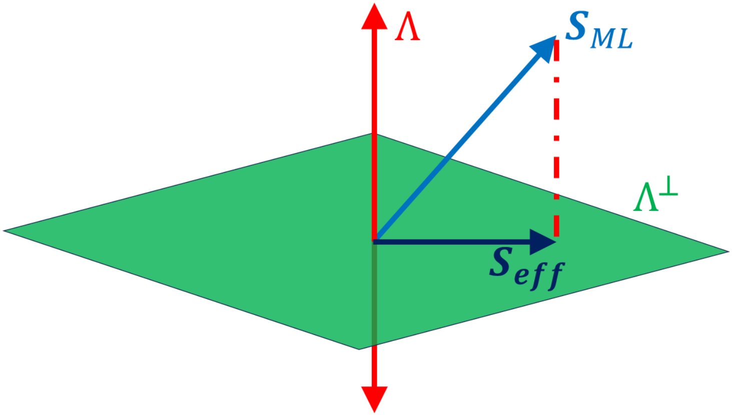
$$\partial \log f_{Y,W,\Delta}(y, w, \delta, \boldsymbol{\beta}, \boldsymbol{\eta}) / \partial \boldsymbol{\eta}$$

- orthogonal complement $\Lambda^{\perp}$

The Semiparametric Recipe



The Semiparametric Recipe



Implementing the SPARCC Estimator

The **SPARCC Estimator** $\widehat{\beta}$ is the solution to

$$\sum_{i=1}^{n} \mathbf{S}_{\mathrm{eff}}(Y_i, W_i, \Delta_i, \beta) = \mathbf{0}$$

Implementing the SPARCC Estimator

The **SPARCC Estimator** $\widehat{\beta}$ is the solution to

$$\sum_{i=1}^{n} \mathbf{S}_{\text{eff}}(Y_i, W_i, \Delta_i, \beta) = \mathbf{0}$$

$\mathbf{S}_{\text{eff}}$ requires $\boldsymbol{\eta} = (f_X, f_C)$

Implementing the SPARCC Estimator

The **SPARCC Estimator** $\widehat{\beta}$ is the solution to

$$\sum_{i=1}^{n} \mathbf{S}_{\text{eff}}(Y_i, W_i, \Delta_i, \beta) = \mathbf{0}$$

$\mathbf{S}_{\text{eff}}$ requires $\boldsymbol{\eta} = (f_X, f_C)$

Can be modeled either **parametrically** or **nonparametrically**

Properties of the SPARCC Estimator

With **parametric** f_X, f_C, $\widehat{\beta}$ is:

✓ **doubly robust**: consistent if one of f_X, f_C is correctly specified

Properties of the SPARCC Estimator

With **parametric** f_X, f_C, $\widehat{\beta}$ is:

- ✓ **doubly robust**: consistent if one of f_X, f_C is correctly specified
- ✓ **semiparametric efficient** if f_X, f_C are *both* correctly specified

Properties of the SPARCC Estimator

With **parametric** f_X, f_C, $\widehat{\beta}$ is:

- ✓ **doubly robust**: consistent if one of f_X, f_C is correctly specified
- ✓ **semiparametric efficient** if f_X, f_C are *both* correctly specified

With **nonparametric** f_X, f_C, $\widehat{\beta}$ is:

- ✓ **consistent**
- ✓ **semiparametric efficient**

Simulation Setup

$$\underbrace{Y}_{\text{outcome}} | X, Z \sim N(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2)$$

Simulation Setup

$$Y | \underbrace{X}_{\text{censored}}, Z \sim N(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2)$$

Simulation Setup

$$Y|X, \underbrace{Z}_{\text{uncensored}} \sim N(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2)$$

Simulation Setup

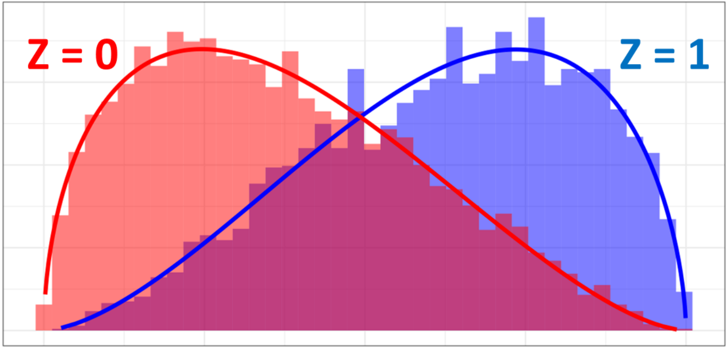
$$Y|X, Z \sim N(\underbrace{\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2}_{\text{parameter of interest: } (\beta_0, \beta_1, \beta_2, \sigma^2)})$$

Simulation Setup: Nuisance Distributions

$$\boldsymbol{\eta} = (f_{X|Z}, f_{C|Z})$$

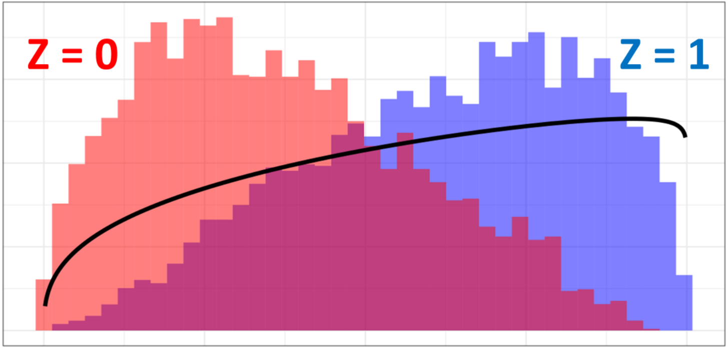
Simulation Setup: Nuisance Distributions

$f_{X|Z}$ correct parametric:



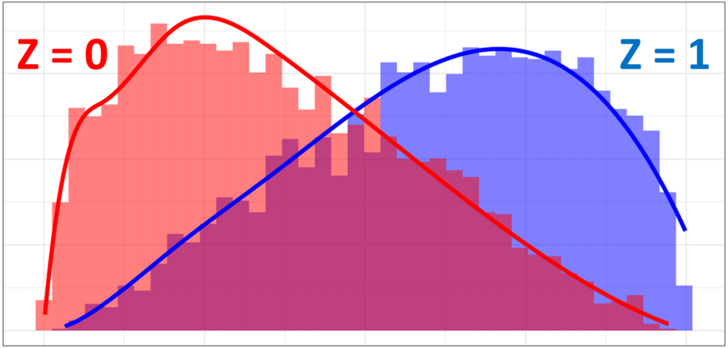
Simulation Setup: Nuisance Distributions

$f_{X|Z}$ incorrect parametric:

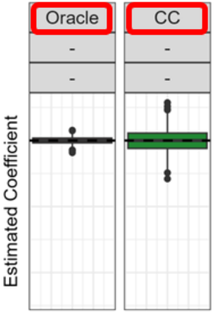


Simulation Setup: Nuisance Distributions

$f_{X|Z}$ nonparametric:

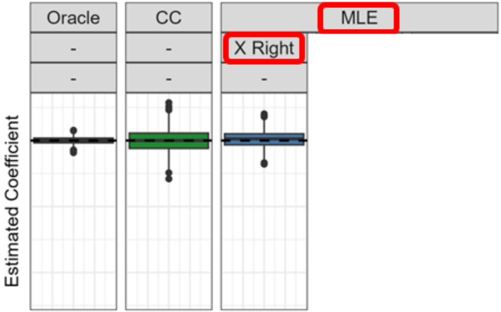


Simulation Results: Robustness



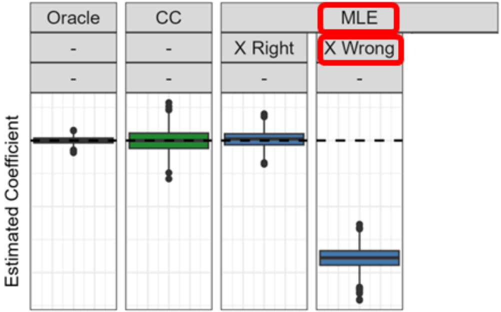
(censoring proportion $q = 0.8$)

Simulation Results: Robustness



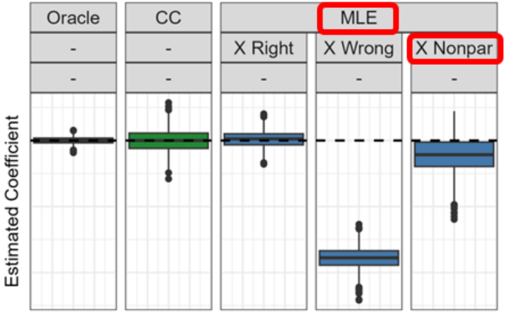
(censoring proportion $q = 0.8$)

Simulation Results: Robustness



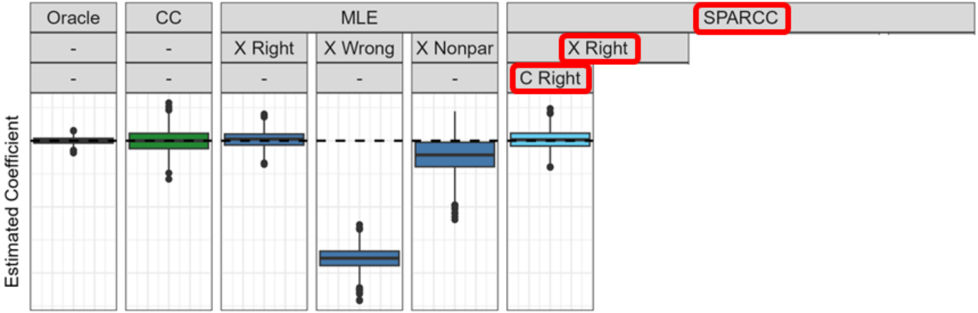
(censoring proportion $q = 0.8$)

Simulation Results: Robustness



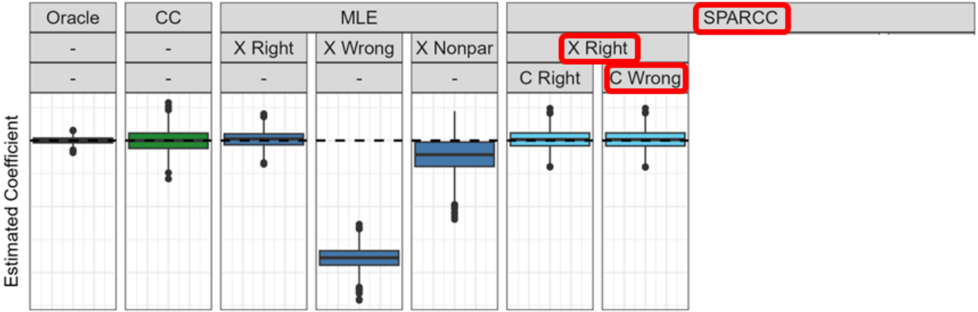
(censoring proportion $q = 0.8$)

Simulation Results: Robustness



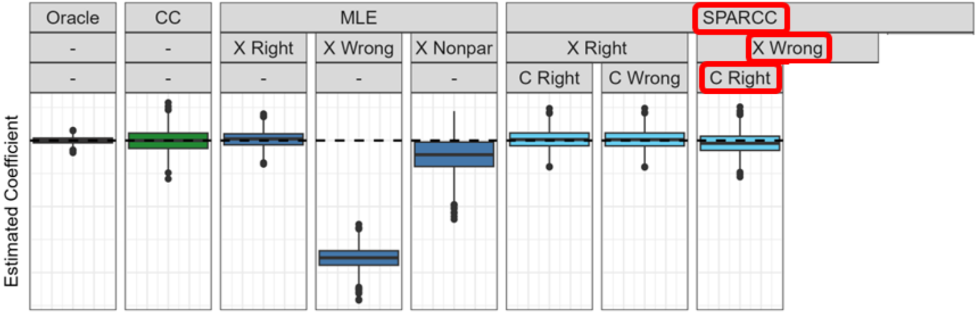
(censoring proportion $q = 0.8$)

Simulation Results: Robustness



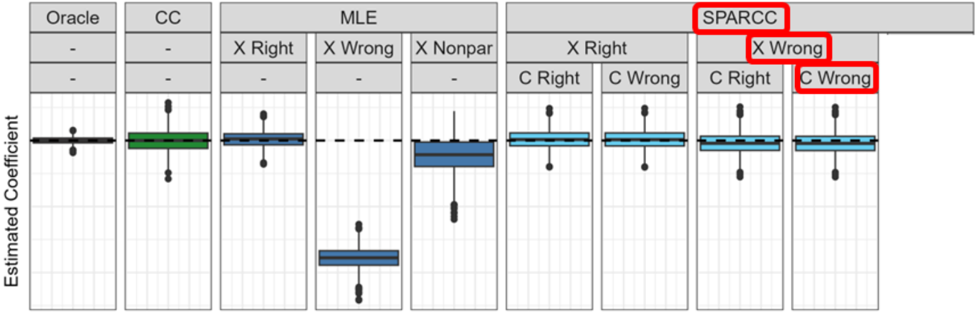
(censoring proportion $q = 0.8$)

Simulation Results: Robustness



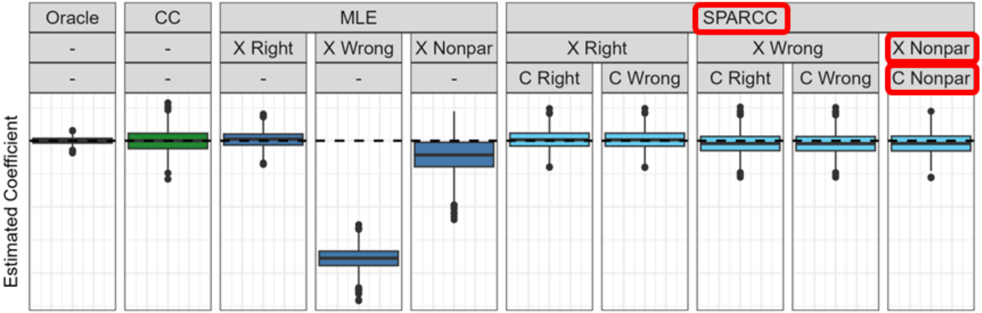
(censoring proportion $q = 0.8$)

Simulation Results: Robustness



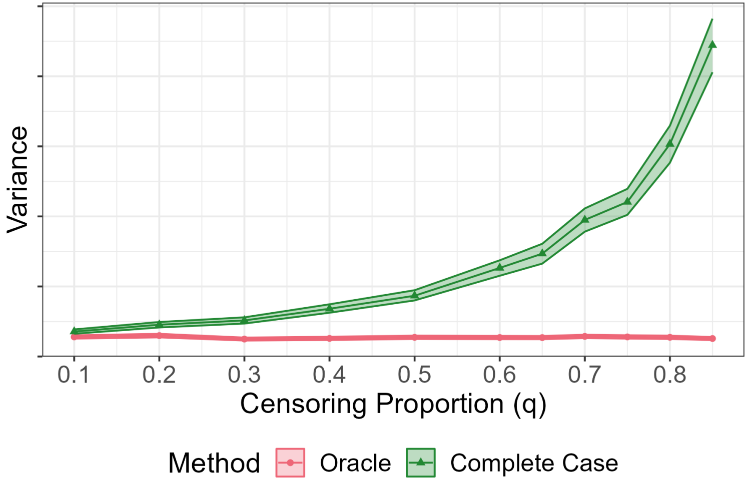
(censoring proportion $q = 0.8$)

Simulation Results: Robustness

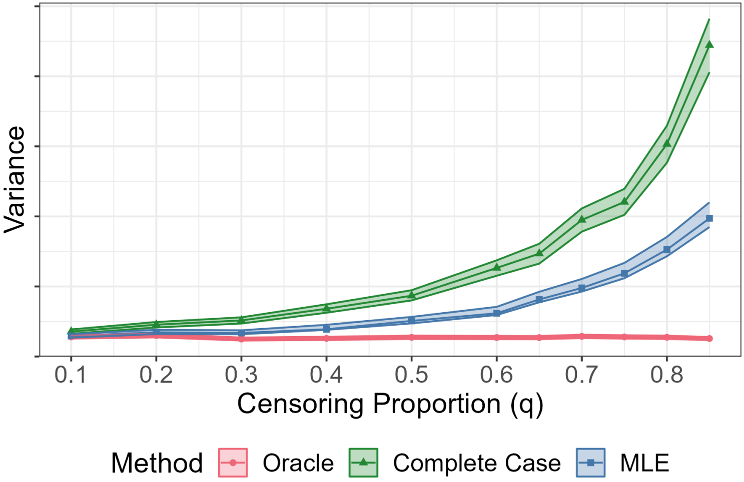


(censoring proportion $q = 0.8$)

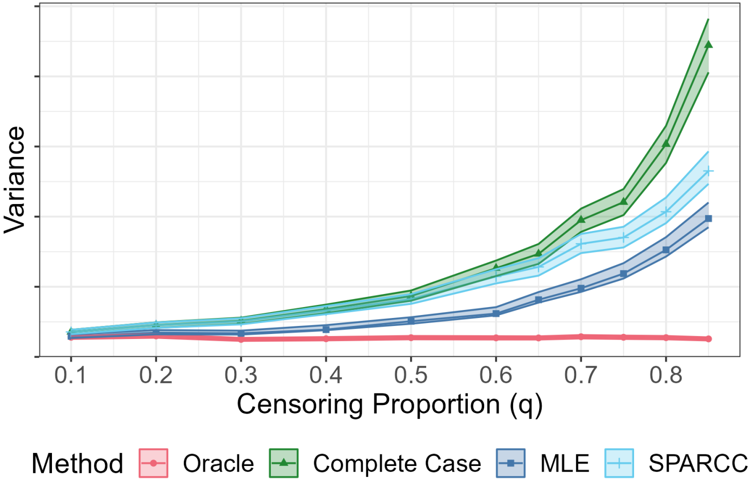
Simulation Results: Efficiency



Simulation Results: Efficiency



Simulation Results: Efficiency

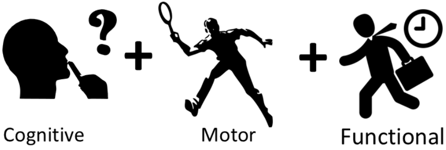


ENROLL-HD Study

- large, observational study of people with Huntington's disease or mutation

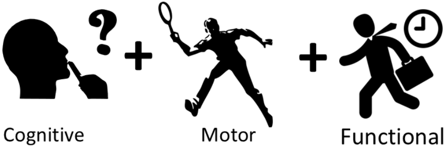
ENROLL-HD Study

- large, observational study of people with Huntington's disease or mutation
- **composite Unified Huntington Disease Rating Scale (cUHDRS) score**



ENROLL-HD Study

- large, observational study of people with Huntington's disease or mutation
- **composite Unified Huntington Disease Rating Scale (cUHDRS) score**



- **CAP** Score: product of age and mutation severity

Huntington's Disease Application

$$\underbrace{Y}_{\text{cUHDRS}} | X, Z \sim N(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2)$$

Huntington's Disease Application

$$Y | \underbrace{X}_{\text{time to diagnosis}}, Z \sim N(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2)$$

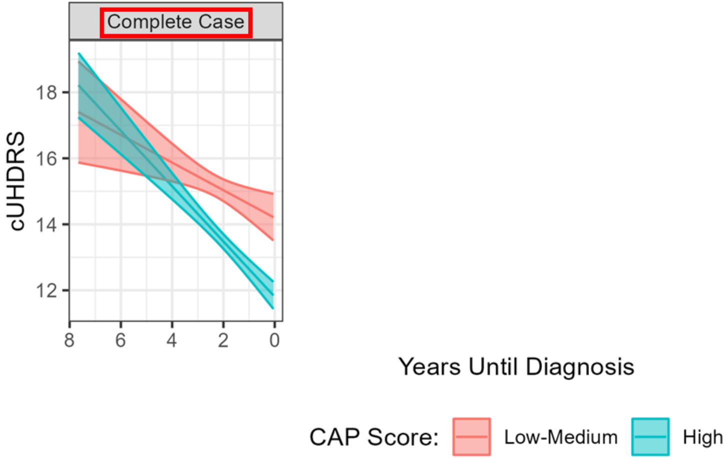
Huntington's Disease Application

$$Y|X, \underbrace{Z}_{\text{CAP group}} \sim N(\beta_0 + \beta_1 X + \beta_2 Z, \sigma^2)$$

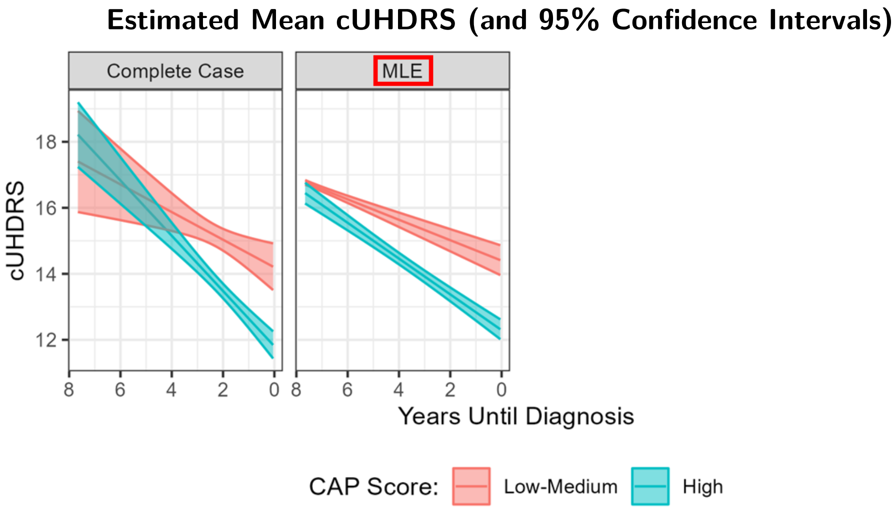
- **sample size**: $n = 4530$
- **censoring rate**: $q = 81.9\%$

Huntington's Disease Results

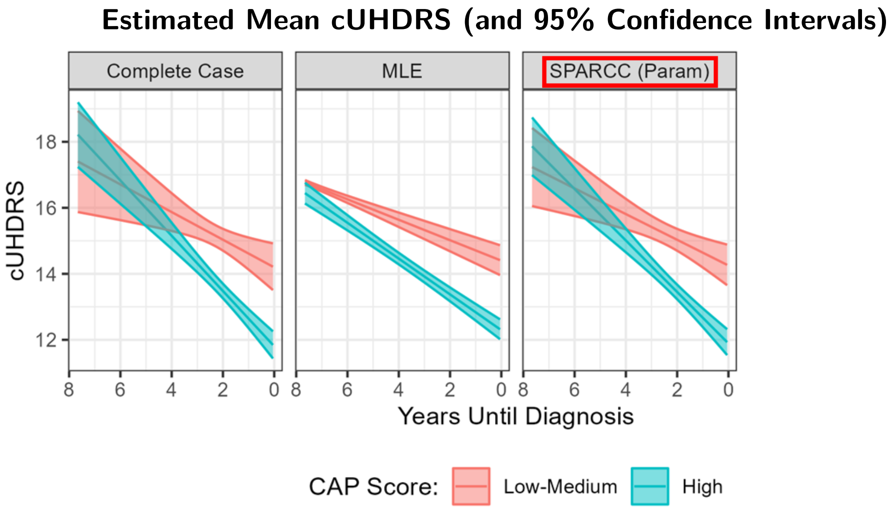
Estimated Mean cUHDRS (and 95% Confidence Intervals)



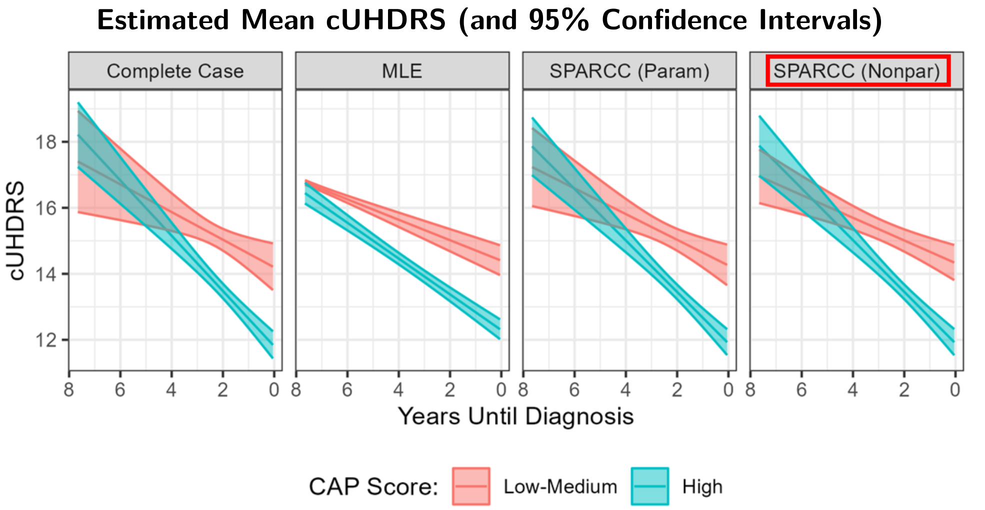
Huntington's Disease Results



Huntington's Disease Results



Huntington's Disease Results



Discussion

Robustness and **efficiency** matter when studying humans

Discussion

Robustness and **efficiency** matter when studying humans

Proposed **SPARCC** estimator has these properties

Discussion

Robustness and **efficiency** matter when studying humans

Proposed **SPARCC** estimator has these properties

Future Work:

- implementation and computation for general models

Discussion

Robustness and **efficiency** matter when studying humans

Proposed **SPARCC** estimator has these properties

Future Work:

- implementation and computation for general models
- longitudinal extension

SPARCC: Semiparametric Censored Covariate Estimation

Paper on arXiv

GitHub R package