

How black-box use of imputation can cause bias

Nicole Erler

Erasmus Medical Center, Rotterdam



Handling Missing Values is Easy!

Functions automatically exclude missing values:

```
## [...]  
## Residual standard error: 2.305 on 69 degrees of freedom  
## (25 observations deleted due to missingness)  
## Multiple R-squared: 0.09255, Adjusted R-squared: 0.02679  
## F-statistic: 1.407 on 5 and 69 DF, p-value: 0.2325
```

Handling Missing Values is Easy!

Functions automatically exclude missing values:

```
## [...]  
## Residual standard error:  2.305 on 69 degrees of freedom  
## (25 observations deleted due to missingness)  
## Multiple R-squared:  0.09255, Adjusted R-squared:  0.02679  
## F-statistic:  1.407 on 5 and 69 DF, p-value:  0.2325
```

Imputation is super easy:

```
library("mice")  
imp <- mice(mydata)
```

However ...

Handling Missing Values Correctly is Not So Easy!

(Imputation) methods makes certain **assumptions**, e.g.:

- **missingness** is M(C)AR

Handling Missing Values Correctly is Not So Easy!

(Imputation) methods makes certain **assumptions**, e.g.:

- **missingness** is M(C)AR
- the incomplete variable has a certain conditional **distribution** (e.g. normal)

Handling Missing Values Correctly is Not So Easy!

(Imputation) methods makes certain **assumptions**, e.g.:

- **missingness** is M(C)AR
- the incomplete variable has a certain conditional **distribution** (e.g. normal)
- all associations are **linear**
 - no interactions
 - no non-linear effects
 - no transformations

Handling Missing Values Correctly is Not So Easy!

(Imputation) methods makes certain **assumptions**, e.g.:

- **missingness** is M(C)AR
- the incomplete variable has a certain conditional **distribution** (e.g. normal)
- all associations are **linear**
 - no interactions
 - no non-linear effects
 - no transformations
- **compatibility** of the imputation models
- **congeniality** (compatibility between analysis and imputation models)

Handling Missing Values Correctly is Not So Easy!

(Imputation) methods makes certain **assumptions**, e.g.:

- **missingness** is M(C)AR
- the incomplete variable has a certain conditional **distribution** (e.g. normal)
- all associations are **linear**
 - no interactions
 - no non-linear effects
 - no transformations
- **compatibility** of the imputation models
- **congeniality** (compatibility between analysis and imputation models)

violation → bias

Several authors have

- investigated robustness to mis-specification (of distribution)
 - in MI using FCS / MICE
 - in joint model MI
- and/or proposed to use
 - Tukey's gh distribution
 - Fleishman polynomials
 - GAMs (in FCS)
 - Doubly-robust weighted estimating equations (instead of MI)

Joint distribution

$$\underbrace{p(y \mid X, b, \theta)}_{\substack{\text{analysis} \\ \text{model}}} \underbrace{p(X \mid \theta)}_{\substack{\text{imputation} \\ \text{part}}} \underbrace{p(b \mid \theta)}_{\substack{\text{random} \\ \text{effects}}} \underbrace{p(\theta)}_{\text{priors}}$$

Joint distribution

$$\underbrace{p(y \mid X, b, \theta)}_{\text{analysis model}} \underbrace{p(X \mid \theta)}_{\text{imputation part}} \underbrace{p(b \mid \theta)}_{\text{random effects}} \underbrace{p(\theta)}_{\text{priors}}$$

Imputation part

$$p(\overbrace{x_1, \dots, x_p, X_{\text{compl.}}}^X \mid \theta) = p(x_1 \mid X_{\text{compl.}}, \theta) \\ p(x_2 \mid X_{\text{compl.}}, x_1, \theta) \\ p(x_3 \mid X_{\text{compl.}}, x_1, x_2, \theta) \\ \dots$$

Fully Bayesian Analysis & Imputation

Joint distribution

$$\underbrace{p(y \mid X, b, \theta)}_{\text{analysis model}} \underbrace{p(X \mid \theta)}_{\text{imputation part}} \underbrace{p(b \mid \theta)}_{\text{random effects}} \underbrace{p(\theta)}_{\text{priors}}$$

Imputation part

$$p(\overbrace{x_1, \dots, x_p, X_{\text{compl.}}}^X \mid \theta) = \begin{aligned} &p(x_1 \mid X_{\text{compl.}}, \theta) \\ &p(x_2 \mid X_{\text{compl.}}, x_1, \theta) \\ &p(x_3 \mid X_{\text{compl.}}, x_1, x_2, \theta) \\ &\dots \end{aligned}$$

Software

Implemented in the **R** package **JointAI**

Imputation in MICE

$$p(x_1 \mid y, X_{\text{compl.}}, x_2, x_3, x_4, \dots, \theta)$$

$$p(x_2 \mid y, X_{\text{compl.}}, x_1, x_3, x_4, \dots, \theta)$$

$$p(x_3 \mid y, X_{\text{compl.}}, x_1, x_2, x_4, \dots, \theta)$$

...

Imputation in JointAI

$$p(y \mid X_{\text{compl.}}, x_1, x_2, x_3, \dots, \theta)$$

$$p(x_1 \mid X_{\text{compl.}}, \theta)$$

$$p(x_2 \mid X_{\text{compl.}}, x_1, \theta)$$

$$p(x_3 \mid X_{\text{compl.}}, x_1, x_2, \theta)$$

...

Imputation in MICE

$$p(x_1 \mid y, X_{\text{compl.}}, x_2, x_3, x_4, \dots, \theta)$$

$$p(x_2 \mid y, X_{\text{compl.}}, x_1, x_3, x_4, \dots, \theta)$$

$$p(x_3 \mid y, X_{\text{compl.}}, x_1, x_2, x_4, \dots, \theta)$$

...

Imputation in JointAI

$$p(y \mid X_{\text{compl.}}, x_1, x_2, x_3, \dots, \theta)$$

$$p(x_1 \mid X_{\text{compl.}}, \theta)$$

$$p(x_2 \mid X_{\text{compl.}}, x_1, \theta)$$

$$p(x_3 \mid X_{\text{compl.}}, x_1, x_2, \theta)$$

...

No issues with

- complex outcomes, e.g.:
 - multi-level
 - survival
- congeniality
- compatibility

Imputation in MICE

$$p(x_1 \mid y, X_{\text{compl.}}, x_2, x_3, x_4, \dots, \theta)$$

$$p(x_2 \mid y, X_{\text{compl.}}, x_1, x_3, x_4, \dots, \theta)$$

$$p(x_3 \mid y, X_{\text{compl.}}, x_1, x_2, x_4, \dots, \theta)$$

...

Imputation in JointAI

$$p(y \mid X_{\text{compl.}}, x_1, x_2, x_3, \dots, \theta)$$

$$p(x_1 \mid X_{\text{compl.}}, \theta)$$

$$p(x_2 \mid X_{\text{compl.}}, x_1, \theta)$$

$$p(x_3 \mid X_{\text{compl.}}, x_1, x_2, \theta)$$

...



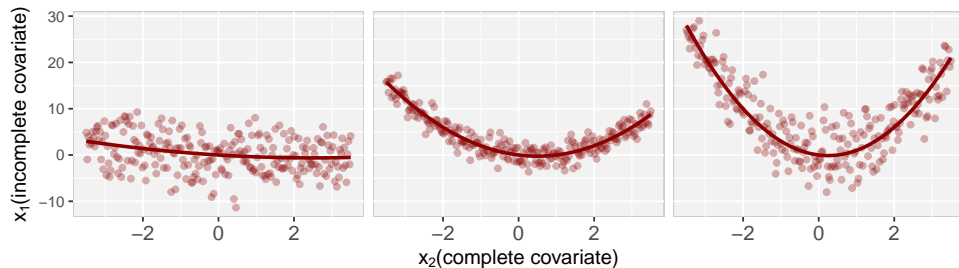
Potential mis-specification of

- functional form
- conditional distribution
- M(C)AR

Simulation Study: Quadratic Effect

Analysis model: $y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$

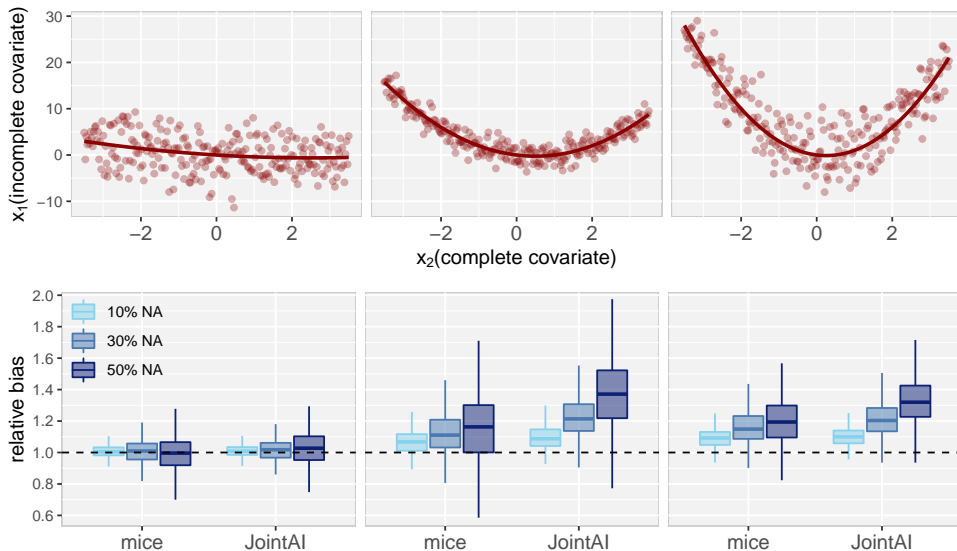
Quadratic association between covariates: $x_1 \sim \alpha_0 + \alpha_1 x_2 + \alpha_2 x_2^2 + \dots$



Simulation Study: Quadratic Effect

Analysis model: $y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$

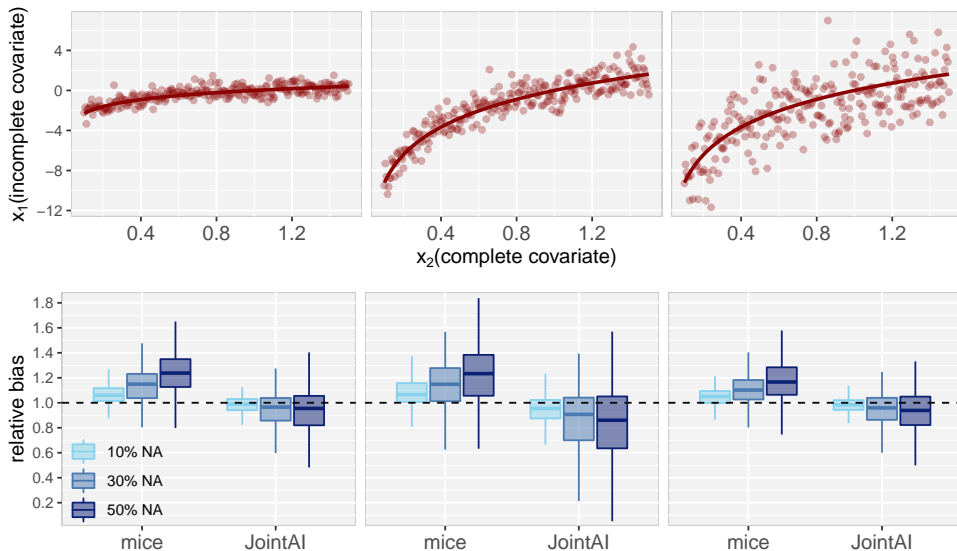
Quadratic association between covariates: $x_1 \sim \alpha_0 + \alpha_1 x_2 + \alpha_2 x_2^2 + \dots$



Simulation Study: Logarithmic Effect

Analysis model: $y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$

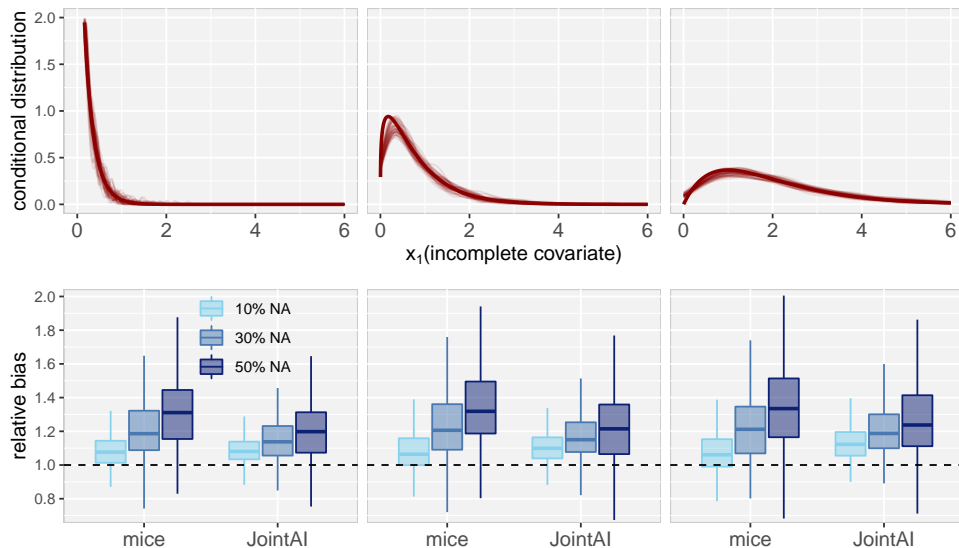
Log-association between covariates: $x_1 \sim \alpha_0 + \alpha_1 \log(x_2) + \dots$



Simulation Study: Gamma distribution

Analysis model: $y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$

Gamma-distributed covariate: $x_1 \mid x_2, x_3, \dots \sim \text{Ga}()$



We need more flexible imputation models!

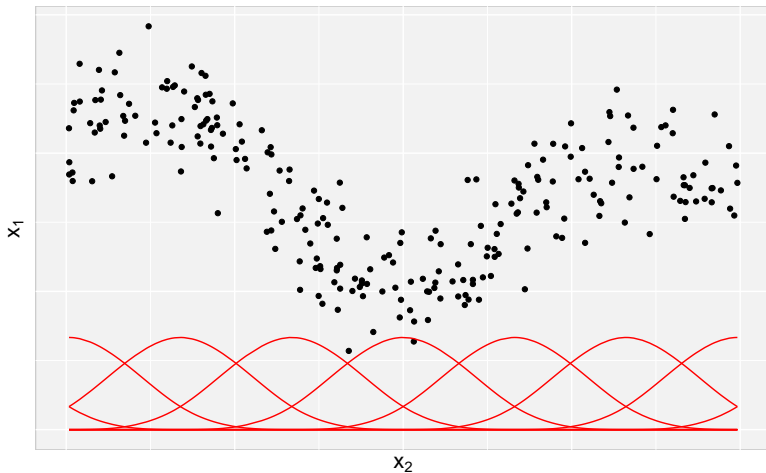
Ideally: models that fit (almost) any distribution / association structure.

Ideas:

- flexible **association** structure: **penalized splines**
- flexible residual **distribution**: **mixture of Polya-Trees**

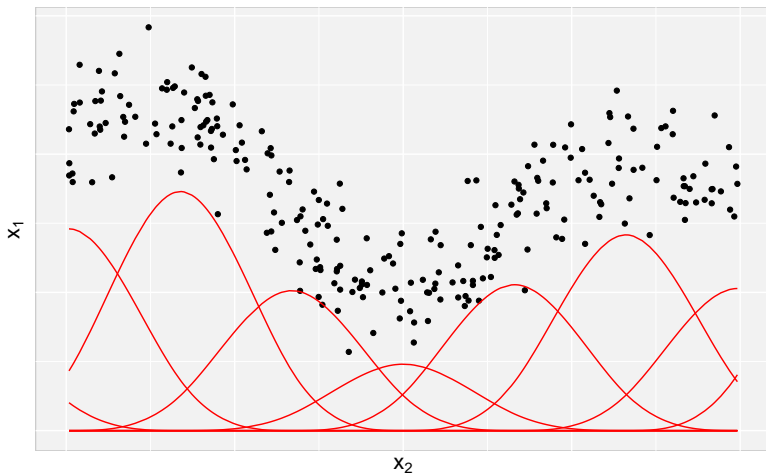
Bayesian P-Splines

Instead of $\beta_1 x_2$ we use $\sum_{\ell=1}^d \beta_{\ell} B_{\ell}(x_1)$:



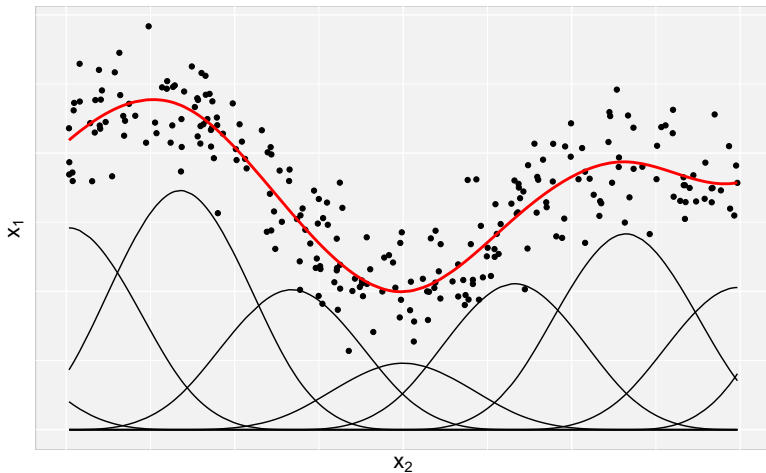
Bayesian P-Splines

Instead of $\beta_1 x_2$ we use $\sum_{\ell=1}^d \beta_{\ell} B_{\ell}(x_1)$:



Bayesian P-Splines

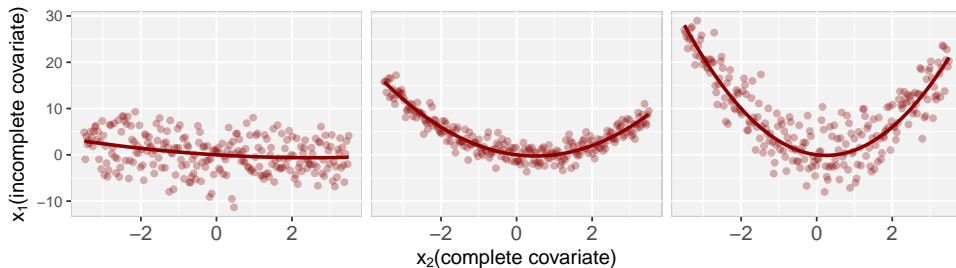
Instead of $\beta_1 x_2$ we use $\sum_{\ell=1}^d \beta_{\ell} B_{\ell}(x_1)$:



Simulation: Bayesian P-Splines

Analysis model: $y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$

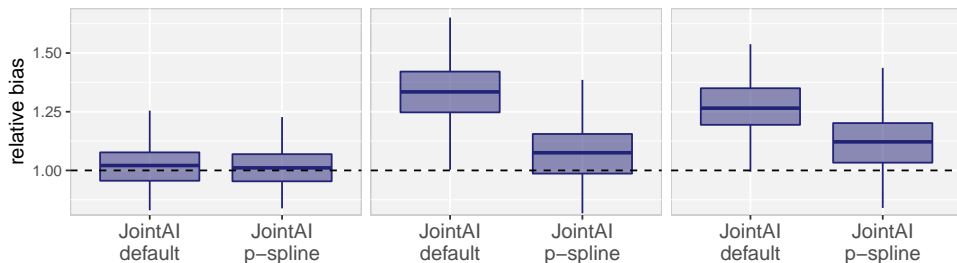
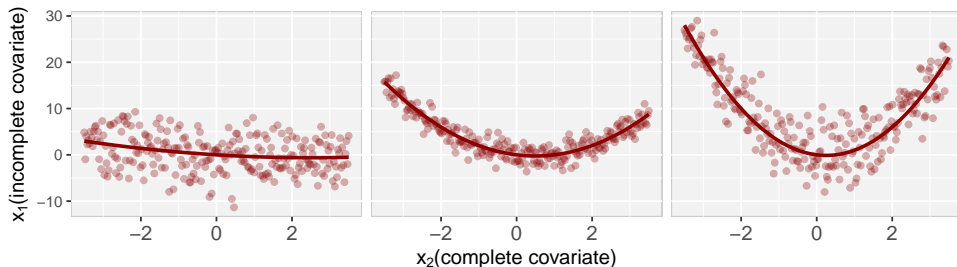
Quadratic association between covariates: $x_1 \sim \alpha_0 + \alpha_1 x_2 + \cancel{\alpha_2 x_2} + \dots$



Simulation: Bayesian P-Splines

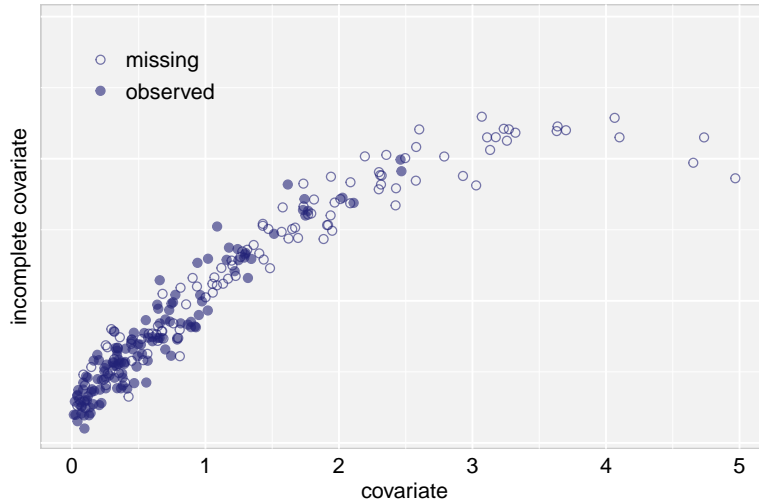
Analysis model: $y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$

Quadratic association between covariates: $x_1 \sim \alpha_0 + \alpha_1 x_2 + \alpha_2 x_2^2 + \dots$



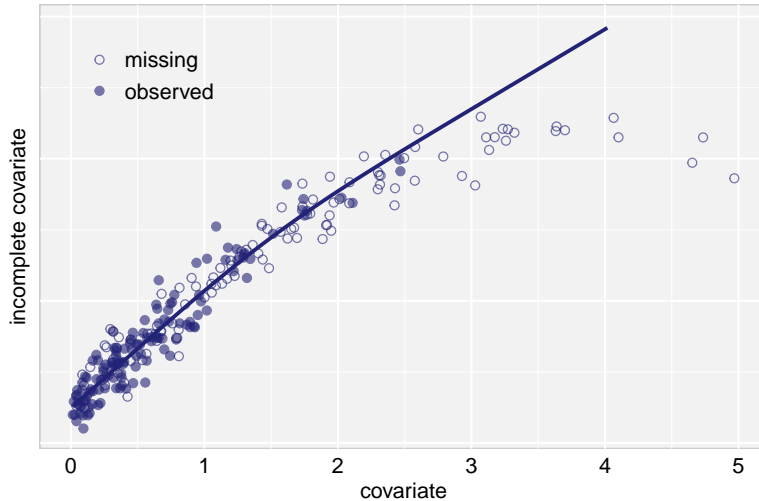
Simulation: Bayesian P-Splines

Potential Issue:

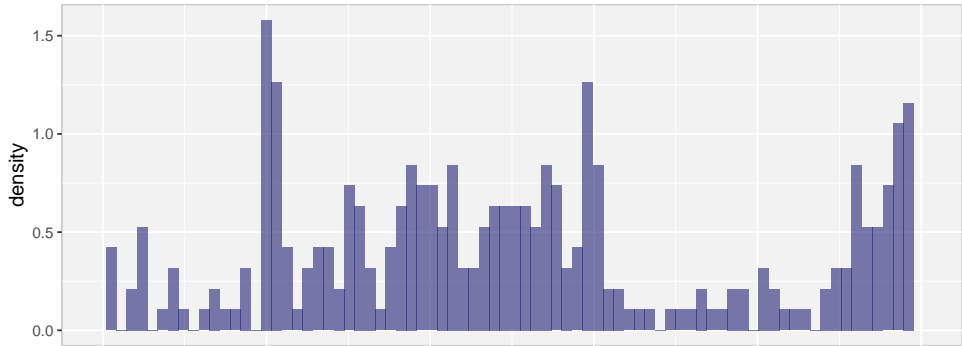


Simulation: Bayesian P-Splines

Potential Issue:

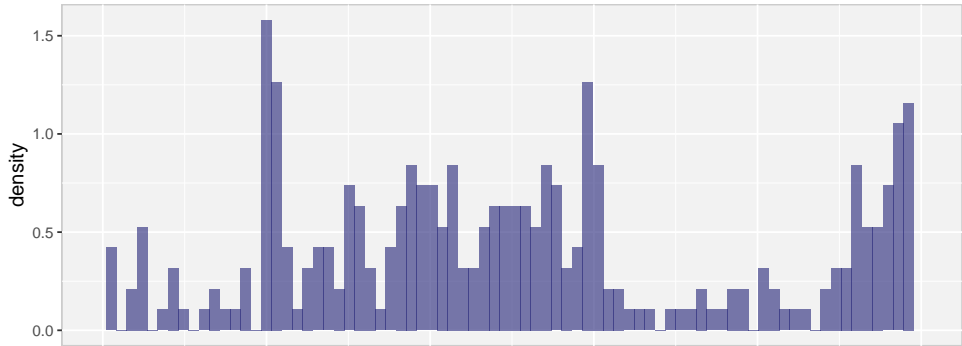


Mixture of Polya Trees



$\text{Beta}(a_0, a_0)$

Mixture of Polya Trees

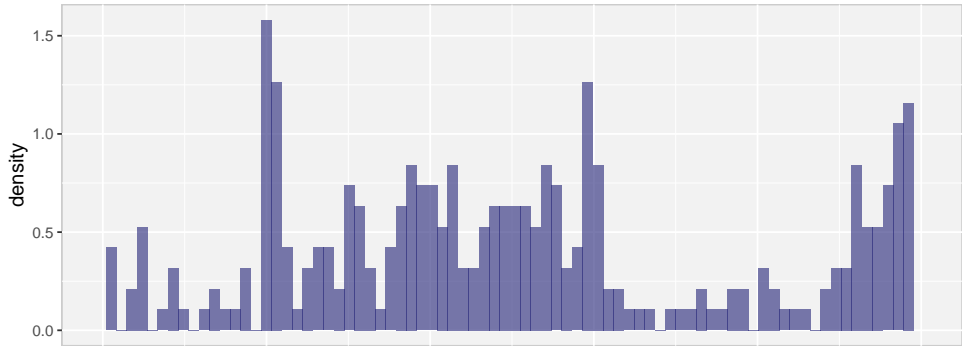


$\text{Beta}(a_0, a_0)$

$\text{Beta}(a_1, a_1)$

$\text{Beta}(a_1, a_1)$

Mixture of Polya Trees



$\text{Beta}(a_0, a_0)$

$\text{Beta}(a_1, a_1)$

$\text{Beta}(a_1, a_1)$

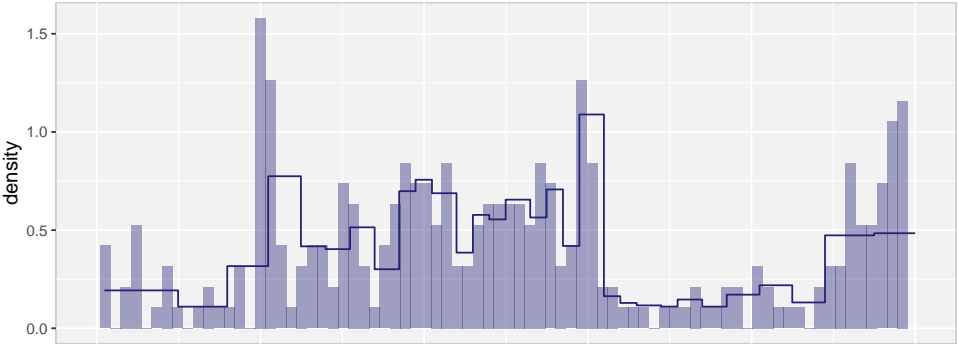
$\text{Beta}(a_2, a_2)$

$\text{Beta}(a_2, a_2)$

$\text{Beta}(a_2, a_2)$

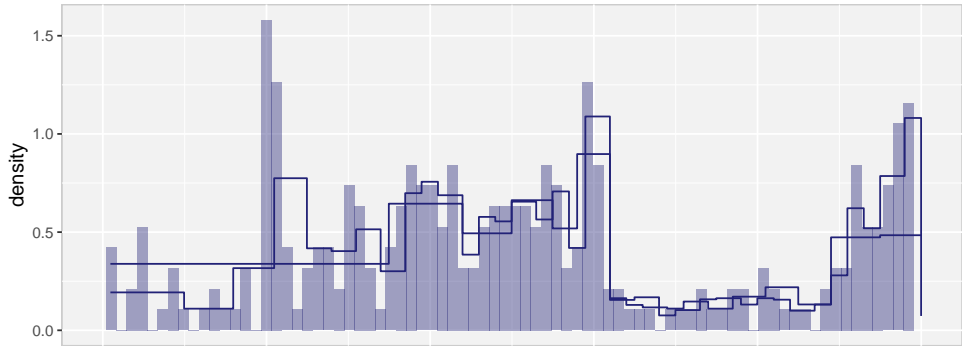
$\text{Beta}(a_2, a_2)$

Mixture of Polya Trees



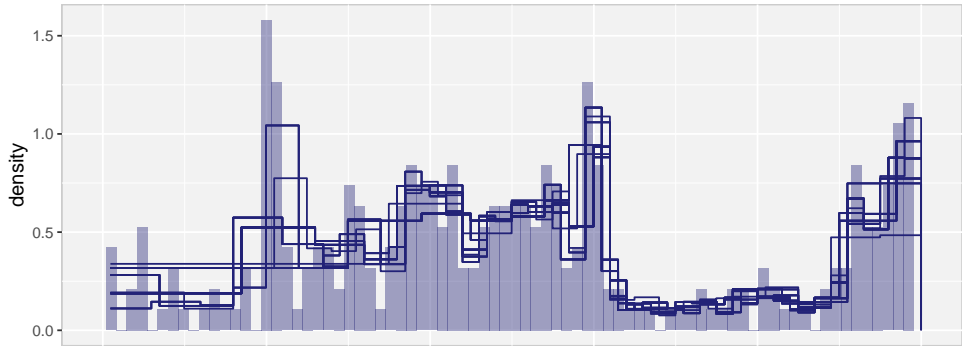
Beta(a ₀ , a ₀)							
Beta(a ₁ , a ₁)				Beta(a ₁ , a ₁)			
Beta(a ₂ , a ₂)		Beta(a ₂ , a ₂)		Beta(a ₂ , a ₂)		Beta(a ₂ , a ₂)	
Beta(a ₃ , a ₃)	Beta(a ₃ , a ₃)					Beta(a ₃ , a ₃)	Beta(a ₃ , a ₃)

Mixture of Polya Trees



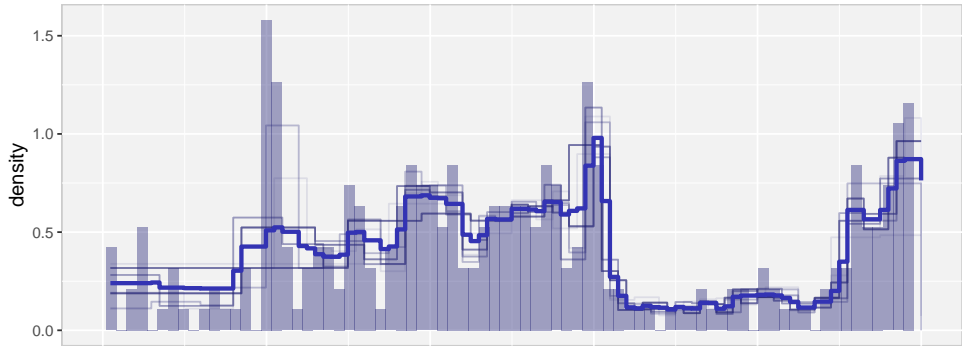
Beta(a_0, a_0)							
Beta(a_1, a_1)				Beta(a_1, a_1)			
Beta(a_2, a_2)		Beta(a_2, a_2)		Beta(a_2, a_2)		Beta(a_2, a_2)	
Beta(a_3, a_3)	Beta(a_3, a_3)					Beta(a_3, a_3)	Beta(a_3, a_3)

Mixture of Polya Trees



Beta(a_0, a_0)							
Beta(a_1, a_1)				Beta(a_1, a_1)			
Beta(a_2, a_2)		Beta(a_2, a_2)		Beta(a_2, a_2)		Beta(a_2, a_2)	
Beta(a_3, a_3)	Beta(a_3, a_3)					Beta(a_3, a_3)	Beta(a_3, a_3)

Mixture of Polya Trees



Beta(a_0, a_0)							
Beta(a_1, a_1)				Beta(a_1, a_1)			
Beta(a_2, a_2)		Beta(a_2, a_2)		Beta(a_2, a_2)		Beta(a_2, a_2)	
Beta(a_3, a_3)	Beta(a_3, a_3)					Beta(a_3, a_3)	Beta(a_3, a_3)

Potential / probable issues for practice:

- flexible fit needs observed data everywhere
- computational time

Potential / probable issues for practice:

- flexible fit needs observed data everywhere
- computational time

Ideas:

- check first if simple model fits, e.g. posterior predictive checks
 - χ^2 type of tests
 - Kolmogorov-Smirnoff test?
 - discordance tests?

Potential / probable issues for practice:

- flexible fit needs observed data everywhere
- computational time

Ideas:

- check first if simple model fits, e.g. posterior predictive checks
 - χ^2 type of tests
 - Kolmogorov-Smirnoff test?
 - discordance tests?
- feasibility checks before running the complex model

- **assumptions** of imputation models can easily be **violated** ➔ **bias**
- **more flexible** imputation models are **needed**
- semi- / **non-parametric (Bayesian) methods** can offer a **solution**
- more **flexibility** ➔ more **complexity** ➔ need for **guidance** tools

Thank you for your attention.



n.erler@erasmusmc.nl



[N_Erler](#)



[NErler](#)

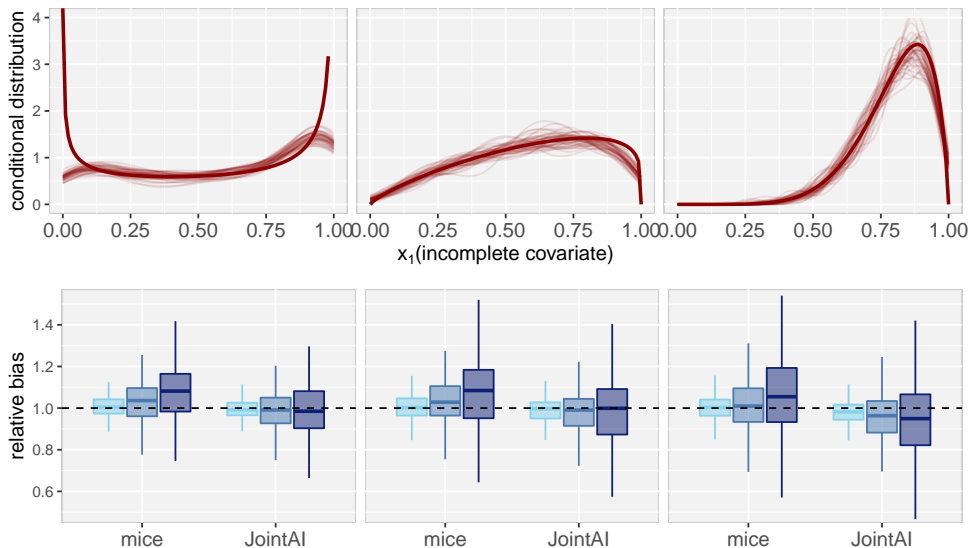


www.nerler.com

Simulation Study: Beta Distribution

Analysis model: $y \sim \beta_0 + \beta_1 x_1 + \beta_2 + \dots$

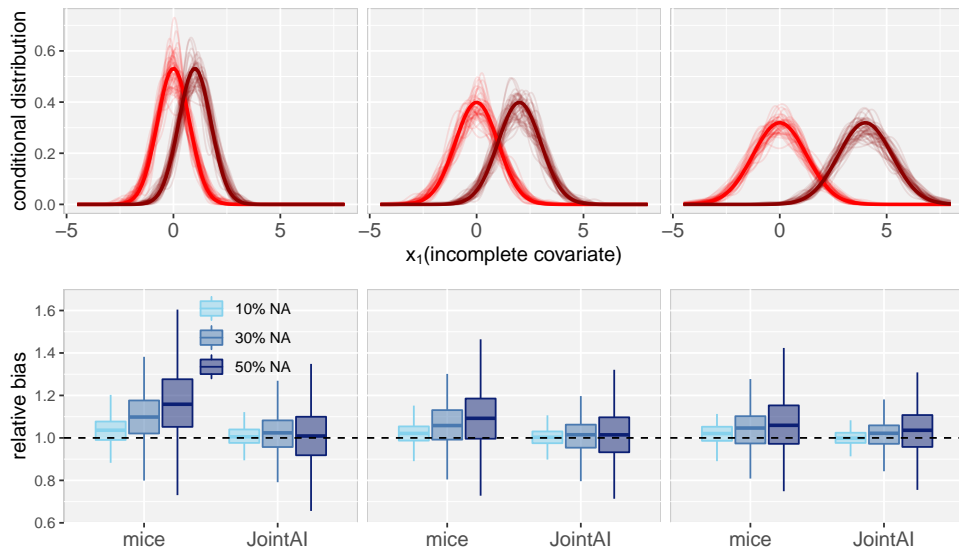
Beta-distributed covariate: $x_1 \mid x_2, x_3, \dots \sim \text{Be}()$



Simulation Study: Reverting the sequence

Analysis model: $y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$

Exclusion of important predictor: $x_1 \sim \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3 + \dots$



Simulation Study: Ignoring an Interaction

Analysis model: $y \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$

Interaction between covariates: $x_1 \sim \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3 + \cancel{\alpha_3 x_2 x_3} + \dots$

