Imputation of Incomplete Covariates in Longitudinal Data

Can Bayesian non-parametric methods prevent model-misspecification?

Nicole Erler and Dimitris Rizopoulos

Erasmus Medical Center, Rotterdam

15 July 2019

Motivation

What are risk factors for diabetic retinopathy?

important predictors:

- blood pressure
- haemoglobin A1c (HbA_{1c})

other covariates:

- age at baseline
- gender
- diabetes duration
- smoking history & status

Motivation

Challenge:

Missing values

retinopathy grade:	43%
blood pressure:	20%
Hb_{A1c}:	20%
diabetes duration:	11%
smoking history:	33%
smoking status:	28%

Motivation

Challenge:

Missing values

retinopathy grade:	43%
blood pressure:	20%
Hb_{A1c}:	20%
diabetes duration:	11%
smoking history:	33%
smoking status:	28%

Solution?

- Multiple Imputation
  - MICE / FCS
  - Joint Model (e.g. multivariate normal)
- Fully Bayesian
- ...

Fully Bayesian Analysis & Imputation

Joint distribution

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

Fully Bayesian Analysis & Imputation

Joint distribution

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

Imputation part

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

Fully Bayesian Analysis & Imputation

Joint distribution

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

Imputation part

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

Software

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

Handling Missing Values

Assumptions about

association structure	➡	linear, additive
conditional distribution	➡	normal (for continuous)
missingness process	➡	ignorable

Handling Missing Values

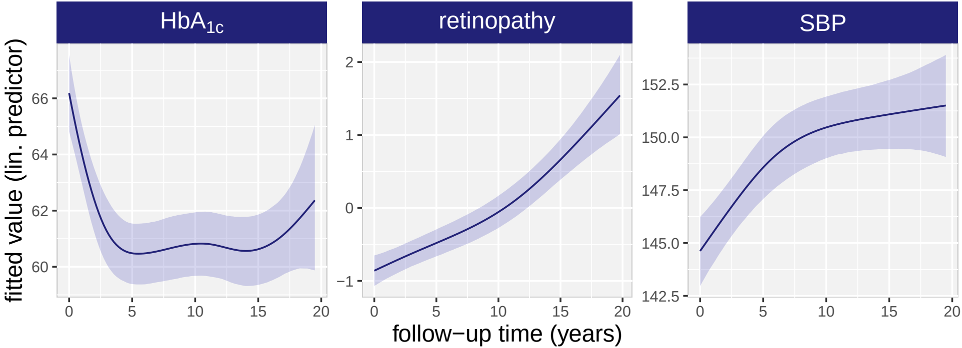
Assumptions about

association structure	➡	linear, additive
conditional distribution	➡	normal (for continuous)
missingness process	➡	ignorable

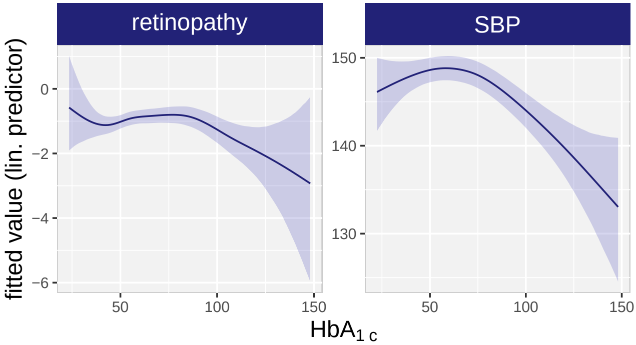
Violation of the implied assumptions may result in **bias**!

Real Data

Non-linear evolutions over time



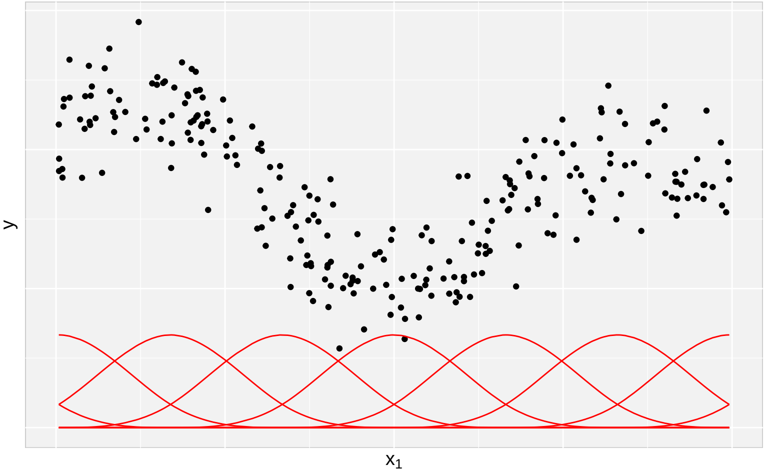
Non-linear associations among variables



Bayesian P-Splines

Instead of $y \sim \beta_0 + \beta_1 x_1 + \ldots$ we assume

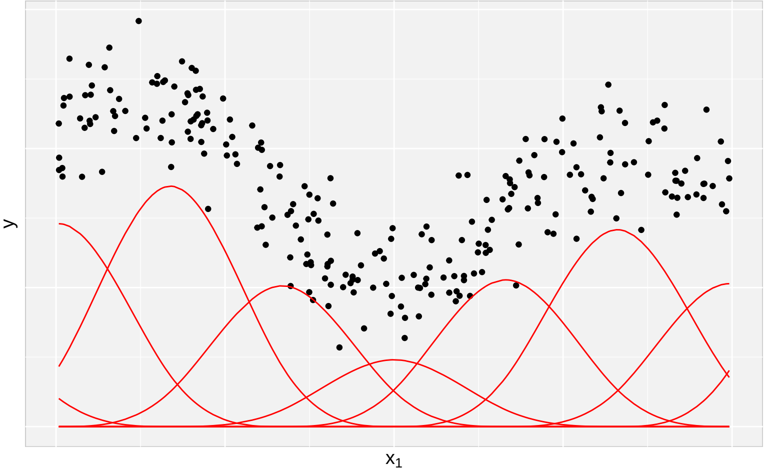
$$y \sim \beta_0 + \sum_{\ell=1}^{d} \beta_\ell B_\ell(x_1) + \ldots$$



Bayesian P-Splines

Instead of $y \sim \beta_0 + \beta_1 x_1 + \ldots$ we assume

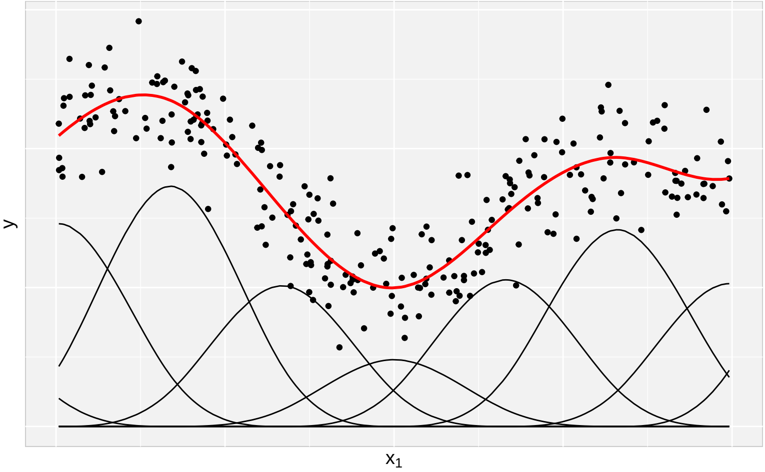
$$y \sim \beta_0 + \sum_{\ell=1}^{d} {\color{red}\beta_\ell} B_\ell(x_1) + \ldots$$



Bayesian P-Splines

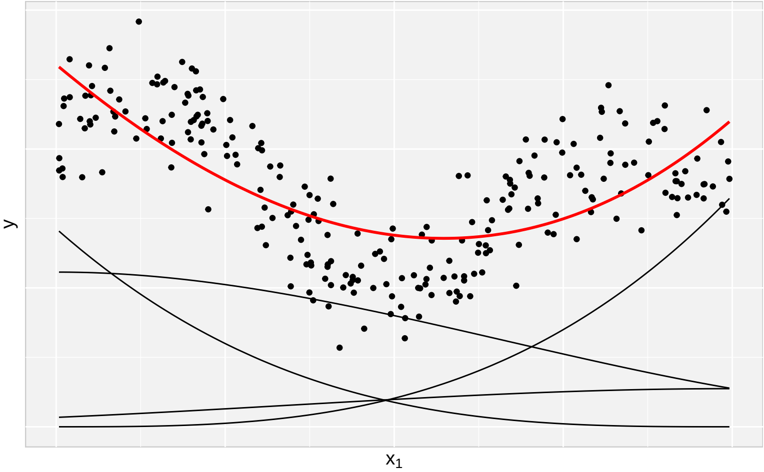
Instead of $y \sim \beta_0 + \beta_1 x_1 + \ldots$ we assume

$$y \sim \beta_0 + \sum_{\ell=1}^{d} \beta_\ell B_\ell(x_1) + \ldots$$



How many B_ℓ's do we need?

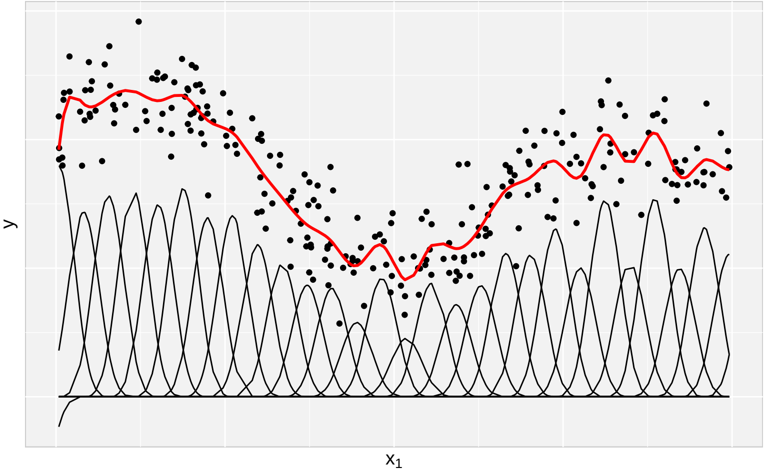
$$y \sim \beta_0 + \sum_{\ell=1}^{d=4} \beta_\ell B_\ell(x_1) + \dots$$



Bayesian P-Splines

How many B_ℓ's do we need?

$$y \sim \beta_0 + \sum_{\ell=1}^{d=30} \beta_\ell B_\ell(x_1) + \ldots$$



Bayesian P-Splines

Idea: Use many functions but **restrict neighboring β's** to be similar:

$$(\beta_1, \ldots, \beta_d) \sim MVN(\mathbf{0}, 1/\sigma^2 \mathbf{D}^T\mathbf{D}),$$

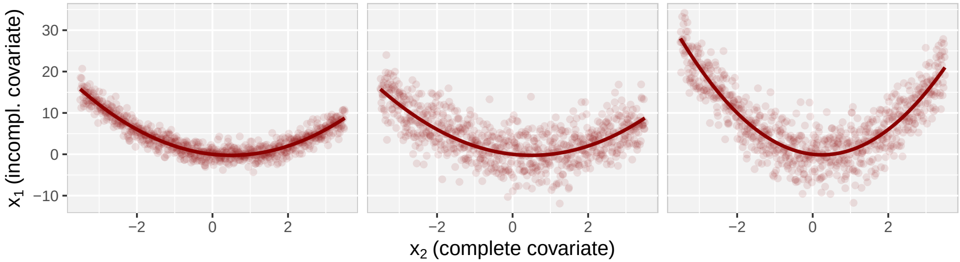
with **penalty matrix D**, for example:

$$\mathbf{D} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 1 & -2 & 1 & 0 & 0 & \cdots \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ \cdots & 0 & 0 & 1 & -2 & 1 & 0 \\ \cdots & 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix}$$

Simulation

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

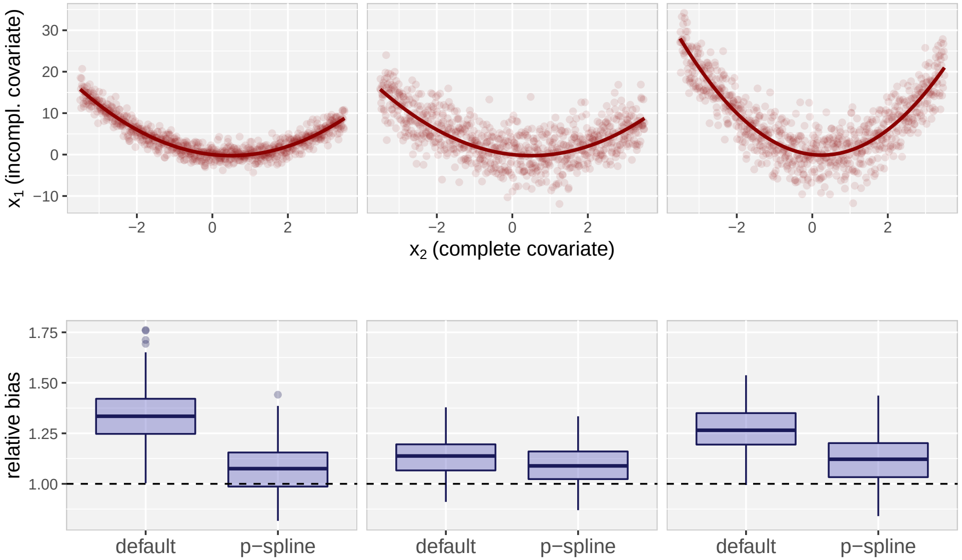
Quadratic association between covariates: $x_1 \sim \alpha_0 + \alpha_1 x_2 + \alpha_2 x_2^2 + \ldots$



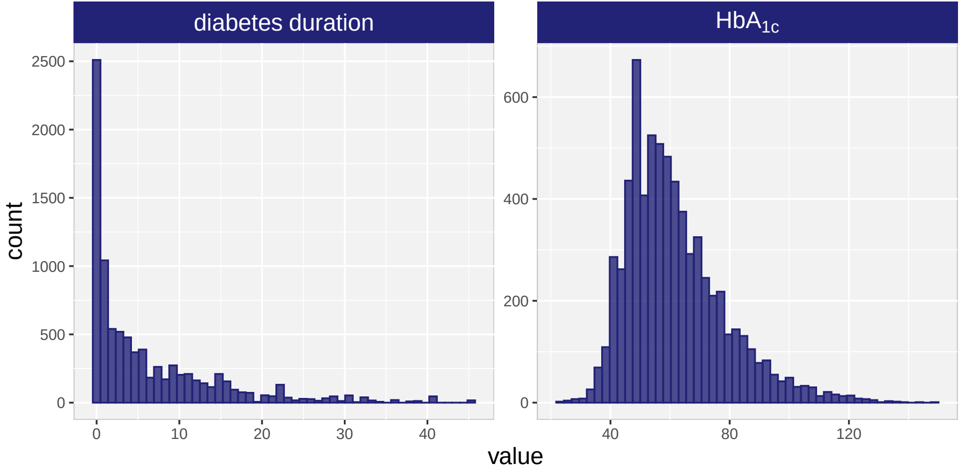
Simulation

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

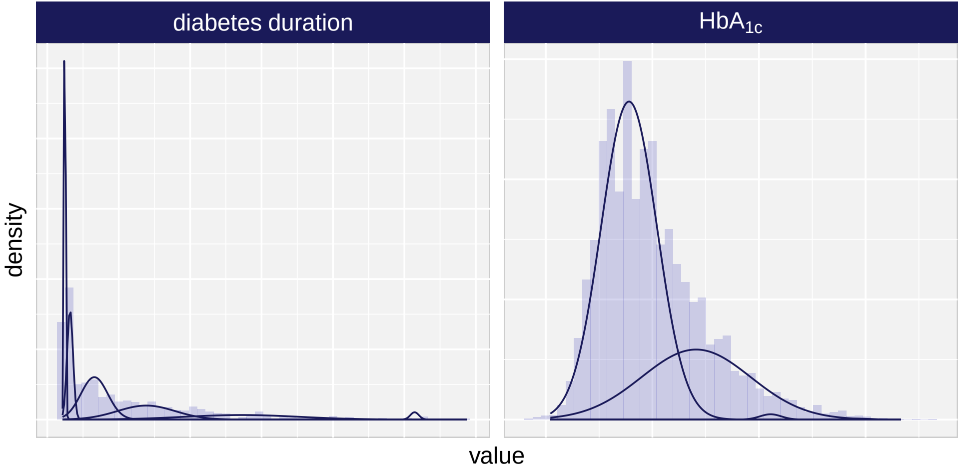
Quadratic association between covariates: $x_1 \sim \alpha_0 + \alpha_1 x_2 + \alpha_2 x_2^2 + \ldots$



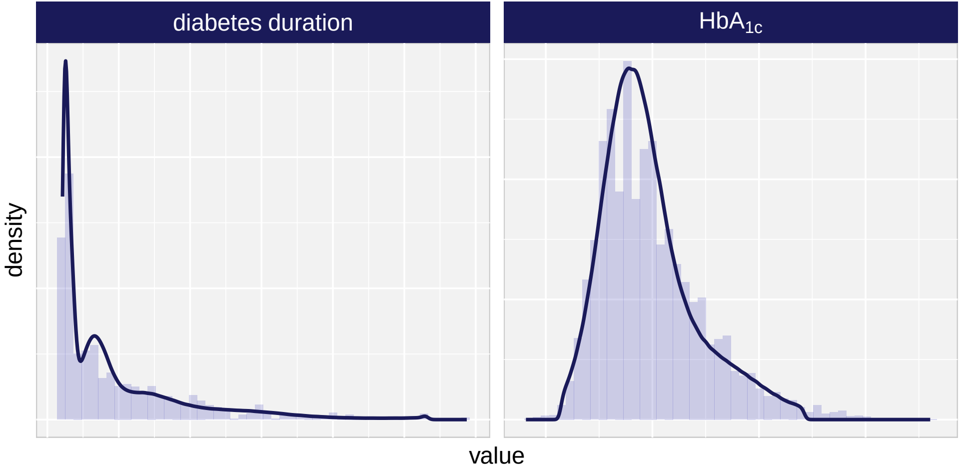
Non-normal continuous distributions



Mixture of normal distributions



Mixture of normal distributions



Dirichlet Process Mixture Model

$$x_{1i} \mid \theta_i \sim F(\theta_i)$$

$$\theta_i \mid G \sim G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

$$G \mid \alpha, G_0 \sim \underset{\substack{\downarrow \\ \text{stick-breaking construction}}}{DP}(\alpha, G_0)$$

e.g. $x_{1i} \sim N(\mu_k, \sigma_k^2)$

Dirichlet Process Mixture Model

$$x_{1i} \mid \theta_i \sim F(\theta_i)$$

$$\theta_i \mid G \sim G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

$$G \mid \alpha, G_0 \sim \underset{\substack{\downarrow \\ \text{stick-breaking construction}}}{DP}(\alpha, G_0)$$

e.g. ~~$x_{1i} \sim N(\mu_k, \sigma_k^2)$~~

$$x_i \sim N(\eta_i + \underset{\substack{\downarrow \\ p(\mu_k)}}{\mu_k}, \underset{\substack{\downarrow \\ p(\sigma_k^2)}}{\sigma_k^2}), \qquad \text{with } \eta_i = \alpha_1 x_{2i} + \alpha_2 x_{3i} + \dots$$

Dirichlet Process Mixture Model

$$x_{1i} \mid \theta_i \sim F(\theta_i)$$

$$\theta_i \mid G \sim G = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k^*}$$

$$G \mid \alpha, G_0 \sim \underset{\substack{\downarrow \\ \text{stick-breaking construction}}}{DP}(\alpha, G_0)$$

e.g. ~~$x_{1i} \sim N(\mu_k, \sigma_k^2)$~~

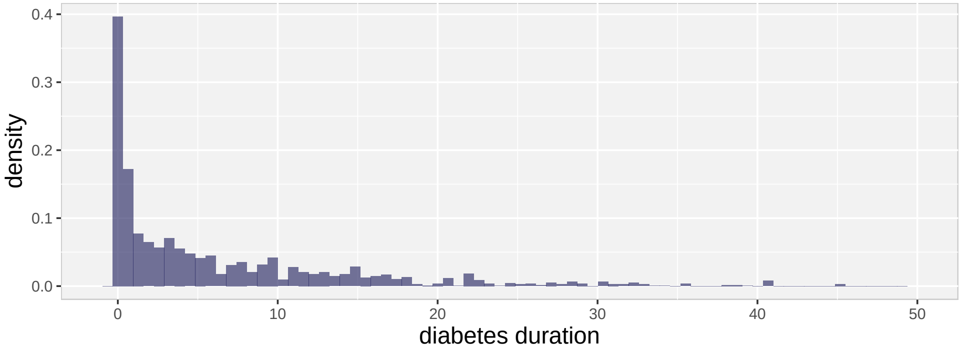
$$x_i \sim N(\eta_i + \underset{\substack{\downarrow \\ p(\mu_k)}}{\mu_k}, \underset{\substack{\downarrow \\ p(\sigma_k^2)}}{\sigma_k^2}),$$

with $\eta_i = \alpha_1 x_{2i} + \alpha_2 x_{3i} + \ldots$

⬇

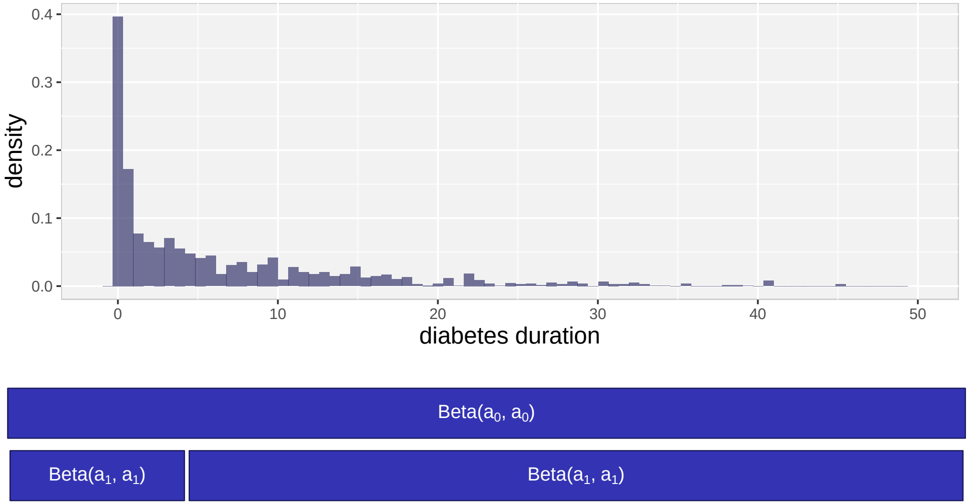
very flexible ☺ ➡ **little contribution** ☹

Mixture of Polya Trees

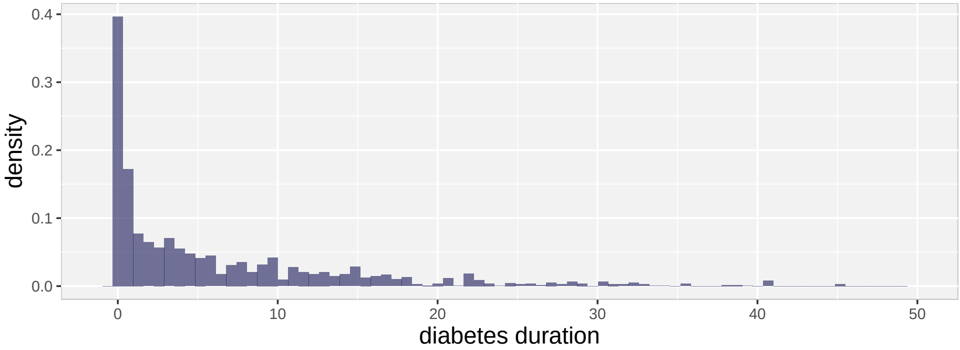


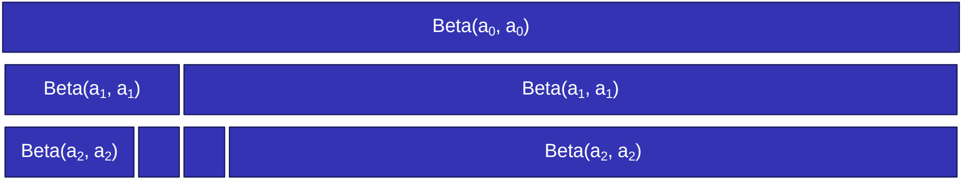
Beta(a_0, a_0)

Mixture of Polya Trees

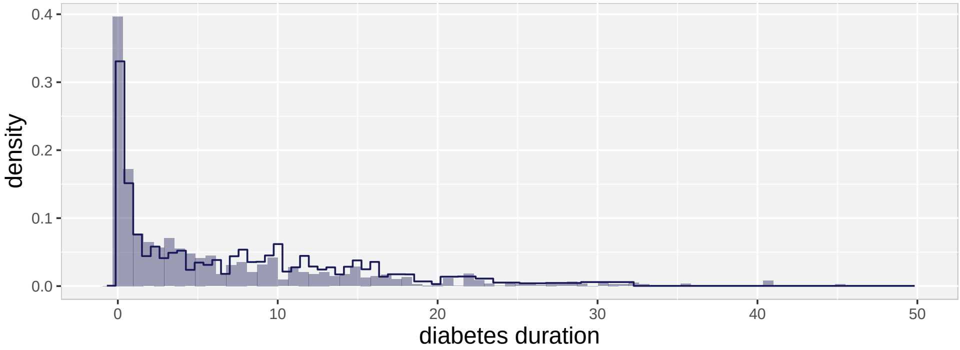


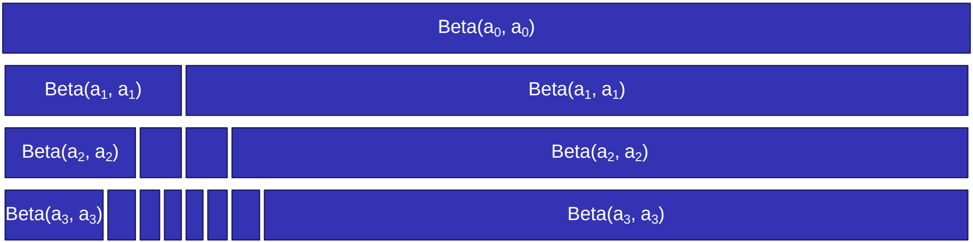
Mixture of Polya Trees



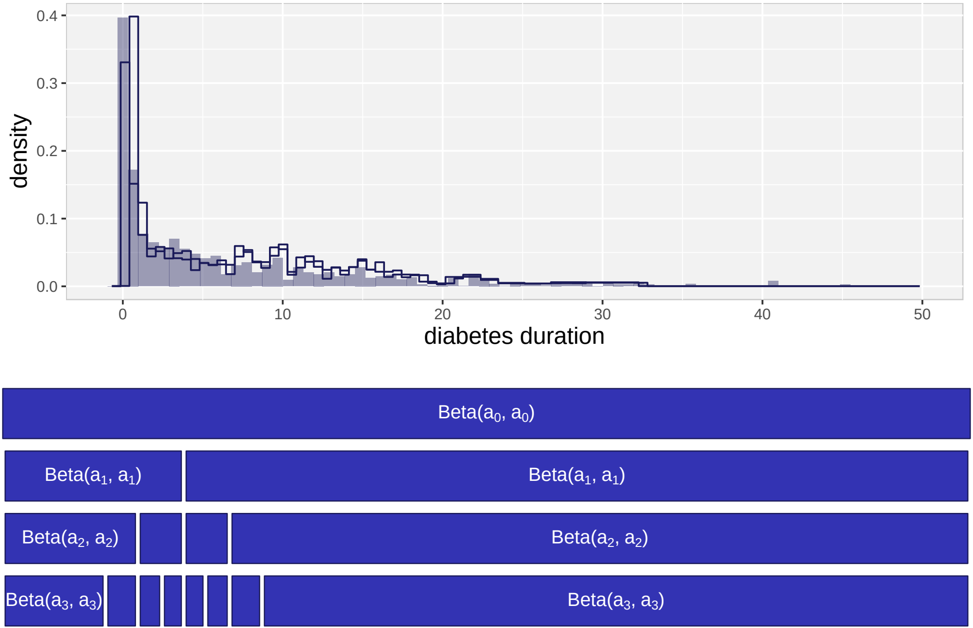


Mixture of Polya Trees

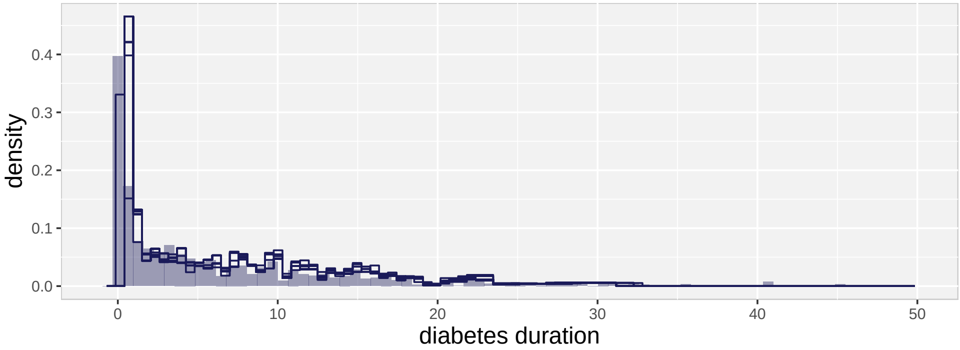


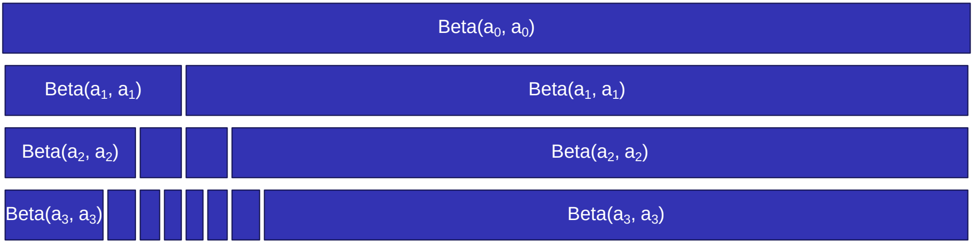


Mixture of Polya Trees

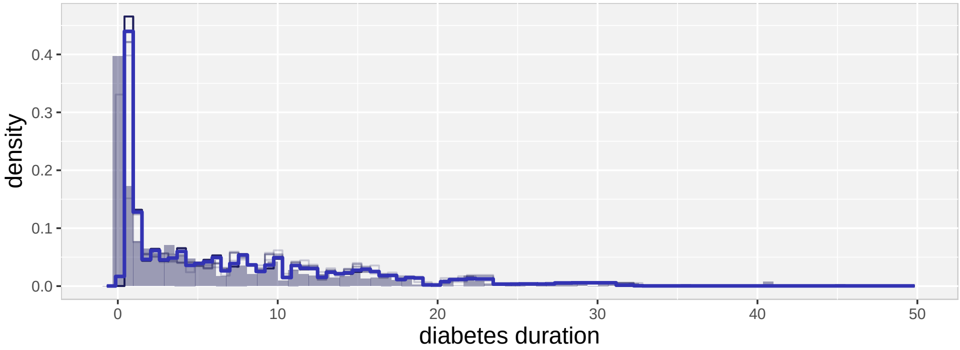


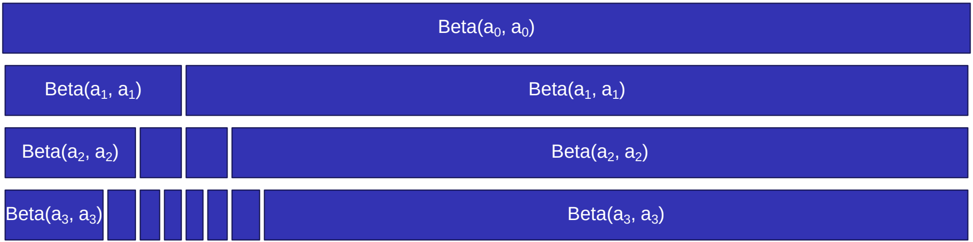
Mixture of Polya Trees





Mixture of Polya Trees

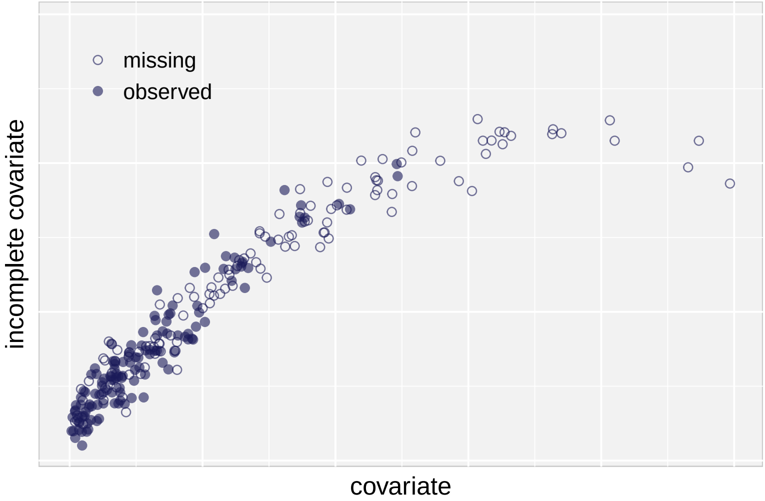




Practical Issues & Ideas

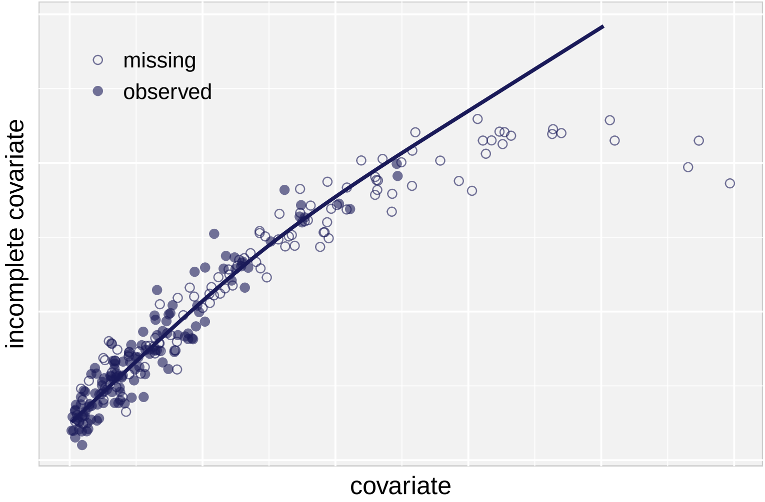
- flexible fit needs observed data everywhere

- flexible fit needs observed data everywhere



Practical Issues & Ideas

- flexible fit needs observed data everywhere



Practical Issues & Ideas

- flexible fit needs observed data everywhere
- computational time

Practical Issues & Ideas

- flexible fit needs observed data everywhere
- computational time

Ideas:

- posterior predictive checks
  - χ^2 type of tests
  - Kolmogorov-Smirnoff test?
  - discordance tests?
- feasibility checks before running the model?

Preliminary Conclusion

Can Bayesian non-parametric methods ~~prevent~~ reduce model-misspecification?

Thank you for your attention.

n.erler@erasmusmc.nl
N_Erler
NErler
www.nerler.com