Bayesian Hierarchical Temporal Modeling and Targeted Learning with Application to Reproductive Health

Candidate: Herb Susmann
Committee: Leontine Alkema (Chair), Laura Balzer, Antoine Chambaz

June 27, 2022

UMassAmherst

Follow along

http://herbsusmann.com/defense

Background

- The international community has set ambitious goals for improvement in global health.
- Where is improvement needed?
  - Chapters 1 and 2: contributions related to statistical estimation and projection of global health indicators, with a focus on family planning.
- Which interventions are effective in improving health outcomes?
  - Chapter 3: methods for estimating the effect of interventions on family planning outcomes.

Outline

1. Chapter 1: Temporal models for demographic and global health outcomes in multiple populations

2. Chapter 2: Flexible Modeling of Transition Processes with B-splines

3. Chapter 3: Automatic Bayesian Targeted Likelihood Estimation of Marginal Structural Models

Outline

1. Chapter 1: Temporal models for demographic and global health outcomes in multiple populations

2. Chapter 2: Flexible Modeling of Transition Processes with B-splines

3. Chapter 3: Automatic Bayesian Targeted Likelihood Estimation of Marginal Structural Models

Background

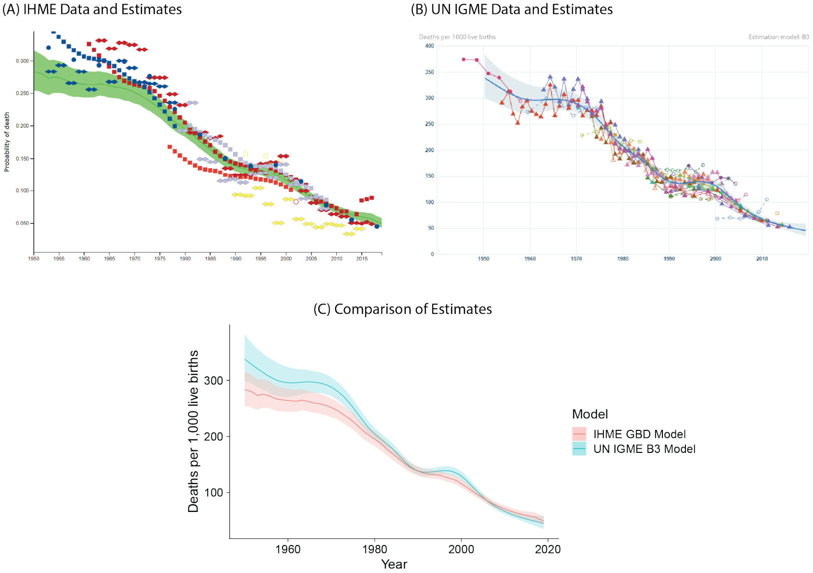
- There is interest in modeling demographic and health indicators in order to measure progress towards international goals.
  - Example: Under-5 Mortality Rate
- Data availability and quality are varied.
  - Some countries have high quality U5MR data from vital registration systems, in other countries data may only come from surveys.
- Many statistical models have been created to provide estimates and projections.
- Comparing across models can be difficult.
- **This chapter:** an overarching model class called *Temporal Models for Multiple Populations* (TMMPs).

Background

- Published in *International Statistical Review*:
  - Susmann, Herbert, Monica Alexander, and Leontine Alkema. "Temporal Models for Demographic and Global Health Outcomes in Multiple Populations: Introducing a New Framework to Review and Standardise Documentation of Model Assumptions and Facilitate Model Comparison." *International Statistical Review* (2022).

Under-5 Mortality Rate (U5MR) Models

Under-five Mortality Rate Estimates in Senegal, 1950-2019

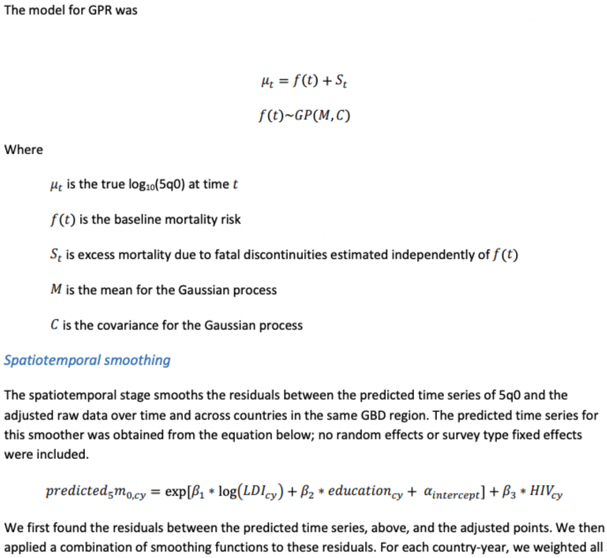


(A) IHME Data and Estimates

(B) UN IGME Data and Estimates

(C) Comparison of Estimates

A glance at the IHME GBD model...



The model for GPR was

$$\mu_t = f(t) + S_t$$

$$f(t) \sim GP(M, C)$$

Where

μ_t is the true $\log_{10}$(5q0) at time t

$f(t)$ is the baseline mortality risk

S_t is excess mortality due to fatal discontinuities estimated independently of $f(t)$

M is the mean for the Gaussian process

C is the covariance for the Gaussian process

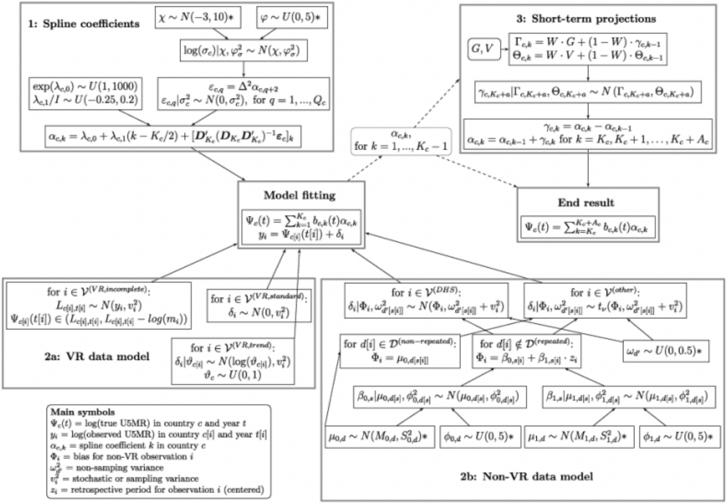
Spatiotemporal smoothing

The spatiotemporal stage smooths the residuals between the predicted time series of 5q0 and the adjusted raw data over time and across countries in the same GBD region. The predicted time series for this smoother was obtained from the equation below; no random effects or survey type fixed effects were included.

$$predicted_5 m_{0,cy} = \exp[\beta_1 * \log(LDI_{cy}) + \beta_2 * education_{cy} + \alpha_{intercept}] + \beta_3 * HIV_{cy}$$

We first found the residuals between the predicted time series, above, and the adjusted points. We then applied a combination of smoothing functions to these residuals. For each country-year, we weighted all

A glance at the UN IGME model...



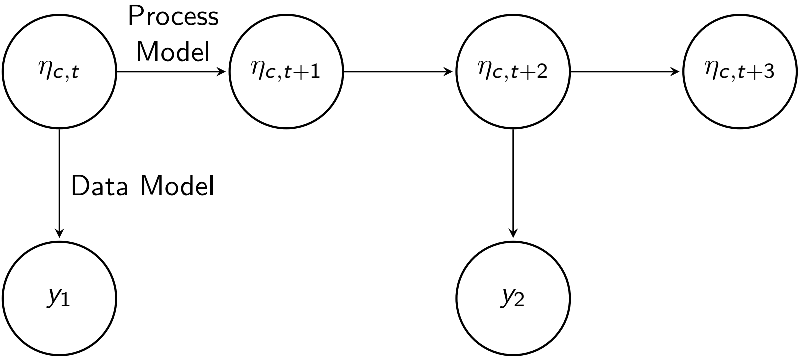
1: Spline coefficients
$\chi \sim N(-3, 10)*$
$\varphi \sim U(0, 5)*$
$\log(\sigma_c)|\chi, \varphi_\sigma^2 \sim N(\chi, \varphi_\sigma^2)$
$\exp(\lambda_{c,0}) \sim U(1, 1000)$
$\lambda_{c,1}/I \sim U(-0.25, 0.2)$
$\varepsilon_{c,q} = \Delta^2 \alpha_{c,q+2}$
$\varepsilon_{c,q}|\sigma_c^2 \sim N(0, \sigma_c^2)$, for $q = 1, ..., Q_c$
$\alpha_{c,k} = \lambda_{c,0} + \lambda_{c,1}(k - K_c/2) + [\boldsymbol{D}'_{K_c}(\boldsymbol{D}_{K_c}\boldsymbol{D}'_{K_c})^{-1}\boldsymbol{\varepsilon}_c]_k$
$\alpha_{c,k}$,
for $k = 1, ..., K_c - 1$
3: Short-term projections
G, V
$\Gamma_{c,k} = W \cdot G + (1 - W) \cdot \gamma_{c,k-1}$
$\Theta_{c,k} = W \cdot V + (1 - W) \cdot \Theta_{c,k-1}$
$\gamma_{c,K_c+a}|\Gamma_{c,K_c+a}, \Theta_{c,K_c+a} \sim N(\Gamma_{c,K_c+a}, \Theta_{c,K_c+a})$
$\gamma_{c,k} = \alpha_{c,k} - \alpha_{c,k-1}$
$\alpha_{c,k} = \alpha_{c,k-1} + \gamma_{c,k}$ for $k = K_c, K_c + 1, \ldots, K_c + A_c$
Model fitting
$\Psi_c(t) = \sum_{k=1}^{K_c} b_{c,k}(t)\alpha_{c,k}$
$y_i = \Psi_{c[i]}(t[i]) + \delta_i$
End result
$\Psi_c(t) = \sum_{k=K_c}^{K_c+A_c} b_{c,k}(t)\alpha_{c,k}$
for $i \in \mathcal{V}^{(VR,incomplete)}$:
$L_{c[i],t[i]} \sim N(y_i, v_i^2)$
$\Psi_{c[i]}(t[i]) \in (L_{c[i],t[i]}, L_{c[i],t[i]} - log(m_i))$
for $i \in \mathcal{V}^{(VR,standard)}$:
$\delta_i \sim N(0, v_i^2)$
2a: VR data model
for $i \in \mathcal{V}^{(VR,trend)}$:
$\delta_i|\vartheta_{c[i]} \sim N(\log(\vartheta_{c[i]}), v_i^2)$
$\vartheta_c \sim U(0, 1)$
for $i \in \mathcal{V}^{(DHS)}$:
$\delta_i|\Phi_i, \omega_{d'[s[i]]}^2 \sim N(\Phi_i, \omega_{d'[s[i]]}^2 + v_i^2)$
for $i \in \mathcal{V}^{(other)}$:
$\delta_i|\Phi_i, \omega_{d'[s[i]]}^2 \sim t_\nu(\Phi_i, \omega_{d'[s[i]]}^2 + v_i^2)$
for $d[i] \in \mathcal{D}^{(non-repeated)}$:
$\Phi_i = \mu_{0,d[s[i]]}$
for $d[i] \notin \mathcal{D}^{(repeated)}$:
$\Phi_i = \beta_{0,s[i]} + \beta_{1,s[i]} \cdot z_i$
$\omega_{d'} \sim U(0, 0.5)*$
$\beta_{0,s}|\mu_{0,d[s]}, \phi_{0,d[s]}^2 \sim N(\mu_{0,d[s]}, \phi_{0,d[s]}^2)$
$\beta_{1,s}|\mu_{1,d[s]}, \phi_{1,d[s]}^2 \sim N(\mu_{1,d[s]}, \phi_{1,d[s]}^2)$
$\mu_{0,d} \sim N(M_{0,d}, S_{0,d}^2)*$
$\phi_{0,d} \sim U(0, 5)*$
$\mu_{1,d} \sim N(M_{1,d}, S_{1,d}^2)*$
$\phi_{1,d} \sim U(0, 5)*$
2b: Non-VR data model
Main symbols
$\Psi_c(t)$ = log(true U5MR) in country c and year t
y_i = log(observed U5MR) in country $c[i]$ and year $t[i]$
$\alpha_{c,k}$ = spline coefficient k in country c
Φ_i = bias for non-VR observation i
$\omega_{d'}^2$ = non-samping variance
v_i^2 = stochastic or sampling variance
z_i = retrospective period for observation i (centered)

FIG. 3. *Model overview. This chart summarizes the model used to estimate the U5MR. In the center is the description of the "Model fitting" part, where $\Psi_c(t)$ refers to the true U5MR on the log-scale, which was modeled with a Bayesian penalized spline regression model, as summarized in block 1 (see Section 3.1). The models for the error term δ_i for observed log(U5MR) are described separately for VR and non-VR data in blocks 2a and 2b (see Section 3.2). Short-term projections are summarized in block 3 (see Section 3.3).*

Modeling Framework

- Let $\eta_{c,t}$ be the true value of the indicator in country c at time t $(c = 1, \ldots, C,\ t = 1, \ldots T)$.
- Observed data $y_i,\ i = 1, \ldots, n$ with associated properties $c[i]$, $t[i]$, ...
- *Process model* describes evolution of $\eta_{c,t}$.
  - Covariates
  - Systematic trends
- *Data model* describes relationship between y_i and $\eta_{c[i],t[i]}$.

Modeling Framework



Data Model Examples

Examples of data models:

- Normal:

$$y_i|\eta_{c[i],t[i]}, \sigma_i^2 \sim N(\eta_{c[i],t[i]}, \sigma_i^2)$$

  where $y_i \in \mathbb{R}$ and σ_i^2 is the sampling variance.

- Binomial:

$$y_i|\eta_{c[i],t[i]} \sim \mathrm{Binom}(n_i, \eta_{c[i],t[i]})$$

  where y_i, n_i are integers.

Process Model

$$g_1(\eta_{c,t}) = \underbrace{g_2(X_{c,t}, \boldsymbol{\beta}_c)}_{\text{covariate}} + \underbrace{g_3(t, \eta_{c,s\neq t}, \boldsymbol{\alpha}_c)}_{\text{systematic}} + \underbrace{a_{c,t}}_{\text{offset}} + \underbrace{\epsilon_{c,t}}_{\text{smoothing}}$$

Covariate component

$$g_1(\eta_{c,t}) = \underbrace{g_2(X_{c,t}, \beta_c)}_{\text{covariate}} + \underbrace{g_3(t, \eta_{c,s\neq t}, \alpha_c)}_{\text{systematic}} + \underbrace{a_{c,t}}_{\text{offset}} + \underbrace{\epsilon_{c,t}}_{\text{smoothing}}$$

- Regression function for incorporating covariates.

Systematic component

$$g_1(\eta_{c,t}) = \underbrace{g_2(X_{c,t}, \beta_c)}_{\text{covariate}} + \underbrace{g_3(t, \eta_{c,s\neq t}, \boldsymbol{\alpha}_c)}_{\text{systematic}} + \underbrace{a_{c,t}}_{\text{offset}} + \underbrace{\epsilon_{c,t}}_{\text{smoothing}}$$

- Parametric function for modeling systematic temporal trends.

Offset

$$g_1(\eta_{c,t}) = \underbrace{g_2(X_{c,t}, \beta_c)}_{\text{covariate}} + \underbrace{g_3(t, \eta_{c,s\neq t}, \alpha_c)}_{\text{systematic}} + \underbrace{a_{c,t}}_{\text{offset}} + \underbrace{\epsilon_{c,t}}_{\text{smoothing}}$$

- The offset term incorporates external information, for example from a separate modeling step.

Smoothing Component

$$g_1(\eta_{c,t}) = \underbrace{g_2(X_{c,t}, \beta_c)}_{\text{covariate}} + \underbrace{g_3(t, \eta_{c,s \neq t}, \alpha_c)}_{\text{systematic}} + \underbrace{a_{c,t}}_{\text{offset}} + \underbrace{\epsilon_{c,t}}_{\text{smoothing}}$$

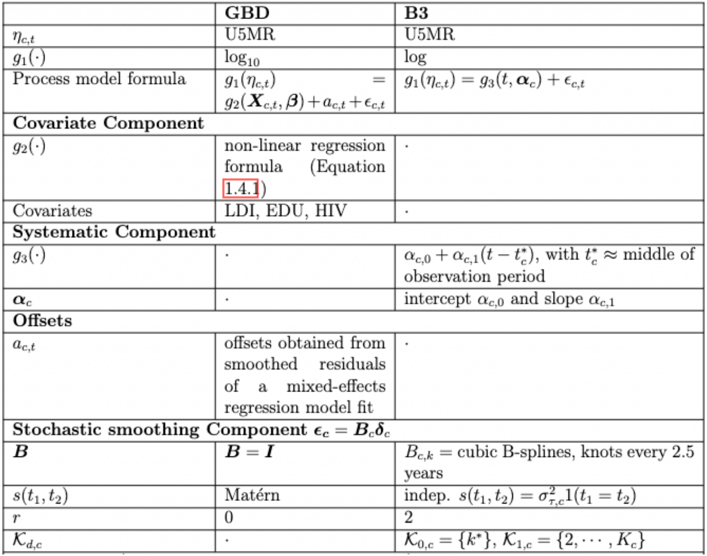
- The smoothing component allows data-driven deviations from the other components, while still enforcing smoothness.
- Many choices B-splines, Gaussian processes, AR(p), RW(p), spatio-temporal smoothing, ...

Hierarchical Modeling

- Each component introduces many country specific parameters that need to be estimated.
- Hierarchical modeling is a way to share information between countries.
- Example: hierarchical model with one level of hierarchy for a country-specific parameter θ_c:

$$\theta_c \mid \theta_w, \sigma_\theta \sim N(\theta_w, \sigma_\theta^2)$$

Comparing the example models...



	GBD	**B3**
$\eta_{c,t}$	U5MR	U5MR
$g_1(\cdot)$	$\log_{10}$	log
Process model formula	$g_1(\eta_{c,t}) = g_2(\boldsymbol{X}_{c,t}, \boldsymbol{\beta}) + a_{c,t} + \epsilon_{c,t}$	$g_1(\eta_{c,t}) = g_3(t, \boldsymbol{\alpha}_c) + \epsilon_{c,t}$
Covariate Component		
$g_2(\cdot)$	non-linear regression formula (Equation 1.4.1)	·
Covariates	LDI, EDU, HIV	·
Systematic Component		
$g_3(\cdot)$	·	$\alpha_{c,0} + \alpha_{c,1}(t - t_c^*)$, with $t_c^* \approx$ middle of observation period
$\boldsymbol{\alpha}_c$	·	intercept $\alpha_{c,0}$ and slope $\alpha_{c,1}$
Offsets		
$a_{c,t}$	offsets obtained from smoothed residuals of a mixed-effects regression model fit	·
Stochastic smoothing Component $\boldsymbol{\epsilon}_c = \boldsymbol{B}_c\boldsymbol{\delta}_c$		
$\boldsymbol{B}$	$\boldsymbol{B} = \boldsymbol{I}$	$B_{c,k}$ = cubic B-splines, knots every 2.5 years
$s(t_1, t_2)$	Matérn	indep. $s(t_1, t_2) = \sigma^2_{\tau,c} 1(t_1 = t_2)$
r	0	2
$\mathcal{K}_{d,c}$	·	$\mathcal{K}_{0,c} = \{k^*\}, \mathcal{K}_{1,c} = \{2, \cdots, K_c\}$

Contributions

- A model class, Temporal Models for Multiple Populations (TMMPs), that encompasses many existing demographic and health models.
    - Model class makes a clear distinction between the *process model* and the *data model*.
    - Process model is split into building blocks: covariates, systematic trends, offsets, and smoothing components.
- Detailed description of six existing models using TMMP notation, and templates provided for documenting additional models as TMMPs.

Outline

1. Chapter 1: Temporal models for demographic and global health outcomes in multiple populations

2. Chapter 2: Flexible Modeling of Transition Processes with B-splines

3. Chapter 3: Automatic Bayesian Targeted Likelihood Estimation of Marginal Structural Models

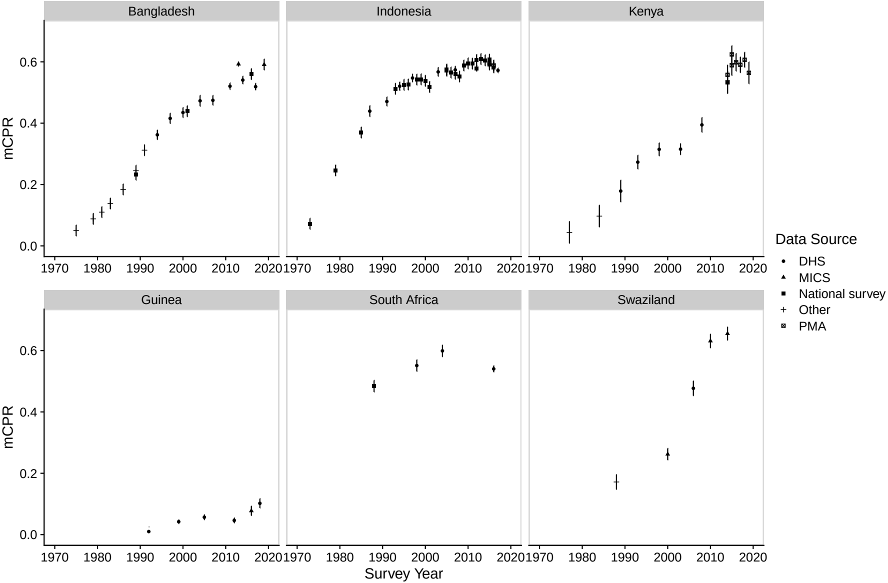
Background

- Some indicators have been observed to evolve similarly across populations.
    - They tend to follow a *transition* between stable states.
- Classic example: demographic transition.
    - Transition from high total fertility rate and high under-5 mortality to low fertility, low mortality.
- Existing statistical models for estimating and projecting trends in these indicators draw on these patterns.
- **This chapter:** We propose a new type of model, called *B-spline Transition Models*, for flexibly estimating indicators that follow transitions.

Case Study

- Modern Contraceptive Prevalence Rate (mCPR) for married or in-union women: proportion of married or in-union women of reproductive age using (or with partner using) a modern contraceptive method.
- Transition: low to high mCPR.
- Existing model: Family Planning Estimation Model (FPEM, Cahill et al. 2018).
- Goal: estimate and project mCPR in countries from 1970-2030.
- Dataset aggregated by United Nations Population Division (UNPD) from surveys conducted by governments or international organizations.

Case Study


Bangladesh
Indonesia
Kenya
Guinea
South Africa
Swaziland
mCPR
0.0
0.2
0.4
0.6
1970
1980
1990
2000
2010
2020
Survey Year
Data Source
● DHS
▲ MICS
■ National survey
+ Other
⊠ PMA

Transition Models

- **Our contribution:** a model class for indicators that follow a transition.
- *Transition Models* have a process model given by

$$g_1(\eta_{c,t}) = \underbrace{g_3(t, \boldsymbol{\eta}_{c,s\neq t}, \boldsymbol{\alpha}_c)}_{\text{systematic}} + \underbrace{\epsilon_{c,t}}_{\text{smoothing}} .$$

- The systematic component has the following form:

$$g_3(t, \boldsymbol{\eta}_{c,s\neq t}, \boldsymbol{\alpha}_c) = \begin{cases} \Omega_c, & t = t_c^*, \\ g_1(\eta_{c,t-1}) + f(\eta_{c,t-1}, P_c, \boldsymbol{\beta}_c), & t > t_c^*, \\ g_1(\eta_{c,t+1}) - f(\eta_{c,t+1}, P_c, \boldsymbol{\beta}_c), & t < t_c^*, \end{cases}$$

  where $\boldsymbol{\alpha}_c = \{\Omega_c, P_c, \boldsymbol{\beta}_c\}$.
- The function f is called the *transition function*.

FPEM Example

- The Family Planning Estimation Model (FPEM) is an example of a Transition Model (Cahill et al., 2018).
- Because $\eta_{c,t} \in (0,1)$, FPEM process model uses a logit transform:

$$\text{logit}(\eta_{c,t}) = g_3(t, \boldsymbol{\eta}_{c,s\neq t}, \boldsymbol{\alpha}_c) + \epsilon_{c,t}.$$

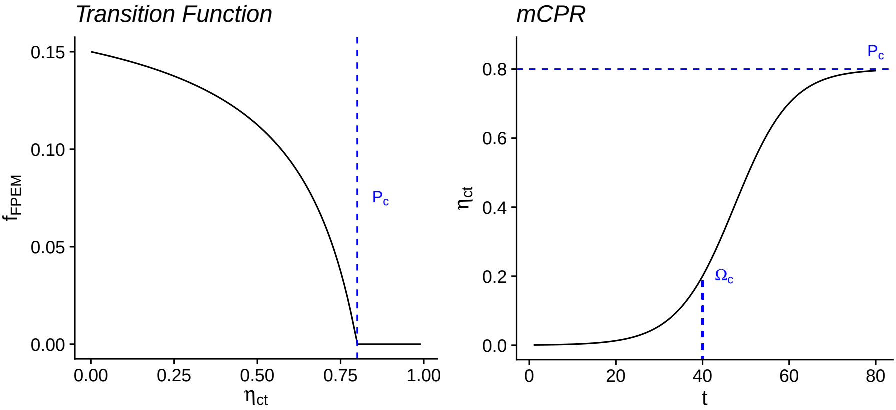
- The FPEM transition function was chosen such that mCPR follows a logistic growth curve:

$$f(\eta_{c,t-1}, P_c, \boldsymbol{\beta}_c) = \begin{cases} \frac{(\eta_{c,t-1}-P_c)\omega_c}{P_c(\eta_{c,t-1}-1)}, & \eta_{c,t-1} < P_c, \\ 0, & \text{otherwise.} \end{cases}$$

where $\boldsymbol{\beta}_c = \{\omega_c\}$, and the parameters can be interpreted as

  - ω_c: rate parameter,
  - P_c: asymptote parameter.

FPEM Example



B-spline Transition Model

- **Our contribution:** estimate the transition function f while making weaker functional form assumptions.
- Approach: estimate f using B-splines.

B-spline Example

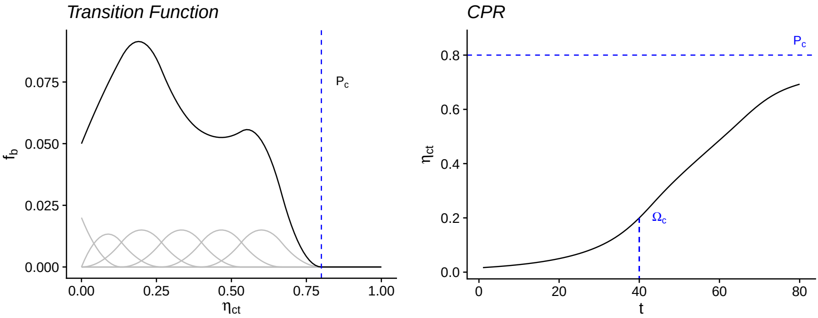
B-spline Transition Model

- Define a transition function f_b as:

$$f_b(\eta_{c,t}, P_c, \boldsymbol{\beta}_c) = \sum_{j=1}^{J} \underbrace{h_j(\beta_{c,j})}_{\text{coefficient}} \cdot \underbrace{B_j(\eta_{c,t}/P_c)}_{\text{basis function}},$$

  where P_c is an asymptote parameter.
- Flexibility of f_b can be tuned through the spline degree and number and positioning of knots.

Example B-spline Transition Function



Sharing information on shape of transition function

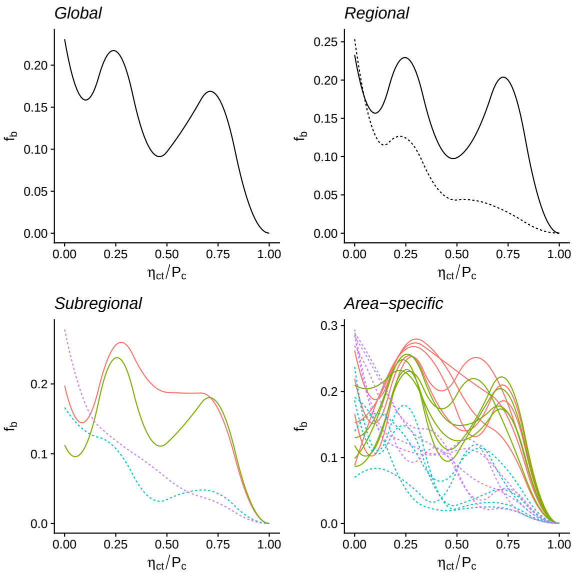
- Varying data availability across countries.
- We would like to share information about the transition between countries.
- Spline coefficients $\beta_{c,j}$ are nested within sub-regions, regions, and world.
- Hierarchical model on the spline coefficients $\beta_{c,j}$ for $j = 1, \dots, J$:

$$\beta_{c,j} \mid \beta^{(s)}_{s[c],j}, \sigma^{(c)}_{\beta,j} \sim N\left(\beta^{(s)}_{s[c],j}, \left(\sigma^{(c)}_{\beta,j}\right)^2\right),$$

$$\beta^{(s)}_{s,j} \mid \beta^{(r)}_{r[s],j}, \sigma^{(s)}_{\beta,j} \sim N\left(\beta^{(r)}_{r[s],j}, \left(\sigma^{(s)}_{\beta,j}\right)^2\right),$$

$$\beta^{(r)}_{r,j} \mid \beta^{(w)}_{j}, \sigma^{(r)}_{\beta,j} \sim N\left(\beta^{(w)}_{j}, \left(\sigma^{(r)}_{\beta,j}\right)^2\right).$$

Sharing information on shape of transition function



Smoothing component

- Recall the process model has two components:

$$g_1(\eta_{c,t}) = \underbrace{g_3(t, \boldsymbol{\eta}_{c,s\neq t}, \boldsymbol{\alpha}_c)}_{\text{systematic}} + \underbrace{\epsilon_{c,t}}_{\text{smoothing}} .$$

- Smoothing component: AR(1) process of the form

$$\epsilon_{c,t} | \epsilon_{c,t-1}, \tau, \rho \sim N(\rho * \epsilon_{c,t-1}, \tau^2)$$

Smoothing component

Data Model: connection to observed data

- Let y_i, $i = 1, \ldots, n$ be the observed mCPR for country $c[i]$ and year $y[i]$ from data source $d[i]$.
- For each observation we have an estimate s_i^2 of the sampling error.
- We also expect each data source to have additional non-sampling error $\sigma^2_{d[i]}$.
- Truncated normal data model:

$$y_i | \eta_{c[i],t[i]}, \sigma^2_{d[i]} \sim N_{(0,1)} \left(\eta_{c[i],t[i]}, s_i^2 + \sigma^2_{d[i]} \right).$$

Computation

- Model fit with full Bayesian inference
- Implementation in Stan, including a fast B-spline algorithm in C++

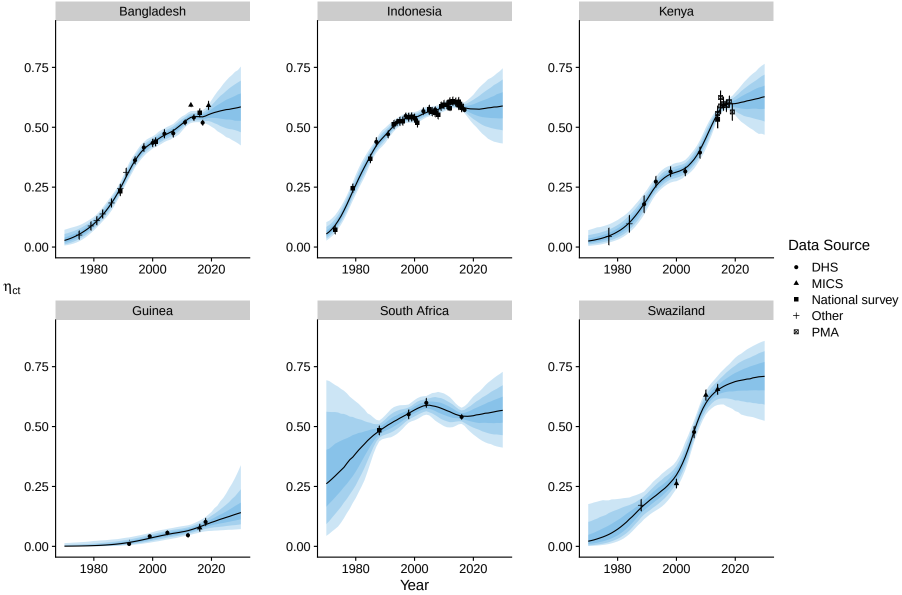
Choosing a spline specification

Validation exercise: hold out all observations after a cutoff year $L = 2010$.

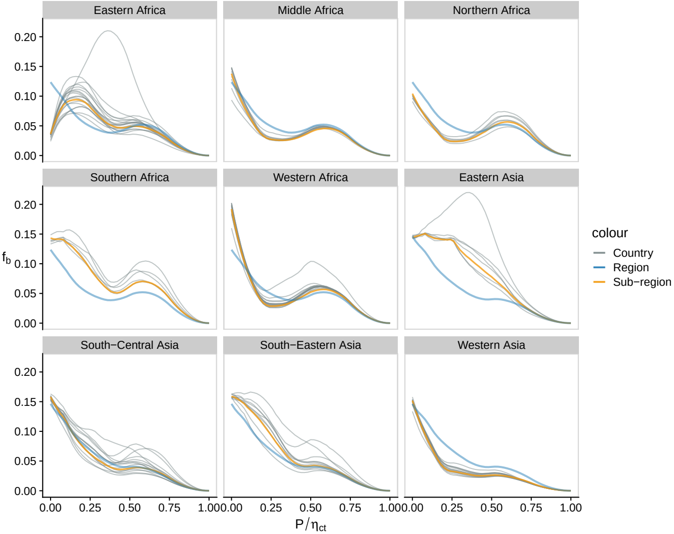
	95% UI				Error	
	% Below	% Included	% Above	CI Width ×100	ME ×100	MAE ×100
Model Check 2 ($L = 2010$), $n = 133$						
B-spline ($d = 2,\ K = 5$)	3.76%	94.7%	1.5%	32.0	-1.670	4.64
B-spline ($d = 2,\ K = 7$)	6.02%	91.7%	2.26%	31.5	-1.260	4.68
B-spline ($d = 3,\ K = 5$)	3.76%	94.7%	1.5%	32.4	-1.630	4.48
B-spline ($d = 3,\ K = 7$)	3.76%	94%	2.26%	31.6	-0.965	4.57

95% UI: 95% uncertainty interval. ME: median error. MAE: median absolute error. Measures calculated using the last held-out observation within each area.

Illustrative Fits from B-spline Model



Trends can be seen in regional and subregional transition functions



Comparison to a logistic type model

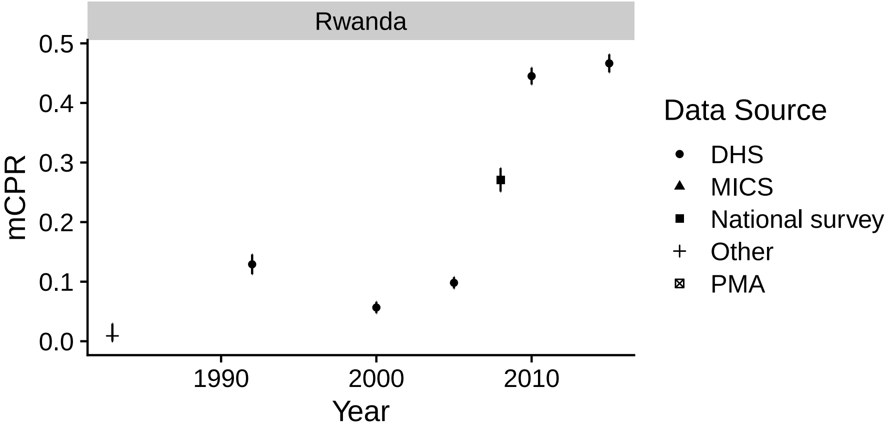
Validation exercise: hold out all observations after a cutoff year $L = 2010$.

	95% UI				Error	
	% Below	% Included	% Above	CI Width ×100	ME ×100	MAE ×100
Model Check 2 ($L = 2010$), $n = 133$						
B-spline ($d = 2$, $K = 5$)	3.76%	94.7%	1.5%	32.0	-1.670	4.64
Logistic	6.77%	92.5%	0.752%	32.7	-2.850	4.82

95% UI: 95% uncertainty interval. ME: median error. MAE: median absolute error.

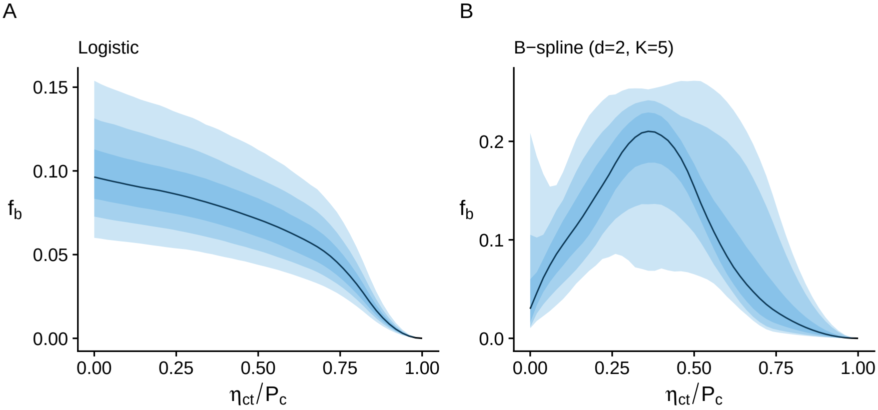
Measures calculated using the last held-out observation within each area.

Detailed Example: Rwanda



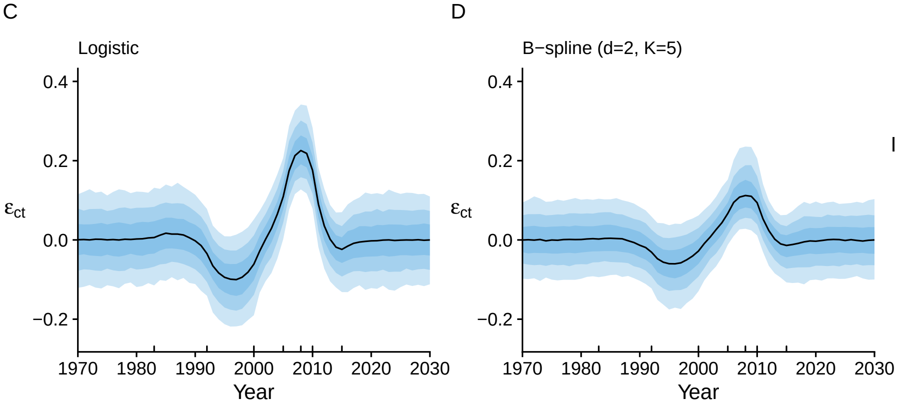
Detailed Example: Rwanda

Transition Functions



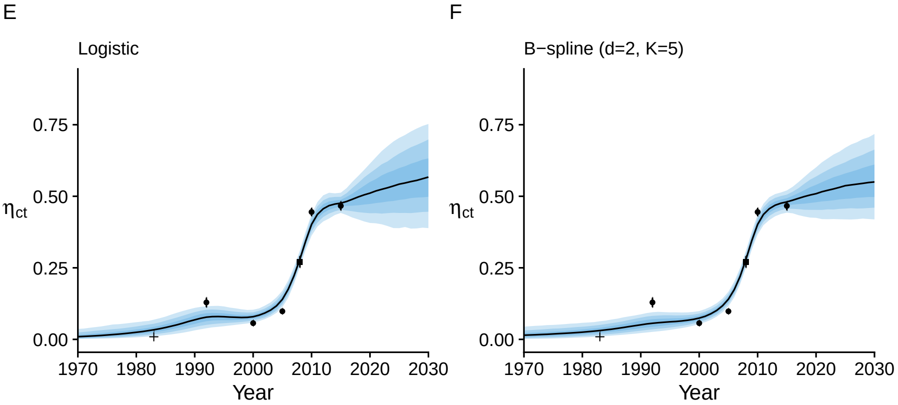
Detailed Example: Rwanda

Smoothing component



Detailed Example: Rwanda

Modern Contraceptive Prevalence Rate



Contributions

- Subclass of *Transition Models* for indicators that follow transitions.
- B-spline Transition Model: flexible modelling approach based on B-splines.
- Generated estimations and projections of mCPR in countries from 1970-2030.
- Found systematically different transitions in countries across regions.
- Flexible model framework that can be easily extended to new settings and use cases.

Outline

1. Chapter 1: Temporal models for demographic and global health outcomes in multiple populations

2. Chapter 2: Flexible Modeling of Transition Processes with B-splines

3. Chapter 3: Automatic Bayesian Targeted Likelihood Estimation of Marginal Structural Models

Background

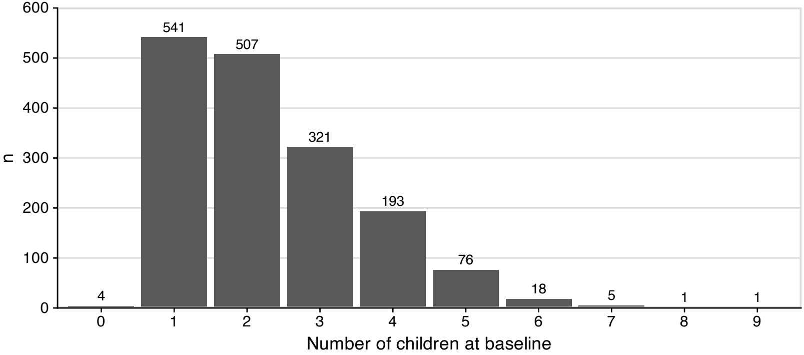
- Which interventions are effective in improving health outcomes?
- *Marginal Structural Models* provide one a way to summarize how the effect of an intervention on an outcome changes within subgroups.
- **This Chapter**: We introduce a novel targeted Bayesian estimator for the parameter of a Marginal Structural Model in a general setting.

Motivating Example

- Randomized field experiment conducted in Lilongwe, Malawi, to investigate effect of family planning intervention on contraceptive use (Karra et al., 2020, 2022).
- Intervention: broad-based intervention including information package and counseling.
- Outcome: contraceptive use two years after intervention.

Scientific question

- Scientific question: does the treatment effect differ depending on number of children at baseline?
- Potential example of *Treatment Effect Modification*.
- Marginal distribution of number of children:



Observed data

- For each participant, we have:
    - X: set of 11 covariates measured at baseline, including number of children X_c;
    - A: indicator of randomization into intervention group;
    - Y: indicator of contraceptive use at endline.
- Let $O_1, \ldots, O_n$ be n i.i.d. draws of the generic variable $O = (X, A, Y)$ from the law P_0 of the experiment.

Conditional Average Treatment Effect

- *Conditional Average Treatment Effect* (CATE):

$$\begin{aligned}\Psi_P^{\mathrm{CATE}}(x) &= \mathbb{E}_P[Y \mid A = 1, X = x] - \mathbb{E}_P[Y \mid A = 0, X = x], \\ &= \bar{Q}_P^{(1)}(x) - \bar{Q}_P^{(0)}(x)\end{aligned}$$

- Causally identifiable under “standard causal assumptions” (consistency, positivity, no unmeasured confounders).

An example of a Marginal Structural Model

- Approach: summarize the relationship between potential *treatment effect modifiers* (X_c) and conditional treatment effects ($\Psi_P^{\mathrm{CATE}}(X)$) using a user-supplied working model.
- For instance, let $B(P) \in \mathbb{R}^2$ be the solution to the following optimisation problem:

$$B(P) = \underset{\boldsymbol{\beta} \in \mathbb{R}^2}{\arg\min}\, \mathbb{E}_P \left[\left(\Psi_P^{\mathrm{CATE}}(X) - (1, X_c)\boldsymbol{\beta} \right)^2 \right].$$

An example of a Marginal Structural Model

$$B(P) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\arg\min}\, \mathbb{E}_P\left[\left(\boxed{\Psi_P^{\mathrm{CATE}}(X)} - (1, X_c)\boldsymbol{\beta}\right)^2\right]$$

conditional average treatment effect

An example of a Marginal Structural Model

$$B(P) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\arg\min}\,\mathbb{E}_P\left[\left(\Psi_P^{\mathrm{CATE}}(X) - \underbrace{\boxed{(1, X_c)\beta}}_{\text{linear working model}}\right)^2\right]$$

An example of a Marginal Structural Model

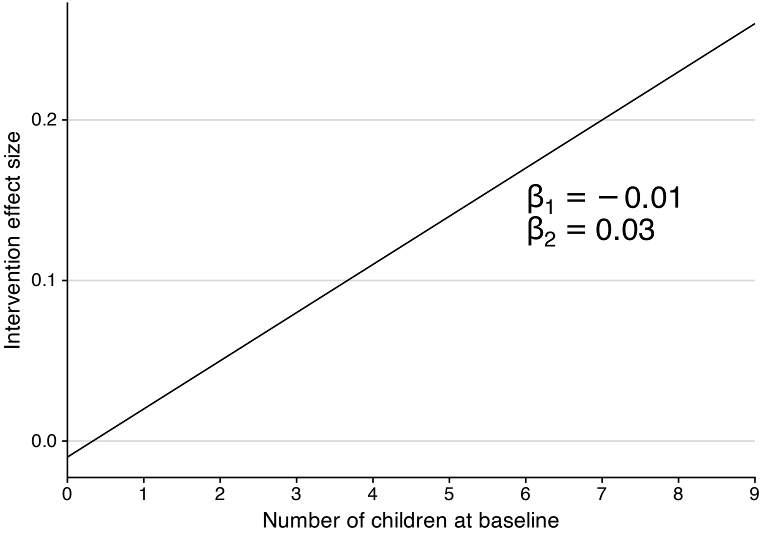
$$B(P) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\arg\min} \underbrace{\boxed{\mathbb{E}_P\left[\left(\Psi_P^{\mathrm{CATE}}(X) - (1, X_c)\beta\right)^2\right]}}_{\text{squared-error risk}}$$

An example of a Marginal Structural Model

$$\boxed{B(P)} = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\arg\min}\, \mathbb{E}_P\left[\left(\Psi_P^{\mathrm{CATE}}(X) - (1, X_c)\beta\right)^2\right]$$

defined in terms of P

What a plot of the results will look like



General Setting for MSMs

- Observed data: $O_1, \ldots, O_n$ i.i.d. copies of a generic variable $O \sim P_0$.
- Assume that P_0 is in a non-parametric statistical model $\mathcal{M}$.
  - The more we know about the law P_0, the smaller the model $\mathcal{M}$.
- Assume that $O = (Z, X)$ for variables $Z \in \mathcal{Z}$, $X \in \mathcal{X}$.
- For all $P \in \mathcal{M}$, let $\Psi_P : \mathcal{X} \to \mathbb{R}$ be a functional summary of P with argument X.
- For the motivating example:
  - $O = (Y, A, X)$, $Z = (Y, A)$, $X = X$.
  - $\Psi_P(X) = \mathbb{E}_P[Y \mid A = 1, X] - \mathbb{E}_P[Y \mid A = 0, X]$, the CATE
    $= \bar{Q}_P^{(1)}(X) - \bar{Q}_P^{(0)}(X)$.

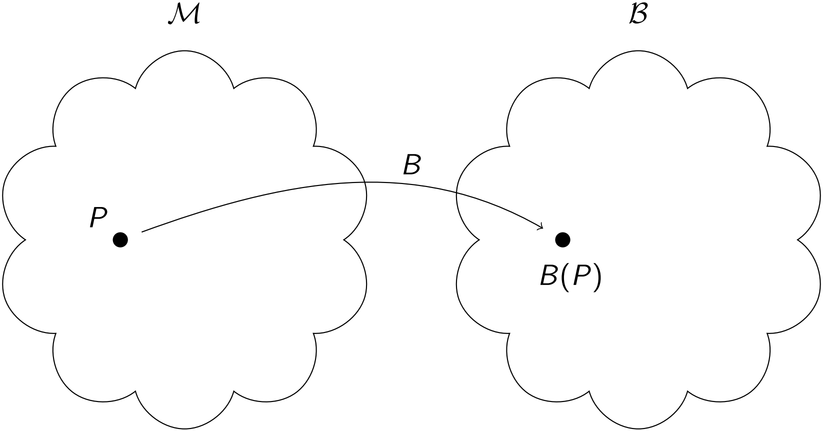
Marginal Structural Models

- Idea: approximate Ψ_P using a user-supplied working model.
- Working model: set $\{m_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathcal{B}\}$ of functions $m_{\boldsymbol{\beta}} : \mathcal{X} \to \mathbb{R}$ with $\mathcal{B}$ a parameter of dimension p.
- Loss function: $L_m(\Psi_P(X), \boldsymbol{\beta})(X)$.
- Define the parameter of interest B as the solution to the optimization problem:

$$B(P) = \underset{\boldsymbol{\beta} \in \mathcal{B}}{\arg\min}\, \mathbb{E}_P \left[L_m(\Psi_P(X), \boldsymbol{\beta})(X) \right].$$

- The combination of working model and loss function is called a *Marginal Structural Model* (Robins et al., 2000; van der Laan and Rose, 2011).
- Causally identifiable under same assumptions as Ψ_P.
- **Our contribution:** a general framework for MSMs.

Marginal Structural Models



Danger! Infinite dimensional space visualized in two dimensions!

Semi-parametric inference

- Goal: estimate $\beta_0 := B(P_0)$.
- What is the semi-parametric efficiency bound for estimating β_0?
- We can write any regular estimator of β_0 as:

$$\hat{\beta}_n = \beta_0 + \frac{1}{n}\sum_{i=1}^{n} IC_{P_0}(O_i) + o_p(n^{-1/2}),$$

where IC_{P_0} is called an *influence function* of the parameter B at P_0.

- The influence function with the smallest variance is called the *efficient influence function* (EIF), which we denote $D^*(P_0)$.
- The semi-parametric efficiency bound for estimating β_0 is given by $\mathrm{var}_{P_0}(D^*(P_0)(O))$

Efficient Influence Function of $B(P)$

Our contribution: we derived the EIF for the MSM parameter $P \mapsto B(P)$ in a general setting.

Theorem (Efficient Influence Function)

(Simplified) The target functional $P \mapsto B(P)$ is pathwise differentiable at every $P \in \mathcal{M}$, with an efficient influence function $D^*(P)$ given by

$$D^*(P)(O) = M^{-1}\left[D_1^*(P)(O) + D_2^*(P)(X)\right],$$

where $D_1^*(P), D_2^*(P) \in L_0^2(P)$ are given by

$$\begin{aligned} D_1^*(P)(O) &= \nabla \dot{L}(\Psi_P(X), B(P))(X) \times \Delta^*(P)(O), \\ D_2^*(P)(X) &= \dot{L}(\Psi_P(X), B(P))(X), \end{aligned}$$

and the normalizing matrix M is given by

$$M = -\mathbb{E}_P\left[\ddot{L}_m(\Psi_P(X), B(P))(X)\right].$$

Efficient Influence Function of $B(P)$

For our motivating example:

Theorem

(Simplified) The target functional $P \mapsto B(P)$ is pathwise differentiable at every $P \in \mathcal{M}$, with an efficient influence function $D^*(P)$ given by

$$D^*(P)(X, A, Y) = M^{-1} \left\{ D_1^*(P)(X, A, Y) + D_2^*(P)(X) \right\},$$

where

$$D_1^*(P)(X, A, Y) = \left\{ \frac{\mathbb{I}(A = 1)}{P[A = 1|X])} - \frac{\mathbb{I}(A = 0)}{P[A = 0|X]} \right\} (Y - \bar{Q}_P^{(A)}(X))(1, X)^\top,$$

$$D_2^*(P)(X) = (\Psi_P(X) - B(P)^\top (1, X)^\top)(1, X)^\top,$$

and the normalizing matrix M is given by

$$M = -\mathbb{E}_P \left[(1, X)^\top (1, X) \right].$$

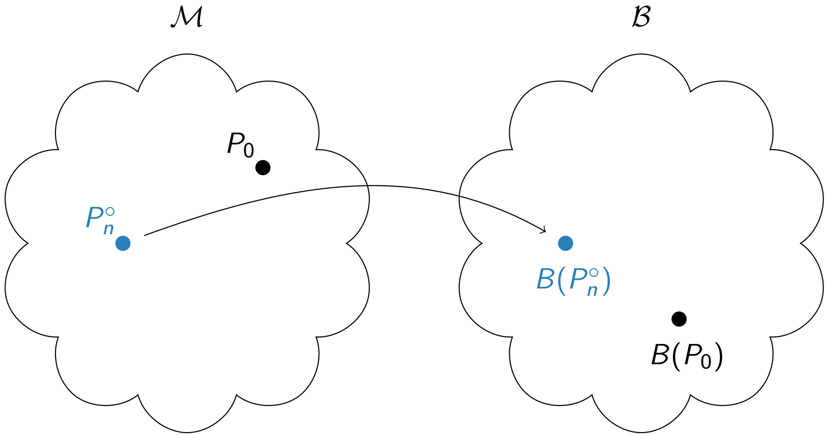
Targeted Minimum Loss-Based Estimation

- It turns out we can construct an estimator that achieves this efficiency bound!
    - *Targeted Minimum Loss-Based Estimation (TMLE)* (van der Laan and Rose, 2011, 2018)
- Suppose we have an initial estimator P_n° of the pieces of P_0 relevant to the MSM parameter $B(P_0)$.
- We can then form a plug-in estimator

$$\hat{\beta}^{\text{plug}-\text{in}} = B(P_n^\circ).$$

- The plug-in estimator is biased!

Plug-in estimation is biased



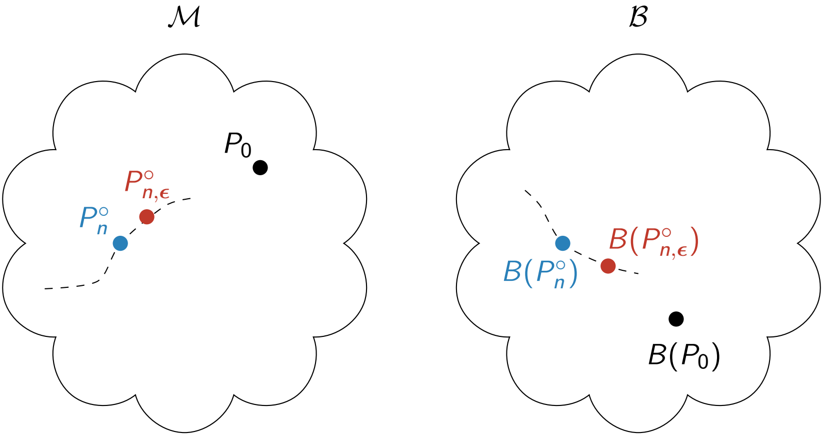
Targeted Minimum Loss-Based Estimation

- Targeted Minimum Loss-Based Estimation (TMLE): plug-in estimator of the form

$$\hat{\boldsymbol{\beta}}^{TMLE} = B(P_n^{\circ}(\boldsymbol{\epsilon}_n^*))$$

where $\{P_n^{\circ}(\boldsymbol{\epsilon}) : \boldsymbol{\epsilon} \in \mathbb{R}^p\} \subset \mathcal{M}$ is a *fluctuation* of an initial estimator P_n° of the pieces of P_0 relevant to $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}_n^*$ is chosen by minimising the empirical risk induced by a well-chosen loss function $\mathcal{L}$.

TMLE: update initial estimate in direction of truth



Bayesian TMLE

- Can we make this procedure Bayesian?
- Core idea: some choices of TMLE loss function $\mathcal{L}$ can be interpreted as defining a likelihood for the data O conditional on the parameter ϵ under the fluctuation submodel.
- We can then use Bayesian inference to estimate ϵ! (Diaz et al., 2011; Díaz et al., 2020)

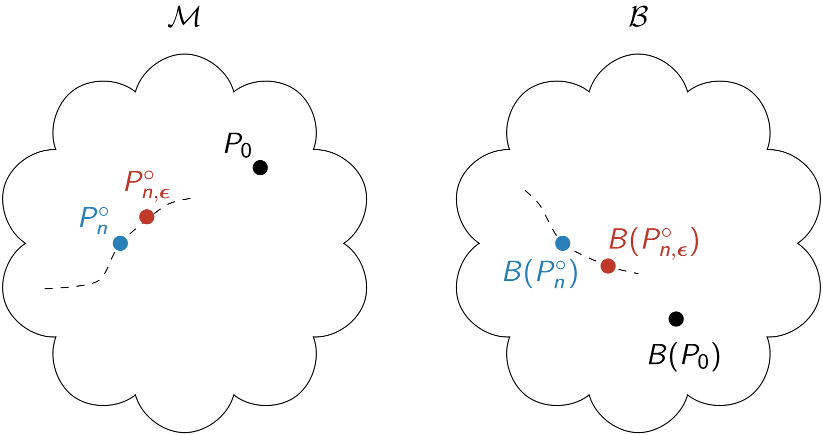
Bayesian TMLE

- Basic application of Bayes rule: posterior distribution of $\boldsymbol{\epsilon}$ is given by

$$\Pi_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon} \mid O_1, \ldots, O_n) \propto \pi_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) \prod_{i=1}^{n} p_n^{\circ}(O_i \mid \boldsymbol{\epsilon})$$

  where $\pi_{\boldsymbol{\epsilon}}$ is a prior distribution for $\boldsymbol{\epsilon}$ and $p_n^{\circ}(O \mid \boldsymbol{\epsilon})$ is the likelihood of O under $P_n^{\circ}(\boldsymbol{\epsilon})$.
- Once we have a posterior distribution for $\boldsymbol{\epsilon}$ we can map it to a posterior distribution for $\boldsymbol{\beta}$.

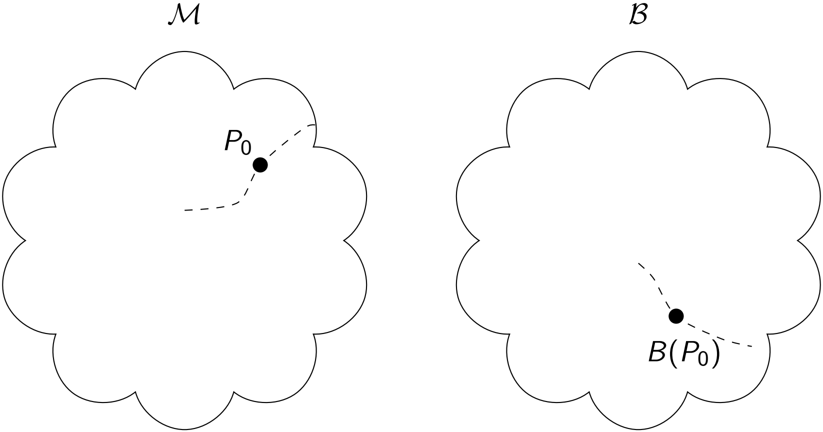
Bayesian TMLE



Bernstein von-Mises

- Desired result: the posterior distribution for β converges to a normal distribution centered on the frequentist TMLE with variance given by the variance of the efficient influence function.
- **Our contribution:** We prove an *oracular* version that provides conditions under which the posterior distribution based on fluctuation of P_0 will converge to the optimal distribution.

Bayesian TMLE



Bernstein von-Mises

- Let $p_n^0(O \mid \boldsymbol{\epsilon})$ be the likelihood of the submodel fluctuating P_0.
- Key conditions:
  - The gradient satisfies

$$\frac{\partial}{\partial \boldsymbol{\epsilon}} \log p_n^0(O|\boldsymbol{\epsilon})\bigg|_{\boldsymbol{\epsilon}=0} = D^*(P_0)(O).$$

  - The Hessian satisfies

$$P_0\left[\frac{\partial^2}{\partial \boldsymbol{\epsilon}^2} \log p_n^0(O|\boldsymbol{\epsilon})\bigg|_{\boldsymbol{\epsilon}=0}\right] = P_0[D^*(P_0)D^*(P_0)^\top].$$

Bernstein von-Mises

Theorem (Oracular Bernstein von-Mises)

(Simplified) Let $N(\mu, \Sigma)$ denote the multivariate normal distribution with mean vector μ and covariance matrix Σ. Then, under certain assumptions,

$$\|\Pi^0_{\boldsymbol{\beta}}(\cdot \mid O_1, \ldots, O_n) - N\left(\Delta^0_n, P_0[D^*(P_0)D^*(P_0)^\top]\right)\|_1 = o_P(1)$$

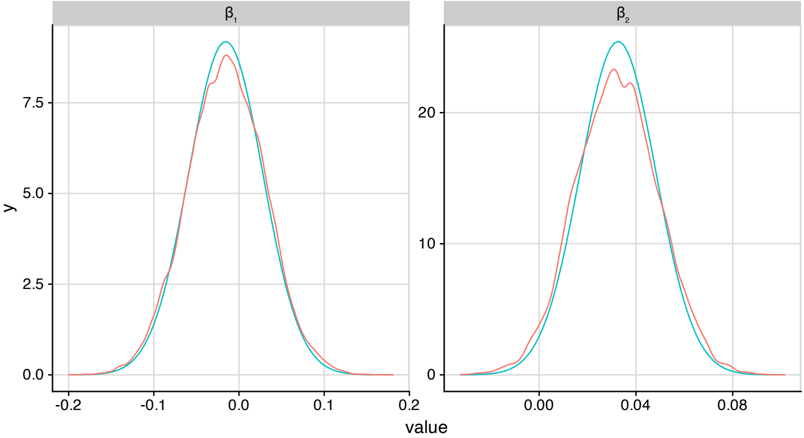
where

$$\Delta^0_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} P_0[\lambda^*(P_0)]^{-1} D^*(P_0)(O_i). \tag{1}$$

Universal Algorithm

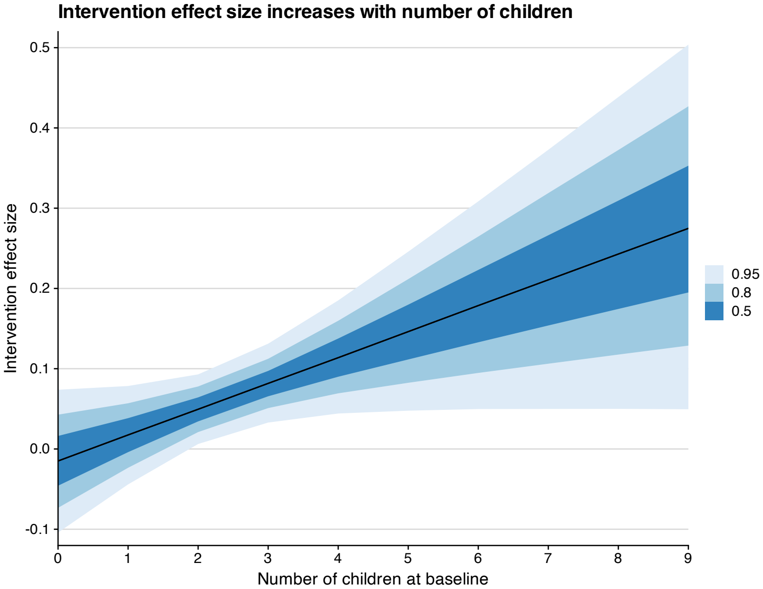
- In practice, users may want to try several working models $\{m_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \mathcal{B}\}$ and loss functions L_m.
- We could anticipate this and choose several working models and loss functions and hand-code the required derivatives. But what if a user wants to use something we haven't implemented?
- An alternative is to use *automatic differentiation* to compute the required derivatives automatically.
- **Our contribution:** We implemented a universal algorithm in Julia that uses auto-differentiation to automatically adapt the fluctuation model and efficient influence function to arbitrary well-chosen working models and loss functions.

Motivating Example: Results



Posterior density (red) and a normal density (blue) centered on the frequentist MLE with variance given by the estimated variance of the efficient influence function.

Motivating Example: Results



Contributions

- Definition of MSMs in a general setting.
- Derivation of efficient influence function for general MSM parameters.
- Novel Bayesian TMLE for MSMs.
- Universal algorithm implemented in Julia using autodifferentiation.
- Application to estimate relationship between effect of intervention on contraceptive use with number of children as an effect modifier in a randomized field experiment.

Future Work

- Strengthening Bernstein von-Mises result
- Developing methods for choosing between multiple working models.

Summary

Where is improvement needed?

- Chapter 1: Temporal Models for Multiple Populations
- Chapter 2: B-spline Transition Model

Which interventions are effective in improving health outcomes?

- Chapter 3: Bayesian targeted learning for Marginal Structural Models

Acknowledgements

Questions?

References I

N. Cahill, E. Sonneveldt, J. Stover, M. Weinberger, J. Williamson, C. Wei, W. Brown, and L. Alkema. Modern contraceptive use, unmet need, and demand satisfied among women of reproductive age who are married or in a union in the focus countries of the Family Planning 2020 initiative: a systematic analysis using the Family Planning Estimation Tool. *The Lancet*, 391(10123):870–882, Mar. 2018. ISSN 0140-6736. doi: 10.1016/S0140-6736(17)33104-5. URL http://www.sciencedirect.com/science/article/pii/S0140673617331045.

I. Diaz, A. E. Hubbard, and M. J. van der Laan. Targeted bayesian learning. In *Targeted Learning*, pages 475–493. Springer, 2011.

I. Díaz, O. Savenkov, and H. Kamel. Nonparametric targeted bayesian estimation of class proportions in unlabeled data. *Biostatistics*, 2020.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis (3rd ed)*. Chapman and Hall/CRC, 2013.

A. Gelman, A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák. Bayesian workflow. *arXiv preprint arXiv:2011.01808*, 2020.

M. Karra, D. Canning, et al. The effect of improved access to family planning on postpartum women: protocol for a randomized controlled trial. *JMIR research protocols*, 9(8):e16697, 2020.

References II

M. Karra, D. Maggio, M. Guo, B. Ngwira, and D. Canning. The causal effect of a family planning intervention on women&#x2019;s contraceptive use and birth spacing. *Proceedings of the National Academy of Sciences*, 119(22):e2200279119, 2022. doi: 10.1073/pnas.2200279119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2200279119.

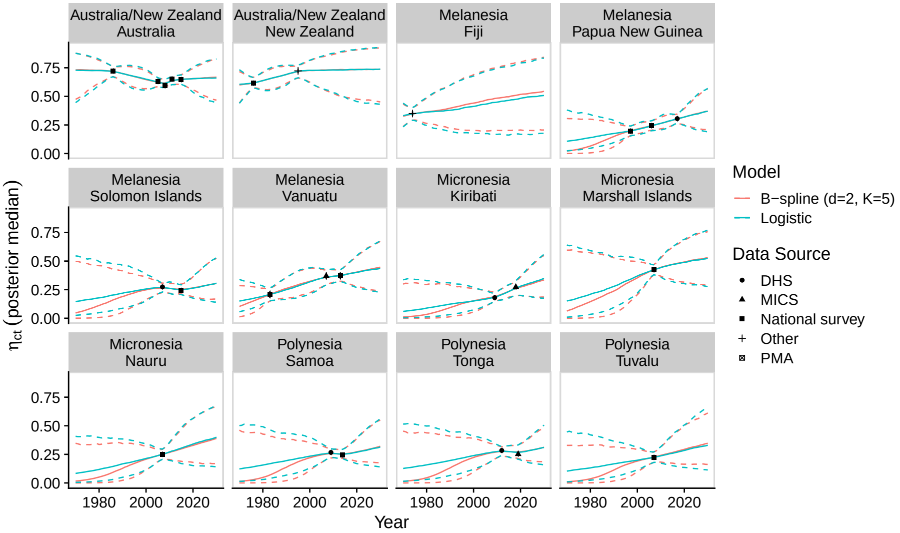
J. M. Robins, M. A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology, 2000.

H. Susmann, M. Alexander, and L. Alkema. Temporal models for demographic and global health outcomes in multiple populations: Introducing a new framework to review and standardize documentation of model assumptions and facilitate model comparison, 2021.

M. J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

M. J. van der Laan and S. Rose. *Targeted learning in data science: causal inference for complex longitudinal studies*. Springer, 2018.

Difference between B-splines and logistic model



Estimation

- Let's analyze the properties of the plug-in estimator.
- We can write

$$
\begin{aligned}
\sqrt{n}\,(\beta_n^\circ - \beta_0) = & \underbrace{\sqrt{n}(P_n - P_0)D^*(P_0)}_{\rightsquigarrow N(0, P_0 D^*(P_0)^2)} \\
& - \underbrace{\sqrt{n}P_n D^*(P_n^\circ)}_{\text{bias term}} \\
& + \underbrace{\sqrt{n}(P_n - P_0)(D^*(P_n^\circ) - D^*(P_0)) + o_p(1)}_{\text{negligible}}.
\end{aligned}
$$

- We want to construct an estimator with a bias term of zero.

A glimpse at how TMLE works

- The fluctuation and loss function are chosen to satisfy (among other things) a key property:

$$D^*(P_n^\circ) \in \mathrm{Span}\left(\frac{\partial}{\partial \boldsymbol{\epsilon}}\mathcal{L}\left(P_n^\circ(\boldsymbol{\epsilon})\right)\Big|_{\boldsymbol{\epsilon}=0}\right).$$

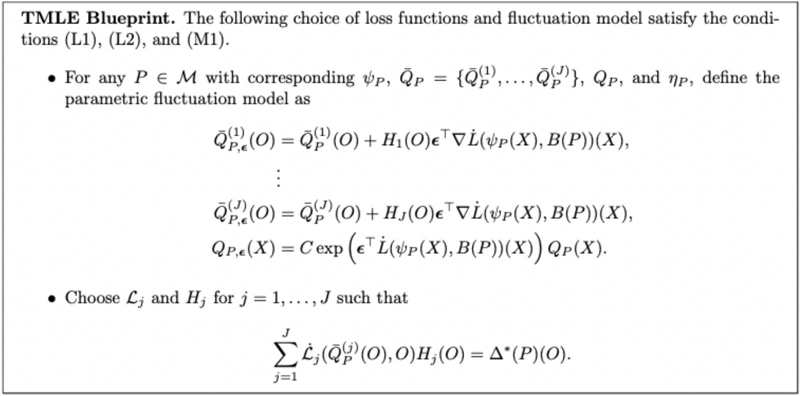
- Importantly, the TMLE solves the EIF of the target parameter:

$$\mathbb{E}_{P_n}[D^*(P_n^\circ(\boldsymbol{\epsilon}_n^*))(O)] = 0.$$

- Under certain conditions, $\hat{\beta}^{TMLE}$ is asymptotically normal and efficient.

Blueprint for fluctuation model

How do we choose the form of the fluctuation model $P_n^\circ(\epsilon)$? **Our contribution:** a blueprint for the fluctuation model.



TMLE Blueprint. The following choice of loss functions and fluctuation model satisfy the conditions (L1), (L2), and (M1).

- For any $P \in \mathcal{M}$ with corresponding ψ_P, $\bar{Q}_P = \{\bar{Q}_P^{(1)}, \ldots, \bar{Q}_P^{(J)}\}$, Q_P, and η_P, define the parametric fluctuation model as

$$\begin{aligned}
\bar{Q}_{P,\boldsymbol{\epsilon}}^{(1)}(O) &= \bar{Q}_P^{(1)}(O) + H_1(O)\boldsymbol{\epsilon}^\top \nabla \dot{L}(\psi_P(X), B(P))(X), \\
&\vdots \\
\bar{Q}_{P,\boldsymbol{\epsilon}}^{(J)}(O) &= \bar{Q}_P^{(J)}(O) + H_J(O)\boldsymbol{\epsilon}^\top \nabla \dot{L}(\psi_P(X), B(P))(X), \\
Q_{P,\boldsymbol{\epsilon}}(X) &= C \exp\left(\boldsymbol{\epsilon}^\top \dot{L}(\psi_P(X), B(P))(X)\right) Q_P(X).
\end{aligned}$$

- Choose $\mathcal{L}_j$ and H_j for $j = 1, \ldots, J$ such that

$$\sum_{j=1}^{J} \dot{\mathcal{L}}_j(\bar{Q}_P^{(j)}(O), O) H_j(O) = \Delta^*(P)(O).$$

Back to the motivating example

- First, we need to find the parts of P relevant to B and D^*.
- Recall the definition of $B(P)$ and Ψ_P^{CATE}:

$$\begin{aligned} B(P) &= \underset{\beta \in \mathcal{B}}{\arg\min}\, \mathbb{E}_P \left[\Psi_P^{\mathrm{CATE}}(X) - (1, X)\beta\right] \\ \Psi_P^{\mathrm{CATE}}(x) &= \bar{Q}_P^{(1)}(x) - \bar{Q}_P^{(0)}(x) \\ &= \mathbb{E}_P[Y \mid A = 1, X = x] - \mathbb{E}_P[Y \mid A = 0, X = x] \end{aligned}$$

- In addition, $D^*(P)$ depends on $g_P(a, x) = P(A = a|X = x)$.
- The relevant parts of P are therefore Q_P (the marginal distribution of X), $\bar{Q}_P^{(1)}$, $\bar{Q}_P^{(0)}$, and g_P.

Back to the motivating example

- Suppose we have initial estimators of each part of P_0 relevant to B and D^*.
    - To estimate Q_{P_0}, the marginal distribution of X under P_0, we use the empirical distribution of X, which we call Q_n°.
    - To estimate $\bar{Q}_{P_0}^{(0)}$, $\bar{Q}_{P_0}^{(1)}$, and g_{P_0}, we use estimators $\bar{Q}_n^{\circ,(0)}$, $\bar{Q}_n^{\circ,(1)}$, and g_n°.
- Let $P_n^\circ \in \mathcal{M}$ be any law such that its relevant features coincide with $\{Q_n^\circ, \bar{Q}_n^{\circ,(0)}, \bar{Q}_n^{\circ,(1)}, g_n^\circ\}$.

Back to the motivating example

- Now we need to build a fluctuation submodel of each of the parts of P relevant to B.
- Fluctuation indexed by $\boldsymbol{\epsilon} \in \mathbb{R}^p$:

$$\bar{Q}_{n,\boldsymbol{\epsilon}}^{\circ,(1)}(x) = \bar{Q}_n^{\circ,(1)}(x) + \frac{1}{g_P^{\circ}(1,x)}\boldsymbol{\epsilon}^{\top}(1,X)^{\top}$$

$$\bar{Q}_{n,\boldsymbol{\epsilon}}^{\circ,(0)}(x) = \bar{Q}_n^{\circ,(0)}(x) - \frac{1}{g_P^{\circ}(0,x)}\boldsymbol{\epsilon}^{\top}(1,X)^{\top}$$

$$Q_{n,\boldsymbol{\epsilon}}^{\circ}(x) \propto \exp(\boldsymbol{\epsilon}^{\top} D_{2,\mathrm{CATE}}^{*}(P_n^{\circ})(x))Q_n^{\circ}(x)$$

- Negative log-likelihood loss function:

$$\mathcal{L}(\bar{Q}_{n,\boldsymbol{\epsilon}}^{\circ,(0)}, \bar{Q}_{n,\boldsymbol{\epsilon}}^{\circ,(1)}, Q_{n,\boldsymbol{\epsilon}}^{\circ}, O) = \left(Y - \bar{Q}_{n,\boldsymbol{\epsilon}}^{\circ,(A)}\right)^2 - \log Q_{n,\boldsymbol{\epsilon}}^{\circ}(X).$$