Fully-Bayesian Imputation for Transit Panel Survey Data

Spencer Aeschliman and **Amanda Stathopoulos**
Department of Civil & Environmental Engineering,
Northwestern University

Modeling Mobility Conference
September 16th, 2025

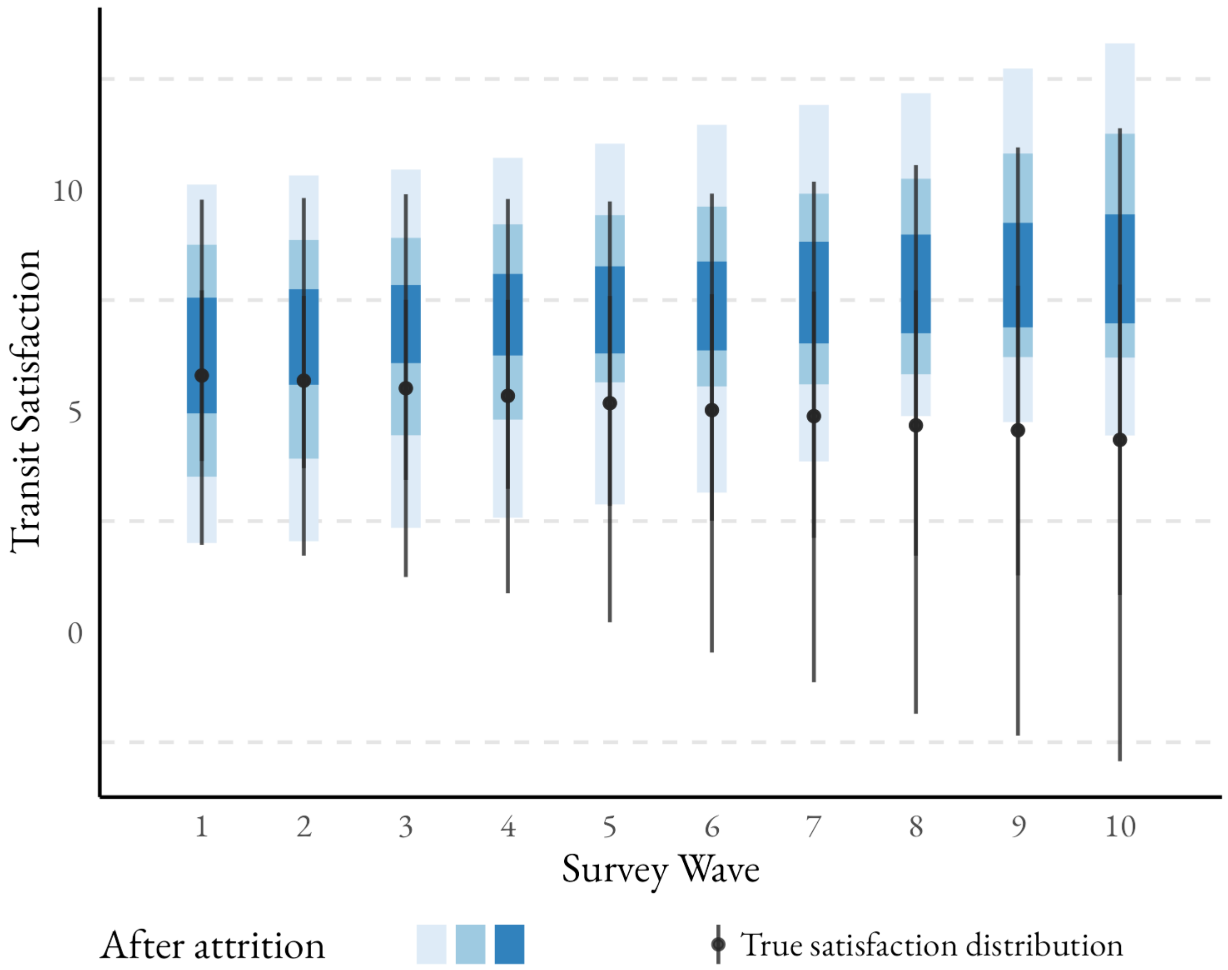
Preview

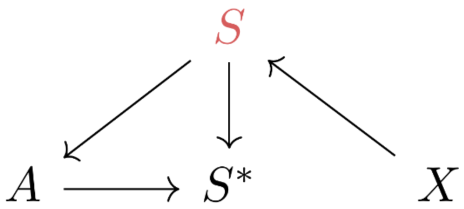
- We can avoid throwing away panel data by using **all** the information we have
- We can do so on our own terms
- Learn through proximity
- Trade off individual and global data trajectories
- Build in uncertainty naturally

Data and Motivation

- Regional Transportation Authority
  - Oversees transit agencies in the greater Chicago area
- Multiple waves of an RTA customer panel survey
- 3,617 unique respondents
- But only 464 complete cases (**12.8%**)
- Outcome variable: **overall satisfaction with the Chicago-area transit service** in each wave

Complete-case analysis





Attrition is related to low satisfaction

Imputation Model

Posterior Likelihood Priors

Hierarchical Model: $p(\theta, \phi \mid y) \propto \prod_i \prod_t \underline{p(y_{it} \mid m_{it}; \theta, \phi)}\, p(\theta \mid \phi)\, p(\phi)$

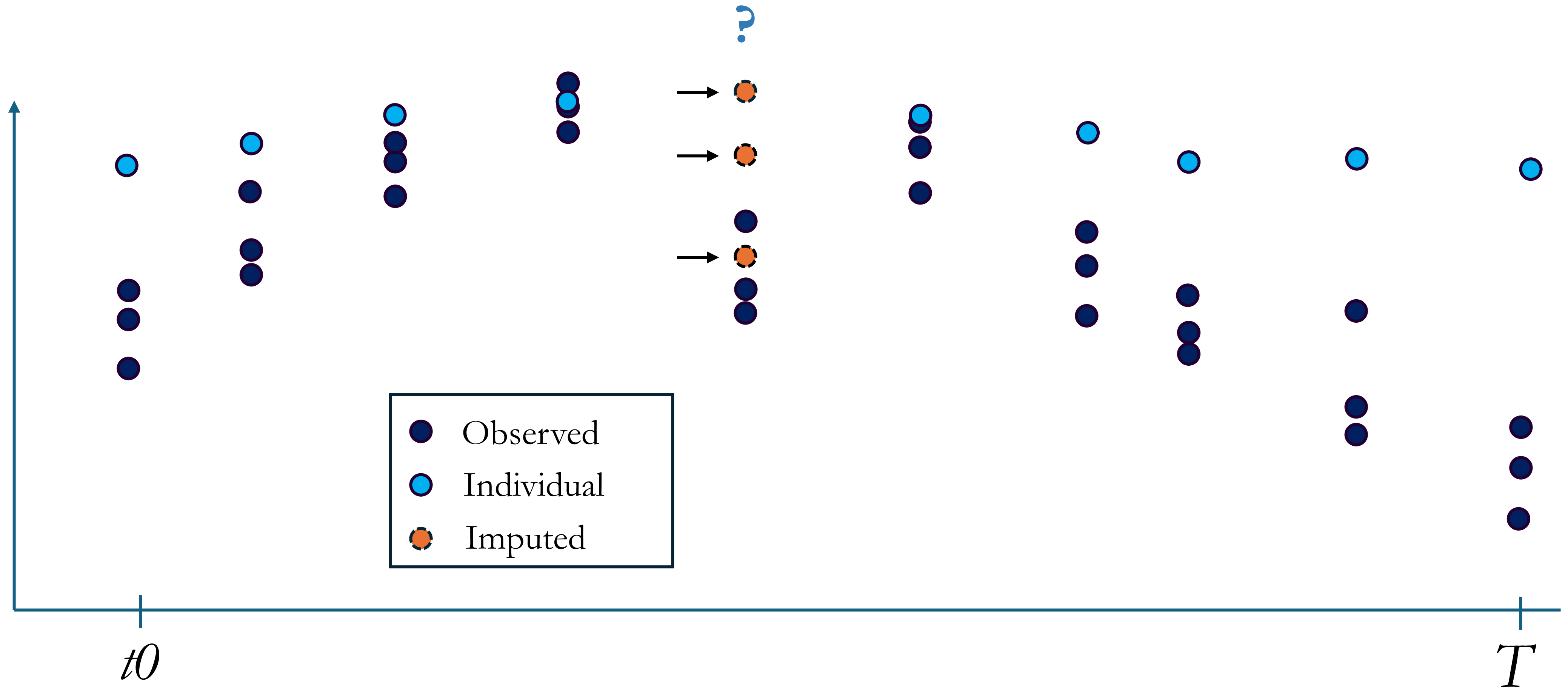
Let: $\boldsymbol{y^{obs}}$ be the observed outcomes, $\boldsymbol{y^{mis}}$ be the missing outcomes

$$p\left(\theta, \phi, y^{mis} \mid y^{obs}\right) \propto p\left(y^{obs} \mid \theta, \phi\right) \underbrace{\boldsymbol{p\left(y^{mis} \mid \theta, \phi\right)}}_{\text{Prior on the missing values}} p(\theta \mid \phi)\, p(\phi)$$

$p(\,y^{mis} \mid \theta, \phi\,)$

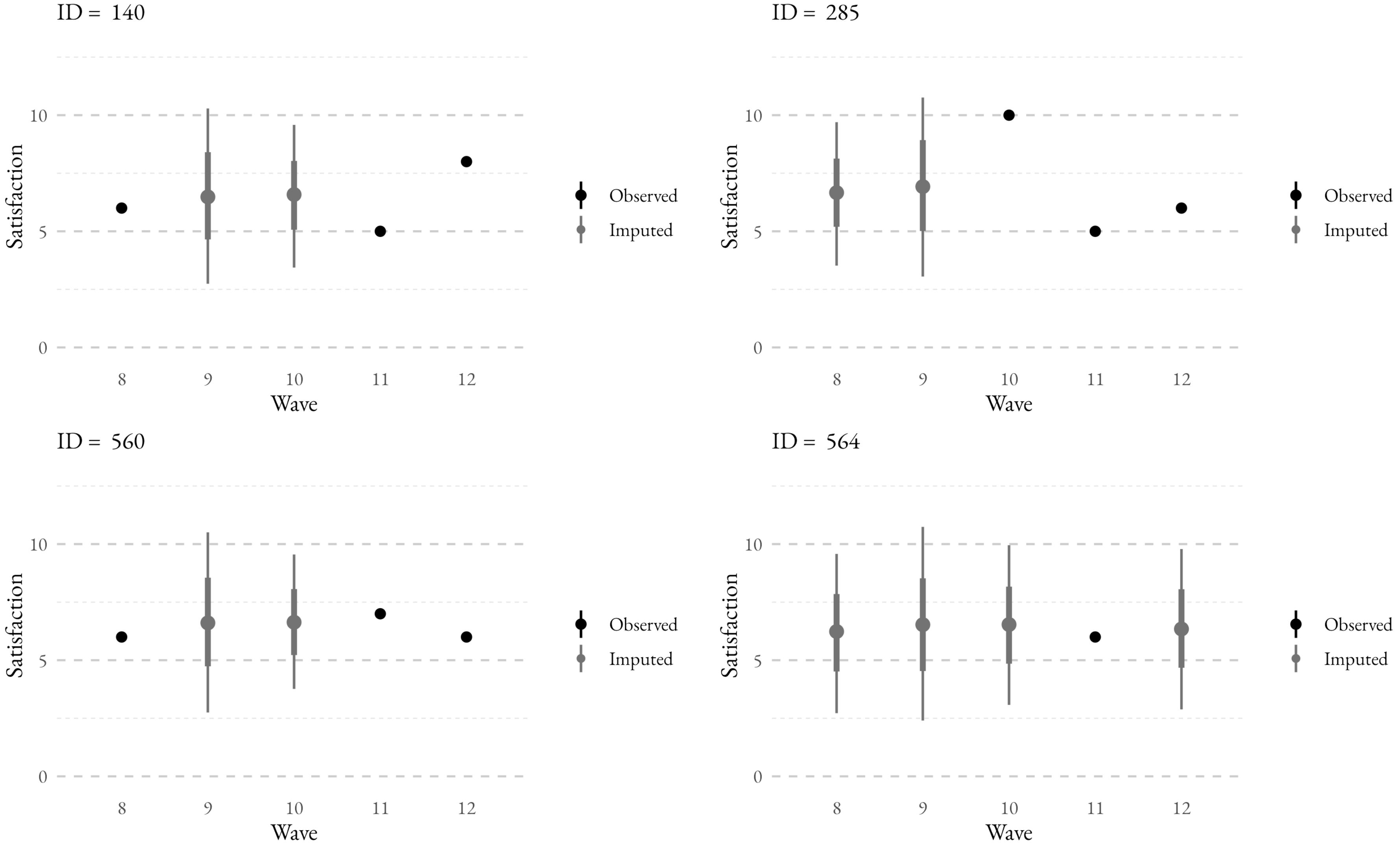
- Take advantage of the **hierarchical and temporal structure**
- Learn from nearby observations:
  - Observed values at the same time from different people
  - Observed values at different times within the same person
- Be flexible

$p(y^{mis} \mid \theta, \phi)$: learn from nearby observations

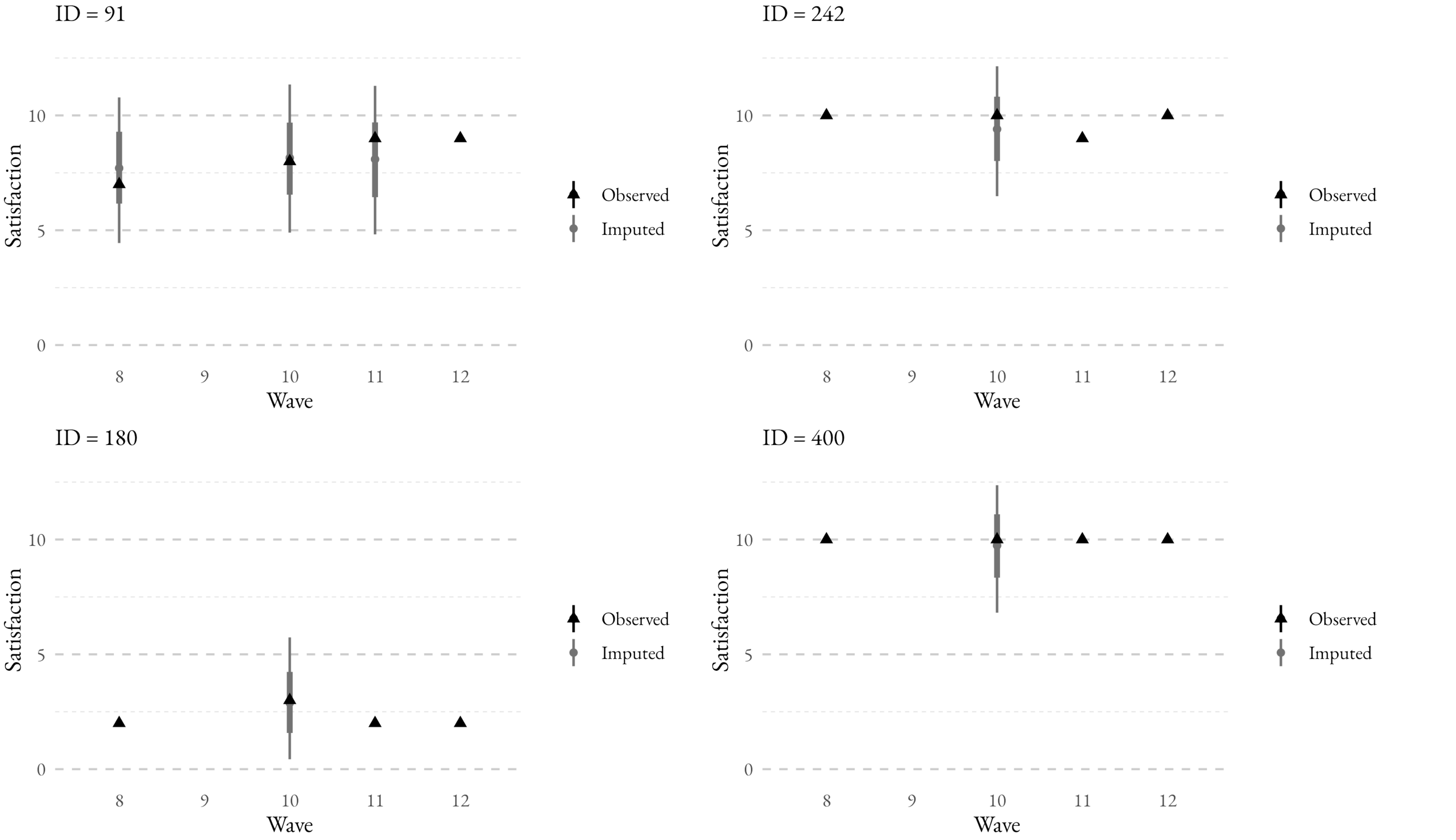


Demonstration

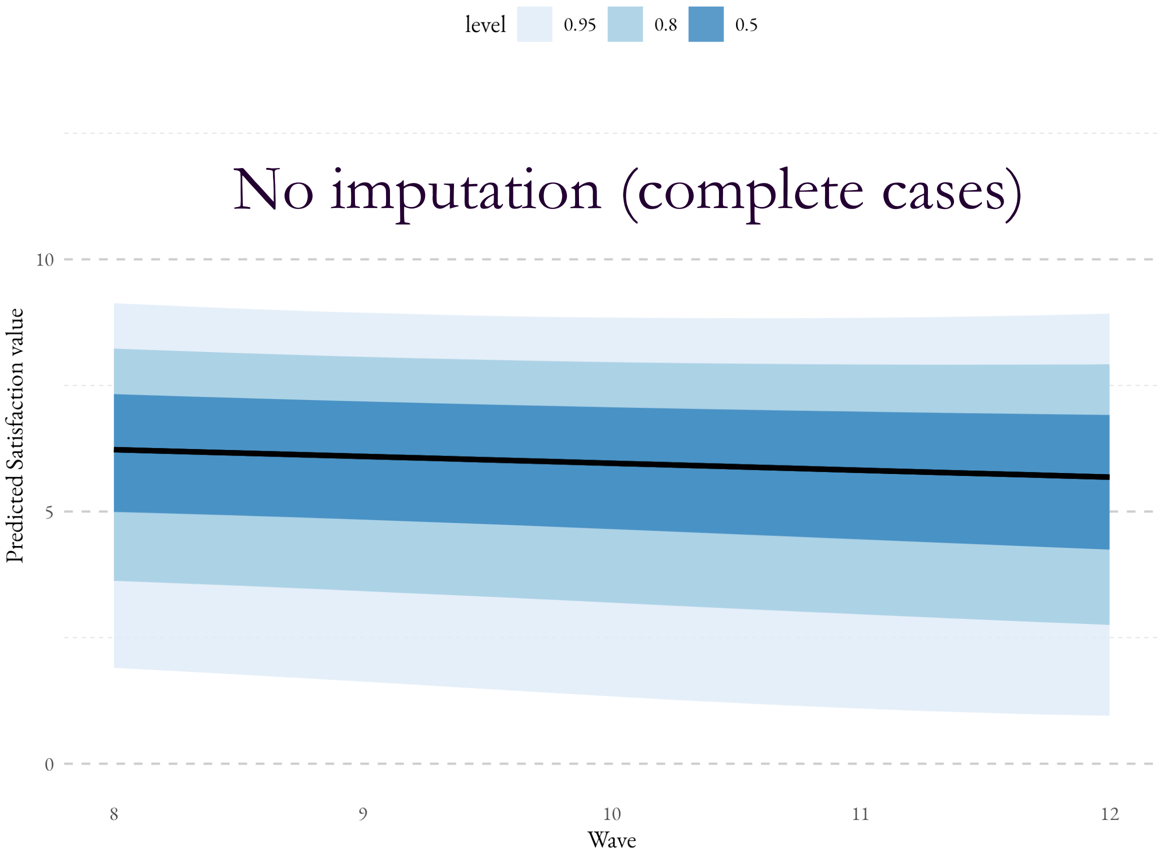
Imputation examples

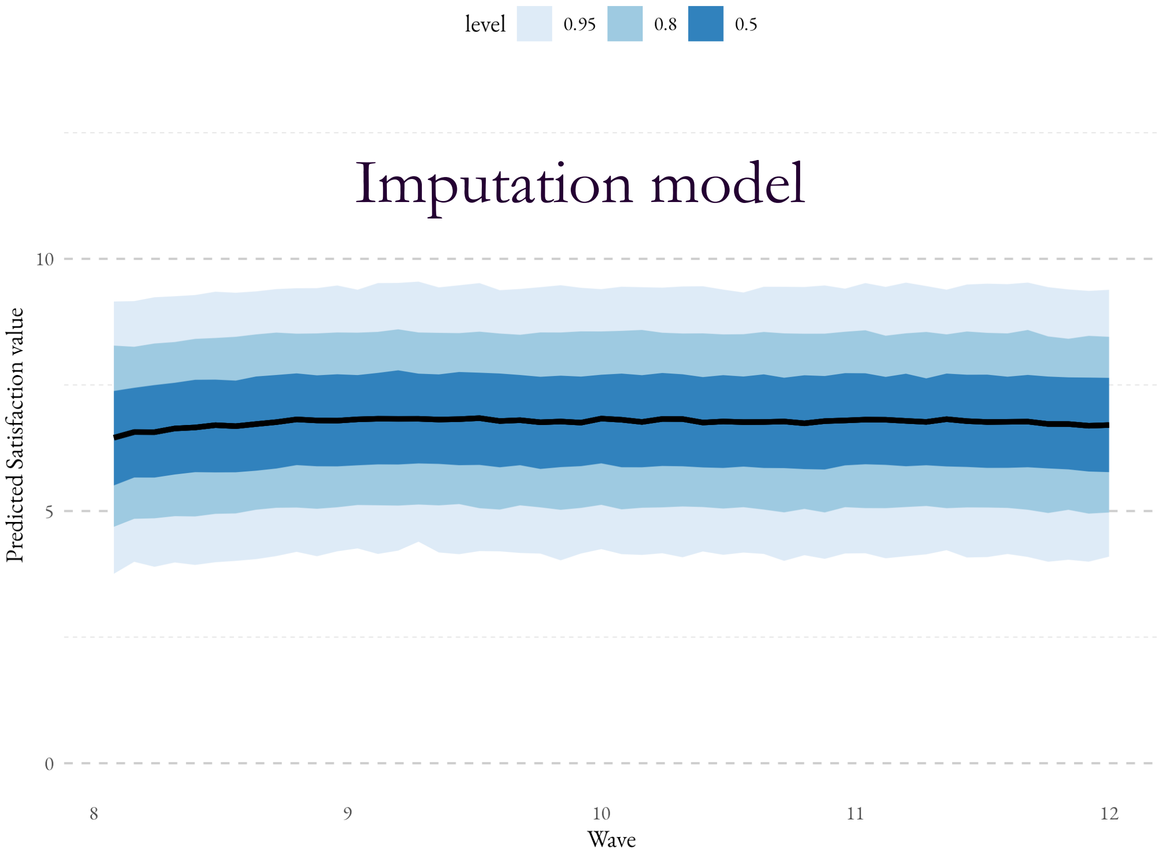


Imputation examples – validation



Prediction Comparison





Benefits and Drawbacks

- Don't need to rely on complete-case data for modeling panel data
- Naturally captures uncertainty in estimates and predictions
- Flexible to different trajectory shapes
- General: can be used to imputed multiple types of variables in the same model
- **But…**
- Relies on our distributional assumptions for the missing data, $p(\, y^{mis} \mid \theta, \phi\,)$
  - However, we can test our model with posterior predictive checks and holdout samples
- Doesn't inform us about the reasons for data missingness

Thank you!

Spencer Aeschliman: spencer.aeschliman@northwestern.edu

Amanda Stathopoulos: a-stathopoulos@northwestern.edu

This presentation is based on work supported by the NSF GRFP grant no. DGE-2234667

In collaboration with Regional Transportation Authority

Bonus Slides

Some options

- Complete case analysis
- Poststratification
- **Imputation**
- Explicit selection modeling
- A combination of the above

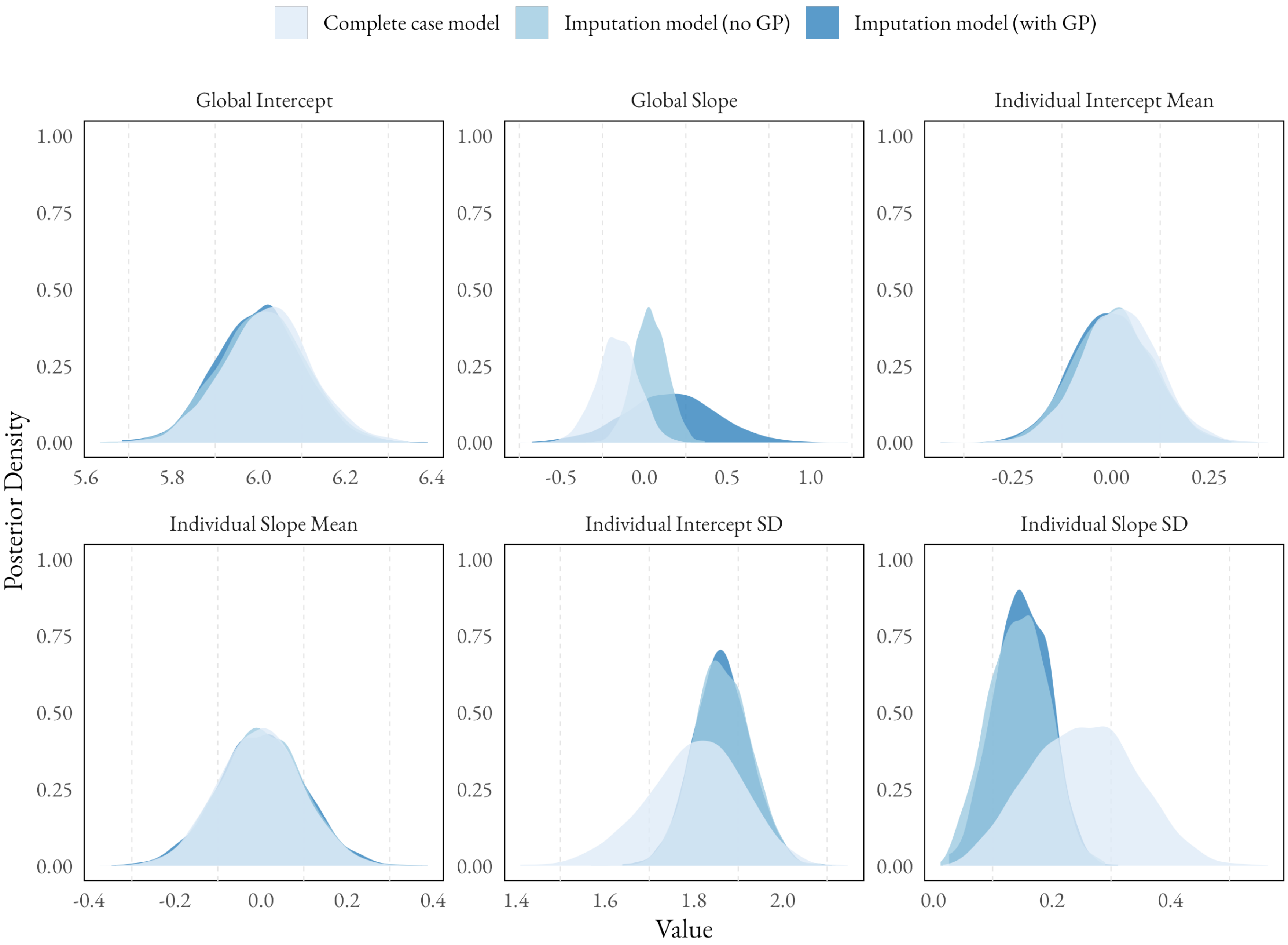
Inference with Imputed Data: The Allure of Making Stuff Up

Charles F. Manski, *Northwestern University*

Submitted October 28, 2022; Accepted August 1, 2023.

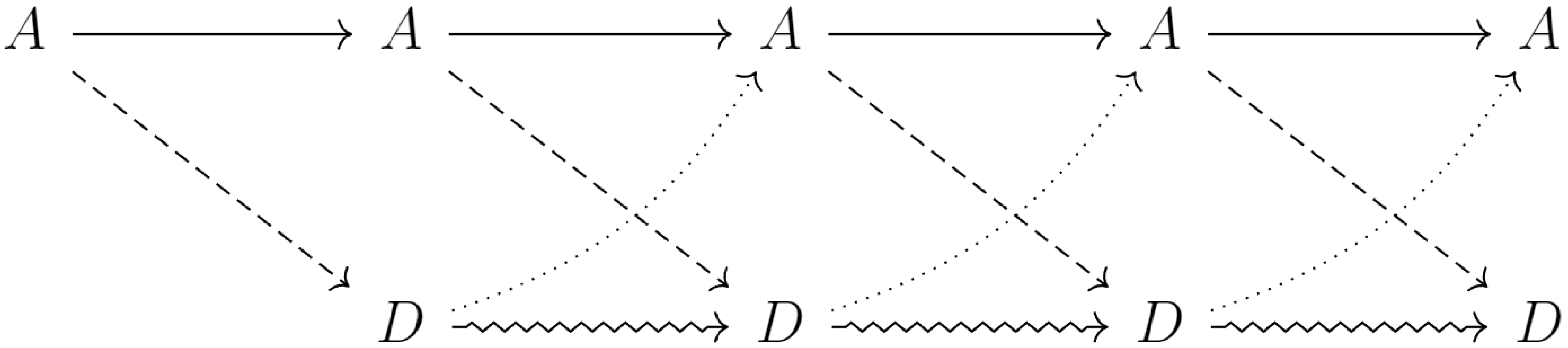
Journal of Labor Economics, volume 43, number S1, April 2025.

Slopes and intercepts



Participation modeling

- At each given time, potential respondent has probability of *answering* or *declining*
- Could model *answering* as a repeated binary choice (See Hensher, 1987)
- Or: model attrition as a time-to-event ("survival")
- Or: both



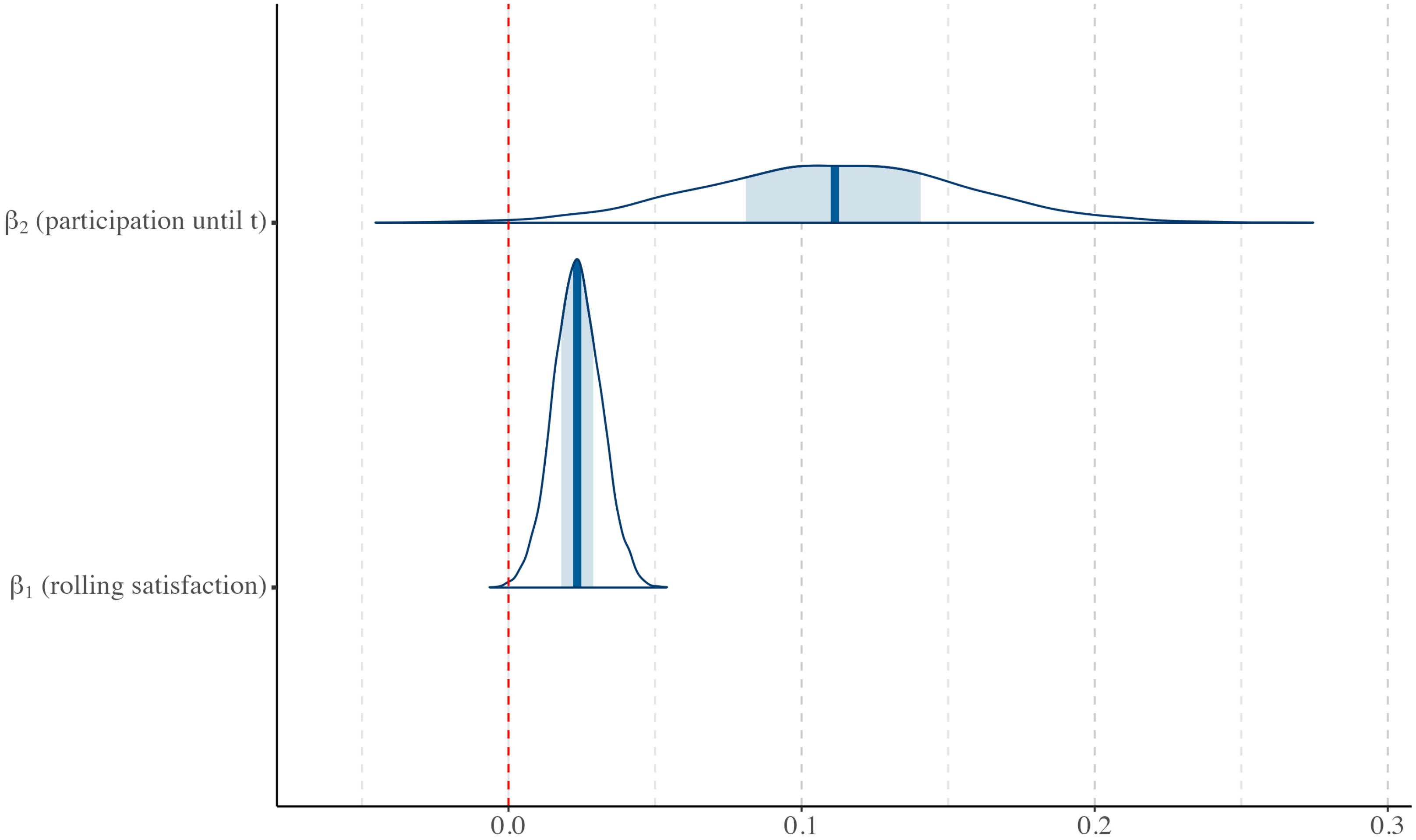
Hierarchical Choice Model Framing

- Let *declining* be the reference
- Standard RUM framing: $U_{it} = \beta_i X_{it} + \varepsilon_{it},$ where $\varepsilon_{it} \sim_{iid} Gumbel$

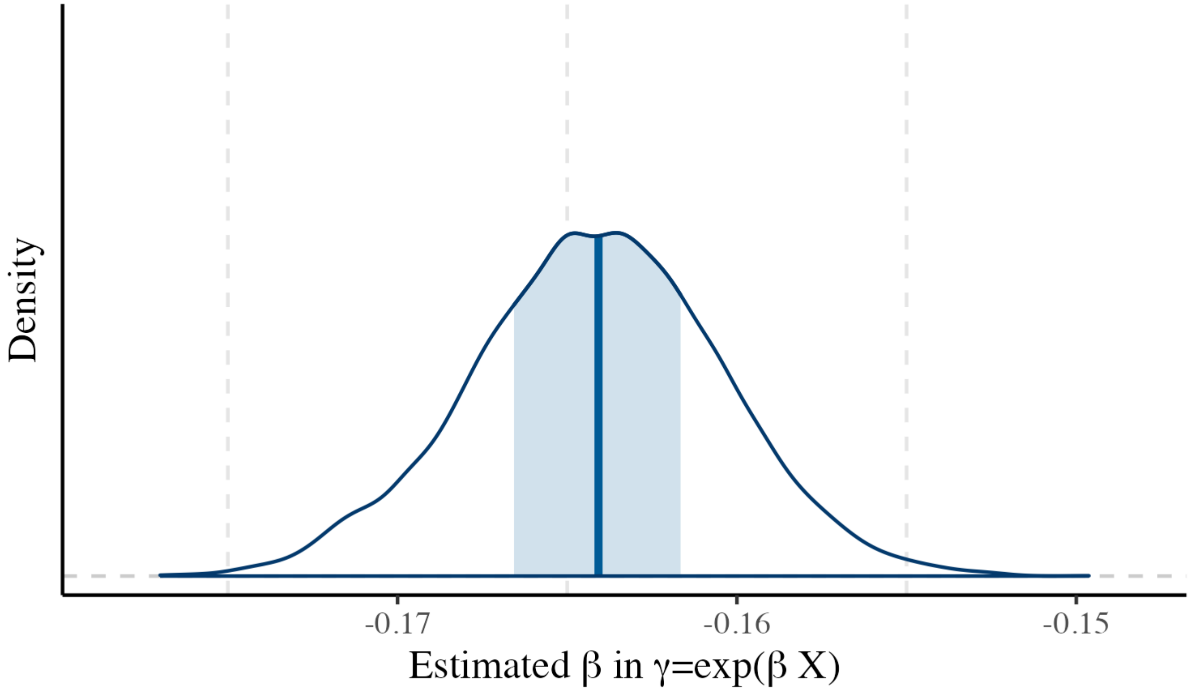
Likelihood: $$p_i(y_i \mid \beta_i) = \prod_t \left(\frac{1}{1 + \exp(\beta_i X_{it})} \right)$$

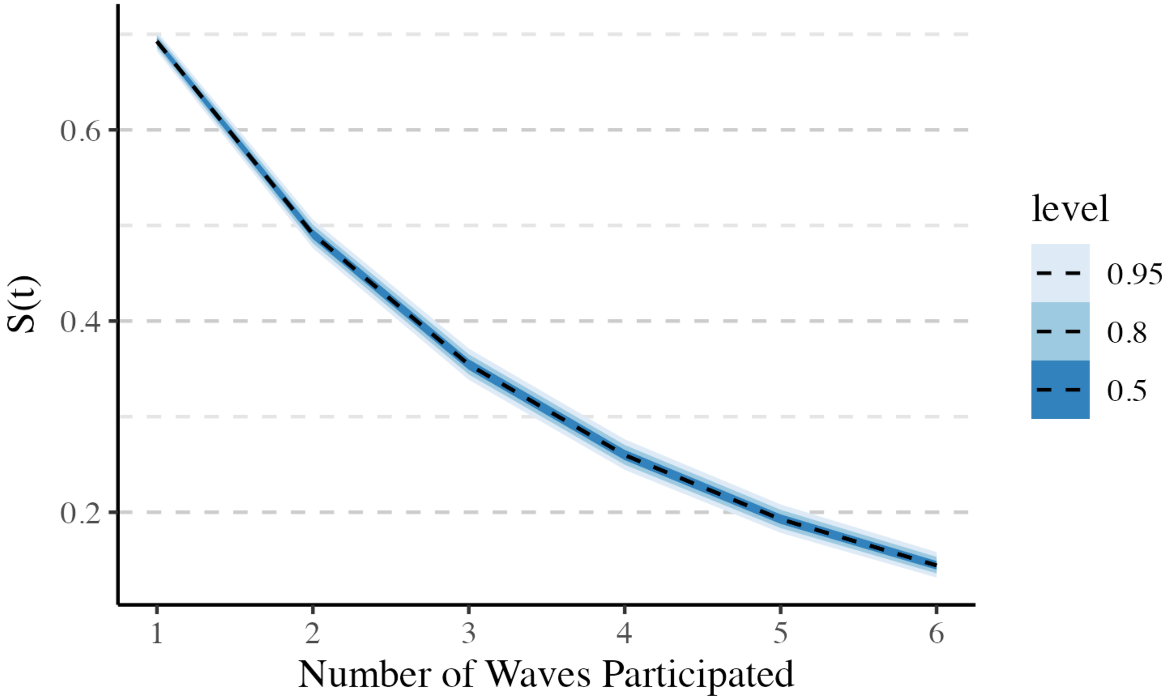
Posterior: $$p(\beta, \mu, \tau \mid \mathrm{y}) = \prod_i \prod_t \underbrace{\left(\frac{1}{1 + \exp(\beta_i X_{it})} \right)}_{\text{Logit kernel}} \underbrace{p(\beta_i \mid \mu, \tau)}_{\text{Mixing distribution}} p(\mu) p(\tau)$$

Choice parameters



Survival model parameters





Hierarchical Model

$$y_{it} \sim N(m_{it}, \sigma)$$

$$m_{it} = \alpha_i + \delta_i t + \sum_k \beta_k X_{itk}$$

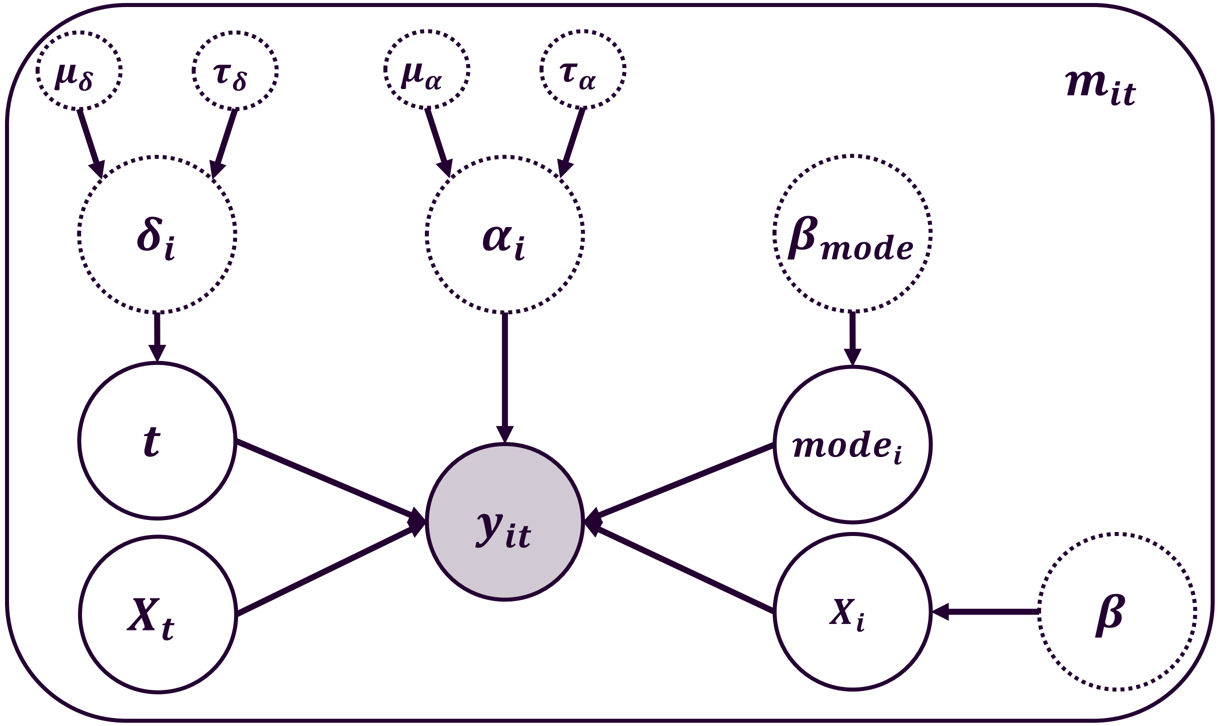
$$\left.\begin{aligned}
&\alpha_i \sim N(\mu_\alpha, \tau_\alpha) \\
&\delta_i \sim N(\mu_\delta, \tau_\delta) \\
&\beta_k \sim N(0,1) \\
&\sigma \sim logNormal(0,1)
\end{aligned}\right\} \text{Priors, } \boldsymbol{\theta}$$

$$\left.\begin{aligned}
&\mu_\alpha, \mu_\delta \sim N(0,1) \\
&\tau_\alpha, \tau_\delta \sim logNormal(0,1)
\end{aligned}\right\} \text{Hyper-priors, } \boldsymbol{\phi}$$

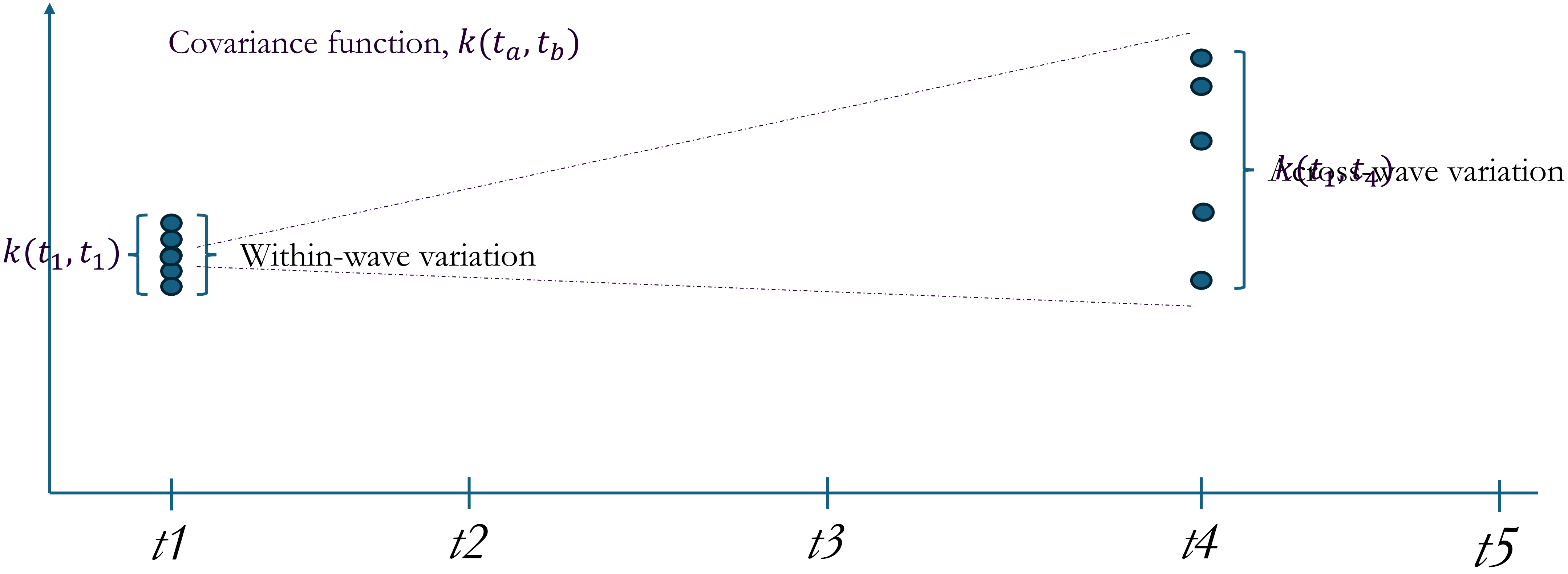
$$p(\theta, \phi \mid y) \propto \prod_i \prod_t p(y_{it} \mid m_{it}; \theta, \phi) p(\theta \mid \phi) p(\phi)$$

$\boldsymbol{y_{it}}$: Person *i*'s satisfaction at time *t*
$\boldsymbol{m_{it}}$: Linear model for satisfaction
$\boldsymbol{\alpha_i}$: Individual (varying) intercept
$\boldsymbol{\delta_i}$: Individual (varying) slope (on *time*)
$\boldsymbol{\beta_k}$: Other coefficients

Hierarchical Model



Adding Explicit Temporal Dependence



Gaussian Process

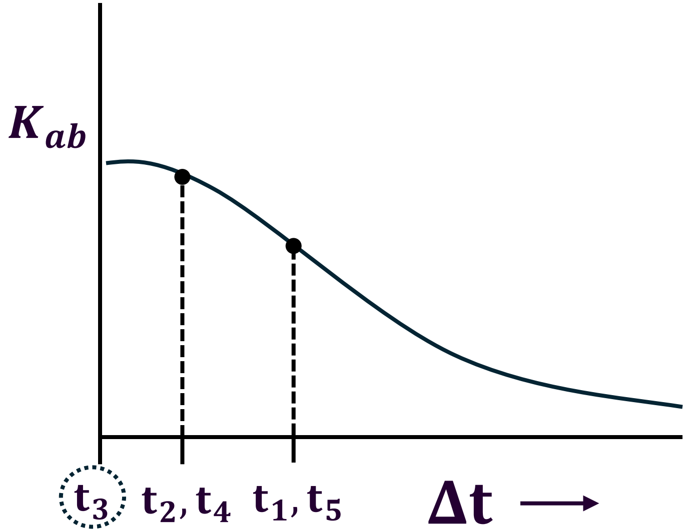
$$p(\theta, \phi, y^{mis} \mid y^{obs}) \propto p(y^{obs} \mid \theta, \phi)\underline{p(y^{mis} \mid \theta, \phi)}p(\theta \mid \phi)p(\phi)$$

Now:

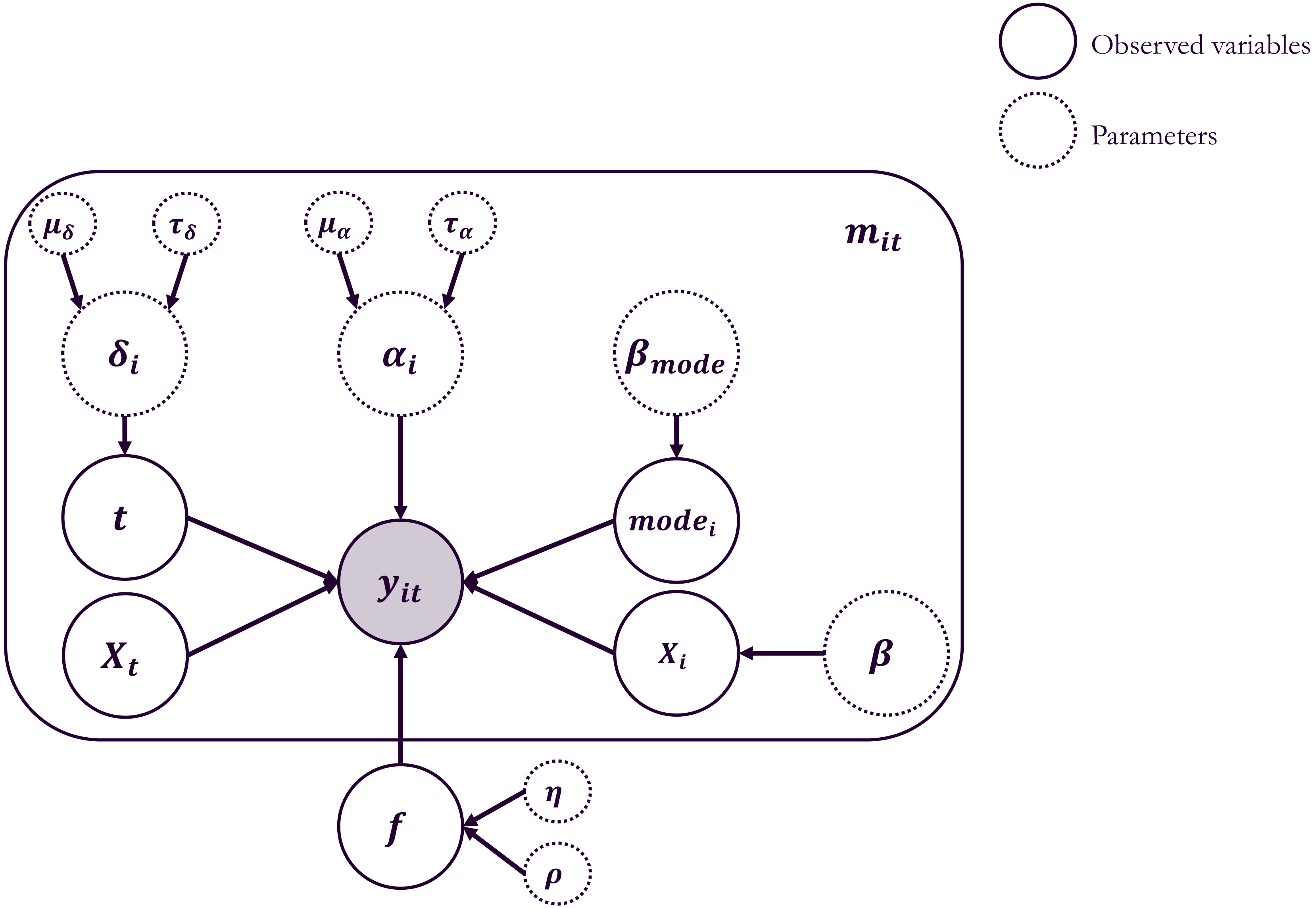
$$y_{it}^{mis} \mid y_i^{obs}, \theta, \phi \sim N\left(m_{it} + \boldsymbol{f}(\boldsymbol{y_i^{obs}}), \sigma\right)$$

$$f(y_i^{obs}) \sim GP(0, K)$$

$$K_{ab}(t_a, t_b \mid \eta, \rho) = \eta^2 \, exp\left(-\frac{1}{2\rho}(t_a - t_b)^2\right)$$

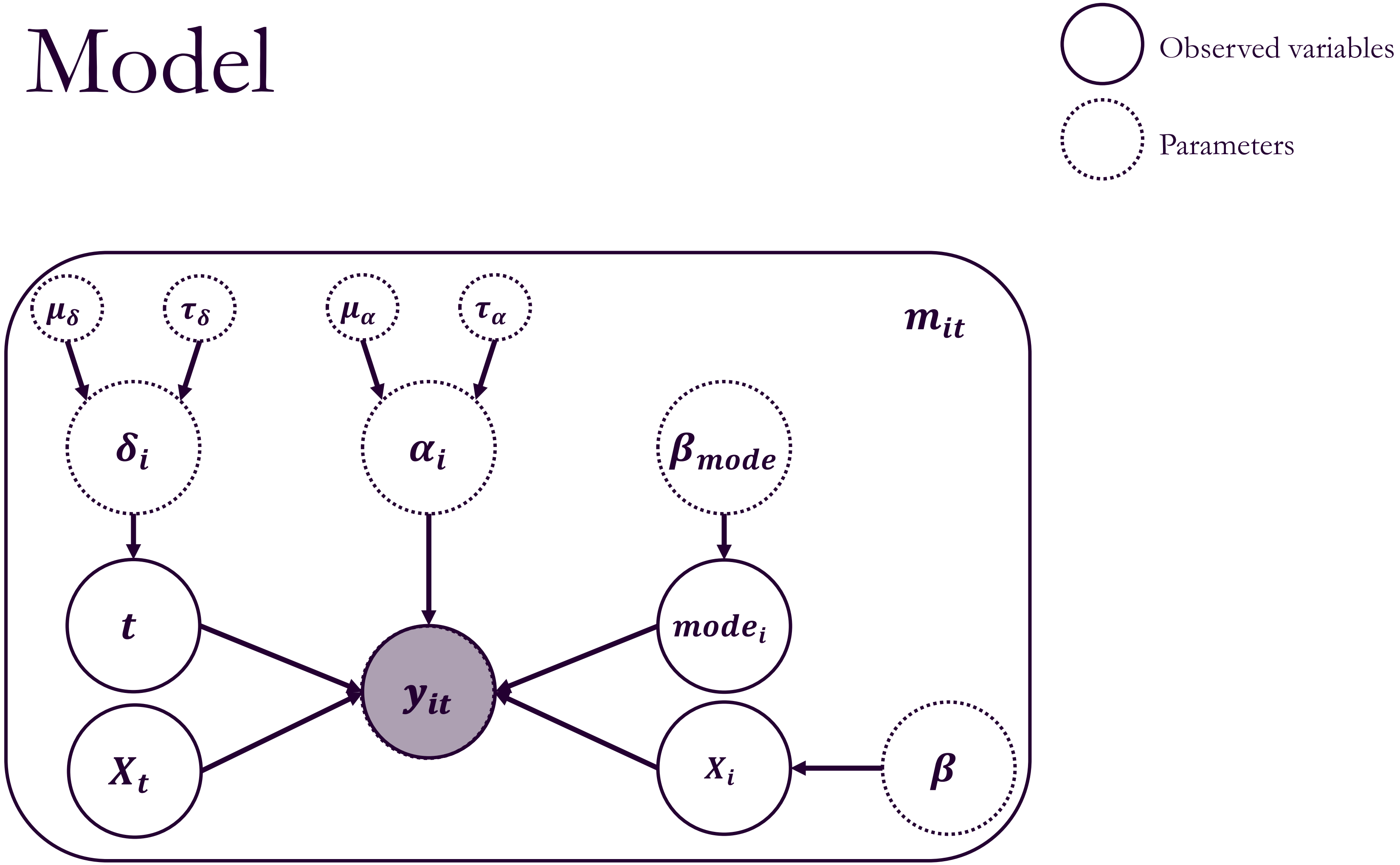


Full Model



Observed variables
Parameters
μ_δ
τ_δ
μ_α
τ_α
m_{it}
δ_i
α_i
β_{mode}
t
$mode_i$
y_{it}
X_t
X_i
β
f
η
ρ

Imputation Model



Imputing Covariates