Appendix for "A Semiparametric Bayesian Approach to Dropout in Longitudinal Studies with Auxiliary Covariates"

Tianjian Zhou*, Michael J. Daniels† and Peter Müller‡

A.1 The Schizophrenia Clinical Trial Dataset Details

Figure A.1 shows individual trajectories and mean responses over time for the three treatment arms.

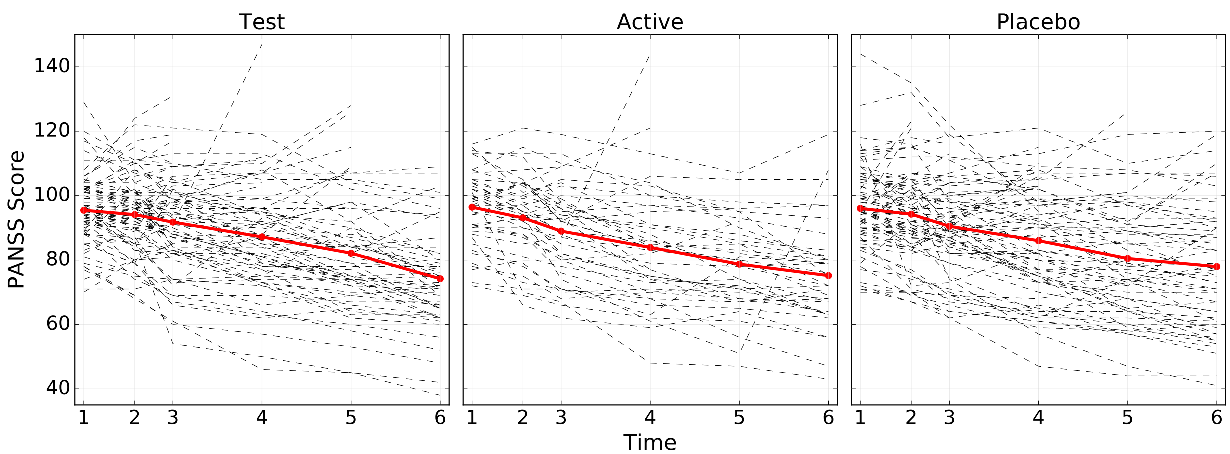

Figure A.1: Trajectories of individual responses (dashed black lines) and mean responses (thick red lines) over time for the active control, placebo and test drug arms.

Table A.1 shows detailed dropout rates for each dropout pattern.

*Department of Public Health Sciences, The University of Chicago
†Department of Statistics, University of Florida. Email: mdaniels@stat.ufl.edu
‡Department of Mathematics, The University of Texas at Austin

	$S_i = 2$	$S_i = 3$	$S_i = 4$	$S_i = 5$	Overall
Test	4.9 (3.7)	12.3 (9.9)	8.6 (8.6)	7.4 (7.4)	33.3 (29.6)
Active	2.2 (2.2)	4.4 (2.2)	8.9 (6.7)	4.4 (4.4)	20.0 (15.6)
Placebo	3.8 (3.8)	5.1 (5.1)	11.5 (11.5)	5.1 (5.1)	25.6 (25.6)

Table A.1: Dropout rates (%) for different dropout patterns in the three treatment arms, with informative dropout rates in parentheses.

A.2 Prior Details

The standardized values for $\boldsymbol{v}$, y_{j-1}, j and s are calculated by

$$\underline{v}_{iq} = \frac{v_{iq} - \text{mean}(v_{\cdot q})}{\text{sd}(v_{\cdot q})}, \qquad \underline{y}_{i,j-1} = \frac{y_{i,j-1} - \text{mean}(y_{\cdot,j-1})}{\text{sd}(y_{\cdot,j-1})},$$
$$\underline{j}_i = \frac{j_i - \min(j_\cdot)}{\max(j_\cdot) - \min(j_\cdot)}, \qquad \underline{s}_i = \frac{s_i - \min(s_\cdot)}{\max(s_\cdot) - \min(s_\cdot)}.$$

We then consider the parameters in the covariance functions (5). We put inverse Gamma priors on κ_0^2 and κ^2,

$$\kappa_0^2 \sim \text{IG}(\lambda_1^{\kappa_0}, \lambda_2^{\kappa_0}), \quad \kappa^2 \sim \text{IG}(\lambda_1^{\kappa}, \lambda_2^{\kappa}).$$

For simplicity, we fix the length scales γ_{v0}, γ_{s0}, γ_y, γ_v, γ_j and γ_s. For example, in practice, we set $\gamma_{v0}^2 = Q$ to introduce moderate correlation between the initial responses of two subjects with similar $\boldsymbol{V}$'s; we set $\gamma_y = \gamma_v = Q + 1$ to introduce moderate correlation between the subsequent responses of two subjects with similar Y_{j-1}'s and $\boldsymbol{V}$'s and to let the effect of the lag-1 response to be roughly equal to an auxiliary covariate; we set $\gamma_j = 5$ to introduce strong correlation between the subsequent responses of one subject measured at two different time points; we set $\gamma_{s0} = \gamma_s = 5$ to introduce strong correlation between the responses of two subjects with the same Y_{j-1}'s and $\boldsymbol{V}$'s but are in two different patterns. We also fix $\tilde{\kappa}_0^2$ and $\tilde{\kappa}^2$ at small values, e.g. $\tilde{\kappa}_0^2 = \tilde{\kappa}^2 = 0.01$.

Next, we consider the parameters in the mean functions (4). We allow the regression coefficients of the auxiliary covariates to vary by pattern. However, it is typical to have sparse patterns. As a result, we consider an informative prior that assumes regression coefficients for neighboring patterns to be similar. In particular, we specify AR(1) type priors on $\boldsymbol{\beta}_{0s}$ and $\boldsymbol{\beta}_s$. For $\boldsymbol{\beta}_s$, we assume

$$\boldsymbol{\beta} \sim N\left[X_\beta \tilde{\boldsymbol{\beta}}, \sigma_\beta^2 \Sigma_\beta(\rho)\right],$$

where

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_3 \\ \vdots \\ \boldsymbol{\beta}_J \end{pmatrix}, \qquad X_\beta = \begin{pmatrix} I \\ I \\ \vdots \\ I \end{pmatrix},$$

and

$$\Sigma_\beta(\rho) = \frac{1}{1-\rho^2} \begin{pmatrix} I & \rho I & \cdots & \rho^{J-2} I \\ \rho I & I & \cdots & \rho^{J-3} I \\ \vdots & \vdots & & \vdots \\ \rho^{J-2} I & \rho^{J-3} I & \cdots & I \end{pmatrix}.$$

The prior on $\boldsymbol{\beta}$ introduces three unknown hyperparameters $\tilde{\boldsymbol{\beta}}$, σ_β^2 and ρ. We specify diffuse normal, inverse Gamma and uniform priors, respectively,

$$\tilde{\boldsymbol{\beta}} \sim N(\mathbf{0}, \delta_\beta^2 I), \quad \sigma_\beta^2 \sim \text{IG}(\lambda_1^\beta, \lambda_2^\beta), \quad \rho \sim \text{Unif}(0,1).$$

Similarly, for $\boldsymbol{\beta}_{0s}$,

$$\boldsymbol{\beta}_0 \sim N\left[X_\beta \tilde{\boldsymbol{\beta}}_0, \sigma_{\beta_0}^2 \Sigma_\beta(\rho_0)\right], \quad \text{with hyper-priors}$$
$$\tilde{\boldsymbol{\beta}}_0 \sim N(\mathbf{0}, \delta_{\beta_0}^2 I), \quad \sigma_{\beta_0}^2 \sim \text{IG}(\lambda_1^{\beta_0}, \lambda_2^{\beta_0}), \quad \rho_0 \sim \text{Unif}(0,1).$$

The time/pattern specific intercepts are given conditional autoregressive (CAR) type priors (De Oliveira, 2012; Banerjee et al., 2014) as we expect them to be similar for neighboring patterns/times. Let $\boldsymbol{b}_0 = (b_{12},\ b_{13},\ \ldots,\ b_{1J})$ and $\boldsymbol{b} = (b_{22};\ b_{23},\ b_{33};\ \ldots\ ;\ b_{2J},\ \ldots,\ b_{JJ})$. The potential neighbors of b_{js} are $\{b_{j-1,s}, b_{j+1,s}, b_{j,s-1}, b_{j,s+1}\}$. Denote by $\mathcal{N}_{js}^b = \{(j', s') : b_{j's'}$ is neighbor of $b_{js}\}$ and $N_{js}^b = |\mathcal{N}_{js}^b|$ which is the number of neighbors of b_{js}. The CAR type prior assigns conditional priors on b_{js} given its neighbors, and under several regularity conditions the conditionals indicate a joint distribution. In particular, we assume

$$b_{js} \mid b_{-js} \sim N\left(\tilde{b} + \sum_{j's' \in \mathcal{N}_{js}^b} \frac{\gamma_b}{N_{js}^b}\left(b_{j's'} - \tilde{b}\right), \frac{\sigma_b^2}{N_{js}^b}\right),$$

which induces a joint prior on $\boldsymbol{b}$ of the form

$$\boldsymbol{b} \sim N\left(\mathbf{1}\tilde{b}, \sigma_b^2 (I - \gamma_b W_b)^{-1} \mathcal{N}_b\right),$$

where

$$(W_b)_{jsj's'} = \begin{cases} 1/N_{js}^b, & \text{if } (j,s) \text{ and } (j',s') \text{ are neighbors;} \\ 0, & \text{otherwise,} \end{cases}$$

$\mathcal{N}_b = \text{diag}(1/N_{js}^b)$, $\tilde{b}$ is a mean parameter for $\boldsymbol{b}$, σ_b^2 is a variance parameter and γ_b is a spatial dependence parameter. Let $\left(e_1^b\right)^{-1}$ and $\left(e_2^b\right)^{-1}$ denote the max and min eigenvalues of W_b. To guarantee that $I - \gamma_b W_b$ is positive definite, γ_b is required to belong to (e_2^b, e_1^b). Furthermore, it is not unreasonable to assume the spatial correlation is positive, i.e. $0 < \gamma_b < e_1^b$. We put hyper-priors on $\tilde{b}$, σ_b^2 and γ_b,

$$\tilde{b} \sim N(0, \delta_b^2), \quad \sigma_b^2 \sim \text{IG}(\lambda_1^b, \lambda_2^b), \quad \gamma_b \sim \text{Unif}(0, e_1^b).$$

Similarly, for $\boldsymbol{b}_0$, we assume

$$\boldsymbol{b}_0 \sim N\left(\mathbf{1}\tilde{b}_0, \sigma_{b_0}^2 (I - \gamma_{b_0} W_{b_0})^{-1} \mathcal{N}_{b_0}\right); \quad \text{with hyper-priors}$$
$$\tilde{b}_0 \sim N(0, \delta_{b_0}^2), \quad \sigma_{b_0}^2 \sim \text{IG}(\lambda_1^b, \lambda_2^b), \quad \gamma_{b_0} \sim \text{Unif}(0, e_1^{b_0}).$$

The time/pattern specific lag-1 coefficients are given CAR type priors similar to the priors on b_{js} for the same reason. Let $\boldsymbol{\psi} = (\psi_{22}; \psi_{23}, \psi_{33}; \ldots; \psi_{2J}, \ldots, \psi_{JJ})$. We assume

$$\boldsymbol{\psi} \sim N\left(\mathbf{1}\tilde{\psi}, \sigma_\psi^2 (I - \gamma_\psi W_\psi)^{-1} \mathcal{N}_\psi\right); \quad \text{with hyper-priors}$$
$$\tilde{\psi} \sim N(1, \delta_\psi^2), \quad \sigma_\psi^2 \sim \text{IG}(\lambda_1^\psi, \lambda_2^\psi), \quad \text{and} \;\; \gamma_\psi \sim \text{Unif}(0, e_1^\psi).$$

A.3 MCMC Implementation Details

We introduce some notation as follows. First considering the responses. Denote by N_s the number of subjects having dropout pattern s, $s = 2, \ldots, J$. Let $\boldsymbol{y}_{js}$ denote the subjects' responses at time j in pattern s, and $\bar{Y}_{js}$ denote the subjects' histories through the first j times in pattern s, i.e.

$$\boldsymbol{y}_{js} = (y_{1js}, y_{2js}, \ldots, y_{N_s,j,s})^T,$$
$$\bar{Y}_{js} = (\boldsymbol{y}_{1s}, \boldsymbol{y}_{2s}, \ldots, \boldsymbol{y}_{js}).$$

Let $\boldsymbol{y}_{\text{vec0}}$ denote the initial responses (with no past) for all subjects, and $\boldsymbol{y}_{\text{vec}}$ denote the subsequent responses (with measured pasts) for all subjects,

$$\boldsymbol{y}_{\text{vec0}} = \left(\boldsymbol{y}_{12}^T, \boldsymbol{y}_{13}^T, \ldots, \boldsymbol{y}_{1J}^T\right)^T$$
$$\boldsymbol{y}_{\text{vec}} = \left(\boldsymbol{y}_{22}^T, \boldsymbol{y}_{23}^T, \boldsymbol{y}_{33}^T, \ldots, \boldsymbol{y}_{2J}^T, \ldots, \boldsymbol{y}_{JJ}^T\right)^T.$$

We then consider the means and covariate matrices for the responses. Let $\boldsymbol{a}_{js}$ denote the vector of random variables (we abuse notation slightly, let $\boldsymbol{a}_{js}$ include $\bar{Y}_{j-2,s}\boldsymbol{\phi}_{js}$ when $j \geq 2$, to simplify notation),

$$\boldsymbol{a}_{js} = \begin{cases} \left(a_0(\boldsymbol{v}_{1s}, s), \ldots, a_0(\boldsymbol{v}_{N_s,s}, s)\right)^T, & \text{if } j = 1; \\ \left(a(y_{1,j-1,s}, \boldsymbol{v}_{1s}, j, s), \ldots, a(y_{N_s,j-1,s}, \boldsymbol{v}_{N_s,s}, j, s)\right)^T + \bar{Y}_{j-2,s}\boldsymbol{\phi}_{js}, & \text{if } j \geq 2, \end{cases}$$

where $\boldsymbol{v}_{is}$ is the vector of auxiliary covariates for subject i in pattern s. The vector $\boldsymbol{a}_{js}$ is the mean of $\boldsymbol{y}_{js}$. Let $\boldsymbol{a}_0$ and $\boldsymbol{a}$ denote the vector of random variables,

$$\boldsymbol{a}_0 = \left(\boldsymbol{a}_{12}^T, \boldsymbol{a}_{13}^T, \ldots, \boldsymbol{a}_{1J}^T\right)^T$$
$$\boldsymbol{a} = \left(\boldsymbol{a}_{22}^T, \boldsymbol{a}_{23}^T, \boldsymbol{a}_{33}^T, \ldots, \boldsymbol{a}_{2J}^T, \ldots, \boldsymbol{a}_{JJ}^T\right)^T.$$

Denote by

$$\Sigma_{y_0} = \text{diag}\left(\sigma_{12}^2 I_{N_2}, \ldots, \sigma_{1J}^2 I_{N_J}\right),$$
$$\Sigma_y = \text{diag}\left(\sigma_{22}^2 I_{N_2}, \sigma_{23}^2 I_{N_3}, \sigma_{33}^2 I_{N_3}, \ldots, \sigma_{2J}^2 I_{N_J}, \ldots, \sigma_{JJ}^2 I_{N_J}\right).$$

Thus, the **likelihoods for the initial responses $\boldsymbol{y}_{\text{vec0}}$ and subsequent responses $\boldsymbol{y}_{\text{vec}}$** are

$$\boldsymbol{y}_{\text{vec0}} \mid \boldsymbol{a}_0, \Sigma_{y_0} \sim N(\boldsymbol{a}_0, \Sigma_{y_0}),$$
$$\boldsymbol{y}_{\text{vec}} \mid \boldsymbol{a}, \Sigma_y \sim N(\boldsymbol{a}, \Sigma_y).$$

Next, we consider the priors for $\boldsymbol{a}_0$ and $\boldsymbol{a}$. Denote by

$$\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0, \boldsymbol{b}_0),$$
$$\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{b}, \boldsymbol{\psi}, \boldsymbol{\phi}),$$

where $\boldsymbol{\phi} = (\boldsymbol{\phi}_{33}; \boldsymbol{\phi}_{34}, \boldsymbol{\phi}_{44}; \ldots; \boldsymbol{\phi}_{3J}, \ldots, \boldsymbol{\phi}_{JJ})$.

Let D_0 and D denote the exponential distance matrices for $\boldsymbol{a}_0$ and $\boldsymbol{a}$ (abuse notation slightly, we use D_0 and D to denote the matrices and $D_0(a;b)$ and $D(a;b)$ to denote the distance functions),

$$D_0 = D_0(V_{\text{vec0}}, \boldsymbol{s}_{\text{vec0}}; V_{\text{vec0}}, s_{\text{vec0}}),$$
$$D = D(\boldsymbol{y}_{\text{lag}}, V_{\text{vec}}, \boldsymbol{j}_{\text{vec}}, \boldsymbol{s}_{\text{vec}}; \boldsymbol{y}_{\text{lag}}, V_{\text{vec}}, \boldsymbol{j}_{\text{vec}}, \boldsymbol{s}_{\text{vec}}),$$

with

$$[D_0]_{ijs,i'j's'} = D_0(\boldsymbol{v}_{is}, s; \boldsymbol{v}_{i's'}, s'),$$
$$[D]_{ijs,i'j's'} = D(y_{i,j-1,s}, \boldsymbol{v}_{is}, j, s; y_{i',j'-1,s'}, \boldsymbol{v}_{i's'}, j', s').$$

Here V_{vec0} is the matrix of auxiliary covariates corresponding to $\boldsymbol{y}_{\text{vec0}}$, $\boldsymbol{s}_{\text{vec0}}$ is the vector of patterns corresponding to $\boldsymbol{y}_{\text{vec0}}$, $\boldsymbol{y}_{\text{lag}}$ is the vector of lag-1 responses corresponding to $\boldsymbol{y}_{\text{vec}}$, V_{vec} is the matrix of auxiliary covariates corresponding to $\boldsymbol{y}_{\text{vec}}$, and $\boldsymbol{j}_{\text{vec}}$ and $\boldsymbol{s}_{\text{vec}}$ are the vectors of times and patterns corresponding to $\boldsymbol{y}_{\text{vec}}$.

We have

$$\boldsymbol{a}_0 \mid \boldsymbol{\theta}_0, \kappa_0^2 \sim N(X_{\theta_0}\boldsymbol{\theta}_0, \kappa_0^2 D_0 + \tilde{\kappa}_0^2 I),$$
$$\boldsymbol{a} \mid \boldsymbol{\theta}, \kappa^2 \sim N(X_\theta \boldsymbol{\theta}, \kappa^2 D + \tilde{\kappa}^2 I),$$

where X_{θ_0} and X_θ are the design matrices corresponding to Equation (4).

Denote by $C_0 = \kappa_0^2 D_0 + \tilde{\kappa}_0^2 I$ and $C = \kappa^2 D + \tilde{\kappa}^2 I$. Integrating out $\boldsymbol{a}_0$ and $\boldsymbol{a}$, the **(marginal) likelihoods** become

$$\boldsymbol{y}_{\text{vec0}} \mid \boldsymbol{\theta}_0, \Sigma_{y_0}, \kappa_0^2 \sim N(X_{\theta_0}\boldsymbol{\theta}_0, \Sigma_{y_0} + C_0),$$
$$\boldsymbol{y}_{\text{vec}} \mid \boldsymbol{\theta}, \Sigma_y, \kappa^2 \sim N(X_\theta \boldsymbol{\theta}, \Sigma_y + C).$$

Update $\boldsymbol{a}_0$ and $\boldsymbol{a}$. It is not unusual to integrate out $\boldsymbol{a}_0$ and $\boldsymbol{a}$ for posterior inference on Gaussian process. However, we find that including $\boldsymbol{a}_0$ and $\boldsymbol{a}$ in the posterior inference would improve the mixing of the Markov chain. Therefore, we update $\boldsymbol{a}_0$ and $\boldsymbol{a}$ at each iteration.

1. The likelihood and prior for $\boldsymbol{a}_0$ are

$$\boldsymbol{y}_{\text{vec0}} \mid \boldsymbol{a}_0, \Sigma_{y_0} \sim N(\boldsymbol{a}_0, \Sigma_{y_0}),$$
$$\boldsymbol{a}_0 \mid \boldsymbol{\theta}_0, \kappa_0^2 \sim N(X_{\theta_0}\boldsymbol{\theta}_0, C_0),$$

which lead to the posterior

$$\begin{aligned}\boldsymbol{a}_0 \mid \boldsymbol{\theta}_0, \kappa_0^2, \Sigma_{y_0}, \boldsymbol{y}_{\text{vec0}} &\sim N(\boldsymbol{a}_0^*, \Sigma_{a_0}^*), \text{ where} \\ \Sigma_{a_0}^* &= [C_0^{-1} + \Sigma_{y_0}^{-1}]^{-1}, \\ \boldsymbol{a}_0^* &= \Sigma_{a_0}^*[C_0^{-1} X_{\theta_0} \boldsymbol{\theta}_0 + \Sigma_{y_0}^{-1} \boldsymbol{y}_{\text{vec0}}].\end{aligned}$$

2. The likelihood and prior for $\boldsymbol{a}$ are

$$\begin{aligned}\boldsymbol{y}_{\text{vec}} \mid \boldsymbol{a}, \Sigma_y &\sim N(\boldsymbol{a}, \Sigma_y), \\ \boldsymbol{a} \mid \boldsymbol{\theta}, \kappa^2 &\sim N(X_\theta \boldsymbol{\theta}, C),\end{aligned}$$

which lead to the posterior

$$\begin{aligned}\boldsymbol{a} \mid \boldsymbol{\theta}, \kappa^2, \Sigma_y, \boldsymbol{y}_{\text{vec}} &\sim N(\boldsymbol{a}^*, \Sigma_a^*), \text{ where} \\ \Sigma_a^* &= [C^{-1} + \Sigma_y^{-1}]^{-1}, \\ \boldsymbol{a}^* &= \Sigma_a^*[C^{-1} X_\theta \boldsymbol{\theta} + \Sigma_y^{-1} \boldsymbol{y}_{\text{vec}}].\end{aligned}$$

Update κ_0^2 and κ^2. 1. The likelihood and prior for κ_0^2 are

$$\begin{aligned}\boldsymbol{a}_0 \mid \boldsymbol{\theta}_0, \kappa_0^2 &\sim N(X_{\theta_0} \boldsymbol{\theta}_0, \kappa_0^2 \boldsymbol{D}_0 + \tilde{\kappa}_0^2 I), \\ \kappa_0^2 &\sim \text{IG}(\lambda_1^{\kappa_0}, \lambda_2^{\kappa_0}).\end{aligned}$$

The posterior for κ_0^2 is

$$p(\kappa_0^2 \mid \boldsymbol{\theta}_0, \boldsymbol{a}_0) \propto p_N(\boldsymbol{a}_0 \mid X_{\theta_0} \boldsymbol{\theta}_0, \kappa_0^2 \boldsymbol{D}_0 + \tilde{\kappa}_0^2 I) \cdot p_{\text{IG}}(\kappa_0^2 \mid \lambda_1^{\kappa_0}, \lambda_2^{\kappa_0}),$$

where $p_N(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma)$ represents (multivariate) normal density at $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ, and $p_{\text{IG}}(x \mid a, b)$ represents inverse gamma density at x with shape parameter a and rate parameter b. We use Metropolis-Hastings step to update κ_0^2.
2. The likelihood and prior for κ^2 are

$$\begin{aligned}\boldsymbol{a} \mid \boldsymbol{\theta}, \kappa^2 &\sim N(X_\theta \boldsymbol{\theta}, \kappa^2 \boldsymbol{D} + \tilde{\kappa}^2 I), \\ \kappa^2 &\sim \text{IG}(\lambda_1^{\kappa}, \lambda_2^{\kappa}).\end{aligned}$$

The posterior for κ^2 is

$$p(\kappa^2 \mid \boldsymbol{\theta}, \boldsymbol{a}) \propto p_N(\boldsymbol{a} \mid X_\theta \boldsymbol{\theta}, \kappa^2 \boldsymbol{D} + \tilde{\kappa}^2 I) \cdot p_{\text{IG}}(\kappa^2 \mid \lambda_1^{\kappa}, \lambda_2^{\kappa}).$$

We use Metropolis-Hastings step to update κ^2.

Update Σ_{y_0} and Σ_y. The likelihood and prior for σ^2_{js} are

$$\boldsymbol{y}_{js} \mid \boldsymbol{a}_{js}, \sigma^2_{js} \sim N(\boldsymbol{a}_{js}, \sigma^2_{js} I),$$
$$\sigma^2_{js} \mid \lambda_\sigma, \nu_\sigma \sim \mathrm{IG}(\lambda_\sigma, \lambda_\sigma \nu_\sigma).$$

The posterior for σ^2_{js} is

$$\sigma^2_{js} \mid \lambda_\sigma, \nu_\sigma, \boldsymbol{a}_{js}, \boldsymbol{y}_{js} \sim \mathrm{IG}\left(\lambda_\sigma + \frac{N_s}{2},\ \lambda_\sigma \nu_\sigma + \frac{RSS_{js}}{2}\right),$$

where $RSS_{js} = \|\boldsymbol{y}_{js} - \boldsymbol{a}_{js}\|_2^2$.

There are two hyperparameters related to σ^2_{js}: λ_σ and ν_σ. Their conditional posteriors are

$$p(\lambda_\sigma \mid \{\sigma^2_{js}\}, \nu_\sigma) \propto \frac{(\nu_\sigma \lambda_\sigma)^{(2+J)(J-1)\lambda_\sigma/2}}{\Gamma(\lambda_\sigma)^{(2+J)(J-1)/2}} \prod_{j,s} \left(\sigma^2_{js}\right)^{-(\lambda_\sigma - 1)} \cdot$$
$$\exp\left(-\sum_{j,s} \frac{\nu_\sigma}{\sigma^2_{js}} \lambda_\sigma\right) \exp\left(-\frac{1}{\lambda_\sigma - 2}\right),$$

and

$$\nu_\sigma \mid \{\sigma^2_{js}\}, \lambda_\sigma \sim \mathrm{Gamma}\left(\frac{(2+J)(J-1)}{2}\lambda_\sigma + 1,\ \sum_{j,s} \frac{\lambda_\sigma}{\sigma^2_{js}} + 1\right).$$

We use Metropolis-Hastings step to update λ_σ.

Update $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$. We integrate out $\boldsymbol{a}_0$ and $\boldsymbol{a}$ to update $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$. The likelihoods become

$$\boldsymbol{y}_{\mathrm{vec}0} \mid \boldsymbol{\theta}_0, \Sigma_{y_0}, \kappa_0^2 \sim N(X_{\theta_0}\boldsymbol{\theta}_0, \Sigma_{y_0} + C_0),$$
$$\boldsymbol{y}_{\mathrm{vec}} \mid \boldsymbol{\theta}, \Sigma_y, \kappa^2 \sim N(X_\theta \boldsymbol{\theta}, \Sigma_y + C).$$

1. For $\boldsymbol{\theta}_0$, the prior is

$$\boldsymbol{\theta}_0 \mid \tilde{\boldsymbol{\beta}}_0, \sigma^2_{\beta_0}, \rho_0, \tilde{b}_0, \sigma^2_{b_0}, \gamma_{b_0} \sim N(\tilde{\boldsymbol{\theta}}_0, \Sigma_{\theta_0}),$$

where $\tilde{\boldsymbol{\theta}}_0 = (X_\beta \tilde{\boldsymbol{\beta}}_0, \mathbf{1}\tilde{b}_0)$, and

$$\Sigma_\theta = \mathrm{diag}\big(\sigma^2_{\beta_0}\Sigma_\beta(\rho_0), \sigma^2_{b_0}(I - \gamma_{b_0} W_{b_0})^{-1}\mathcal{N}_{b_0}\big).$$

Thus, the posterior of $\boldsymbol{\theta}_0$ is

$$\begin{aligned}
\boldsymbol{\theta}_0 \mid \boldsymbol{y}_{\text{vec}0}, \ldots &\sim N(\boldsymbol{\theta}_0^*, \Sigma_{\theta_0}^*), \quad \text{where} \\
\Sigma_{\theta_0}^* &= \left[\Sigma_{\theta_0}^{-1} + X_{\theta_0}^T (\Sigma_{y_0} + C_0)^{-1} X_{\theta_0}\right]^{-1}, \\
\boldsymbol{\theta}_0^* &= \Sigma_{\theta_0}^* \left[\Sigma_{\theta_0}^{-1} \tilde{\boldsymbol{\theta}}_0 + X_{\theta_0}^T (\Sigma_{y_0} + C_0)^{-1} \boldsymbol{y}_{\text{vec}0}\right].
\end{aligned}$$

2. For $\boldsymbol{\theta}$, the prior is

$$\boldsymbol{\theta} \mid \tilde{\boldsymbol{\beta}}, \sigma_\beta^2, \rho, \tilde{b}, \sigma_b^2, \gamma_b, \tilde{\psi}, \sigma_\psi^2, \gamma_\psi, \sigma_\phi^2 \sim N(\tilde{\boldsymbol{\theta}}, \Sigma_\theta),$$

where $\tilde{\boldsymbol{\theta}} = (X_\beta \tilde{\boldsymbol{\beta}}, \mathbf{1}\tilde{b}, \mathbf{1}\tilde{\psi}, \mathbf{0})$, and

$$\Sigma_\theta = \text{diag}\left(\sigma_\beta^2 \Sigma_\beta(\rho), \sigma_b^2 (I - \gamma_b W_b)^{-1} \mathcal{N}_b, \sigma_\psi^2 (I - \gamma_\psi W_\psi)^{-1} \mathcal{N}_\psi, \sigma_\phi^2 I\right).$$

Thus, the posterior of $\boldsymbol{\theta}$ is

$$\begin{aligned}
\boldsymbol{\theta} \mid \boldsymbol{y}_{\text{vec}}, \ldots &\sim N(\boldsymbol{\theta}^*, \Sigma_\theta^*), \quad \text{where} \\
\Sigma_\theta^* &= \left[\Sigma_\theta^{-1} + X_\theta^T (\Sigma_y + C)^{-1} X_\theta\right]^{-1}, \\
\boldsymbol{\theta}^* &= \Sigma_\theta^* \left[\Sigma_\theta^{-1} \tilde{\boldsymbol{\theta}} + X_\theta^T (\Sigma_y + C)^{-1} \boldsymbol{y}_{\text{vec}}\right].
\end{aligned}$$

Hyperparameters related to $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. There are three hyperparameters related to $\boldsymbol{\beta}$: $\tilde{\boldsymbol{\beta}}$, σ_β^2 and ρ. The conditional posteriors are as follows.

1. Conditional posterior of $\tilde{\boldsymbol{\beta}}$:

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} \mid \boldsymbol{\beta}, \sigma_\beta^2, \rho &\sim N(\tilde{\boldsymbol{\beta}}^*, \Sigma_{\tilde{\boldsymbol{\beta}}}^*), \quad \text{where} \\
\Sigma_{\tilde{\boldsymbol{\beta}}}^* &= \left[\frac{1}{\delta_\beta^2} I + \frac{1}{\sigma_\beta^2} X_\beta' \Sigma_\beta(\rho)^{-1} X_\beta\right]^{-1}, \\
\tilde{\boldsymbol{\beta}}^* &= \Sigma_{\tilde{\boldsymbol{\beta}}}^* \left[\frac{1}{\sigma_\beta^2} X_\beta' \Sigma_\beta(\rho)^{-1} \boldsymbol{\beta}\right].
\end{aligned}$$

2. Conditional posterior of σ_β^2:

$$\sigma_\beta^2 \mid \boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}, \rho \sim \text{IG}\left[\lambda_1^\beta + \frac{(J-1)Q}{2},\ \lambda_2^\beta + \frac{1}{2}(\boldsymbol{\beta} - X_\beta \tilde{\boldsymbol{\beta}})' \Sigma_\beta(\rho)^{-1} (\boldsymbol{\beta} - X_\beta \tilde{\boldsymbol{\beta}})\right].$$

3. Conditional posterior of ρ:

$$
\begin{aligned}
& p(\rho \mid \boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}, \sigma_\beta^2) \\
& \propto \det[\sigma_\beta^{-2}\Sigma_\beta(\rho)^{-1}]^{1/2} \exp\left[-\frac{1}{2\sigma_\beta^2}(\boldsymbol{\beta} - X_\beta\tilde{\boldsymbol{\beta}})'\Sigma_\beta(\rho)^{-1}(\boldsymbol{\beta} - X_\beta\tilde{\boldsymbol{\beta}})\right] \\
& \propto (1-\rho^2)^{Q/2} \exp\left[-\frac{1}{2\sigma_\beta^2}\left(\rho^2 R_{\beta 1} - 2\rho R_{\beta 2}\right)\right],
\end{aligned}
$$

where

$$
R_{\beta 1} = \sum_{s=3}^{J-1} \|\boldsymbol{\beta}_s - \tilde{\boldsymbol{\beta}}\|_2^2, \qquad R_{\beta 2} = \sum_{s=3}^{J} (\boldsymbol{\beta}_s - \tilde{\boldsymbol{\beta}})'(\boldsymbol{\beta}_{s-1} - \tilde{\boldsymbol{\beta}}).
$$

We use the following properties to derive the conditional posterior of ρ. The inverse and determinant of $\Sigma_\beta(\rho)$ are

$$
\Sigma_\beta(\rho)^{-1} = \begin{pmatrix}
I & -\rho I & & & & \\
-\rho I & (1+\rho^2)I & -\rho I & & & \\
 & -\rho I & (1+\rho^2)I & -\rho I & & \\
 & & -\rho I & \ddots & \ddots & \\
 & & & \ddots & (1+\rho^2)I & -\rho I \\
 & & & & -\rho I & I
\end{pmatrix},
$$

and $\det[\Sigma_\beta(\rho)^{-1}] = (1-\rho^2)^Q$, respectively. To update $\tilde{\boldsymbol{\beta}}$ and σ_β^2, we use regular Gibbs steps. To update ρ, given $\{\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}}, \sigma_\beta^2\}$ we can easily evaluate its posterior on the $[0,1]$ grid, and sample from it.

Similarly, there are three hyperparameters related to $\boldsymbol{\beta}_0$: $\tilde{\boldsymbol{\beta}}_0$, $\sigma_{\beta_0}^2$ and ρ_0. Their conditional posteriors have exactly the same form as those for $\tilde{\boldsymbol{\beta}}$, σ_β^2 and ρ.

Hyperparameters related to $\boldsymbol{b}$ and $\boldsymbol{b}_0$. There are three hyperparameters related to $\boldsymbol{b}$: $\tilde{b}$, σ_b^2 and γ_b. The conditional posteriors are as follows.

1. Conditional posterior of $\tilde{b}$:

$$
\begin{aligned}
\tilde{b} \mid \boldsymbol{b}, \sigma_b^2, \gamma_b &\sim N(\tilde{b}^*, \delta_{\tilde{b}}^{*2}), \quad \text{where} \\
\delta_{\tilde{b}}^{*2} &= \left[\frac{1}{\delta_{\tilde{b}}^2} + \frac{1}{\sigma_b^2}\mathbf{1}^T\mathcal{N}_b^{-1}(I - \gamma_b W_b)\mathbf{1}\right]^{-1}, \\
\tilde{b}^* &= \delta_{\tilde{b}}^{*2}\left[\frac{1}{\sigma_b^2}\mathbf{1}^T\mathcal{N}_b^{-1}(I - \gamma_b W_b)\boldsymbol{b}\right].
\end{aligned}
$$

2. Conditional posterior of σ_b^2:

$$\sigma_b^2 \mid \boldsymbol{b}, \tilde{b}, \gamma_b \sim \text{IG}\left[\lambda_1^b + \frac{\dim(\boldsymbol{b})}{2},\ \lambda_2^b + \frac{1}{2}(\boldsymbol{b} - \mathbf{1}\tilde{b})'\mathcal{N}_b^{-1}(I - \gamma_b W_b)(\boldsymbol{b} - \mathbf{1}\tilde{b})\right].$$

3. Conditional posterior of γ_b:

$$p(\gamma_b \mid \boldsymbol{b}, \tilde{b}, \sigma_b^2) \propto \det(I - \gamma_b W_b)^{1/2} \cdot \exp\left[\gamma_b \cdot \frac{1}{2\sigma_b^2}(\boldsymbol{b} - \mathbf{1}\tilde{b})'\mathcal{N}_b^{-1}W_b(\boldsymbol{b} - \mathbf{1}\tilde{b})\right].$$

To update $\tilde{b}$ and σ_b^2, we use regular Gibbs steps. To update γ_b, given $\{\boldsymbol{b}, \tilde{b}, \sigma_b^2\}$ we can easily evaluate its posterior on the $[0, 1]$ grid, and sample from it. To facilitate computation, we can calculate $\det(I - \gamma_b W_b)^{1/2}$ on the $[0, 1]$ grid, save the values and use it at each iteration.

Similarly, there are three hyperparameters related to $\boldsymbol{b}_0$: $\tilde{b}_0$, $\sigma_{b_0}^2$ and γ_{b_0}. Their conditional posteriors have exactly the same form as those for $\tilde{b}$, σ_b^2 and γ_b.

Hyperparameters related to $\boldsymbol{\psi}$. There are three hyperparameters related to $\boldsymbol{\psi}$: $\tilde{\psi}$, σ_ψ^2 and γ_ψ. The conditional posteriors are as follows.

1. Conditional posterior of $\tilde{\psi}$:

$$\begin{aligned}\tilde{\psi} \mid \boldsymbol{\psi}, \sigma_\psi^2, \gamma_\psi &\sim N(\tilde{\psi}^*, \delta_{\tilde{\psi}}^{*2}), \quad \text{where} \\ \delta_{\tilde{\psi}}^{*2} &= \left[\frac{1}{\delta_\psi^2} + \frac{1}{\sigma_\psi^2}\mathbf{1}'\mathcal{N}_\psi^{-1}(I - \gamma_\psi W_\psi)\mathbf{1}\right]^{-1}, \\ \tilde{\psi}^* &= \delta_{\tilde{\psi}}^{*2}\left[\frac{1}{\delta_\psi^2} \cdot 1 + \frac{1}{\sigma_\psi^2}\mathbf{1}'\mathcal{N}_\psi^{-1}(I - \gamma_\psi W_\psi)\boldsymbol{\psi}\right].\end{aligned}$$

2. Conditional posterior of σ_ψ^2:

$$\sigma_\psi^2 \mid \boldsymbol{\psi}, \tilde{\psi}, \gamma_\psi \sim \text{IG}\left[\lambda_1^\psi + \frac{\dim(\boldsymbol{\psi})}{2}, \lambda_2^\psi + \frac{1}{2}(\boldsymbol{\psi} - \mathbf{1}\tilde{\psi})'\mathcal{N}_\psi^{-1}(I - \gamma_\psi W_\psi)(\boldsymbol{\psi} - \mathbf{1}\tilde{\psi})\right].$$

3. Conditional posterior of γ_ψ:

$$p(\gamma_\psi \mid \boldsymbol{\psi}, \tilde{\psi}, \sigma_\psi^2) \propto \det(I - \gamma_\psi W_\psi)^{1/2} \cdot \exp\left[\gamma_\psi \cdot \frac{1}{2\sigma_\psi^2}(\boldsymbol{\psi} - \mathbf{1}\tilde{\psi})'\mathcal{N}_\psi^{-1}W_\psi(\boldsymbol{\psi} - \mathbf{1}\tilde{\psi})\right].$$

Hyperparameters related to $\boldsymbol{\phi}$. There is one hyperparameter related to $\boldsymbol{\phi}$: σ_ϕ^2. The conditional posterior is

$$\sigma_\phi^2 \mid \boldsymbol{\phi} \sim \text{IG}\left[\lambda_1^\phi + \frac{1}{2}\dim(\boldsymbol{\phi}),\ \lambda_2^\phi + \frac{1}{2}\boldsymbol{\phi}^T\boldsymbol{\phi}\right].$$

Update intermittent missing responses. The focus of our method is dealing with monotone missing data. Sometimes there are (typically few) intermittent missing responses, and we impute it under the partial ignorability assumption (Harel and Schafer, 2009). Suppose y_{ijs} is missing. Its conditional distribution is

$$p\left(y_{ijs} \mid y_{-ijs}, \boldsymbol{\pi}\right) \propto p\left(\boldsymbol{y}_{\text{vec0}}, \boldsymbol{y}_{\text{vec}} \mid \boldsymbol{\pi}\right),$$

We use a Metropolis-Hastings step to update y_{ijs}. We use a symmetric normal proposal distribution, $y_{ijs}^{\text{pro}} \sim N\left(y_{ijs}^{\text{cur}},\ 0.5 \times \text{sd}(\boldsymbol{y}_{\text{vec0}}, \boldsymbol{y}_{\text{vec}})\right)$.

A.4 G-computation Implementation Details

The steps for conducting the G-computation for our setting are summarized in Algorithm A.1.

Algorithm A.1 G-computation

for l in $1, \ldots, L$ **do**
  for m in $1, \ldots, M$ **do**
    1. Draw $\boldsymbol{V}^* = \boldsymbol{v}^* \sim p(\boldsymbol{v}^* \mid \boldsymbol{\eta}^{(l)})$
    2. Draw $S^* = s^* \sim p(s^* \mid \boldsymbol{v}^*, \boldsymbol{\varphi}^{(l)})$
    3. Draw $\bar{\boldsymbol{Y}}_s^* = \bar{\boldsymbol{y}}_s^* \sim p(\bar{\boldsymbol{y}}_s^* \mid s^*, \boldsymbol{v}^*, \boldsymbol{\pi}^{(l)})$
    4. Draw $\tilde{\boldsymbol{Y}}_s^* = \tilde{\boldsymbol{y}}_s^* \sim p(\tilde{\boldsymbol{y}}_s^* \mid \bar{\boldsymbol{y}}_s^*, s^*, \boldsymbol{v}^*, \boldsymbol{\omega}_E^{(l)})$
    5. Set $\boldsymbol{Y}^{*(m,l)} = (\bar{\boldsymbol{Y}}_s^*, \tilde{\boldsymbol{Y}}_s^*)$
  end for
end for
return $(1/ML) \cdot \sum_{m,l} t\left[\boldsymbol{Y}^{*(m,l)}\right]$

Next, we describe in detail how to draw the pseudo responses using Gaussian process prediction rule, i.e. steps 3 and 4 in Algorithm A.1. We generally use a superscript $*$ to denote the pseudo subject and response.

Observed response. To draw a vector of pseudo observed responses $\bar{\boldsymbol{Y}}_s^* = \bar{\boldsymbol{y}}_s^*$ from $p(\bar{\boldsymbol{y}}_s^* \mid s^*, \boldsymbol{v}^*, \boldsymbol{\pi})$, we do the following.

1. Draw y_1^* from $p(y_1^* \mid s^*, \boldsymbol{v}^*, \boldsymbol{\pi})$. Consider the joint distribution of $a_{1s^*}^* = a_0(\boldsymbol{v}^*, s^*)$ and the training data points $\boldsymbol{y}_{\text{vec0}}$,

$$\begin{pmatrix} \boldsymbol{y}_{\text{vec0}} \\ a_{1s^*}^* \end{pmatrix} \sim N\left[\begin{pmatrix} X_{\theta_0}\boldsymbol{\theta}_0 \\ \mu_{1s^*}^* \end{pmatrix}, \begin{pmatrix} \Sigma_{y0} + C_0 & C_{1s^*}^* \\ C_{1s^*}^{*T} & C_{1s^*}^{**} \end{pmatrix}\right],$$

where

$$\begin{aligned} \mu_{1s^*}^* &= \mu_0(\boldsymbol{v}^*, s^*), \\ C_{1s^*}^* &= C_0(V_{\text{vec0}}, \boldsymbol{s}_{\text{vec0}}; \boldsymbol{v}^*, s^*), \\ C_{1s^*}^{**} &= C_0(\boldsymbol{v}^*, s^*; \boldsymbol{v}^*, s^*). \end{aligned}$$

The predictive distribution for $a_{1s^*}^*$ is thus

$$a_{1s^*}^* \mid \boldsymbol{y}_{\text{vec0}}, \boldsymbol{\pi} \sim N\big[\mu_{1s^*}^* + C_{1s^*}^{*T}(\Sigma_{y0} + C_0)^{-1}(\boldsymbol{y}_{\text{vec0}} - X_{\theta_0}\boldsymbol{\theta}_0), \\ C_{1s^*}^{**} - C_{1s^*}^{*T}(\Sigma_{y0} + C_0)^{-1}C_{1s^*}^*\big],$$

and we can draw

$$y_1^* \mid a_{1s^*}^* \sim N(a_{1s^*}^*, \sigma_{1s^*}^2).$$

Integrating out $a_{1s^*}^*$, the above two steps simplify to

$$\begin{aligned} y_1^* \mid \boldsymbol{y}_{\text{vec0}}, \boldsymbol{\pi} &\sim N(\check{\mu}_{1s^*}^*, \check{\sigma}_{1s^*}^2), \quad \text{where} \\ \check{\mu}_{1s^*}^* &= \mu_{1s^*}^* + C_{1s^*}^{*T}(\Sigma_{y0} + C_0)^{-1}(\boldsymbol{y}_{\text{vec0}} - X_{\theta_0}\boldsymbol{\theta}_0), \\ \check{\sigma}_{1s^*}^2 &= C_{1s^*}^{**} - C_{1s^*}^{*T}(\Sigma_{y0} + C_0)^{-1}C_{1s^*}^* + \sigma_{1s^*}^2. \end{aligned}$$

2. Draw y_j^* from $p(y_j^* \mid \bar{\boldsymbol{y}}_{j-1}^*, s^*, \boldsymbol{v}^*, \boldsymbol{\pi})$, $(1 < j \le s^*)$. The joint distribution of $a_{js^*}^* = a(y_{j-1}^*, \boldsymbol{v}^*, j, s^*) + \bar{\boldsymbol{y}}_{j-2}^{*T}\boldsymbol{\phi}_{js^*}$ and the training data points $\boldsymbol{y}_{\text{vec}}$ is

$$\begin{pmatrix} \boldsymbol{y}_{\text{vec}} \\ a_{js^*}^* \end{pmatrix} \sim N\left[\begin{pmatrix} X_{\theta}\boldsymbol{\theta} \\ \mu_{js^*}^* + \bar{\boldsymbol{y}}_{j-2}^{*T}\boldsymbol{\phi}_{js^*} \end{pmatrix}, \begin{pmatrix} \Sigma_y + C & C_{js^*}^* \\ C_{js^*}^{*T} & C_{js^*}^{**} \end{pmatrix}\right],$$

where

$$\begin{aligned} \mu_{js^*}^* &= \mu(y_{j-1}^*, \boldsymbol{v}^*, j, s^*), \\ C_{js^*}^* &= C(\boldsymbol{y}_{\text{lag}}, V_{\text{vec}}, \boldsymbol{j}_{\text{vec}}, \boldsymbol{s}_{\text{vec}}; y_{j-1}^*, \boldsymbol{v}^*, j, s^*), \\ C_{js^*}^{**} &= C(y_{j-1}^*, \boldsymbol{v}^*, j, s^*; y_{j-1}^*, \boldsymbol{v}^*, j, s^*). \end{aligned}$$

The predictive distribution for $a^*_{js^*}$ is thus

$$a^*_{js^*} \mid \bar{\boldsymbol{y}}^*_{j-1}, \boldsymbol{y}_{\text{vec}}, \boldsymbol{\pi} \sim N\big[\mu^*_{js^*} + \bar{\boldsymbol{y}}^{*T}_{j-2}\boldsymbol{\phi}_{js^*} + C^{*T}_{js^*}(\Sigma_y + C)^{-1}(\boldsymbol{y}_{\text{vec}} - X_\theta\boldsymbol{\theta}), \\ C^{**}_{js^*} - C^{*T}_{js^*}(\Sigma_y + C)^{-1}C^*_{js^*}\big],$$

and we can draw

$$y^*_j \mid a^*_{js^*} \sim N(a^*_{js^*}, \sigma^2_{js^*}).$$

Integrating out $a^*_{js^*}$, the above two steps simplify to

$$\begin{aligned} y^*_j \mid \bar{\boldsymbol{y}}^*_{j-1}, \boldsymbol{y}_{\text{vec}}, \boldsymbol{\pi} &\sim N(\breve{\mu}^*_{js^*}, \breve{\sigma}^2_{js^*}), \quad \text{where} \\ \breve{\mu}^*_{js^*} &= \mu^*_{js^*} + \bar{\boldsymbol{y}}^{*T}_{j-2}\boldsymbol{\phi}_{js^*} + C^{*T}_{js^*}(\Sigma_y + C)^{-1}(\boldsymbol{y}_{\text{vec}} - X_\theta\boldsymbol{\theta}), \\ \breve{\sigma}^2_{js^*} &= C^{**}_{js^*} - C^{*T}_{js^*}(\Sigma_y + C)^{-1}C^*_{js^*} + \sigma^2_{js^*}. \end{aligned}$$

Missing response. To draw a pseudo response $Y^*_j = y^*_j$ from the extrapolation distribution $p(y^*_j \mid \bar{\boldsymbol{y}}^*_{j-1}, s^*, \boldsymbol{v}^*, \boldsymbol{\omega})$ $(j > s^*)$, do the following.

(I) Under MAR,

$$\begin{aligned} p(y^*_j \mid \bar{\boldsymbol{y}}^*_{j-1}, \boldsymbol{v}^*, S = s^*, \boldsymbol{\omega}) &= p(y^*_j \mid \bar{\boldsymbol{y}}^*_{j-1}, \boldsymbol{v}^*, S \geq j, \boldsymbol{\omega}) \\ &= \sum_{k=j}^{J} \alpha_{kj} p(y^*_j \mid \bar{\boldsymbol{y}}^*_{j-1}, \boldsymbol{v}^*, S = k, \boldsymbol{\omega}), \end{aligned} \tag{1}$$

where

$$\begin{aligned} \alpha_{kj} &= \alpha_{kj}(\bar{\boldsymbol{y}}^*_{j-1}, \boldsymbol{v}^*) = p(S = k \mid \bar{\boldsymbol{y}}^*_{j-1}, \boldsymbol{v}^*, S \geq j) \\ &= \frac{p(\bar{\boldsymbol{y}}^*_{j-1} \mid \boldsymbol{v}^*, S = k)\; p(S = k \mid \boldsymbol{v}^*, S \geq j)}{\sum_{k=j}^{J} p(\bar{\boldsymbol{y}}^*_{j-1} \mid \boldsymbol{v}^*, S = k)\; p(S = k \mid \boldsymbol{v}^*, S \geq j)}, \quad k = j, \ldots, J \end{aligned}$$

The above expression can be calculated by

$$p(\bar{\boldsymbol{y}}^*_{j-1} \mid \boldsymbol{v}^*, S = k) = p_k(y^*_1 \mid \boldsymbol{v}^*) \cdot \prod_{l=2}^{j-1} p_k(y^*_l \mid \bar{\boldsymbol{y}}^*_{l-1}, \boldsymbol{v}^*)$$

where

$$\begin{aligned} p_k(y^*_1 \mid \boldsymbol{v}^*) &= p_N\big(y^*_1 \mid \breve{\mu}^*_{1k}, \breve{\sigma}^2_{1k}\big), \\ p_k(y^*_l \mid \bar{\boldsymbol{y}}^*_{l-1}, \boldsymbol{v}^*) &= p_N\big(y^*_l \mid \breve{\mu}^*_{lk}, \breve{\sigma}^2_{lk}\big), \end{aligned}$$

and

$$
\begin{aligned}
&p(S = k \mid \boldsymbol{v}^*, S \geq j) \\
&= p(S = k \mid \boldsymbol{v}^*, S \geq k) \cdot \prod_{l=j}^{k-1} p(S \geq l+1 \mid \boldsymbol{v}^*, S \geq l) \\
&= p(S = k \mid \boldsymbol{v}^*, S \geq k) \cdot \prod_{l=j}^{k-1} \left[1 - p(S = l \mid \boldsymbol{v}^*, S \geq l)\right].
\end{aligned}
$$

To sample from (1), after calculating $(\alpha_{jj}, \ldots, \alpha_{Jj})$, we can draw $K = k$ with probability α_{kj}, and sample $Y_j^* = y_j^*$ from $p_k(y_j^* \mid \bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*, \boldsymbol{\omega})$.

(II) Under NFD.

(II-1) For $j = s^* + 1$,

$$
[Y_j \mid \bar{\boldsymbol{Y}}_{j-1}, S = j-1, \boldsymbol{V}, \boldsymbol{\omega}] \overset{\mathrm{d}}{=} \left[Y_j + \tau_j \mid \bar{\boldsymbol{Y}}_{j-1}, S \geq j, \boldsymbol{V}, \boldsymbol{\omega}\right].
$$

We first sample from $p_{\geq j}(y_j^* \mid \bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*, \boldsymbol{\omega})$. Then, we apply the location shift (9) with

$$
\tau_j(\bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*) = \tilde{\tau} \cdot \Delta_j(\bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*),
$$

where $\Delta_j(\bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*)$ is chosen to be the standard deviation of $p_{j-1}(y_j^* \mid \bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*, \boldsymbol{\omega})$ under MAR, i.e. $p_{\geq j}(y_j^* \mid \bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*, \boldsymbol{\omega})$. We have

$$
p_{\geq j}(y_j^* \mid \bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*, \boldsymbol{\omega}) = \sum_{k=j}^{J} \alpha_{kj} N(\check{\mu}_{jk}^*, \check{\sigma}_{jk}^2).
$$

The standard deviation of this normal mixture is given by

$$
\Delta_j(\bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*) = \sqrt{\sum_{k=j}^{J} \alpha_{kj} \check{\sigma}_{jk}^2 + \sum_{k=j}^{J} \alpha_{kj} \check{\mu}_{jk}^{*2} - \left(\sum_{k=j}^{J} \alpha_{kj} \check{\mu}_{jk}^*\right)^2}.
$$

(II-2) For $j > s^* + 1$,

$$
\begin{aligned}
p(y_j^* \mid \bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*, S = s^*, \boldsymbol{\omega}) &= p(y_j^* \mid \bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*, S \geq j-1, \boldsymbol{\omega}) \\
&= \sum_{k=j-1}^{J} \alpha_{k,j-1} p(y_j^* \mid \bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*, S = k, \boldsymbol{\omega}),
\end{aligned} \tag{2}
$$

where

$$\begin{aligned}\alpha_{k,j-1} &= \alpha_{k,j-1}(\bar{\boldsymbol{y}}^*_{j-1}, \boldsymbol{v}^*) = p(S = k \mid \bar{\boldsymbol{y}}^*_{j-1}, \boldsymbol{v}^*, S \geq j-1) \\ &= \frac{p(\bar{\boldsymbol{y}}^*_{j-1} \mid \boldsymbol{v}^*, S = k)\, p(S = k \mid \boldsymbol{v}^*, S \geq j-1)}{\sum_{k=j-1}^{J} p(\bar{\boldsymbol{y}}^*_{j-1} \mid \boldsymbol{v}^*, S = k)\, p(S = k \mid \boldsymbol{v}^*, S \geq j-1)}, \quad k = j-1, \ldots, J.\end{aligned}$$

To sample from (2), after calculating $(\alpha_{j-1,j-1}, \ldots, \alpha_{J,j-1})$, we can draw $K = k$ with probability $\alpha_{k,j-1}$.

(II-2a) If $k = j - 1$, draw again $K' = k'$ with probability $\alpha_{k',j-1}/(1 - \alpha_{j-1,j-1})$ for $k' = j, \ldots, J$. Then, sample $Y^*_j = y^*_j$ from $p_{k'}(y^*_j \mid \bar{\boldsymbol{y}}^*_{j-1}, \boldsymbol{v}^*, \boldsymbol{\omega})$, and apply the location shift (9).

(II-2b) If $k \in \{j, \ldots, J\}$, sample $Y^*_j = y^*_j$ from $p_k(y^*_j \mid \bar{\boldsymbol{y}}^*_{j-1}, \boldsymbol{v}^*, \boldsymbol{\omega})$.

The steps for sampling the pseudo response $\boldsymbol{Y}^* = \boldsymbol{y}^*$ from $p(\boldsymbol{y}^* \mid s^*, \boldsymbol{v}^*, \boldsymbol{\omega})$ are summarized in Algorithm A.2.

Algorithm A.2 Draw $\boldsymbol{Y}^* = \boldsymbol{y}^*$ from $p(\boldsymbol{y}^* \mid s^*, \boldsymbol{v}^*, \boldsymbol{\omega})$

```
Draw $Y_1^* = y_1^* \sim N(\check{\mu}_{1s^*}^*, \check{\sigma}_{1s^*}^2)$
**for** $j$ in $2, \ldots, s^*$ **do**
    Draw $Y_j^* = y_j^* \sim N(\check{\mu}_{js^*}^*, \check{\sigma}_{js^*}^2)$
**end for**
**if** MAR **then**
    **for** $j$ in $s^* + 1, \ldots, J$ **do**
        Calculate $\boldsymbol{\alpha}_j(\bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*) = (\alpha_{jj}, \ldots, \alpha_{Jj})$
        Draw $K = k \sim$ Categorical$[(j, \ldots, J); \boldsymbol{\alpha}_j]$
        Draw $y_j^* \sim N(\check{\mu}_{jk}^*, \check{\sigma}_{jk}^2)$
    **end for**
**else if** NFD **then**
    Set $j = s^* + 1$
    Calculate $\boldsymbol{\alpha}_j(\bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*) = (\alpha_{jj}, \ldots, \alpha_{Jj})$
    Draw $K = k \sim$ Categorical$[(j, \ldots, J); \boldsymbol{\alpha}_j]$
    Calculate $\tau_j(\bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*) = \tilde{\tau} \cdot \Delta_j(\bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*)$
    Draw $y_j^* \sim N(\check{\mu}_{jk}^* + \tau_j, \check{\sigma}_{jk}^2)$
    **for** $j$ in $s^* + 2, \ldots, J$ **do**
        Calculate $\boldsymbol{\alpha}_{j-1}(\bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*) = (\alpha_{j-1,j-1}, \ldots, \alpha_{J,j-1})$
        Draw $K = k \sim$ Categorical$[(j-1, \ldots, J); \boldsymbol{\alpha}_{j-1}]$
        **if** $k = j - 1$ **then**
            Calculate $\boldsymbol{\alpha}_j' = (\alpha_{j,j-1}, \ldots, \alpha_{J,j-1})/(1 - \alpha_{j-1,j-1})$
            Draw $K' = k' \sim$ Categorical$[(j, \ldots, J); \boldsymbol{\alpha}_j']$
            Calculate $\tau_j(\bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*) = \tilde{\tau} \cdot \Delta_j(\bar{\boldsymbol{y}}_{j-1}^*, \boldsymbol{v}^*)$
            Draw $y_j^* \sim N(\check{\mu}_{jk'}^* + \tau_j, \check{\sigma}_{jk'}^2)$
        **else**
            Draw $y_j^* \sim N(\check{\mu}_{jk}^*, \check{\sigma}_{jk}^2)$.
        **end if**
    **end for**
**end if**
```

A.5 Simulation Details

Prior and hyper-prior parameters. We set the prior and hyper-prior parameters at standard noninformative choices. We generally use $N(0, (\sqrt{30})^2)$ and IG(1, 1) as noninformative normal and inverse-gamma priors, respectively. Since we have standardized the covariates and responses, it is thought unlikely that the regression coefficients would have a scale greater than $\sqrt{30} \approx 5.5$. Table A.2 shows the exact values. We set $\kappa_0^2 \sim \text{IG}(10, 1)$ and $\kappa^2 \sim \text{IG}(10, 1)$ to shrink the semiparametric model towards a simple linear regression model. We also set $\lambda_1^\phi = 30$ and $\lambda_2^\phi = 1$ to shrink $\boldsymbol{\phi}_{js}$ towards 0. Since higher order lag responses are highly correlated with lag-1 responses, shrinking $\boldsymbol{\phi}_{js}$ towards 0 helps us correctly identify the effect of lag-1 responses.

$\lambda_1^{\kappa_0}$	10	$\lambda_1^{\lambda_\sigma}$	1	$\delta_{\beta_0}^2$	30	$\delta_{b_0}^2$	30	δ_ψ^2	30
$\lambda_2^{\kappa_0}$	1	$\lambda_2^{\lambda_\sigma}$	1	$\lambda_1^{\beta_0}$	1	$\lambda_1^{b_0}$	1	λ_1^ψ	1
λ_1^κ	10	$\lambda_1^{\nu_\sigma}$	1	$\lambda_2^{\beta_0}$	1	$\lambda_2^{b_0}$	1	λ_2^ψ	1
λ_2^κ	1	$\lambda_2^{\nu_\sigma}$	1	δ_β^2	30	δ_b^2	30	λ_1^ϕ	30
				λ_1^β	1	λ_1^b	1	λ_2^ϕ	1
				λ_2^β	1	λ_2^b	1	δ_η	0.1

Table A.2: Choices of hyperparameters in the observed data model. These hyperparameters are used for simulations and real data analysis.

Scenario 1. The covariance matrix for generating $\boldsymbol{V}$ is

$$\Sigma_{vv} = \begin{pmatrix} 1.0 & 0.52 & -0.22 & 0.07 \\ 0.52 & 1.0 & -0.23 & -0.02 \\ -0.22 & -0.23 & 1.0 & 0.45 \\ 0.07 & -0.02 & 0.45 & 1.0 \end{pmatrix},$$

which is the correlation matrix of the subjects' numerical auxiliary covariates from the schizophrenia clinical trial dataset.

The parameters for generating S are

$$\boldsymbol{\zeta} = (-4.346, -2.193, -2.606, -2.678)^T,$$

where ζ_s corresponds to the $(s-1)$-th element $(s = 2, \ldots, 5)$, and

$$\boldsymbol{\xi} = \begin{pmatrix} -1.057 & 0.328 & -0.121 & 0.273 \\ -0.826 & 0.128 & 0.525 & -0.781 \\ -0.487 & 0.479 & 0.534 & -0.480 \\ 0.642 & 0.129 & 0.448 & 0.122 \end{pmatrix},$$

where $\boldsymbol{\xi}_s$ corresponds to the $(s-1)$-th column $(s = 2, \ldots, 5)$. These parameters come from fitting the sequential logistic regression model to the test drug arm of the schizophrenia clinical trial dataset and taking posterior mean of each parameter.

The parameters for generating $\bar{\boldsymbol{Y}}_S$ are

$$\{\sigma^2_{js}\} = \begin{pmatrix} 0.232 & 0.221 & & & & \\ 0.365 & 0.243 & 0.196 & & & \\ 0.403 & 0.222 & 0.228 & 0.941 & & \\ 0.438 & 0.228 & 0.225 & 0.213 & 0.284 & \\ 0.335 & 0.192 & 0.265 & 0.140 & 0.167 & 0.160 \end{pmatrix},$$

where σ^2_{js} corresponds to the element in the $(s-1)$-th row and j-th column;

$$(\boldsymbol{b}_0, \boldsymbol{b}) = \begin{pmatrix} 0.069 & -0.191 & & & & \\ 0.507 & 0.219 & 0.302 & & & \\ 0.393 & 0.060 & -0.022 & 0.399 & & \\ 0.798 & 0.048 & -0.051 & 0.051 & 0.362 & \\ 0.384 & -0.107 & -0.250 & -0.367 & -0.250 & -0.321 \end{pmatrix},$$

where b_{js} corresponds to the element in the $(s-1)$-th row and j-th column;

$$\boldsymbol{\beta}_0 = \begin{pmatrix} -0.046 & 0.174 & -0.005 & 0.024 & 0.230 \\ -0.200 & -0.099 & -0.124 & -0.451 & -0.163 \\ -0.315 & -0.191 & -0.104 & 0.140 & 0.032 \\ -0.053 & 0.065 & 0.003 & -0.044 & -0.092 \end{pmatrix},$$

where $\boldsymbol{\beta}_{0s}$ corresponds to the $(s-1)$-th column;

$$\boldsymbol{\beta} = \begin{pmatrix} -0.080 & -0.117 & -0.118 & 0.010 & 0.066 \\ -0.044 & -0.113 & 0.023 & -0.035 & -0.030 \\ -0.109 & -0.020 & -0.014 & -0.022 & 0.056 \\ 0.170 & 0.127 & 0.166 & -0.060 & 0.002 \end{pmatrix},$$

where $\boldsymbol{\beta}_s$ corresponds to the $(s-1)$-th column;

$$(\boldsymbol{\phi}_1) = \begin{pmatrix} 1.078 & & & & \\ 1.088 & 0.938 & & & \\ 0.830 & 0.893 & 0.830 & & \\ 0.637 & 0.877 & 0.907 & 1.065 & \\ 0.881 & 0.871 & 0.842 & 0.929 & 0.943 \end{pmatrix},$$

where ϕ_{1js} corresponds to the element in the $(s-1)$-th row and $(j-1)$-th column;

$$(\boldsymbol{\phi}_2) = \begin{pmatrix} -0.045 & & & \\ 0.040 & -0.025 & & \\ 0.021 & 0.022 & 0.035 & \\ 0.089 & 0.129 & 0.019 & -0.020 \end{pmatrix},$$

where ϕ_{2js} corresponds to the element in the $(s-2)$-th row and $(j-2)$-th column; and

$$(\boldsymbol{\phi}_3) = \begin{pmatrix} 0.011 & & \\ 0.037 & \begin{pmatrix} 0.074 \\ 0.037 \end{pmatrix} & \\ 0.078 & \begin{pmatrix} -0.027 \\ -0.086 \end{pmatrix} & \begin{pmatrix} 0.021 \\ 0.010 \\ -0.009 \end{pmatrix} \end{pmatrix},$$

where $\boldsymbol{\phi}_{3js}$ corresponds to the element in the $(s-3)$-th row and $(j-3)$-th column. These parameters come from fitting the linear regression model to the test drug arm of the schizophrenia clinical trial dataset and taking posterior mean of each parameter.

Scenario 2. We use the same choices of $\boldsymbol{b}_0$, $\boldsymbol{b}$, $\boldsymbol{\phi}_1$, $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi}_3$ as in Scenario 1. We set

$$\Sigma_{vv} = \begin{pmatrix} 1.0 & 0.52 & -0.22 \\ 0.52 & 1.0 & -0.23 \\ -0.22 & -0.23 & 1.0 \end{pmatrix},$$

i.e. the upper left 3×3 submatrix of Σ_{vv} in Scenario 1. We change $\{\sigma^2_{js}\}$, $\boldsymbol{\zeta}$, $\boldsymbol{\xi}$, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}$ to

$$\{\sigma^2_{js}\} = \begin{pmatrix} 0.155 & 0.101 & & & & \\ 0.217 & 0.133 & 0.112 & & & \\ 0.099 & 0.082 & 0.101 & 0.115 & & \\ 0.141 & 0.127 & 0.169 & 0.132 & 0.107 & \\ 0.106 & 0.119 & 0.095 & 0.081 & 0.266 & 0.174 \end{pmatrix},$$

where σ^2_{js} corresponds to the element in the $(s-1)$-th row and j-th column;

$$\boldsymbol{\zeta} = (-3.0, -2.1, -1.6, -1.3)^T,$$

where ζ_s corresponds to the $(s-1)$-th element $(s = 2, \ldots, 5)$, and

$$\boldsymbol{\xi} = \begin{pmatrix} -1.057 & 0.328 & -0.121 & 0.273 \\ -0.826 & 0.128 & 0.525 & -0.781 \\ -0.487 & 0.479 & 0.534 & -0.480 \\ -0.528 & 0.164 & -0.061 & 0.136 \\ -0.413 & 0.064 & 0.263 & -0.390 \\ -0.244 & 0.239 & 0.267 & -0.240 \\ 0.321 & 0.064 & 0.224 & 0.061 \\ -0.528 & 0.164 & -0.061 & 0.136 \\ -0.413 & 0.064 & 0.263 & -0.390 \end{pmatrix},$$

where $\boldsymbol{\xi}_s$ corresponds to the $(s-1)$-th column $(s = 2, \ldots, 5)$.

$$\boldsymbol{\beta}_0 = \begin{pmatrix} -0.530 & -0.508 & -0.561 & -0.507 & -0.525 \\ -0.366 & -0.377 & -0.421 & -0.417 & -0.386 \\ 0.351 & 0.309 & 0.323 & 0.318 & 0.346 \\ 0.283 & 0.291 & 0.282 & 0.277 & 0.275 \\ -0.316 & -0.321 & -0.319 & -0.319 & -0.316 \\ 0.288 & 0.285 & 0.293 & 0.288 & 0.289 \\ 0.033 & 0.030 & 0.033 & 0.020 & 0.033 \\ -0.083 & -0.087 & -0.094 & -0.082 & -0.092 \\ 0.124 & 0.125 & 0.115 & 0.120 & 0.116 \end{pmatrix},$$

where $\boldsymbol{\beta}_{0s}$ corresponds to the $(s-1)$-th column;

$$\boldsymbol{\beta} = \begin{pmatrix} -0.395 & -0.387 & -0.427 & -0.434 & -0.443 \\ 0.320 & 0.337 & 0.339 & 0.317 & 0.338 \\ 0.331 & 0.349 & 0.400 & 0.385 & 0.356 \\ 0.317 & 0.315 & 0.309 & 0.313 & 0.310 \\ 0.354 & 0.355 & 0.342 & 0.354 & 0.349 \\ -0.301 & -0.299 & -0.303 & -0.306 & -0.306 \\ -0.082 & -0.082 & -0.073 & -0.068 & -0.079 \\ -0.077 & -0.088 & -0.082 & -0.085 & -0.081 \\ -0.129 & -0.126 & -0.130 & -0.133 & -0.128 \\ 0.025 & 0.022 & 0.024 & 0.022 & 0.023 \\ -0.021 & -0.020 & -0.020 & -0.022 & -0.024 \\ -0.015 & -0.015 & -0.014 & -0.015 & -0.019 \\ 0.004 & 0.003 & 0.004 & 0.003 & 0.002 \end{pmatrix},$$

where $\boldsymbol{\beta}_s$ corresponds to the $(s-1)$-th column.

Scenario 3. The parameter for generating K is

$$\boldsymbol{\pi} = (0.119, 0.579, 0.001, 0.115, 0.186),$$

which is taken from Linero and Daniels (2015) by fitting the mixture model to the active control arm of the schizophrenia clinical trial dataset.

The parameters for the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{V}$ are specified and generated as follows. Within mixture component k, the joint distribution of $\boldsymbol{Y}$ and $\boldsymbol{V}$ is

$$\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{V} \end{pmatrix} \mid K = k \sim N\left[\boldsymbol{\mu}^{(k)}, \Omega^{(k)}\right],$$

where

$$\begin{aligned}
\boldsymbol{\mu}^{(k)} &= \begin{pmatrix} \boldsymbol{\mu}_y^{(k)} \\ \boldsymbol{0} \end{pmatrix}, \\
\Omega^{(k)} &\sim \mathcal{W}^{-1}\left((\nu - J - Q - 1)\Omega_0^{(k)}, \nu\right), \\
\Omega_0^{(k)} &= \begin{pmatrix} \Sigma_{yy}^{(k)} & \Sigma_{yv}^{(k)} \\ \Sigma_{vy}^{(k)} & \Sigma_{vv} \end{pmatrix}.
\end{aligned}$$

Here $\boldsymbol{\mu}_y^{(k)}$ and $\Omega_0^{(k)}$ correspond to a linear model of $(\boldsymbol{Y} \mid \boldsymbol{V})$, where

$$\begin{aligned}
\boldsymbol{V} \mid K = k &\sim N(\boldsymbol{0}, \Sigma_{vv}), \\
Y_1 \mid \boldsymbol{V}, K = k &\sim N\left(b_1^{(k)} + \boldsymbol{V}^T\boldsymbol{\beta}_0^{(k)},\ \sigma_1^{2(k)}\right), \\
Y_j \mid \bar{\boldsymbol{Y}}_{j-1}, \boldsymbol{V}, K = k &\sim N\left(b_j^{(k)} + \boldsymbol{V}^T\boldsymbol{\beta}^{(k)} + \phi_j^{(k)} Y_{j-1},\ \sigma_j^{2(k)}\right), \quad j = 2, \ldots, J.
\end{aligned}$$

Let $\boldsymbol{b}^{(k)} = (b_1^{(k)}, \ldots, b_J^{(k)})^T$, $B^{(k)} = (\boldsymbol{\beta}_0^{(k)}, \boldsymbol{\beta}^{(k)}, \ldots, \boldsymbol{\beta}^{(k)})$, $\Sigma_0^{(k)} = \mathrm{diag}(\sigma_1^{2(k)}, \ldots, \sigma_J^{2(k)})$,

$$\Phi^{(k)} = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 \\ \phi_2^{(k)} & 0 & 0 & \cdots & 0 \\ 0 & \phi_3^{(k)} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \phi_J^{(k)} & 0 \end{pmatrix},$$

and $\dot{\Phi}^{(k)} = \left(I - \Phi^{(k)}\right)^{-1}$. We have

$$\begin{aligned}
\boldsymbol{\mu}_y^{(k)} &= \dot{\Phi}^{(k)}\boldsymbol{b}^{(k)}, \\
\Sigma_{yy}^{(k)} &= \dot{\Phi}^{(k)} B^{(k)T} \Sigma_{vv} B^{(k)} \dot{\Phi}^{(k)T} + \dot{\Phi}^{(k)} \Sigma_0^{(k)} \dot{\Phi}^{(k)T}, \\
\Sigma_{yv}^{(k)} &= \dot{\Phi}^{(k)} B^{(k)T} \Sigma_{vv}.
\end{aligned}$$

We use the same Σ_{vv} as in Scenario 2. The parameters $\{\boldsymbol{\mu}_y^{(k)}\}$ and $\Sigma_0^{(k)}$ are taken from Linero and Daniels (2015) (after standardization), which are generated by fitting the model to the active control arm of the schizophrenia clinical trial dataset. In particular,

$$\{\boldsymbol{\mu}_y^{(k)}\} = \begin{pmatrix} 0.715 & 0.559 & -0.649 & -0.085 & 0.677 \\ 0.581 & 0.406 & -1.368 & -0.207 & 0.799 \\ 0.329 & 0.175 & -1.404 & -0.851 & 0.944 \\ 0.319 & -0.217 & -1.650 & -1.181 & 1.276 \\ 0.889 & -0.473 & -1.765 & -1.363 & 0.483 \\ -0.664 & -0.593 & -3.195 & -1.562 & 1.081 \end{pmatrix},$$

where $\boldsymbol{\mu}_y^{(k)}$ corresponds to the k-th column. Then, we add the effects of auxiliary covariates by randomly generating $B^{(k)}$ and $\Phi^{(k)}$ (values not shown). Based on $B^{(k)}$, $\Phi^{(k)}$, Σ_{vv} and $\Sigma_0^{(k)}$ we calculate $\Omega_0^{(k)}$. Finally, we generate $\Omega^{(k)} \sim \mathcal{W}^{-1}\left((\nu - J - Q - 1)\Omega_0^{(k)}, \nu\right)$ and get

$$\Omega^{(1)} = \left(\begin{array}{cccccc|ccc} 0.9 & 1.3 & 1.7 & 1.9 & 2.3 & 2.6 & -1.0 & -0.4 & 0.4 \\ 1.3 & 2.2 & 2.9 & 3.4 & 4.2 & 4.9 & -1.6 & -0.4 & 0.9 \\ 1.7 & 2.9 & 4.1 & 4.8 & 5.9 & 7.0 & -2.1 & -0.4 & 1.4 \\ 1.9 & 3.4 & 4.8 & 5.8 & 7.1 & 8.3 & -2.4 & -0.3 & 1.7 \\ 2.3 & 4.2 & 5.9 & 7.1 & 8.8 & 10.4 & -3.0 & -0.4 & 2.2 \\ 2.6 & 4.9 & 7.0 & 8.3 & 10.4 & 12.2 & -3.5 & -0.4 & 2.6 \\ \hline -1.0 & -1.6 & -2.1 & -2.4 & -3.0 & -3.5 & 1.7 & 0.5 & -0.2 \\ -0.4 & -0.4 & -0.4 & -0.3 & -0.4 & -0.4 & 0.5 & 0.7 & -0.1 \\ 0.4 & 0.9 & 1.4 & 1.7 & 2.2 & 2.6 & -0.2 & -0.1 & 1.2 \end{array}\right),$$

$$\Omega^{(2)} = \left(\begin{array}{rrrrrr|rrr}
0.2 & 0.3 & 0.3 & 0.4 & 0.5 & 0.6 & -0.2 & -0.3 & 0.3 \\
0.3 & 0.6 & 0.8 & 1.0 & 1.3 & 1.6 & -0.2 & -0.3 & 0.7 \\
0.3 & 0.8 & 1.2 & 1.5 & 1.9 & 2.4 & -0.3 & -0.2 & 1.0 \\
0.4 & 1.0 & 1.5 & 2.0 & 2.5 & 3.1 & -0.4 & -0.1 & 1.2 \\
0.5 & 1.3 & 1.9 & 2.5 & 3.2 & 4.0 & -0.4 & -0.2 & 1.6 \\
0.6 & 1.6 & 2.4 & 3.1 & 4.0 & 5.0 & -0.4 & -0.2 & 2.1 \\
\hline
-0.2 & -0.2 & -0.3 & -0.4 & -0.4 & -0.4 & 0.5 & 0.1 & 0.1 \\
-0.3 & -0.3 & -0.2 & -0.1 & -0.2 & -0.2 & 0.1 & 0.9 & -0.4 \\
0.3 & 0.7 & 1.0 & 1.2 & 1.6 & 2.1 & 0.1 & -0.4 & 1.2
\end{array}\right),$$

$$\Omega^{(3)} = \left(\begin{array}{rrrrrr|rrr}
1.2 & 1.3 & 1.3 & 1.3 & 1.5 & 1.6 & -0.8 & -0.8 & 0.4 \\
1.3 & 1.5 & 1.6 & 1.7 & 1.9 & 2.1 & -0.9 & -0.7 & 0.6 \\
1.3 & 1.6 & 1.7 & 1.9 & 2.2 & 2.4 & -0.9 & -0.5 & 0.7 \\
1.3 & 1.7 & 1.9 & 2.1 & 2.4 & 2.7 & -0.9 & -0.4 & 0.8 \\
1.5 & 1.9 & 2.2 & 2.4 & 2.9 & 3.3 & -1.1 & -0.4 & 0.9 \\
1.6 & 2.1 & 2.4 & 2.7 & 3.3 & 3.7 & -1.2 & -0.3 & 1.1 \\
\hline
-0.8 & -0.9 & -0.9 & -0.9 & -1.1 & -1.2 & 0.8 & 0.5 & -0.1 \\
-0.8 & -0.7 & -0.5 & -0.4 & -0.4 & -0.3 & 0.5 & 0.9 & -0.1 \\
0.4 & 0.6 & 0.7 & 0.8 & 0.9 & 1.1 & -0.1 & -0.1 & 0.6
\end{array}\right),$$

$$\Omega^{(4)} = \left(\begin{array}{rrrrrr|rrr}
1.0 & 1.3 & 1.5 & 1.7 & 2.0 & 2.2 & -0.9 & -0.7 & 0.5 \\
1.3 & 2.0 & 2.4 & 2.7 & 3.2 & 3.6 & -1.4 & -0.7 & 0.6 \\
1.5 & 2.4 & 2.9 & 3.4 & 4.0 & 4.5 & -1.7 & -0.7 & 0.8 \\
1.7 & 2.7 & 3.4 & 4.0 & 4.7 & 5.4 & -2.0 & -0.7 & 0.9 \\
2.0 & 3.2 & 4.0 & 4.7 & 5.6 & 6.3 & -2.3 & -0.7 & 1.0 \\
2.2 & 3.6 & 4.5 & 5.4 & 6.3 & 7.3 & -2.6 & -0.7 & 1.2 \\
\hline
-0.9 & -1.4 & -1.7 & -2.0 & -2.3 & -2.6 & 1.3 & 0.5 & -0.1 \\
-0.7 & -0.7 & -0.7 & -0.7 & -0.7 & -0.7 & 0.5 & 0.9 & -0.2 \\
0.5 & 0.6 & 0.8 & 0.9 & 1.0 & 1.2 & -0.1 & -0.2 & 0.7
\end{array}\right),$$

$$\Omega^{(5)} = \left(\begin{array}{cccccc|ccc} 0.8 & 1.0 & 1.3 & 1.5 & 1.7 & 2.0 & -0.8 & -0.4 & 0.5 \\ 1.0 & 1.7 & 2.2 & 2.7 & 3.2 & 3.8 & -1.4 & -0.3 & 0.7 \\ 1.3 & 2.2 & 2.9 & 3.7 & 4.3 & 5.1 & -1.8 & -0.2 & 1.0 \\ 1.5 & 2.7 & 3.7 & 4.8 & 5.7 & 6.7 & -2.2 & -0.1 & 1.2 \\ 1.7 & 3.2 & 4.3 & 5.7 & 6.8 & 8.0 & -2.6 & -0.1 & 1.4 \\ 2.0 & 3.8 & 5.1 & 6.7 & 8.0 & 9.5 & -3.1 & -0.0 & 1.7 \\ \hline -0.8 & -1.4 & -1.8 & -2.2 & -2.6 & -3.1 & 1.4 & 0.3 & -0.3 \\ -0.4 & -0.3 & -0.2 & -0.1 & -0.1 & -0.0 & 0.3 & 0.5 & -0.1 \\ 0.5 & 0.7 & 1.0 & 1.2 & 1.4 & 1.7 & -0.3 & -0.1 & 0.7 \end{array}\right),$$

The parameters for generating S are

$$\boldsymbol{\zeta} = (-2.61, -2.75, -2.08, -1.52)^T,$$

where ζ_s corresponds to the $(s-1)$-th element $(s = 2, \ldots, 5)$,

$$\boldsymbol{\psi} = (-0.96, 0.66, 0.78, 0.54)^T,$$

where ψ_s corresponds to the $(s-1)$-th element $(s = 2, \ldots, 5)$, and

$$\boldsymbol{\xi} = \begin{pmatrix} -1.057 & 0.328 & -0.121 & 0.273 \\ -0.826 & 0.128 & 0.525 & -0.781 \\ -0.487 & 0.479 & 0.534 & -0.480 \end{pmatrix},$$

where $\boldsymbol{\xi}_s$ corresponds to the $(s-1)$-th column $(s = 2, \ldots, 5)$. The parameters are chosen to mimic the dropout rate of the real data.

MNAR results. Detailed simulation results for Scenario 3 under MNAR are given in Table A.3.

A.6 The Schizophrenia Clinical Trial Data Analysis Details

Comparison with previous results. Table A.4 shows a comparison of data analysis results with Linero and Daniels (2015) under both the MAR and the mixture of MAR/MNAR assumptions.

Model	E($\tilde{\tau}$)	Bias	CI width	CI coverage	MSE
GP	-0.25	-0.055(0.007)	0.687(0.002)	0.940(0.010)	0.063(0.002)
	0	-0.014(0.007)	0.690(0.002)	0.972(0.007)	0.061(0.002)
	0.25	0.027(0.007)	0.693(0.002)	0.968(0.008)	0.063(0.002)
	0.5	0.069(0.008)	0.699(0.002)	0.946(0.010)	0.068(0.002)
LM	-0.25	-0.042(0.007)	0.725(0.002)	0.961(0.008)	0.066(0.002)
	0	-0.001(0.007)	0.728(0.002)	0.980(0.006)	0.065(0.002)
	0.25	0.042(0.008)	0.734(0.002)	0.972(0.007)	0.068(0.002)
	0.5	0.085(0.008)	0.741(0.002)	0.948(0.010)	0.075(0.002)
LM–	-0.25	-0.047(0.007)	0.751(0.002)	0.972(0.007)	0.068(0.002)
	0	0.015(0.007)	0.761(0.002)	0.987(0.005)	0.068(0.002)
	0.25	0.079(0.007)	0.768(0.002)	0.966(0.008)	0.075(0.002)
	0.5	0.144(0.008)	0.783(0.003)	0.909(0.012)	0.092(0.003)
DPM	-0.25	-0.040(0.007)	0.789(0.002)	0.982(0.006)	0.072(0.002)
	0	-0.008(0.008)	0.792(0.002)	0.984(0.006)	0.071(0.002)
	0.25	0.024(0.008)	0.794(0.002)	0.982(0.006)	0.072(0.002)
	0.5	0.056(0.008)	0.798(0.002)	0.965(0.008)	0.075(0.002)
DPM–	-0.25	-0.052(0.007)	0.703(0.002)	0.958(0.009)	0.065(0.002)
	0	-0.001(0.008)	0.709(0.002)	0.967(0.008)	0.064(0.003)
	0.25	0.050(0.008)	0.716(0.002)	0.947(0.010)	0.066(0.002)
	0.5	0.098(0.008)	0.725(0.002)	0.914(0.013)	0.074(0.003)

Table A.3: Summary of simulation results for Scenario 3 under MNAR. Values shown are averages over repeat sampling, with numerical Monte Carlo standard errors in parentheses. CI width and coverage are based on 95% credible intervals. The values of E($\tilde{\tau}$), -0.25, 0, 0.25 and 0.5, correspond to prior specifications Unif$(-0.75, 0.25)$, Unif$(-0.5, 0.5)$, Unif$(-0.25, 0.75)$ and Unif$(0, 1)$, respectively.

Sensitivity analysis. Figure A.2 shows how inferences on $r_{\mathrm{T}} - r_{\mathrm{P}}$ and $r_{\mathrm{A}} - r_{\mathrm{P}}$ change for different choices of $\tilde{\tau}_{\mathrm{T}}$, $\tilde{\tau}_{\mathrm{A}}$ and $\tilde{\tau}_{\mathrm{P}}$.

Model	MDM	$r_{\mathrm{T}} - r_{\mathrm{P}}$	$r_{\mathrm{A}} - r_{\mathrm{P}}$
NP	MAR	0.6(-5.1, 7.0)	-6.1(-13.9, 1.7)
L & D (2015)	MAR	-1.7(-8.0, 4.8)	-5.4(-12.6, 2.3)
NP	MAR/MNAR	0.9(-5.3, 7.8)	-6.4(-14.3, 1.8)
L & D (2015)	MAR/MNAR	-1.6(-8.4, 5.4)	-6.2(-13.8, 2.0)

Table A.4: Comparison of inference results with Linero and Daniels (2015). NP represents the proposed model, and L & D (2015) represents the model in Linero and Daniels (2015). MDM refers to the missing data mechanism. Values shown are posterior means, with 95% credible intervals in parentheses.

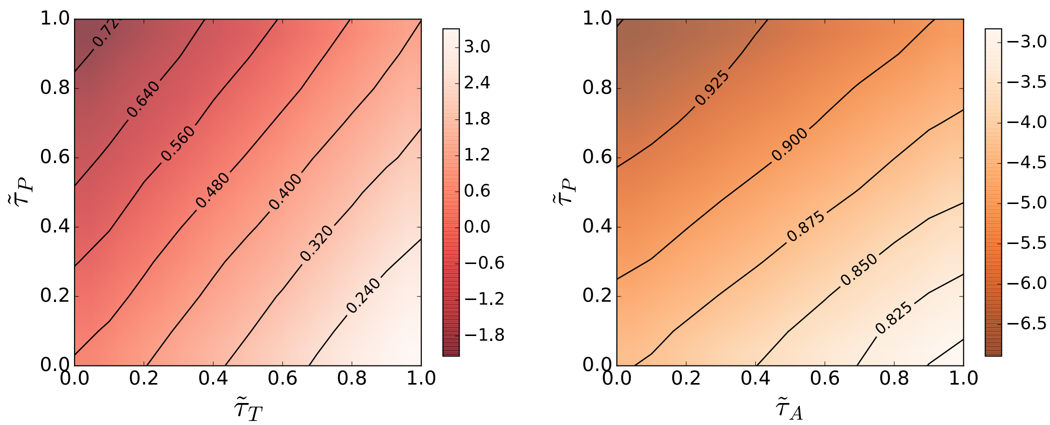

Figure A.2: Contour plots showing inferences on treatment effects $r_{\mathrm{T}} - r_{\mathrm{P}}$ (left) and $r_{\mathrm{A}} - r_{\mathrm{P}}$ (right) for different choices of the sensitivity parameters along the $[0, 1]$ grid. The colors represent posterior means of $r_x - r_{\mathrm{P}}$, where a deeper color indicates larger improvement compared to placebo. The black contour lines show posterior probabilities of $r_x - r_{\mathrm{P}} < 0$.

A.7 Computational Details

We report computational details for the simulation studies and the schizophrenia clinical trial data analysis here.

Chain lengths. For all the simulation studies, we run 15,000 iterations to obtain samples from $p\left(\boldsymbol{\pi} \mid \{\bar{\boldsymbol{y}}_{is_i}, s_i, \boldsymbol{v}_i\}_{i=1}^{N}\right)$ (under the Gaussian process and AR/CAR priors, see Equation 11). We discard the first 5,000 draws as initial burn-in, and keep every 10th iteration. We run 10,000 iterations to sample from $p\left(\boldsymbol{\varphi} \mid \{s_i, \boldsymbol{v}_i\}_{i=1}^N\right)$, discarding the first 5,000 draws and keeping every 5th iteration. Finally, we directly draw 1,000 samples from $p\left(\boldsymbol{\eta} \mid \{\boldsymbol{v}_i\}_{i=1}^N\right)$. As a result, we have $L = 1,000$ posterior draws for $\boldsymbol{\pi}$, $\boldsymbol{\varphi}$ and $\boldsymbol{\eta}$, i.e. $\{\boldsymbol{w}_O^{(l)} = (\boldsymbol{\pi}^{(l)}, \boldsymbol{\varphi}^{(l)}, \boldsymbol{\eta}^{(l)}), l = 1, \ldots, 1000\}$. Next, in the G-computation, for each posterior draw, response values for $M = 1,000$ pseudo subjects are generated.

For the real data analysis, we run 50,000 iterations to sample from $p\left(\boldsymbol{\pi} \mid \{\bar{\boldsymbol{y}}_{is_i}, s_i, \boldsymbol{v}_i\}_{i=1}^N\right)$, discarding the first 10,000 draws and keeping every 40th iteration. In the G-computation, for each posterior draw, response values for $M = 20,000$ pseudo subjects are generated.

Chain mixing and convergence diagnostic. We present some convergence diagnostics of the Markov chains using the R package `coda` (Plummer et al., 2008). Without loss of generality, we use the test drug arm of the schizophrenia clinical trial data as an example.

First, we compute the Geweke diagnostic (Geweke, 1991) for a single Markov chain, which takes the first L_1 draws and the last L_2 draws of the Markov chain and makes a difference of means test for the two parts. If the draws are from the stationary distribution, the difference of the means should have an asymptotically standard normal distribution. By default, we set $L_1 = 0.1 \cdot L$ and $L_2 = 0.5 \cdot L$. As an example, we use the time/pattern specific intercepts $\boldsymbol{b}$ as the test statistics. The Geweke z-score and the corresponding p-values are reported in Table A.5. All p-values are greater than 0.05, indicating no evidence of lack of convergence.

We also compute the Gelman-Rubin diagnostic (Gelman and Rubin, 1992) for multiple Markov chains. We run three chains with different random seeds and compare the draws from the three runs. We calculate the potential scale reduction factor (PSRF, or Gelman-Rubin statistic) for the three chains. The PSRF is a weighted sum of within-chain and between-chain variances. A PSRF close to 1 indicates the three chains are similar to each other, i.e. convergence of the chains to the target distribution. For the multivariate $\boldsymbol{b}$, the multivariate PSRF (Brooks and Gelman, 1998) is 1.08. Figure A.3 shows the traceplot of b_{22} for the three Markov chains. To summarize, there is no strong evidence that the Markov chains are not converging.

Test stat.	z-score	p-value	PSRF	PSRF upper CI
b_{22}	-1.14	0.25	1.01	1.05
b_{23}	-0.74	0.46	1.00	1.00
b_{33}	1.32	0.19	1.00	1.01
b_{24}	1.25	0.21	1.01	1.01
b_{34}	0.26	0.79	1.02	1.07
b_{44}	-0.57	0.57	1.01	1.01
b_{25}	0.52	0.60	1.01	1.02
b_{35}	-1.62	0.11	1.01	1.05
b_{45}	-1.45	0.15	1.01	1.03
b_{55}	-0.07	0.94	1.00	1.01
b_{26}	0.75	0.45	1.01	1.02
b_{36}	-0.28	0.78	1.01	1.05
b_{46}	-1.56	0.12	1.01	1.03
b_{56}	0.67	0.50	1.01	1.03
b_{66}	1.23	0.22	1.01	1.02

Table A.5: Convergence diagnostics. Columns 1 to 4 show the Geweke z-scores, corresponding p-values of the z-scores, potential scale reduction factors (PSRFs) and upper confidence limits of the PSRFs for the test statistics $\boldsymbol{b} = (b_{22}, \ldots, b_{66})$, respectively.

Computing specifications and times. All computations in this paper are conducted using Lonestar 5 at the Texas Advanced Computing Center (TACC). The computations for multiple replicates of the simulated datasets are conducted in parallel using multiple cores and multiple computing nodes. Each computing node is a Xeon E5-2690 v3 (Haswell) with 12 cores per socket (24 cores/node), 2.6 GHz (`https://portal.tacc.utexas.edu/user-guides/lonestar5`).

The average computing times for all model components and all data analysis scenarios are summarized in Table A.6. The time to sample from $p\left(\boldsymbol{\pi} \mid \{\bar{\boldsymbol{y}}_{is_i}, s_i, \boldsymbol{v}_i\}_{i=1}^N\right)$ depends on the number of subjects and on the dropout rates. A scenario where the subjects have lower dropout rates has more observed responses, and thus sampling from $p\left(\boldsymbol{\pi} \mid \{\bar{\boldsymbol{y}}_{is_i}, s_i, \boldsymbol{v}_i\}_{i=1}^N\right)$

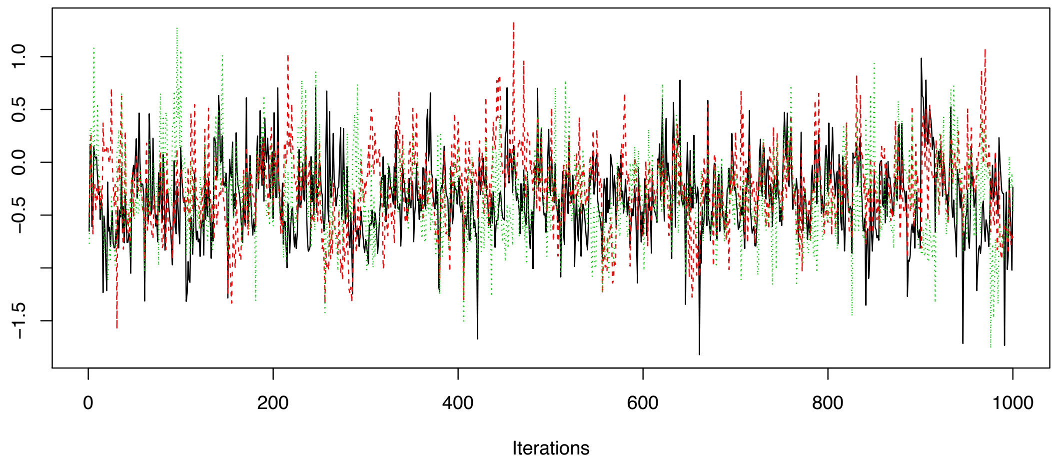

Figure A.3: Traceplot of b_{22} for the three Markov chains in three different colors.

takes longer. Therefore, although simulation scenarios 1, 2 and 3 have the same number of subjects, their times for sampling from $p\left(\boldsymbol{\pi} \mid \{\bar{\boldsymbol{y}}_{is_i}, s_i, \boldsymbol{v}_i\}_{i=1}^{N}\right)$ are different. Table A.6 shows the time for G-computation under MAR. Under MNAR, the time needed for G-computation increases by 1.5 to 2 fold depending on the dropout rates.

	N	J	Time for GP	Time for BART	Time for G-comp.
Simu. 1	200	6	8500	160	9500
Simu. 2	200	6	6000	160	9500
Simu. 3	200	6	7000	160	9500
Test	81	6	4400	60	5500
Active	45	6	1700	40	2800
Placebo	78	6	4000	60	4900

Table A.6: Average computing time (in seconds) for each model component and each data analysis scenario. The values N and J represent the number of subjects and number of time points for the corresponding scenario, respectively. Time for GP, time for BART and time for G-comp. are in short for the times for drawing L_π samples from $p\left(\boldsymbol{\pi} \mid \{\bar{\boldsymbol{y}}_{is_i}, s_i, \boldsymbol{v}_i\}_{i=1}^N\right)$, drawing L_φ samples from $p\left(\boldsymbol{\varphi} \mid \{s_i, \boldsymbol{v}_i\}_{i=1}^N\right)$ and generating M pseudo subjects for L posterior draws under MAR, respectively. For the simulation scenarios, $L_\pi = 15,000$, $L_\varphi = 10,000$, $M = 1,000$ and $L = 1,000$. For the real data analysis, $L_\pi = 50,000$, $L_\varphi = 10,000$, $M = 20,000$ and $L = 1,000$, where the $M = 20,000$ pseudo subjects are drawn using 20 parallel threads (each thread generates 1,000).

References

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2014). *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

Brooks, S. P. and A. Gelman (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics 7*(4), 434–455.

De Oliveira, V. (2012). Bayesian analysis of conditional autoregressive models. *Annals of the Institute of Statistical Mathematics 64*(1), 107–133.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

Geweke, J. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, Volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.

Harel, O. and J. L. Schafer (2009). Partial and latent ignorability in missing-data problems. *Biometrika 96*(1), 37–50.

Linero, A. R. and M. J. Daniels (2015). A flexible Bayesian approach to monotone missing data in longitudinal studies with nonignorable missingness with application to an acute schizophrenia clinical trial. *Journal of the American Statistical Association 110*(509), 45–55.

Plummer, M., N. Best, K. Cowles, and K. Vines (2008). *coda: Output analysis and diagnostics for MCMC*. R package version 0.13-3.