A Bayesian Nonparametic Approach for Estimating Individualized Treatment-Response Curves

Yanbo Xu   YXU68@JHU.EDU
Department of Computer Science

Yanxun Xu   YANXUN.XU@JHU.EDU
Department of Applied Mathematics and Statistics

Suchi Saria   SSARIA@CS.JHU.EDU
Department of Computer Science
Johns Hopkins University

Abstract

We study the problem of estimating the continuous response over time of interventions from *observational time series*—a retrospective dataset where the policy by which the data are generated are unknown to the learner. We are motivated by applications where response varies by individuals and therefore, estimating responses at the individual-level are valuable for personalizing decision-making. We refer to this as the problem of estimating individualized treatment response (ITR) curves. In statistics, G-computation formula (Robins, 1986) has been commonly used for estimating treatment responses from observational data containing sequential treatment assignments. However, past studies have focused predominantly on obtaining point-in-time estimates at the population level. We leverage G-computation formula and develop a novel method based on Bayesian nonparametrics (BNP) that can flexibly model functional data and provide posterior inference over the treatment response curves both at the individual and population level. On a challenging dataset containing time series from patients admitted to a hospital, we estimate treatment responses for treatments used in managing kidney function and show that the resulting fits are more accurate than alternative approaches. Accurate methods for obtaining ITRs from observational data can dramatically accelerate the pace at which personalized treatment plans become possible.

1. Introduction

Accurate models of actions and their effects on the state of the agent are critical for decision-making. Learning of action-effect models is most straightforward from data where the learner can control the choice of actions and observe their responses. But, such data are not always possible to acquire. Alternatively, retrospective data may be available that contain time series generated from observing other agents act. Estimating action-effect models from *observational data*—data where the learner cannot control the actions that are prescribed, and the actions may be prescribed by a mechanism that is not known to the learner—are more challenging. We study an instance of this problem: specifically, we consider the problem of estimating the continuous response over time to an action. We are particularly motivated by applications in medicine where accurate action-effect models for estimating treatment effects can be used for personalizing therapy.

In statistics, the problem of estimating treatment effects from observational data containing sequential treatment assignments has been studied extensively using approaches such as the G-computation formula (Robins, 1986), G-estimation of structural nested models (Robins, 2004), inverse probability of treatment weighted (IPTW) estimation of marginal structure models (van der Laan and Petersen, 2007), doubly robust learning (Tsiatis, 2007; Zhao et al., 2015) with applications to longitudinal data analysis (Hernán et al., 2000), survival analysis (Lunceford et al., 2002), and

adaptive treatment selections in clinical trials (Murphy et al., 2007a,b). A related problem in reinforcement learning is off-policy evaluation where the goal is to estimate the value of a policy (sequence of actions) from data collected by another policy (Sutton et al., 1998). For example, doubly-robust estimators for policy evaluation have been developed for contextual bandits (Dudik et al., 2011) and for sequential decision-making problems (Jiang and Li, 2015). See survey of example techniques in Paduraru et al. (2012). In this paper, we use G-computation formula to adjust for time-varying confounding. However, we depart from the existing literature by using a novel Bayesian nonparametric method so as to (1) flexibly model the longitudinal outcome over time, and (2) characterize heterogeneity in treatment effects across individuals.

Bayesian nonparametric (BNP) methods (Ferguson, 1973; Müller and Mitra, 2013; Müller and Rodriguez, 2013) are gaining popularity in longitudinal data analysis and treatment effect modeling since they are characterized by parameters that live in an infinite-dimensional space, allowing one to flexibly approximate arbitrary distributions. For flexible longitudinal data analysis, Silva (2016) uses Gaussian process to model longitudinal outcome under different levels of interventions. In another example, Chib and Hamilton (2002) uses the Dirichlet Process prior to add flexiblility in representing the outcome and the treatment effects.

A number of related works have focused on heterogeneous treatment effects (HTE) by estimating the effects conditional on covariates defining subpopulations. For example, Tian et al. (2014) and Imai et al. (2013) apply regularized linear regression to select covariates characterizing subpopulations with differential outcomes. Other work use tree structure to partition based on covariates that identify subpopulations with different outcomes (Foster et al., 2010; Su et al., 2009) or different conditional treatment effects (Athey and Imbens, 2015). All of the above-mentioned works focus on obtaining point-in-time estimates. Only recently, Huang et al. (2015) and Xu and Ji (2014) have used parametric models to estimate treatment effects over time.

The proposed method advances state-of-the-art in a number of ways. First, in contrast with pat works that focus on modeling response at a point-in-time, this work obtains the continuous response over time. Further, we obtain these from sparse and irregularly sampled observational data. Second, the proposed BNP model flexibly models variations in treatment effects while borrowing strength across individuals. In applications such as education and healthcare where response across individuals can vary widely, recovering individual level effects are more informative for decision-making. Third, the fully Bayesian approach quantifies uncertainty at the individual-level; this is particularly important for individuals where the estimated effects maybe uncertain due to lack of data. A key practical advantage of using a nonparametric approach is that they often provide better fits to challenging data than can be obtained using parametric model based methods. This is particularly important in our application of estimating treatment response curves for physiologic time series.

2. Longitudinal Treatment Response Model

As a running example, we use the application of estimating the longitudinal model for creatinine (a measurement of kidney function). Specifically, our goal is to obtain an individualized estimate of the effect over time for treatments given for modulating creatinine. We consider the problem of estimating these from sparse, irregularly sampled data such as those in electronic health records (EHRs). There are two key challenges that must be addressed. First, in clinical data contained within EHRs, measurements are often not obtained at regular intervals, and measurement schedules vary across individuals. For example, caregivers may choose to make measurements of creatinine once a day on some patients while multiple times a day on others. When the data are collected at fixed regular intervals, discrete-time approaches that maintain estimates only at specific points-in-time are adequate (e.g. Taubman et al. (2009)). To address this, we will employ functional representations instead (Quintana et al., 2015).

Another key challenge is the presence of *time-varying confounding* (Robins, 1986, 1987). To understand time-varying confounding, first, let us consider the simple example where a treatment

tends to be assigned to sicker patients. Since these patients are also more likely to die, without accounting for this bias, one might assume that this treatment kills people. In the sequential-treatment assignment setting, such confounding occurs because doctors use the measurement of a variable (creatinine) to determine whether or not to treat which affects its value at a subsequent time. To correct for this confounding, our estimation is based on Robins (1986, 1987)'s G-computation formula, a widely used approach in estimating treatment effects from data with time-varying confounding. The key assumption for G-computation formula is that the treatment received at each time was allocated (i.e., ignorable) conditional on the observed past treatment and covariate history. In medicine, often treatment decisions are made based on well-documented clinical history about the individual. Therefore, this assumption is often applicable.

Notation: Assume we have observations $\boldsymbol{Y}_i = \{Y_{ij} : j = 1, .., J_i\}$ from the ith individual at (irregularly-sampled) times $\{t_{i1}, ..., t_{iJ_i}\}$. In addition, we have $\boldsymbol{X}_i$, a $1 \times p$ vector of observed covariates (e.g., age, gender) about this individual. We also have treatments $\boldsymbol{A}_i = \{A_{il} : l = 1, ..., L_i\}$ that were given to patient i at times $\{\tau_{i1}, ..., \tau_{iL_i}\}$, where $A_{il} = d$ for some treatment type $d \in \{1, .., D\}$. We assume that the effect of each treatment type d lasts at most within a window W_d. In our setting, as is the case for many drugs, the effects are transient and last over the course of the disease but do not permanently alter the body chemistry. We use $\mathcal{H}_{ij}$ to denote the set of *active treatments* at t_{ij} i.e., treatments that were given to patient i within their windows prior to time t_{ij} and affect the outcome at t_{ij}. Specifically, $\mathcal{H}_{ij} = \{A_{il} : t_{ij} - W_{A_{il}} \leq \tau_{il} \leq t_{ij}\}$. The value of a measurement $\boldsymbol{Y}_i$ within an interval (t, T) is denoted by $\boldsymbol{Y}_{i,(t,T]}$. The sets of measurements and treatments preceding a time t are denoted by $\boldsymbol{Y}_{i,<t}$ and $\boldsymbol{A}_{i,<t}$ respectively.

Our goal is to obtain posterior inference for the treatment response curves at the individual and population levels, and for the outcomes $\boldsymbol{Y}_{i,>t}$ given any sequence of treatments conditioned upon historical data about the individual and the population. Estimation using G-computation requires the specification of the outcome model for $\boldsymbol{Y}$ as a function of the covariates and treatments (Robins, 1986, 1987). In contrast with prior methods that assume a parametric model for $\boldsymbol{Y}$ at discrete points in time (e.g., Hernán et al. (2000)), we describe below a Bayesian nonparametric model that models uncertainty over the baseline progression of a measurement under no treatment and the response to treatment over time. We tackle the general setting of learning from data with multiple exposures to the same treatment or different treatments under the assumption of additivity of treatment effects.

We model the outcome Y_{ij} using a generalized mixed-effects model combining the baseline progression and the change in response due to prior treatments as follows:

$$Y_{ij}|\boldsymbol{X}_i, \mathcal{H}_{ij} = \underbrace{b(\boldsymbol{X}_i) + \boldsymbol{u}_i(t_{ij})}_{\text{baseline progression}} + \underbrace{\boldsymbol{f}_i(t_{ij}; \mathcal{H}_{ij})}_{\text{treatment response}} + \underbrace{\boldsymbol{\epsilon}_i(t_{ij}; \mathcal{H}_{ij})}_{\text{noise}}, \; j = 1, ..., J_i. \tag{1}$$

We describe each of these components in more details below.

2.1 Modeling Baseline Progression

$b(\boldsymbol{X}_i)$ is the fixed-effects component that captures the dependence of the outcome variable on the observed features $\boldsymbol{X}_i$ (e.g., age, gender, genetic mutation). Here we model $b(\boldsymbol{X}_i)$ as a linear regression, by assuming $b(\boldsymbol{X}_i; \boldsymbol{\beta}_i) = \boldsymbol{X}_i^T \boldsymbol{\beta}_i$. $\boldsymbol{u}_i$ is the random-effects component that models the individual-specific deviations over time in baseline progression from $b(\boldsymbol{X}_i)$. We choose $\boldsymbol{u}_i$ to be generated from a zero-mean Gaussian process with a structured covariance $\mathcal{K}_{ui}(\sigma_{ui}^2, \rho_{ui}) = Cov(\boldsymbol{u}_i(t_{ij}), \boldsymbol{u}_i(t_{ij'})) = \sigma_{ui}^2 \rho_{ui}^{|t_{ij} - t_{ij'}|}$ with $\rho_{ui} \in (0, 1)$; this represents an exponential covariance function. Similar choices were made by Quintana et al. (2015) in their application of modeling functional data. A different choice for both the mean and the covariance kernel can be made depending on the properties of the data; see Schulam and Saria (2015) for a different example of the baseline model for modeling progression in chronic diseases.
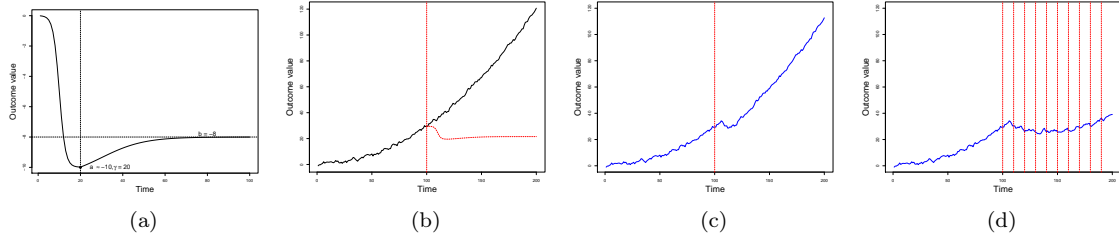
3

Figure 1: An example of decreasing treatment-response curve and an illustration of additive effects

2.2 Modeling Treatment-Response

We focus on the scenario where treatment choices are discrete and assume that the treatment effects are additive. Further, we assume that the effect of each treatment type d lasts at most within window W_d. Given the set of treatments $\mathcal{H}_{ij} = \{A_{il} : t_{ij} - W_{A_{il}} \leq \tau_{il} \leq t_{ij}\}$ preceding time t_{ij}, we formulate the treatment response model as:

$$\boldsymbol{f}_i(t_{ij}; \mathcal{H}_{ij}) = \sum_{l:t_{ij}-W_{A_{il}} \leq \tau_{il} \leq t_{ij}} g_{i,A_{il}}(t_{ij} - \tau_{il}), \tag{2}$$

where $g_{i,A_{il}}(t_{ij} - \tau_{il})$ denotes the response curve of individual i for treatment A_{il} that was given at time τ_{il}. To estimate the cumulative effect at t_{ij}, the response curves from the treatment set $\mathcal{H}_{ij}$ are added. We parameterize the function $g_{id}(t)$ as

$$g_{id}(t) = \begin{cases} b_0 + \alpha_{1_{id}}/[1 + \exp(-\alpha_{2_{id}}(t - \gamma_{id}/2))], & \text{if } 0 \leq t < \gamma_{id} \\ b_{id} + \alpha_0/[1 + \exp(\alpha_{3_{id}}(t - 3\gamma_{id}/2))]. & \text{if } t \geq \gamma_{id}, \end{cases} \tag{3}$$

with the free parameters $\{\alpha_1, \alpha_2, \alpha_3, \gamma, b\}$; here, the collection of individual-specific treatment effect parameters $\alpha_{1_{id}}$'s are short-handed to α_1 and so on.

The motivation of choosing this particular form of the $g_{id}(t)$ function is to obtain a flexible asymmetric "U" shape curve. We concatenate two sigmoid curves and allow the parameters for the two sigmoid functions and the point of switching between the two sigmoids to vary so that it can flexibly capture responses where a marker may either increase or decrease and eventually converges to a stable value. Figure 1 (a) visualizes one example $g(t; \alpha_1 = -10, \alpha_2 = 0.7, \alpha_3 = 0.1, \gamma = 20, b = -8)$, where $\alpha_1 \in \Re$ represents the curve's maximum value and the sign of α_1 determines how the individual responds to the treatment. For example, $\alpha_1 < 0$ if the marker decreases in response, $\alpha_1 > 0$ if the marker increases in response. $\alpha_2 \in (0,1)$ and $\alpha_3 \in (0,1)$ individually models the "steepness" of the two sigmoid curves; $\gamma \in \Re$ denotes the switching point; b denotes the value that the curve stabilizes and is constrained such that $b/g(\gamma) \in (0,1)$. Lastly, to make the $g_{id}(t)$ function well defined, we set $b_0 = -\alpha_{1_{id}}/[1 + \exp(\alpha_{2_{id}}\gamma_{id}/2)]$ for attaining $g_{id}(0) = 0$, and $\alpha_0 = (a_{1_{id}} + 2b_0 - b_{id})/(1 + \exp(-a_{3_{id}}\gamma_{id}/2))$ for attaining a unique peek value at $t = \gamma_{id}$.

Based on the example we give in Figure 1 (a) where the marker's value decreases in response to the treatment, Figures 1 (b-c) illustrate the cumulative effect where the response curves from sequential treatments are added. The black line in Figure 1 (b) denotes the increasing baseline progression under no treatment, and the vertical red line denotes when the treatment was prescribed; the outcome value is reduced after adding the response curve as is shown in Figure 1 (c); and the response to multiple sequential treatments is shown in Figure 1 (d).

2.3 Modeling Noise

We model the noise in two parts: the independent random noise ϵ^0_{ij} for individual i at each time point t_{ij} and the time-dependent random noise $\epsilon'_d(t)$ for treatment d within its effective window W_d.

4

We model this additional noise because adding the effects of the treatments can introduce higher uncertainty and error into the outcome model. The noise is time-dependent because the uncertainty reduces as the time from treatment administration increases and the effects diminish. So we formulate the overall noise as

$$\epsilon_i(t_{ij}; \mathcal{H}_{ij}) = \epsilon_{ij}^0 + \sum_{l: t_{ij} - W_{A_{il}} \leq \tau_{il} \leq t_{ij}} \epsilon'_{A_{il}}(t_{ij} - \tau_{il}), \tag{4}$$

where ϵ_{ij}^0's are i.i.d. Gaussian distributed with mean zero and variance $\sigma_{\epsilon_i}^2$, $\epsilon'_d(t)$'s are jointly Gaussian distributed with mean zero and structured covariance $\mathcal{K}_{\epsilon_d}(\sigma_{\epsilon'_d}^2, \rho_{\epsilon'_d}) = Cov(\epsilon'_d(t), \epsilon'_d(t')) = \sigma_{\epsilon'_d}^2 \rho_{\epsilon'_d}^{|t-t'|}$.

2.4 A Hierarchical Prior for Estimating Individualized Treatment Response (ITR) Curves

The baseline progression and the treatment response curves for individuals can be more flexibly modeled with nonparametric hierarchical models. This allows the model to estimate the parameters at the individual level while borrowing strength across individuals. Bayesian nonparametric approaches such as Dirichlet process (DP) and DP mixture have been widely used in clustering time-series data. For example, Ren et al. (2015) have applied DP priors in latent factor models to cluster multiple housing price data streams. Nieto-Barajas et al. (2014) uses a generalization of DP mixture—Possion-DP priors—in linear dynamic models to group stock exchange data. We use the DP mixture to cluster the parameters for the fix-effects component of the baseline progression and the treatment response curves, and use the DP mixture of Gaussian processes to cluster the random-effects component for the baseline progression. The DP mixture of GPs has been used to identify and group the sub-divisions in each individual's gene expression (Hensman et al., 2015) or disease trajectories (Ross and Dy, 2013). We use DP mixture of GPs to group individuals' baseline deviations on their similarity in the GP's kernel parameters. Alternative solutions can use a hierarchical GP (Hensman et al., 2013, 2015), where a single GP is used to model the behavior of the population and a fix number of other GPs are used to model the group-specific deviations from the population. Others have proposed methods for transfer across tasks (e.g., Li et al. (2009), Ammar et al. (2015)) though often assuming that the responses are modeled at discrete times.

2.4.1 BACKGROUND ON THE DIRICHLET PROCESS MIXTURE

We briefly describe the DP and the DP mixture (DPM). Ferguson (1973) introduced the DP prior as a probability distribution on an infinite dimensional measurable space of probability measures. The stick-breaking construction by Sethuraman (1994) provides an intuitive and interpretable representation of the DP. Let G_0 be a known distribution and let $M > 0$ be a positive constant. Then we say $G \sim DP(G_0, M)$ provided

$$G(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{\theta_k}(\cdot), \ \theta_k \overset{iid}{\sim} G_0,$$

where $\delta_{\theta_k}(\cdot)$ defines point mass at θ_k and ω_k's are defined as $\omega_k = V_k \prod_{r=1}^{k-1}(1 - V_r)$, $V_k \sim Beta(1, M)$. Thus G is a random distribution that is discrete with probability one. G_0 is the base or centering distribution since $E(G) = G_0$.

The discrete nature of the DP makes it inappropriate for modeling continuous data where units within a partition share similar rather than the same parameter. Therefore, DPM extends DP by introducing a continuous kernel centered at θ_k instead of a point mass δ_{θ_k}. Let $y_1, y_2, \ldots$ be i.i.d. samples and $f(\cdot|\theta)$ be a parametric density function, we can write the stick-breaking construction of the DPM as

$$y_i \mid (\omega_k), (\theta_k) \sim \sum_{k=1}^{\infty} \omega_k f(\cdot|\theta_k), \ \theta_k \sim G_0.$$
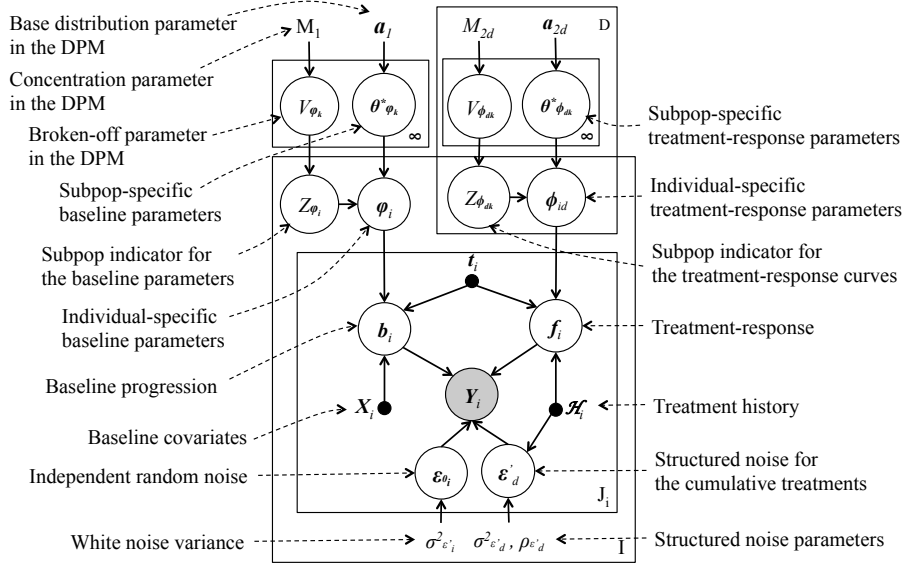
5

Figure 2: Graphical representation of the hierarchical treatment response model. Hidden variables are circled; observed outcome variables are shaded; observed input variables are filled.

2.4.2 Hierarchical Individualized Treatment-Response (ITR) model

We leverage the DPM prior to cluster both the baseline progression and the treatment response parameters—while allowing individual-specific variability—and obtain a hierarchical treatment-response model as shown in Figure 2. Specifically, let $\boldsymbol{b}_i$ denote the sum of the fixed-effects component $b(\boldsymbol{X}_i)$ and the random-effects component $\boldsymbol{u}_i$ in the baseline progression. Then based on the description in Section 2.1, $\boldsymbol{b}_i$ follows the distribution

$$p(\boldsymbol{b}_i|\boldsymbol{\varphi}_i) = \mathcal{N}(\boldsymbol{X}_i^T\boldsymbol{\beta}_i, \mathcal{K}_{ui}), \tag{5}$$
$$\mathcal{K}_{ui}(t_{ij}, t_{ij'}; \sigma_{ui}^2, \rho_{ui}) = \sigma_{ui}^2 \rho_{ui}^{|t_{ij}-t_{ij'}|},$$

where $\boldsymbol{\varphi}_i = \{\boldsymbol{\beta}_i, \sigma_{ui}^2, \rho_{ui}\}$ denotes all the individual-specific baseline progression parameters. We put a DPM prior on $\boldsymbol{\varphi}_i$'s. As shown in Figure 2, Z_{φ_i} is a discrete latent variable that indicates the mixture component associated with individual i. $\boldsymbol{\varphi}_i$ is sampled from a multivariate normal distribution centered at the parameters $\boldsymbol{\theta}_{\varphi_k}^*$ associated with the mixture component specified by Z_{φ_i} (we put superscript $*$ to denote the unique component parameters). The hyperparameter M_1 controls the degree of clustering and generates the V_{φ_k}'s and the associated mixture component weights. The distributions for the remainder of the parameters in the DPM prior for the baseline model are specified in Eq. (6). For parameters that lie in the real-space we assume they are sampled from a Gaussian distribution. For parameters that are constrained (e.g., $\sigma_{ui}^2 \in (0, +\infty)$), we transform the support of these variables into the real-space first and posit Gaussian priors on the transformed variables. This requires a calculation of the additional Jacobian adjustment $|\det J(T^{-1}(y))|$ for each transformation $y = T(x)$ (Olive, 2014). More details are in Appendix A. The parameters $\beta_0, \kappa_0, \theta_0, \boldsymbol{S}_0, \mu_{\rho_0'}, \sigma_{\rho_0'}^2, \mu_{\sigma_0'}, \sigma_{\sigma_0'}^2$ in the base-distribution G_{10} serve as hyperparameters $\boldsymbol{a}_1$ in Figure 2. These are selected to place uninformative priors; we discuss this in more detail in Section 4.

$$p(\boldsymbol{\beta}_i) = \mathcal{N}(\boldsymbol{\beta}_i; \boldsymbol{\beta}_{b_i}, \Sigma_{b_i}), \tag{6}$$
$$p(\sigma_{ui}^2) = \mathcal{N}(\log(\sigma_{ui}^2); \mu_{\sigma_{ui}'}, \sigma_{\sigma_{u0}'}^2)/\sigma_{ui}^2,$$
$$p(\rho_{ui}) = \mathcal{N}(\text{logit}(\rho_{ui}); \mu_{\rho_{ui}'}, \sigma_{\rho_{u0}'}^2)/(1 - \rho_{ui})^2,$$
$$G_{10}(\boldsymbol{\beta}_{b_i}, \Sigma_{b_i}, \mu_{\sigma_{ui}'}, \mu_{\rho_{ui}'}) = \text{NIW}(\boldsymbol{\beta}_{b_i}, \Sigma_{b_i}; \beta_0, \kappa_0, \theta_0, \boldsymbol{S}_0)\mathcal{N}(\mu_{\sigma_{ui}'}; \mu_{\sigma_0'}, \sigma_{\sigma_0'}^2)\mathcal{N}(\mu_{\rho_{ui}'}; \mu_{\rho_0'}, \sigma_{\rho_0'}^2).$$

For the treatment response model, $\boldsymbol{f}_i$ is defined in Eq. (2). We let $\boldsymbol{\phi}_{id} = \{\alpha_{1_{id}}, \alpha_{2_{id}}, \alpha_{3_{id}}, \gamma_{id}, b_{id}\}$ denote the individual-specific treatment parameters in the $g_{id}(t)$ function in Eq. (3). Similarly, we put DPM priors on $\boldsymbol{\phi}_{id}$'s. Then $\boldsymbol{\phi}_{id}$ is sampled from a multivariate normal distribution centered at the parameters $\boldsymbol{\theta}^*_{\phi_{id}}$ associated with the mixture component specified by $Z_{\phi_{id}}$. The hyperparameter M_{2d} generates the broken-off portions V_{ϕ_k}'s and controls the degree of clustering. The distributions for the remainder of the parameters in the DPM prior for the treatment response model are specified in Eq. (7). For those constrained parameters (e.g. $\alpha_{2_{id}} \in (0,1)$), we still transform them into the real-space first and posit Gaussian priors on the transformed variables. The parameters $\boldsymbol{\mu}_{d_0}$ and $\boldsymbol{D}_{d_0}$ in the base-distribution G_{20} serve as hyperparamters $\boldsymbol{a}_{2d}$ in Figure 2. These are also selected to place broad priors; we discuss this in more detail in Section 4.

$$p(\boldsymbol{\phi}_{id}) = \mathcal{N}(T^{-1}(\boldsymbol{\phi}_{id}); \boldsymbol{\mu}_{\phi'_{id}}, \boldsymbol{D}_{\phi'_0})|\zeta_{id}|/(1-\alpha_{2_{id}})^4, \ G_{2d0}(\boldsymbol{\mu}_{\phi'_{id}}) = \mathcal{N}(\boldsymbol{\mu}_{d_0}, \boldsymbol{D}_{d_0}), \tag{7}$$

where $\zeta_{id} = -1/g(\gamma_{id})(1 - b/g(\gamma_{id}))^2$ is from the Jacobian adjustment for the transformed $\boldsymbol{\phi}_{id}$'s.

For the noise model, based on the description in Section 2.3, $\boldsymbol{\epsilon}_i$ follows the distribution

$$p(\boldsymbol{\epsilon}_i | \sigma^2_{\epsilon_i}, \boldsymbol{\sigma}^2_{\epsilon'}, \boldsymbol{\rho}_{\epsilon'}) = \mathcal{N}(\boldsymbol{0}, \sigma^2_{\epsilon_i}\boldsymbol{I}_{J_i} + \mathcal{K}_{\epsilon'_i}), \tag{8}$$

$$\mathcal{K}_{\epsilon'_i}(t_{ij}, t_{ij'}; \boldsymbol{\sigma}^2_{\epsilon'}, \boldsymbol{\rho}_{\epsilon'}) = \begin{cases} \sum_l \sigma^2_{\epsilon'_{A_{il}}} \rho^{|t_{ij}-t_{ij'}|}_{\epsilon'_{A_{il}}} & , \ \forall l \text{ s.t. } \tau_{il} \leq t_{ij}, t_{ij'} \leq \tau_{il} + W_{A_{il}} \\ 0 & , \text{ otherwise.} \end{cases}$$

To complete the prior specification ,we put inverse gamma (IG) on the variance parameter $\sigma^2_{\epsilon_i}$, and transform the constrained parameters $\sigma^2_{\epsilon'_d}$ and $\rho_{\epsilon'_d}$ into the real-space then posit Gaussian priors on the transformed variables.

$$p(\sigma^2_{\epsilon_i}) = \text{IG}(s_\epsilon, \nu), \tag{9}$$
$$p(\sigma^2_{\epsilon'_d}) = \mathcal{N}(\log(\sigma^2_{\epsilon'_d}); \mu_{\epsilon_1}, \sigma^2_{\epsilon_1})/\sigma^2_{\epsilon'_d},$$
$$p(\rho_{\epsilon'_d}) = \mathcal{N}(\text{logit}(\rho_{\epsilon'_d}); \mu_{\epsilon_2}, \sigma^2_{\epsilon_2})/(1-\rho_{\epsilon'_d})^2.$$

3. Inference

We use MCMC to approximate the posterior inference of our model. Consider the joint posterior

$$p(\boldsymbol{u}, \boldsymbol{f}, \boldsymbol{\varphi}, \boldsymbol{\phi}, \sigma^2_\epsilon, \boldsymbol{\sigma}^2_{\epsilon'}, \boldsymbol{\rho}_{\epsilon'} \mid \boldsymbol{Y}) \tag{10}$$

$$\propto \left\{ \prod_{i=1}^I p(\boldsymbol{Y}_i, \boldsymbol{b}_i, \boldsymbol{f}_i \mid \boldsymbol{\varphi}_i, \boldsymbol{\phi}_i, \sigma^2_{\epsilon_i}, \boldsymbol{\sigma}^2_{\epsilon'}, \boldsymbol{\rho}_{\epsilon'}) \right\} p(\boldsymbol{\varphi})p(\boldsymbol{\phi})p(\boldsymbol{\sigma}^2_{\epsilon_i})p(\boldsymbol{\sigma}^2_{\epsilon'})p(\boldsymbol{\rho}_{\epsilon'})$$

$$= \left\{ \prod_{i=1}^I p(\boldsymbol{Y}_i \mid \boldsymbol{b}_i, \boldsymbol{f}_i, \sigma^2_{\epsilon_i}, \boldsymbol{\sigma}^2_{\epsilon'}, \boldsymbol{\rho}_{\epsilon'})p(\boldsymbol{b}_i \mid \boldsymbol{\varphi}_i)p(\boldsymbol{f}_i \mid \boldsymbol{\phi}_i) \right\} p(\boldsymbol{\varphi})p(\boldsymbol{\phi})p(\boldsymbol{\sigma}^2_\epsilon)p(\boldsymbol{\sigma}^2_{\epsilon'})p(\boldsymbol{\rho}_{\epsilon'}),$$

the first term in the product is $\mathcal{N}(\boldsymbol{X}_i\boldsymbol{b}_i + \boldsymbol{f}_i, \sigma^2_{\epsilon_i}\boldsymbol{I}_{J_i} + \mathcal{K}_{\epsilon'_i})$ with $\mathcal{K}_{\epsilon'_i}$ specified in Eq. (8), the second term is defined in Eq. (5), the third term is deterministic and specified in Eq. (2). We factorize the remaining terms as

$$\left\{ \prod_{i=1}^I \left\{ p(\boldsymbol{\varphi}_i) \prod_{d=1}^D p(\boldsymbol{\phi}_{id}) \right\} \right\} \left\{ \prod_{i=1}^I p(\sigma^2_{\epsilon_i}) \right\} \left\{ \prod_{d=1}^D p(\sigma^2_{\epsilon'_d})p(\rho_{\epsilon'_d}) \right\},$$

with each distribution specified in Eq. (6, 7 & 9) respectively.

For the infinite-dimensional DPM priors on $\boldsymbol{\varphi}_i$ and $\boldsymbol{\phi}_{id}$, we use a truncated stick-breaking process that was developed by Ishwaran and James (2001) to approximate. Ishwaran and James (2001)

7

justify that the truncated process greatly reduces computations and can closely approximate a full Dirichlet process when the truncation level is relatively large to the number of the observations. For our data that contains 123 patients, we choose the truncation level to be 20. This results at a small approximation error of 2.7×10^{-6} based on Theorem 2 in Ishwaran and James (2001). Given the truncation approximation, it allows us to use standard MCMC algorithms to update the parameters in the finite-dimension space. Particularly, we use Gibbs sampler to update the DPM parameters because for all of the variables, due to our use of conjugate priors, their full conditional distributions (conditioned upon all the other variables in the model) can be derived in closed form. Specifically, the form of this distribution for the component indicator variables Z_{φ_i} and $Z_{\phi_{id}}$ is Dirichlet; for the component-level variables $\beta_{b_k}^*$, $\Sigma_{b_k}^*$ is NIW, for $\mu_{\sigma'_{u_k}}^*$, $\mu_{\rho'_{u_k}}^*$ and $\mu_{\phi_{d_k}}^*$ is Gaussian; for V_{φ_k} and $V_{\phi_{dk}}$ is Beta; for the concentration parameters M_1 and M_{2d} is Gamma. In addition, we also use Gibbs sampler to update $\boldsymbol{\beta}_i$, $\boldsymbol{b}_i$ and $\sigma_{\epsilon_i}^2$. The form of these conditional distributions are all Gaussian. See more details in Appendix B.1 and B.2.

For the remaining variables, the conditional distributions cannot be derived in closed form. Thus we use a Metropolis-Hastings sampler. Particularly, for the unconstrained variables $\alpha_{1_{id}}$ and γ_{id}, we choose the proposal distribution to be normal with standard deviation 0.3. For the positive variables σ_{ui}^2 and $\sigma_{\epsilon'_d}^2$, we choose the *truncated normal* (dividing the p.d.f of normal by its c.d.f.) with standard deviation 0.3. For variables ρ_{ui}, $\alpha_{2_{id}}, \alpha_{3_{id}}$ and $b_{id}/g(\gamma_{id})$ that are constrained in $(0, 1)$, we choose normal with standard deviation 0.15 to propose a new sample and then reflect it by 0 or 1 to map it back onto its support. See Appendix B.3 for details. In our experiments discussed in Section 4, these proposal distributions performed well (with acceptance rates between 0.14 - 0.22 across the chains).

4. Numerical Analysis

We evaluate the proposed model on the task of estimating treatment response curves using an observational health dataset. Specifically, we focus on the population of patients with sepsis, a deathly adverse event. We estimate the responses for different treatments used to manage creatinine, a measure used for monitoring kidney deterioration that is a common symptom of sepsis. Intermittent hemodialysis (IHD) and continuous renal replacement therapy (CRRT), the latter prescribed at three different levels, are the four main treatment choices for managing kidney function.

We fit the models on electronic health record data from patients admitted to the Beth Israel Deaconess Medical Center in Boston. The data are publicly available in the MIMIC-II Clinical Database (Saeed et al., 2002). To identify those with sepsis, we used the criteria described in Henry et al. (2015). We use expert guidance to set the window sizes to be sufficiently long to capture the full response curve. For example, the effect of IHD typically occurs over a few hours and therefore, we set $W_{IHD} = 1$ hr. Decisions about CRRT prescriptions are made on a daily, rather than hourly basis, and therefore we set $W_{CRRT} = 24$ hrs.

The creatinine data contains timeseries from 123 individuals with average duration of 20.75 days and a total of 6,992 observations. These were obtained from the MIMIC2 database. Series containing atleast 50 creatinine measurements were included, within which we observe a total of 56, 107, 1238 and 289 instances of IHD and CRRT with the three dose levels of < 500 ml/hr, $= 500$ ml/hr and > 500 ml/hr. Data were standardized by the population mean 2.75 and standard deviation 17.32.

Baselines. We refer to the individualized treatment response model as ITR, and compare its performance to three baselines. First, we evaluate against what we refer to as the *pop* model, which implements estimating treatment responses at the population level and variations across individuals are not accounted for. This is an instance of ITR where the baseline progression and the treatment response (transformed) parameters are drawn uniformly from a broad prior. To evaluate the extent to which individualizing the treatment response estimates is important, we also compare against a second baseline where the model parameters are drawn independently from a broad prior so that each individual samples it's own set of parameters. We refer to this as the *individual* model. At last, we compare against a third baseline where the model parameters are drawn from a DP instead of a

DPM. This allows treatment responses to vary by subgroups but there is no explicit representation for differences across individuals within a subgroup. Hereon, we refer to this as the *sub-pop* baseline.

Experimental setup. We use prediction error on a held-out test set to compare the proposed model to alternatives. We use the first 50 observations of the measurements for training and the remainder for measuring prediction error in the test set. Predictions of the creatinine measurements are made under the treatment strategy prescribed in the test set. We run 4 randomly initialized chains each for 5000 iterations, and report the root mean squared error (RMSE) averaged across the last 2500 iterations for all 4 chains.

We assume $b(\boldsymbol{X}_i)$ to be a linear model, i.e., $b(\boldsymbol{X}_i) = \boldsymbol{X}_i \boldsymbol{\beta}_i$, and assign a non-informative normal-inverse-Wishart (NIW) prior for the $\boldsymbol{\beta}$'s as $\beta^0 = \mathbf{0}, \kappa^0 = 1, \nu^0 = p + 2, \boldsymbol{S}^0 = I_p$. We condition upon only $p = 2$ observed covariates, age and admitted weight, which along with the individual-specific intercept turn out to be sufficient to estimate the patient's initial value of creatinine. The longitudinal measurements of creatinine are then modeled by the time-dependent progression $\boldsymbol{u}_i(t_{ij})$ and the parametric treatment-response function $\boldsymbol{f}(t_{ij}; \mathcal{H}_{ij})$ that were defined in Section 2. Within the treatment response model, the prior for the change point γ is set to be the normal distribution with a wide-range and mean chosen by the domain expert based on the expected duration over which the treatment takes effect. Specifically, these were set as N(1 hr,100 hrs) for IHD, N(10 hrs, 100 hrs) for all three levels of CRRT. Similarly, for the maximum effect, the priors were chosen to be normal with the mean guided by the domain expert. Specifically, for α_1's, the means were respectively set to be -2, -1, -1.5, -2 and variances to be 4 for IHD and the three levels of CRRT. Because the long-term effect b should be bounded by the peak effect in the curve, that is $b/g(\gamma) \in (0, 1)$, so we put priors directly on this constrained ratio instead of b. The base distribution for the transformed ratio logit($b/g(\gamma)$)'s and the two transformed "steepness" parameters logit(α_2)'s and logit(α_3)'s are both set to be non-informative, i.e., normal with mean logit(0.5) and variance 4 (which covers the range from (0.02, 0.98)). Finally, the prior for the noise variance $\sigma_{\epsilon i}^2$ is set to be IG(1, 1). For the window W, we set it to be 24 hours for IHD and CRRT.
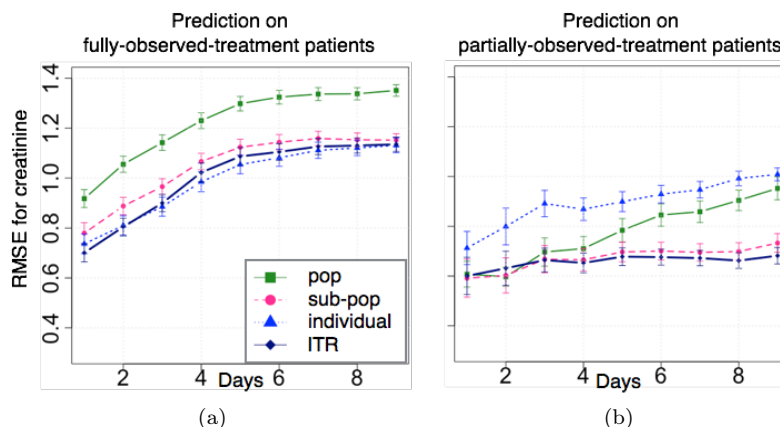

Figure 3: Prediction RMSE on creatinine for 10 future days (error bars denoting the standard errors that were calculated across the prediction errors from the 123 individuals).

Results. In Figure 3, we report held-out prediction error. We report errors averaged within a day for 10 future days following the time of prediction. To evaluate the flexibility and expressivity of ITR on modeling the treatment response, the left panel reports the errors from those patients whose treatment types prescribed in the test data were all observed in the training data, and the right panel reports the errors from those of whom at least one treatment type in the test data was not observed in the training data. In Figure 3a, we see that ITR's performance is as good as the *individual* model. It outperforms the *sub-pop* model significantly because it is more expressive since it allows individual-level heterogeneity and borrowing strength across individuals in the same group.

In Figure 3b, we see the *individual* model performs the worst, even worse than *pop* model, because the lack of subgroup structure makes the *individual* model statistically less efficient. Overall, ITR performs similar to the subpop model when little data is available on the individual, and as more data begins to accrue its performance improves. When sufficient data are available to estimate individual-specific parameters reliably, the *individual* model begins to perform similar to ITR.

In Figure 4, we show trajectory predictions for a randomly chosen patient (ID number 43) under ITR and the *pop* model. Starting with the left panel, the first 4 black points are observations in the training set and the rest red points are unseen measurements in the test set. Prescriptions of IHD are shown with vertical dashed lines. The ribbons denote the 95% confidence interval for the prediction. We see that ITR has a better short-term prediction with tighter confidence interval than *pop*. From left to right, as we observe more evidence (including both the exposure of creatinine measurements and the prescribed treatment strategy), both models predict better with higher confidence whereas *pop* is less accurate and suffers from high variance compared to the estimates obtained from ITR.
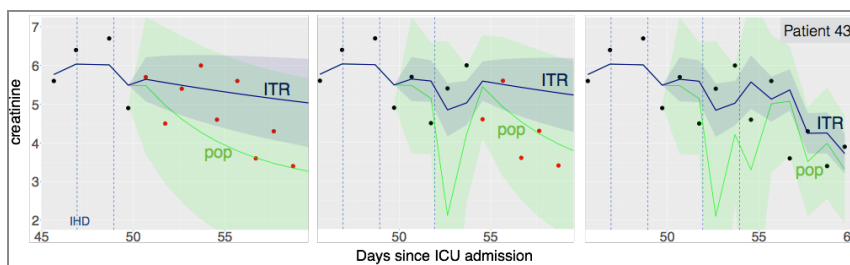

Figure 4: Comparison of ITR vs. *pop* on predicting creatinine measurements for an example patient: black and red points are observed and unseen measurements respectively. IHD prescription appears as vertical dashed lines.

In Figure 5, we show the distribution over the individual-specific response curves for IHD and CRRT at the three different dose levels. These were obtained by averaging the function estimated for each individual from the last iteration in each chain. As is clear, there is significant treatment heterogeneity across all treatments.
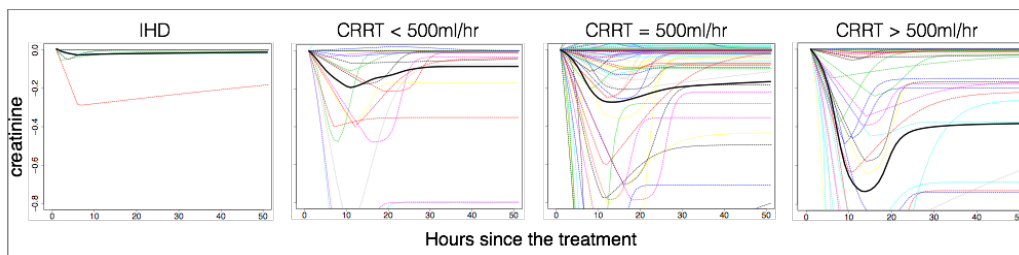

Figure 5: Treatment-response curves

5. Conclusion

In this paper, we have developed a novel Bayesian nonparametric method for estimating treatment response curves from sparse observational time series. We leverage hierarchical priors that allow individual-specific estimates while borrowing strength across individuals. Notably, we maintain the full posterior rather than just point estimates. We demonstrate significant gains in performance for modeling creatinine and effects of treatments used for managing kidney function. As future work, we plan to evaluate these models on other subpopulations with MIMIC and test sensitivity to different modeling choices. Access to accurate models for estimating treatments responses at the individual are critical for designing new personalized treatments.

References

Haitham Bou Ammar, Eric Eaton, José Marcio Luna, and Paul Ruvolo. Autonomous cross-domain knowledge transfer in lifelong policy gradient reinforcement learning. In *Proc. of IJCAI*, 2015.

Susan Athey and Guido W Imbens. Machine learning methods for estimating heterogeneous causal effects. *stat*, 1050:5, 2015.

Siddhartha Chib and Barton H Hamilton. Semiparametric bayes analysis of longitudinal data treatment models. *Journal of Econometrics*, 110(1):67–89, 2002.

Miroslav Dudik, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *ICML*, 2011.

Thomas S Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

J Foster, J Taylor, and S Ruberg. Subgroup identification from randomized clinical data. *Statistics in Medicine*, 30:2867–2880, 2010.

Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science Translational Medicine*, 7(299):299ra122–299ra122, 2015.

James Hensman, Neil D Lawrence, and Magnus Rattray. Hierarchical bayesian modelling of gene expression time series across irregularly sampled replicates and clusters. *BMC bioinformatics*, 14 (1):252, 2013.

James Hensman, Magnus Rattray, and Neil D Lawrence. Fast nonparametric clustering of structured time-series. *IEEE transactions on pattern analysis and machine intelligence*, 37(2):383–393, 2015.

Miguel Ángel Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men. *Epidemiology*, 11(5): 561–570, 2000.

Biwei Huang, Kun Zhang, and Bernhard Schölkopf. Identification of time-dependent causal model: A gaussian process treatment. In *the 24th International Joint Conference on Artificial Intelligence., Machine Learning TrackBuenos, Argentina*, pages 3561–3568, 2015.

Kosuke Imai, Marc Ratkovic, et al. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453), 2001.

Nan Jiang and Lihong Li. Doubly robust off-policy evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*, 2015.

Hui Li, Xuejun Liao, and Lawrence Carin. Multi-task reinforcement learning in partially observable stochastic environments. *The Journal of Machine Learning Research*, 10:1131–1186, 2009.

Jared K Lunceford, Marie Davidian, and Anastasios A Tsiatis. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics*, 58(1):48–57, 2002.

Peter Müller and Riten Mitra. Bayesian nonparametric inference–why and how. *Bayesian Analysis*, 8(2):269–302, 2013.

Peter Müller and Abel Rodriguez. Nonparametric bayesian inference. *IMS-CBMS Lecture Notes. IMS*, 270, 2013.

Susan A Murphy, Linda M Collins, and A John Rush. Customizing treatment to the patient: adaptive treatment strategies. *Drug and alcohol dependence*, 88(Suppl 2):S1–3, 2007a.

Susan A Murphy, Kevin G Lynch, David Oslin, James R McKay, and Tom TenHave. Developing adaptive treatment strategies in substance abuse research. *Drug and alcohol dependence*, 88: S24–S30, 2007b.

Luis E Nieto-Barajas, Alberto Contreras-Cristán, et al. A bayesian nonparametric approach for time series clustering. *Bayesian Analysis*, 9(1):147–170, 2014.

David J Olive. *Statistical theory and inference*. Springer, 2014.

Cosmin Paduraru, Doina Precup, Joelle Pineau, and Gheorghe Comanici. A study of off-policy learning in computational sustainability. In *Proceedings of the 10th European Workshop on Reinforcement Learning*. Citeseer, 2012.

Fernando A Quintana, Wesley O Johnson, Elaine Waetjen, and Ellen Gold. Bayesian nonparametric longitudinal data analysis. *Journal of the American Statistical Association*, 2015. doi: 10.1080/ 01621459.2015.1076725.

You Ren, Emily B Fox, and Andrew Bruce. Achieving a hyperlocal housing price index: Overcoming data sparsity by bayesian dynamical modeling of multiple data streams. *arXiv preprint arXiv:1505.01164*, 2015.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period - application to control of the healthy worker survivor effect. *Mathematical Modeling*, 7(9): 1393–1512, 1986.

James M Robins. Addendum to "a new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect". *Computers & Mathematics with Applications*, 14(9):923–945, 1987.

James M Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the second seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004.

James C Ross and Jennifer G Dy. Nonparametric mixture of gaussian processes with constraints. In *ICML (3)*, pages 1346–1354, 2013.

Mohammed Saeed, C Lieu, G Raber, and RG Mark. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. In *Computers in Cardiology, 2002*, pages 641–644. IEEE, 2002.

P.F. Schulam and S. Saria. A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In *Advances in Neural Information Processing Systems*, pages 748–756, 2015.

Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

Ricardo Silva. Observational-interventional priors for dose-response learning. *arXiv preprint arXiv:1605.01573*, 2016.

Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158, 2009.

Richard S Sutton, Doina Precup, and Satinder P Singh. Intra-option learning about temporally abstract actions. In *ICML*, volume 98, pages 556–564, 1998.

Sarah L Taubman, James M Robins, Murray A Mittleman, and Miguel A Hernán. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *International Journal of Epidemiology*, 38(6):1599–1611, 2009.

Lu Tian, Ash A Alizadeh, Andrew J Gentles, and Robert Tibshirani. A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532, 2014.

Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer, 2007.

Mark J van der Laan and Maya L Petersen. Causal effect models for realistic individualized treatment and intention to treat rules. *International Journal of Biostatistics*, 3(1):3, 2007.

Yanxun Xu and Yuan Ji. A latent gaussian process model with application to monitoring clinical trials. *arXiv preprint arXiv:1403.7853*, 2014.

Ying-Qi Zhao, Donglin Zeng, Eric Benjamin Laber, Rui Song, Ming Yuan, and Michael Rene Kosorok. Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151–168, 2015.

Appendix A. Transformation of Constrained Variables

The treatment-response curves were characterized using a parametric form containing constrained variables (e.g., $\alpha_2, \alpha_3 \in (0,1)$). To simplify inference, we transform the support of these variables such that they live in the real space $\mathbb{R}$ and posit (Gaussian) priors on these transformed variables. Given a random variable $X \in \mathbb{R}^d$ with continuous probability density function $f_X(x)$ and support $\mathcal{X} = \mathrm{supp}(f_X(x))$, we can define a random variable $Y \in \mathbb{R}^d$ such that $Y = T(X)$ with support $\mathcal{Y} = \mathrm{supp}(f_Y(y))$ and a one-to-one differentiable function $T : \mathcal{X} \to \mathcal{Y}$. Then based on Olive (2014), Y has the probability density function

$$f_Y(y) = f_X(T^{-1}(y))|\det J_{T^{-1}}(y)|,$$

where the adjustment term is the absolute determinant of the Jacobian:

$$J_{T^{-1}}(Y) = \begin{pmatrix} \frac{\partial T_1^{-1}}{\partial y_1} & \cdots & \frac{\partial T_1^{-1}}{\partial y_d} \\ \vdots & & \vdots \\ \frac{\partial T_d^{-1}}{\partial y_1} & \cdots & \frac{\partial T_d^{-1}}{\partial y_d} \end{pmatrix}$$

Let us first consider the univariate variable $\alpha_2 \in (0,1)$ from the g function we defined in Section 2.2. We transform it to be $\alpha_2' = \mathrm{logit}(\alpha_2)$, and posit a Gaussian prior on it. That is, $\alpha_2' \sim \mathcal{N}(\alpha_2'; \mu_{\alpha_2'}, \sigma_{\alpha_2'}^2)$. The Jacobian adjustment is calculated as $|\det J(\mathrm{logit}(\alpha_2))| = 1/\alpha_2(1 - \alpha_2)$. Thus we get the probability density function

$$p(\alpha_2) = \mathcal{N}(\mathrm{logit}(\alpha_2); \mu_{\alpha_2'}, \sigma_{\alpha_2'}^2)/\alpha_2(1 - \alpha_2).$$

Now let us consider the multivariate variable $\phi = \{\alpha_1, \alpha_2, \alpha_3, \gamma, b : \alpha_1 \in \mathbb{R}, \alpha_2, \alpha_3 \in (0,1), \gamma \in \mathbb{R}, b/g(\gamma) \in (0,1)\}$ in the g function. We define a transformation

$$\phi' = T^{-1}(\phi) = \{\alpha_1, \mathrm{logit}(\alpha_2), \mathrm{logit}(\alpha_3), \gamma, \mathrm{logit}(b/g(\gamma))\},$$

where $g(\gamma) = \alpha_1(\exp(\alpha_2\gamma/2) - 1)/(\exp(\alpha_2\gamma/2) + 1)$. Since the support of ϕ' is $\mathbb{R}^d$, we can posit a diagonal Gaussian prior $\phi' \sim \mathcal{N}(\phi'; \mu_{\phi'}, D_{\phi'})$, and calculate the Jacobian

$$J_{T^{-1}}(\phi) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1/\alpha_2(1 - \alpha_2) & 0 & 0 & 0 \\ 0 & 0 & 1/\alpha_3(1 - \alpha_3) & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ \frac{b\zeta}{\alpha_1} & \frac{b\gamma\zeta\exp(\alpha_2\gamma/2)}{\exp(\alpha_2\gamma)-1} & 0 & \frac{b\alpha_2\zeta\exp(\alpha_2\gamma/2)}{\exp(\alpha_2\gamma)-1} & \zeta \end{pmatrix}.$$

Here, $\zeta = g(\gamma)/b(g(\gamma) - b)$. Thus we obtain the adjustment $|\det J_{T^{-1}}(\phi)| = |\zeta|/\alpha_2\alpha_3(1 - \alpha_2)(1 - \alpha_3)$ and reach at the probability density function

$$p(\phi) = \mathcal{N}(T^{-1}(\phi); \mu_{\phi'}, D_{\phi'})|\zeta|/\alpha_2\alpha_3(1 - \alpha_2)(1 - \alpha_3).$$

We also have constrained parameters in the exponential kernels: $\sigma^2 \in \mathbb{R}^+$ and $\rho \in (0,1)$ (to be more precise, σ_{ui}^2, ρ_{ui} and $\sigma_{\epsilon_d'}^2, \rho_{\epsilon_d'}$). We again define transformations $\sigma'^2 = \log(\sigma^2)$ and $\rho' = \mathrm{logit}(\rho)$ and posit Gaussian priors on them. Thereafter, we get the densities $p(\sigma^2) = \mathcal{N}(\log(\sigma^2); \mu_{\sigma'}, \sigma_{\sigma'}^2)/\sigma^2$ and $p(\rho) = \mathcal{N}(\mathrm{logit}(\rho); \mu_{\rho'}, \sigma_{\rho'}^2)/\rho(1 - \rho)$ respectively.

Appendix B. Posterior inference for the individualized treatment response model

B.1 Blokced Gibbs Sampler for the DPM

We first summarize the blocked Gibbs sampler for general DPMs, and then apply it specifically to φ_i and ϕ_i in our model.

Given a sufficiently large K, the mixture component parameters $\boldsymbol{\theta}^* = \{\theta_1^*, ..., \theta_K^*\}$, the stick breaking variables $\boldsymbol{V} = \{V_1, ..., V_{K-1}, V_K = 1\}$ and the component indicators $\boldsymbol{Z} = \{Z_1, ..., Z_N\}$ for the N observations $\boldsymbol{o} = \{o_1, ..., o_N\}$, the truncated stick-breaking representation of DPM is written as follows.

$$p(o_n|\pi_k, \theta_k^*) = \sum_{k=1}^{K} \pi_k p(o_n|\theta_k^*) \tag{11}$$

$$\pi_k = V_k \prod_{j<k}(1 - V_j), \text{ for } k = 1, ..., K$$

$$V_k \sim \text{Beta}(1, M), \text{ for } k = 1, ..., K - 1$$

Then the blocked Gibbs sampler is formulated by the following steps.

1. Independently sample θ_k^* from $p(\theta_k^*|\boldsymbol{Z}, \boldsymbol{V}, \boldsymbol{o}) \propto G_0(\theta_k^*) \prod_{n=1}^{N} p(o_n|\theta_k^*)\mathbf{1}_{\{Z_n=k\}}$;

2. Independently sample v_k from $p(V_k|\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{o}) = \text{Beta}(1 + n_k, M + \sum_{j=k+1}^{K} n_j)$, where n_j is the number of observations in cluster j;

3. Independently sample z_n from $p(Z_n = k|\boldsymbol{V}, \boldsymbol{\theta}, \boldsymbol{o}) = \pi_k p(o_n|\theta_k)$, where $\pi_k = V_k \prod_{j<k}(1 - V_j)$.

Note that step 1 can be derived in closed form if the base distribution G_0 is chosen to be conjugate—a choice we make in (6). Now let us specify the samplers for the DPM parameters in our model.

We first describe the steps of sampling DP mixtures for $\boldsymbol{\varphi}_i$'s. Suppose K_1 is the truncation level we assume for the baseline progression. Denote the mixture component hyperparameters as $\boldsymbol{\theta}_\varphi^* = \{\boldsymbol{\beta}_b^*, \Sigma_b^*, \boldsymbol{\mu}_{\sigma_u'}^*, \boldsymbol{\mu}_{\rho_u'}^*\}$, where $\boldsymbol{\beta}_b^* = \{\beta_{b_1}^*, ..., \beta_{b_{K_1}}^*\}$, $\Sigma_b^* = \{\Sigma_{b_1}^*, ..., \Sigma_{b_{K_1}}^*\}$, $\boldsymbol{\mu}_{\sigma_u'}^* = \{\mu_{\sigma_{u_1}'}^*, ..., \mu_{\sigma_{u_{K_1}}'}^*\}$, $\boldsymbol{\mu}_{\rho_u'}^* = \{\mu_{\rho_{u_1}'}^*, ..., \mu_{\rho_{u_{K_1}}'}^*\}$. Further, the stick breaking variables $\boldsymbol{V}_\varphi = \{V_{\varphi_1}, ..., V_{\varphi_{K_1-1}}, V_{\varphi_{K_1}} = 1\}$ and the component indicators $\boldsymbol{Z}_\varphi = \{Z_{\varphi_1}, ..., Z_{\varphi_I}\}$ for the parameters $\boldsymbol{\varphi} = \{\boldsymbol{\varphi}_1, .., \boldsymbol{\varphi}_I\}$.

1. Independently sample $\boldsymbol{\beta}_{b_k}^*, \Sigma_{b_k}^*$ from

$$p(\boldsymbol{\beta}_{b_k}^*, \Sigma_{b_k}^*|\boldsymbol{Z}_\varphi, \boldsymbol{V}_\varphi, \boldsymbol{\varphi}) = \text{NIW}(\boldsymbol{\beta}_{b_k}^*, \Sigma_{b_k}^*|\boldsymbol{m}_k, \kappa_k, \nu_k, \boldsymbol{S}_k)$$

$$\boldsymbol{m}_k = \frac{\kappa_0\beta_0 + \sum_{i=1}^{I}\boldsymbol{\beta}_i\mathbf{1}_{\{Z_{\varphi_i}=k\}}}{\kappa_k}$$

$$\kappa_k = \kappa_0 + \sum_{i=1}^{I}\mathbf{1}_{\{Z_{\varphi_i}=k\}}$$

$$\nu_k = \nu_0 + \sum_{i=1}^{I}\mathbf{1}_{\{Z_{\varphi_i}=k\}}$$

$$\boldsymbol{S}_k = \boldsymbol{S}_0 + \sum_{i=1}^{I}\boldsymbol{\beta}_i\boldsymbol{\beta}_i^T\mathbf{1}_{\{Z_{\varphi_i}=k\}} + \kappa_0\beta_0\beta_0^T - \kappa_k\boldsymbol{m}_k\boldsymbol{m}_k^T;$$

2. Independently sample $\boldsymbol{\mu}_{\sigma_{u_k}'}^*$ from

$$p(\boldsymbol{\mu}_{\sigma_{u_k}'}^*|\boldsymbol{Z}_\varphi, \boldsymbol{V}_\varphi, \boldsymbol{\varphi}) = \mathcal{N}(\boldsymbol{\mu}_{\sigma_{u_k}'}^*; m_{\sigma_u'}, s_{\sigma_u'})$$

$$m_{\sigma'} = \frac{\sigma_{\sigma_{u0}'}^2\mu_{\sigma_0'} + \sigma_{\sigma_0'}^2\sum_{i=1}^{I}\log(\sigma_{ui}^2)\mathbf{1}_{\{Z_{\varphi_i}=k\}}}{\sigma_{\sigma_{u0}'}^2 + \sum_{i=1}^{I}\sigma_{\sigma_0'}^2\mathbf{1}_{\{Z_{\varphi_i}=k\}}}$$

$$s_{\sigma_u'} = \frac{\sigma_{\sigma_{u0}'}^2\sigma_{\sigma_0'}^2}{\sigma_{\sigma_{u0}'}^2 + \sum_{i=1}^{I}\sigma_{\sigma_0'}^2\mathbf{1}_{\{Z_{\varphi_i}=k\}}}$$

15

3. Independently sample $\boldsymbol{\mu}^*_{\rho'_{u_k}}$ from

$$p(\boldsymbol{\mu}^*_{\rho'_{u_k}}|\boldsymbol{Z}_\varphi, \boldsymbol{V}_\varphi, \boldsymbol{\varphi}) = \mathcal{N}(\boldsymbol{\mu}^*_{\rho'_{u_k}}; m_{\rho'_u}, s_{\rho'_u})$$

$$m_{\rho'} = \frac{\sigma^2_{\rho'_{u0}}\mu_{\rho'_0} + \sigma^2_{\rho'_0}\sum_{i=1}^{I}\text{logit}(\rho_{ui})\mathbf{1}_{\{Z_{\varphi_i}=k\}}}{\sigma^2_{\rho'_{u0}} + \sum_{i=1}^{I}\sigma^2_{\rho'_0}\mathbf{1}_{\{Z_{\varphi_i}=k\}}}$$

$$s_{\rho'_u} = \frac{\sigma^2_{\rho'_{u0}}\sigma^2_{\rho'_0}}{\sigma^2_{\rho'_{u0}} + \sum_{i=1}^{I}\sigma^2_{\rho'_0}\mathbf{1}_{\{Z_{\varphi_i}=k\}}}$$

4. Independently sample V_{φ_k} from

$$p(V_{\varphi_k}|\boldsymbol{Z}_\varphi, \boldsymbol{\theta}^*_\varphi, \boldsymbol{\varphi}) = \text{Beta}(1 + n_{1k}, M_1 + \sum_{j=k+1}^{K_1} n_{1j}),$$

where n_{1j} is the number of $\boldsymbol{\varphi}_i$'s that were assigned to cluster j;

5. Independently sample Z_{φ_i} from

$$p(Z_{\varphi_i} = k|\boldsymbol{V}_\varphi, \boldsymbol{\theta}^*_\varphi, \boldsymbol{\varphi})$$
$$= \omega_{1k}\mathcal{N}(\boldsymbol{\beta}_i; \boldsymbol{\beta}^*_{b_k}, \Sigma^*_{b_k})\mathcal{N}(\log(\sigma^2_{u_i}); \mu^*_{\sigma'_{u_k}}, \sigma^2_{\sigma'_{u0}})\mathcal{N}(\text{logit}(\rho_{u_i}); \mu^*_{\rho'_{u_k}}, \sigma^2_{\rho'_{u0}})/\sigma^2_{u_i}(1-\rho_{u_i})^2,$$

where $\omega_{1k} = V_{\varphi_k}\prod_{j<k}(1 - V_{\varphi_j})$.

Now we describe the steps of sampling DP mixtures for $\boldsymbol{\phi}_{id}$'s. Let K_{2d} be the truncation level assumed for the DPM prior on the dth treatment-response for $(d = 1, ..., D)$. Denote the mixture component hyperparameters as $\boldsymbol{\theta}^*_{\phi_d} = \{\boldsymbol{\mu}^*_{\phi'_d}\}$, where $\boldsymbol{\mu}^*_{\phi'_d} = \{\boldsymbol{\mu}^*_{\phi'_{d1}}, ..., \boldsymbol{\mu}^*_{\phi_{dK_{2d}}}\}$. Further, the stick breaking variables $\boldsymbol{V}_{\phi_d} = \{V_{\phi_{d1}}, ..., V_{\phi_{dK_{2d}-1}}, V_{\phi_{dK_{2d}}} = 1\}$ and the component indicators $\boldsymbol{Z}_{\phi_d} = \{Z_{\phi_{1d}}, ..., Z_{\phi_{Id}}\}$ for the parameters $\boldsymbol{\phi}_d = \{\boldsymbol{\phi}_{1d}, .., \boldsymbol{\phi}_{Id}\}$.

6. Independently sample $\boldsymbol{\mu}^*_{\phi'_d}$ from

$$p(\boldsymbol{\mu}^*_{\phi'_d}|\boldsymbol{Z}_{\phi_d}, \boldsymbol{V}_{\phi_d}, \boldsymbol{\phi}_d) = \mathcal{N}(\boldsymbol{\mu}^*_{\phi'_d}; \boldsymbol{m}_{\phi'_d}, \boldsymbol{S}_{\phi'_d})$$

$$\boldsymbol{S}_{\phi'_d} = (\boldsymbol{D}^{-1}_{0_d} + \boldsymbol{D}^{-1}_{\phi'_0}\sum_{i=1}^{I}\mathbf{1}_{\{Z_{\phi_{id}}=k\}})^{-1}$$

$$\boldsymbol{m}_{\phi'_d} = \boldsymbol{S}_{\phi'_d}(\boldsymbol{D}^{-1}_{0_d}\boldsymbol{\mu}_{0_d} + \boldsymbol{D}^{-1}_{\phi'_0}\sum_{i=1}^{I}T^{-1}(\boldsymbol{\phi}_{id})\mathbf{1}_{\{Z_{\phi_{id}}=k\}}),$$

where $T^{-1}(\boldsymbol{\phi}_{id}) = \{\alpha_{1_{id}}, \text{logit}(\alpha_{2_{id}}), \text{logit}(\alpha_{3_{id}}), \gamma_{id}, \text{logit}(b_{id}/g(\gamma_{id}))\}$;

7. Independently sample $V_{\phi_{dk}}$ from

$$p(V_{\phi_{dk}}|\boldsymbol{Z}_{\phi_d}, \boldsymbol{\theta}^*_{\phi_d}, \boldsymbol{\phi}_d) = \text{Beta}(1 + n_{2dk}, M_{2d} + \sum_{j=k+1}^{K_{2d}} n_{2dj}),$$

where n_{2dj} is the number of $\boldsymbol{\phi}_{id}$'s that were assigned to cluster j;

8. Independently sample $Z_{\phi_{id}}$ from

$$p(Z_{\phi_{id}} = k|\boldsymbol{V}_{\phi_d}, \boldsymbol{\theta}^*_{\phi_d}, \boldsymbol{\phi}_d) = \omega_{2dk}\mathcal{N}(T^{-1}(\boldsymbol{\phi}_{id}); \boldsymbol{\mu}^*_{\phi'_{dk}}, \boldsymbol{D}_{\phi'_0})|\zeta_{id}|/(1-\alpha_{2_{id}})^4,$$

where $\omega_{2dk} = V_{\phi_{dk}}\prod_{j<k}(1 - V_{\phi_{dj}})$.

B.2 Gibbs Sampler for the Variables with Conjugate Priors

9. Independently sample $\boldsymbol{\beta}_i$ from

$$p(\boldsymbol{\beta}_i | \boldsymbol{\theta}^*_\varphi, Z_{\varphi_i}, \sigma^2_{\epsilon_i}, \boldsymbol{b}_i, \boldsymbol{f}_i) = \mathcal{N}(\boldsymbol{\beta}_i; \boldsymbol{m}_{b_i}, \boldsymbol{S}_{b_i})$$
$$\boldsymbol{S}_{b_i} = (\Sigma^{*-1}_{bZ_{\varphi_i}} + \sigma^{-2}_{\epsilon_i} \boldsymbol{X}_i \boldsymbol{X}_i^T)^{-1}$$
$$\boldsymbol{m}_{b_i} = \boldsymbol{S}_{b_i} (\sigma^{-2}_{\epsilon_i} \boldsymbol{X}_i \boldsymbol{Y}_{b_i} + \Sigma^{*-1}_{bZ_{\varphi_i}} \boldsymbol{\mu}^*_{bZ_{\varphi_i}}),$$

where $\boldsymbol{X}_i = \{\boldsymbol{X}_{i1}, ..., \boldsymbol{X}_{iJ_i}\}^T$ is a $J_i \times p$ matrix, $\boldsymbol{Y}_{b_i} = \boldsymbol{Y}_i - \boldsymbol{b}_i - \boldsymbol{f}_i$ is a $1 \times J_i$ vector and $\boldsymbol{f}_i$ is defined in Eq. (2). $\Sigma^{*-1}_{bZ_{\varphi_i}}$ and $\boldsymbol{\mu}^*_{bZ_{\varphi_i}}$ are sampled in Step 1.

10. Independently sample $\boldsymbol{b}_i$ from

$$p(\boldsymbol{b}_i | \boldsymbol{\beta}_i, \boldsymbol{\varphi}_i, \sigma^2_{\epsilon_i}, \boldsymbol{f}_i) = \mathcal{N}(\boldsymbol{b}_i; \boldsymbol{m}_{u_i}, \boldsymbol{S}_{u_i})$$
$$\boldsymbol{S}_{u_i} = \left(\mathcal{K}_u^{-1}(t_{ij}, t_{ij'}; \boldsymbol{\varphi}_i) + \sigma^{-2}_{\epsilon_i} \boldsymbol{I}_{J_i}\right)^{-1}$$
$$\boldsymbol{m}_{u_i} = \sigma^{-2}_{\epsilon_i} \boldsymbol{S}_{u_i} \boldsymbol{Y}_{u_i},$$

where $\boldsymbol{Y}_{u_i} = \boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_i - \boldsymbol{f}_i$ is a $1 \times J_i$ vector.

11. Independently sample $\sigma^2_{\epsilon_i}$ from

$$p(\sigma^2_{\epsilon_i} | \boldsymbol{\beta}_i, \boldsymbol{b}_i, \boldsymbol{f}'_i) = \mathrm{IG}(s_\epsilon + J_i/2, \nu + \boldsymbol{Y}_{e_i} \boldsymbol{Y}_{e_i}^T/2),$$

where $\boldsymbol{Y}_{e_i} = \boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_i - \boldsymbol{b}_i - \boldsymbol{f}'_i$ is a $1 \times J_i$ vector and the auxiliary variable $\boldsymbol{f}'_i$ is sampled from

$$p(\boldsymbol{f}'_i | \boldsymbol{\beta}_i, \boldsymbol{\phi}_i, \boldsymbol{\sigma}^2_f, \boldsymbol{\rho}^2_f, \sigma^2_{\epsilon_i}, \boldsymbol{b}_i) = \mathcal{N}(\boldsymbol{f}'_i; \boldsymbol{m}_{f_i}, \boldsymbol{S}_{f_i})$$
$$\boldsymbol{S}_{f_i} = \left(\mathcal{K}_f^{-1}(t_{ij}, t_{ij'}; \boldsymbol{\sigma}^2_f, \boldsymbol{\rho}^2_f) + \sigma^{-2}_{\epsilon_i} \boldsymbol{I}_{J_i}\right)^{-1}$$
$$\boldsymbol{m}_{f_i} = \boldsymbol{S}_{f_i} \left(\sigma^{-2}_{\epsilon_i} \boldsymbol{Y}_{f_i} + \mathcal{K}_f^{-1}(t_{ij}, t_{ij'}; \boldsymbol{\sigma}^2_f, \boldsymbol{\rho}^2_f) m(\boldsymbol{t}_i; \boldsymbol{\phi}_i)\right),$$

where $\boldsymbol{Y}_{f_i} = \boldsymbol{Y}_i - \boldsymbol{X}_i \boldsymbol{\beta}_i - \boldsymbol{b}_i$ is a $1 \times J_i$ vector.

12. Sample M_1 from

$$p(M_1 | \eta_1, k_1) \sim \frac{c_1 + k_1 - 1}{c_1 + k_1 - 1 + I(d_1 - \log(\eta_1))} \mathrm{Gamma}(c_1 + k_1, d_1 - \log(\eta_1))$$
$$+ \frac{I(d_1 - \log(\eta_1))}{c_1 + k_1 - 1 + I(d_1 - \log(\eta_1))} \mathrm{Gamma}(c_1 + k_1 - 1, d_1 - \log(\eta_1)),$$

where the auxiliary variable $\eta_1 \sim \mathrm{Beta}(M_1 + 1, I)$, the prior for M_1 is $\mathrm{Gamma}(c_1, d_1)$, and k_1 is the current cluster number for $\boldsymbol{\varphi}_i$'s.

13. Independently sample M_{2d} from

$$p(M_{2d} | \eta_{2d}, k_{2d}) \sim \frac{c_{2d} + k_{2d} - 1}{c_{2d} + k_{2d} - 1 + I(d_{2d} - \log(\eta_{2d}))} \mathrm{Gamma}(c_{2d} + k_{2d}, d_{2d} - \log(\eta_{2d}))$$
$$+ \frac{I(d_{2d} - \log(\eta_{2d}))}{c_{2d} + k_{2d} - 1 + I(d_{2d} - \log(\eta_{2d}))} \mathrm{Gamma}(c_{2d} + k_{2d} - 1, d_{2d} - \log(\eta_{2d})),$$

where the auxiliary variable $\eta_{2d} \sim \mathrm{Beta}(M_{2d} + 1, I)$, the prior for M_{2d} is $\mathrm{Gamma}(c_{2d}, d_{2d})$, and k_{2d} is the current cluster number for $\boldsymbol{\phi}_{id}$'s.

B.3 Metropolis-Hastings Sampler in the Non-Conjugate Case

We use blocked Metropolis-Hastings to sample the remaining parameters i.e., parameters for which we cannot obtain the conditional distributions in closed-form: $\sigma_{u_i}^2$, ρ_{u_i}, $\sigma_{\epsilon'_d}^2$, $\rho_{\epsilon'_d}$, and ϕ_{id}'. Specifically, for a variable $\boldsymbol{x}$, we propose a candidate value $\boldsymbol{x}^{\mathrm{cand}}$ from a proposal distribution $p(\boldsymbol{x}^{\mathrm{cand}}|\boldsymbol{x})$ and accept the candidate with probability

$$\min\{1, \frac{\pi(\boldsymbol{x}^{\mathrm{cand}})p(\boldsymbol{x}|\boldsymbol{x}^{\mathrm{cand}})}{\pi(\boldsymbol{x})p(\boldsymbol{x}^{\mathrm{cand}}|\boldsymbol{x})}\},$$

where $\pi(\cdot)$ is the full joint posterior defined in Eq. (10). Below, we choose different proposal distributions for x for the following three different types of support.

- For $x \in \mathbb{R}$, we propose new sampler from $\mathcal{N}(x, 0.3^2)$, which is a symmetric proposal distribution.

- For $x \in \mathbb{R}^+$, we propose new sampler from $\mathcal{N}(x, 0.3^2)/\Phi(x, 0.3^2)$, where Φ is the CDF of the normal distribution. This is not a symmetric proposal distribution.

- For $x \in (0, 1)$, we propose a new sampler from $\mathcal{N}(x, 0.15^2)$ and "reflect" it by 0 or 1 to make it fall back in $(0, 1)$. This is so-called "reflected normal", and the reflection can be done multiple times if needed. It is still a symmetric proposal distribution.

We experimented with a few different choice of values for the variance parameter in the proposal distribution. The values selected above yielded reasonable acceptance rates in the range of $0.14 - 0.22$.

In detail, the sampling for the remainder of the parameters proceeds as follows.

14. Propose $\sigma_{u_i}^{2\mathrm{cand}} \sim \mathcal{N}(\sigma_{u_i}^2, 0.3^2)/\Phi(\sigma_{u_i}^2, 0.3^2)$, $\rho_{u_i}^{\mathrm{cand}} \sim \mathcal{N}(\rho_{u_i}, 0.15^2)$ an reflect $\rho_{u_i}^{\mathrm{cand}}$ into $(0, 1)$. We accept the proposal with probability of $\min\{1, \frac{\pi(\sigma_{u_i}^{2\mathrm{cand}}, \rho_{u_i}^{\mathrm{cand}})\Phi(\sigma_{u_i}^2)}{\pi(\sigma_{u_i}^2, \rho_{u_i})\Phi(\sigma_{u_i}^{2\mathrm{cand}})}\}$, where $\pi(\sigma_{u_i}^2, \rho_{u_i})$ is

$$\mathcal{N}\big(\boldsymbol{Y}_{u_i}; \boldsymbol{0}, \mathcal{K}_u(\boldsymbol{\beta}_i, \sigma_{u_i}^2, \rho_{u_i}) + \sigma_{\epsilon_i}^2 \boldsymbol{I}_{J_i}\big)\mathcal{N}\big(\log(\sigma_{u_i}^2); \mu_{\sigma'_{ui}}, \sigma_{\sigma'_{u0}}^2\big)\mathcal{N}\big(\mathrm{logit}(\rho_{u_i}); \mu_{\rho'_{ui}}, \sigma_{\rho'_{u0}}^2\big)/\sigma_{u_i}^2(1 - \rho_{u_i})^2.$$

15. Propose $\boldsymbol{\sigma}_{\epsilon'}^{2\mathrm{cand}} \sim \mathcal{N}(\boldsymbol{\sigma}_{\epsilon'}^2, 0.3^2\boldsymbol{I}_D)/\Phi(\boldsymbol{\sigma}_{\epsilon'}^2, 0.3^2\boldsymbol{I}_D)$ and $\boldsymbol{\rho}_{\epsilon'}^{\mathrm{cand}} \sim \mathcal{N}(\boldsymbol{\rho}_{\epsilon'}, 0.15^2\boldsymbol{I}_D)$ and reflect $\boldsymbol{\rho}_{\epsilon'}^{\mathrm{cand}}$ into $(0, 1)^D$. We accept the proposal with probability of $\min\{1, \frac{\pi(\boldsymbol{\sigma}_{\epsilon'}^{2\mathrm{cand}}, \boldsymbol{\rho}_{\epsilon'}^{\mathrm{cand}})\Phi(\boldsymbol{\sigma}_{\epsilon'}^2, 0.3^2\boldsymbol{I}_D)}{\pi(\boldsymbol{\sigma}_{\epsilon'}^2, \boldsymbol{\rho}_{\epsilon'})\Phi(\boldsymbol{\sigma}_{\epsilon'}^{2\mathrm{cand}}, 0.3^2\boldsymbol{I}_D)}\}$, where $\pi(\boldsymbol{\sigma}_{\epsilon'}^2, \boldsymbol{\rho}_{\epsilon'})$ is

$$\prod_{i=1}^I \mathcal{N}\big(\boldsymbol{Y}_{e_i}; m(\boldsymbol{t}_i, \boldsymbol{\phi}_i), \mathcal{K}_{\epsilon'}(\boldsymbol{\sigma}_{\epsilon'}^2, \boldsymbol{\rho}_{\epsilon'}) + \sigma_{\epsilon_i}^2 \boldsymbol{I}_{J_i}\big) \prod_{d=1}^D \mathcal{N}\big(\log(\sigma_{\epsilon'_d}^2); \mu_{\epsilon'_1}, \sigma_{\epsilon'_1}^2\big)$$
$$\mathcal{N}\big(\mathrm{logit}(\rho_{\epsilon'_d}); \mu_{\epsilon'_2}, \sigma_{\epsilon'_2}^2\big)/\sigma_{\epsilon'_d}^2(1 - \rho_{\epsilon'_d})^2.$$

16. Propose $\{\alpha_{1_{id}}^{\mathrm{cand}}, \alpha_{2_{id}}^{\mathrm{cand}}, \alpha_{3_{id}}^{\mathrm{cand}}, \gamma_{id}^{\mathrm{cand}}, b_{id}^{\mathrm{cand}}/g(\gamma_{id}^{\mathrm{cand}})\} \sim \mathcal{N}(\{\alpha_{1_{id}}, \alpha_{2_{id}}, \alpha_{3_{id}}, \gamma_{id}, b_{id}/g(\gamma_{id})\}$, $\mathrm{Diag}(0.3^2, 0.15^2, 0.15^2, 0.3^2, 0.15^2))$, $\alpha_{2_{id}}^{\mathrm{cand}} \sim \mathcal{N}(\alpha_{1_{id}}, 0.3^2)$ and reflect $\alpha_{2_{id}}^{\mathrm{cand}}, \alpha_{3_{id}}^{\mathrm{cand}}$ and $b_{id}^{\mathrm{cand}}/g(\gamma_{id}^{\mathrm{cand}})$ into $(0, 1)$ individually. We accept the proposal with probability of $\min\{1, \frac{\pi(\boldsymbol{\phi}_i^{\mathrm{cand}})}{\pi(\boldsymbol{\phi}_i)}\}$, where $\pi(\boldsymbol{\phi}_i)$ is

$$\mathcal{N}\big(\boldsymbol{Y}_{e_i}; m(\boldsymbol{t}_i, \boldsymbol{\phi}_i), \mathcal{K}_f(\boldsymbol{\sigma}_{\epsilon'}^2, \boldsymbol{\rho}_{\epsilon'}) + \sigma_{\epsilon_i}^2 \boldsymbol{I}_{J_i}\big) \prod_{d=1}^D \mathcal{N}\big(T^{-1}(\phi_{id}); \boldsymbol{\mu}_{\phi'_{id}}, \boldsymbol{D}_{\phi'_0}\big)|\zeta_{id}|/(1 - \alpha_{2_{id}})^4.$$