Published in final edited form as:

J Am Stat Assoc. 2016 ; 111(515): 936–942. doi:10.1080/01621459.2016.1200911.

Comment

Qian Guan, Eric B. Laber, and Brian J. Reich

Department of Statistics, North Carolina State University, Raleigh, NC, USA

1. Introduction

We congratulate Xu, Muller, Wahed, and Thall (hereafter, XMWT) on an interesting and novel article and we thank the Editors for organizing these discussions. The integration of non-parametric Bayesian (NPB) methods and optimal treatment regimes was long overdue; we expect that the work of XMWT will be the first in a long and fruitful line of research. There is currently a great amount of interest in non-parametric methods for estimation of optimal treatment regimes; for example, recently proposed methodology includes kernel-based methods (Zhao et al. 2009, 2012, 2015, 2014; Zhou et al. 2015), nearest-neighbor methods (Zhou and Kosorok 2016), generalized additive models (Moodie, Dean, and Sun 2013), boosting (Kang, Janes, and Huang 2014), and trees (Laber and Zhao 2015; Doove et al. 2015; Zhang et al. 2015). This interest stems in part from a desire to mitigate the risk of model misspecification as it is well-known that the optimal treatment regime can be a highly nonlinear function of patient covariates even under simple generative models (Laber, Linn, and Stefanski 2014; Schulte et al. 2014). NPB methods are in line with this trend and, as we discuss below, may possess a number of advantages over existing methods in the context of policy-search algorithms.

However, nonparametric estimation of an optimal treatment regime often comes at the price of a loss of interpretability within a domain context which can be a major detriment to scientific progress, especially if estimation is done a secondary, hypothesis-generating analysis. In this discussion, we argue that NPB methods can have tremendous value as an engine for policy-search algorithms used to estimate an optimal treatment regime within a prespecified class (Robins, Orellana, and Rotnitzky 2008; Orellana, Rotnitzky, and Robins 2010; Zhao et al. 2012; Zhang et al. 2012, 2013; Zhao et al. 2014, 2015). An advantage of policy-search methods is that the prespecified class can be chosen to ensure parsimony and interpretability, etc. However, applying NPB methods for this purpose is challenging because of the curse of dimensionality, we expect that the authors will have some insight into this issue.

In Section 2, we discuss whether NPB methods are necessary if the primary goal is to evaluate the marginal mean outcome under a small number of fixed regimes and compare the method of XMWT with a flexible accelerated failure time model (Wahed and Thall

CONTACT Eric Laber, eblaber@ncsu.edu, Department of Statistics, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695-8203.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

2013). In Section 3, we describe a schematic for NPB methods for policy-search with sequential transition times. We make concluding remarks and discuss directions for future research in Section 4.

2. Nonparametric Bayesian Methods for Evaluating Fixed Treatment Regimes

XMWT use nonparametric Bayesian (NPB) methods to evaluate a finite set of fixed treatment regimes. This requires estimating a series of densities conditional on a patient's covariate and outcome trajectory. Thus, to estimate the marginal mean outcome under each fixed regime, they must take the intermediate step of performing conditional density estimation; this density can be high-dimensional, especially if one desires to incorporate accumulating longitudinal patient information to individualize multi-stage treatment decisions (see Section 3). Furthermore, as XMWT illustrate, implementation of NPB methods for evaluating fixed treatment regimes is highly non-trivial. In contrast, existing regression-based or inverse-probability weighting estimators of the marginal mean outcome under a fixed treatment regime involve minimal modeling and computational burden (Zhang et al. 2012, 2013). However, inverse-weighting methods can be highly variable especially if the number of treatment combinations is large and the sample size is small; and parametric regression-based estimators can be highly biased if the regression model is misspecified. NPB methods seem to provide a nice compromise in that modeling the conditional densities introduces enough structure to reduce variability yet the class of densities is sufficiently rich to avoid severe misspecification. A similar, but much simpler compromise, is to use a flexible regression-based estimator to estimate the marginal mean outcome under each fixed regime (this approach was advocated by Taylor, Cheng, and Foster 2015).

2.1. A Simple Semiparametric Model

To illustrate the use of a flexible regression-based estimator, we apply a semiparametric accelerated failure time model based on Wahed and Thall (2013); using the notation of XMWT, this model postulates that the transition T (the model is applied separately for each arm of the study) follows the additive model

$$\log(T) = \alpha + \sum_{l=1}^p b_l(x_l) + \varepsilon, \quad (1)$$

where $x = (x_1, \dots, x_p)$ are the covariates, b_l are smooth functions; and ε is an independent error term with smooth density. For binary covariates (e.g., treatment indicators) we take $b_l(x_l) = x_l \beta_l$; for continuous covariates we model b_l using M b-spline basis functions,

$b_l(u) = \sum_{m=1}^M \psi_m(u) \gamma_{lj}$, where $\psi_1, \dots, \psi_M$ are a fixed b-splines basis functions and γ_{lj} are unknown coefficients. The density of ε is assumed to be a finite mixture of J normals,

$f(u) = \sum_{j=1}^J p_j \phi(u; \mu_j, \sigma_j^2)$, where ϕ is the Gaussian density function. For illustration, we fix $J = M = 5$ and use the following priors:

$\alpha, \beta_j \stackrel{\text{iid}}{\sim} \text{Normal}(0, 100^2); \gamma_{lj} \stackrel{\text{iid}}{\sim} \text{Normal}(0, \sigma_2^2); \mu_j \stackrel{\text{iid}}{\sim} \text{Normal}(0, \sigma_3^2); (p_1, \dots, p_J) \sim \text{Dirichlet}(1, \dots, 1); \text{ and } \sigma_k^2 \stackrel{\text{iid}}{\sim} \text{InvGamma}(0.1, 0.1)$. Censoring is handled using standard data augmentation methods (Tanner and Wong 1987), and given the complete imputed dataset all parameters have conjugate full conditional distributions leading to straightforward Gibbs sampling. We generate 5000 MCMC samples and discard the first 2000 as burn-in. For the three-stage clinical trial data described below, this requires around 20 sec on a standard PC (as opposed to approximately 50 min for the fully nonparametric model of XMWT).

We compare this model to the DDP-GP method of XMWT in a suite of simulation experiments. To explore the effects of different types of model misspecification we fit the above model with linear (“Ln”) mean, that is, $M = 1$ and $b_j(x) = x_j$ and nonlinear (“NLn”) mean, $M = 5$ as above, and with Gaussian (“Gs”) residuals, that is, $J = 1$ and $\mu_1 = 0$, and a mixture of J Gaussian (“MGs”) densities as above. This gives four models for comparison in Figures 1, 2, and 3. For all scenarios we generate and fit 100 simulated datasets and report boxplots of the estimated sampling distribution of the difference between causal effect estimation errors as in Figure 4 of XMWT.

2.2. Setting 1: Single-Stage Additive Model

We simulated data as in Simulation 2 in XMWT. The model in (1) is applied with $x = (Z, W, L)$. As in our simplified model (1), the mean under this scenario is a nonlinear function of the covariates and the residual distribution is a mixture of normals that does not depend on the covariates. As expected, the “NLn-MGs” model with nonlinear mean and non-Gaussian errors yields the most accurate estimates (Figure 1, left). Surprisingly, even for the complex data-generation scheme the linear Gaussian model (“LnGs”) outperforms the DDP-GP.

2.3. Setting 2: Single-Stage Nonadditive Model

To explore the extent to which violation of model assumptions degrades performance, we simulated data from an even more pathological case than Setting 1. In this setting, the covariates L and W and the treatment group indicator Z are generated as in Setting 1, but the responses are generated as

$$Y(Z) = P_Z(x)N(4, 1) + [1 - P_Z(x)]N(1, 1),$$

where $\eta = -0.2L + \sqrt{L} - 0.1W$, $\text{logit}[P_0(x)] = \eta - 1$, and $\text{logit}[P_1(x)] = \eta + 2$. This model cannot be written in the form of (1) because the mean is not an additive function of the covariates and the residual distribution depends on the covariates. In this case, the DDP-GP model outperforms all four variations of (1).

2.4. Setting 3: Multistage Linear Model

We also compared the semiparametric model in (1) with the DDP-GP using data simulated from the three-stage clinical trial specified by Simulation 4 in XMWT. The semiparametric model (1) was applied separately of the each arm of the trial. In this setting all log survival time are generated as Gaussian with mean set to a linear combination of the covariates. The best-performing model in Figure 2 is indeed the multivariate regression model “LnGs,”

followed closely by the linear mean model with mixture of Gaussian residuals. The comparison of the DDP-GP to other nonlinear mean models is mixed; the DDP-GP model performs well in some cases and poorly in others.

2.5. Setting 4: Multi-Stage Nonlinear Model

In the final simulation, we considered a three-stage trial with nonlinear mean function by modifying Setting 3's assumptions about the log survival distributions as follows. We add a quadratic term in the first stage, $x_i^{(0,R)} = x_i^{(0,C)} = (1, L_i^2, Z_i^1)$, with $\beta^{(0,R)} = (2, 0.01, 0)$ and $\beta^{(0,C)} = (1.5, 0.02, -0.8)$. We also added a square root transformation (which is valid because the log survival times are positive almost surely) at the second stage

$$x_i^{(R,D)} = \left(1, L_i^2, Z_i^1, \sqrt{\log \left(T_i^{(0,R)} \right)}, Z_i^{2,1} \right)$$

$$x_i^{(C,P)} = \left(1, L_i^2, Z_i^1, \sqrt{\log \left(T_i^{(0,C)} \right)} \right)$$

$$x_i^{(P,D)} = \left(1, L_i^2, Z_i^1, \sqrt{\log \left(T_i^{(0,C)} \right)}, \log \left(T_i^{(C,P)} \right), Z_i^{2,2} \right)$$

with $\beta^{(R,D)} = (-0.5, 0.015, 0.2, 1.0, 0.3)$, $\beta^{(C,P)} = (1.0, 0.03, 1.0, -1.0)$, and $\beta^{(P,D)} = (0.8, 0.01, 1.5, -1.0, 0.7, 0.5)$. As shown in Figure 3, the full “NlnMGs” model consistently outperforms the DDP-GP model, and the multiple linear model “LnGs” also often produces smaller errors than the DDP-GP.

2.6. Summary

While the full BNP model is theoretically appealing, estimating a conditional density that varies over several covariates is a daunting task. These additional simulations illustrate that for the small datasets considered here, quite often simpler models yield faster and more stable estimates than the full BNP model. In some nonlinear and/or non-Gaussian cases, we find that even multiple linear regression outperforms the DDP-GP approach. Therefore, despite the elegance of the full BNP approach, we conclude that there is still room for standard model fitting combined with careful regression diagnostics.

3. Nonparametric Bayesian Methods for Policy-Search

Policy-search methods use a flexible model for the mean under any regime and estimate the optimal regime with the maximizer of this fitted model over a prespecified class of regimes. NPB methods are well-suited to policy-search as they permit flexible modeling of the mean outcome (or other suitable functional) and also allow Bayesian machinery to be leveraged to conduct inference.

To simplify our development, we assume a two-stage decision problem with no censoring; though the transition times are random, possibly outcome-dependent, and not all patients progress through both stages. The extension to censored data is straightforward. We assume that the observed data are $\mathcal{D} = \{(\mathbf{X}_{1,i}, A_{1,i}, T_{1,i}, \mathbf{X}_{2,i}, A_{2,i}, T_{2,i})\}_{i=1}^n$ which comprise n independent, identically distributed trajectories of the form $(\mathbf{X}_1, A_1, Y_1, \mathbf{X}_2, A_2, Y_2)$, where: $\mathbf{X}_1 \in \mathbb{R}^{p_1}$ denotes baseline patient covariates; $A_1 \in \mathcal{A}_1$ denote the first treatment actually received; $T_1 \in \mathbb{R}_+$ denotes the time until death or transition to the second treatment stage; $\mathbf{X}_2 \in \mathbb{R}^{p_2}$ denotes interim covariate information collected during the first treatment period, for example, disease progression, remission, etc.; $A_2 \in \mathcal{A}_2$ denotes the second treatment actually received; and $T_2 \in \mathbb{R}_+$ denotes time until death in the second stage. Define $\mathbf{H}_1 = \mathbf{X}_1$ and $\mathbf{H}_2 = (\mathbf{X}_1^\top, A_1, T_1, \mathbf{X}_2^\top)^\top$ so that $\mathbf{H}_t$ is the information available to the decision maker at time $t = 1, 2$.

A treatment regime is a pair of functions, $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$, one per stage of intervention, so that under $\boldsymbol{\pi}$ a patient presenting with $\mathbf{H}_t = \mathbf{h}_t$ at time t is recommended treatment $\boldsymbol{\pi}_t(\mathbf{h}_t)$. We restrict attention to regimes in a prespecified class Π which might be constrained to ensure parsimony or interpretability; we assume that all regimes are feasible (Schulte et al. 2014) so that any $\boldsymbol{\pi} \in \Pi$ satisfies $\boldsymbol{\pi}_t(\mathbf{h}_t) \in \Psi_t(\mathbf{h}_t)$ with probability one, where $\Psi_t(\mathbf{h}_t) \subseteq \mathcal{A}_t$ is the set of treatments that can be feasibly (ethically) assigned to a subject with $\mathbf{H}_t = \mathbf{h}_t$ for $t = 1, 2$. To define an optimal treatment regime we use the language of potential outcomes (Rubin 1978; Splawa-Neyman, Dabrowska, and Speed 1990). Define $\mathbf{H}_2^*(a_1)$ to be the potential second stage history under first-stage treatment a_1 , $T_2^*(a_1, a_2)$ the potential time in the second stage under treatment sequence (a_1, a_2) . Let $D^*(a_1) = D\{\mathbf{H}_2^*(a_1)\}$ denote the indicator that a patient transitions to the second stage; $T_2^*(a_1, a_2)$ is only defined if $D^*(a_1) = 1$. In concordance with XMWT, the potential outcome of interest is $T_1^*(a_1) + T_2^*(a_1, a_2)\{1 - D^*(a_1)\}$. The expected outcome under a regime $\boldsymbol{\pi}$ is

$$V(\boldsymbol{\pi}) = \mathbb{E} \left(\sum_{(a_1, a_2)} [T_1^*(a_1) + T_2^*(a_1, a_2)\{1 - D^*(a_1)\}] \times 1_{a_1 = \boldsymbol{\pi}_1(\mathbf{H}_1)} 1_{a_2 = \boldsymbol{\pi}_2\{\mathbf{H}_2^*(a_1)\}} \right).$$

The optimal regime, say $\boldsymbol{\pi}^{\text{opt}} \in \Pi$, satisfies $V(\boldsymbol{\pi}^{\text{opt}}) \geq V(\boldsymbol{\pi})$ for all $\boldsymbol{\pi} \in \Pi$.

To express $\boldsymbol{\pi}^{\text{opt}}$ in terms of the data generating model we make the following assumptions which are standard in the context of estimating optimal treatment regimes (Robins 2004; Schulte et al. 2014): (A1) consistency, $\mathbf{H}_2 = \mathbf{H}_2^*(A_1)$, $T_2 = T_2^*(A_1, A_2)$; (A2) sequential ignorability, $\{\mathbf{H}_2^*(a_1), T_2^*(a_1, a_2)\}_{(a_1, a_2) \in \mathcal{A}_1 \times \mathcal{A}_2} \perp A_t | \mathbf{H}_t$; and (A3) positivity, $P(A_t = a_t | \mathbf{H}_t = \mathbf{h}_t) > 0$ for some $\epsilon > 0$, all $a_t \in \mathcal{A}_t(\mathbf{h}_t)$, and almost all $\mathbf{h}_t$ for $t = 1, 2$. In addition, we make the stable unit treatment value assumption (SUTVA). Under the foregoing assumptions, it can be shown that for any $\boldsymbol{\pi} \in \Pi$

$$V(\boldsymbol{\pi}) = \int [t_1 + t_2(1-d)] f_{T_2|H_2, A_2}(t_2|\mathbf{h}_2, a_2) \delta_{\pi_2(\mathbf{h}_2)}(a_2) f_{H_2|A_1, H_1}(\mathbf{h}_2|a_1, \mathbf{h}_1) \delta_{\pi_1(\mathbf{h}_1)}(a_1) f_{H_1}(\mathbf{h}_1) \times d\lambda(t_2, \mathbf{h}_2, \mathbf{h}_1, a_1, a_2),$$

$$(2)$$

where λ is a dominating measure, $\delta_u(\cdot)$ denotes a point mass at u , and $f_{W|Z}(w/z)$ denotes the conditional density of W given Z , and we have used the fact that t_1, d are functions of $\mathbf{h}_2$. Given the expression for $V(\boldsymbol{\pi})$ in (2) a natural approach to estimating $\boldsymbol{\pi}^{\text{opt}}$ is to first construct estimators $\hat{f}_{T_2|\mathbf{H}_2, A_2}$, $\hat{f}_{\mathbf{H}_2|\mathbf{H}_1, A_1}$, and $\hat{f}_{\mathbf{H}_1}$ of $f_{T_2|\mathbf{H}_2, A_2}$, $f_{\mathbf{H}_2|\mathbf{H}_1, A_1}$, and $f_{\mathbf{H}_1}$ and subsequently

$$\begin{aligned} \hat{\boldsymbol{\pi}}_n^{\text{plug-in}} &= \arg \max_{\boldsymbol{\pi} \in \Pi} \int [t_1 + t_2(1-d)] \hat{f}_{T_2|\mathbf{H}_2, A_2}(t_2|\mathbf{h}_2, a_2) \delta_{\pi_2(\mathbf{h}_2)}(a_2) \\ &\quad \hat{f}_{\mathbf{H}_2|\mathbf{H}_1, A_1}(\mathbf{h}_2|a_1, \mathbf{h}_1) \delta_{\pi_1(\mathbf{h}_1)}(a_1) \hat{f}_{\mathbf{H}_1}(\mathbf{h}_1) \\ &\quad \times d\lambda(t_2, \mathbf{h}_2, \mathbf{h}_1, a_1, a_2). \end{aligned} \quad (3)$$

Of course one of the primary difficulties with this approach is estimating the requisite conditional densities which may be high-dimensional (the problem becomes even worse as the number of treatment stages increases and the dimension of the history grows). An appeal of nonparametric Bayesian (NPB) methods is the ability to impose structure on these densities by anchoring them to parsimonious parametric densities through the prior; thus, at least in principle, NPB methods permit as much flexibility as the data can shoulder. One approach would be to use a mixture of Gaussian processes prior anchored to normal (log) linear models, that is, one might postulate the following models

$$\log(T_2) = \mathbf{H}_{2,0}^\top \beta_{2,0} + A_2 \mathbf{H}_{2,1} + \delta_2(\mathbf{H}_2, A_2)$$

$$\mathbf{H}_2 = \Omega_{1,0} \mathbf{H}_1 + A_1 \Omega_{1,1} \mathbf{H}_1 + \delta_1(\mathbf{H}_1, A_1)$$

$$\log(T_1) = \mathbf{H}_{1,0}^\top \beta_{1,0} + A_1 \mathbf{H}_{1,1}^\top \beta_{1,1} + \delta_0(\mathbf{H}_1, A_1),$$

where $\mathbf{H}_{j,0}, \mathbf{H}_{j,1}$ are known features constructed from $\mathbf{H}_j$, $\beta_{2,0}, \beta_{2,1}, \Omega_{1,0}, \Omega_{1,1}, \beta_{1,0}, \beta_{1,1}$ are unknown parameters, and $\delta_2, \delta_1, \delta_0$ are errors with unspecified distributions. The distribution of $\mathbf{H}_1$ can be modeled using standard methods. Using the DDP-GP prior of XMWT for δ_j with a strong prior on the covariance kernel of each Gaussian process being identically zero corresponds to strong weighting on the (log)linear models with normal errors.

A schematic for NPB policy-search is as follows. Postulate a prior $\mathcal{G}$ for $\mathbf{f} = (f_{T_2|H_2, A_2}, \mathbf{H}_2|H_1, A_1, \mathbf{H}_1)$, and let $P(\mathbf{f}|\mathcal{D})$ denote the posterior distribution of $\mathbf{f}$. The an NPB estimator could be taken to be the posterior mode

$$\begin{aligned}\hat{\pi}_n^{\text{NPB-mode}} &= \arg \max_{\pi \in \Pi} \mathbb{E} \{V_{\mathbf{f}}(\pi) | \mathcal{D}\} \\ &= \arg \max_{\pi \in \Pi} \int \left(\int [t_1 + t_2(1-d)] f_{T_2|H_2, A_2} \right. \\ &\quad \times (t_2 | \mathbf{h}_2, a_2) \delta_{\pi_2(\mathbf{h}_2)}(a_2) \\ &\quad \times f_{H_2|A_1, H_1}(\mathbf{h}_2 | a_1, \mathbf{h}_1) \delta_{\pi_1(\mathbf{h}_1)} \\ &\quad \left. \times (a_1) f_{H_1}(\mathbf{h}_1) d\lambda(t_2, \mathbf{h}_2, \mathbf{h}_1, a_1, a_2) \right) dP(\mathbf{f} | \mathcal{D}),\end{aligned}$$

where $V_{\mathbf{f}}(\pi)$ is the marginal mean outcome computed using (2) with density $\mathbf{f}$. An advantage of this form is that it permits: (i) estimation within a prespecified class which may be restricted to be interpretable, respect cost constraints, etc.; (ii) inferences that are notoriously difficult under a frequentist paradigm; and (iii) estimation of regimes that optimize a functional other than a mean, for example, a quantile or measure of variability. For example, to illustrate (ii) suppose Π_1 is a class of regimes composed of decision rules representable as trees and Π_2 is the space of all feasible regimes, then one can compute a posterior credible set for the difference $\max_{\pi \in \Pi_2} V_{\mathbf{f}}(\pi) - \max_{\pi \in \Pi_1} V_{\mathbf{f}}(\pi)$. If this credible interval lies above zero then one may conclude that the class of regimes Π_1 is not sufficiently expressive and opt to enrich this class. In contrast, standard asymptotic approaches for constructing a confidence interval, for example, the bootstrap or series approximations, for this difference cannot be applied without modification because the marginal mean outcome $V(\pi)$ is a nonsmooth functional of the underlying generative distribution (e.g., Van Der Vaart 1991; Hirano and Porter 2009; Laber et al. 2014; Chakraborty, Laber, and Zhao 2014). It would be interesting to investigate the operating characteristics of an NPB approach to these inferential problems.

4. Discussion

NPB methods have great potential for estimation of and inference for optimal treatment regimes. Under suitable identifiability conditions, they can be used to estimate regimes that optimize essentially any smooth functional of the potential outcome distribution. Furthermore, the underlying Bayesian machinery makes it straightforward to construct credible sets for functionals for which standard frequentist inference procedures cannot be applied without modification. However, it is not apparent whether these credible sets will have good frequentist operating characteristics. Investigation of frequentist properties of NPB methods in the context of treatment regimes would be of great interest.

5. Simulation of Comparing Several Models

5.1. Single-Stage Regimes

We simulated the data under the same setting as indicated in simulation 2 in the article. There are two continuous covariates, $x_i = (L_i, W_i)$, for $n = 100$ patients, $i = 1, \dots, n$. We

generated L_i from a mixture of normals, $L_i \sim \frac{1}{2}N(40, 10^2) + \frac{1}{2}N(20, 10^2)$. We generated W_i from a uniform with zero mean and unit standard deviation, $W_i \sim \text{Unif}(-\sqrt{12}, \sqrt{12})$. We generated the treatment indicators $Z_i \in \{0 = \text{control}, 1 = \text{experimental}\}$ using the modified logistic regression model

$$p(Z_i=1|L_i, W_i) = \begin{cases} 0.05 & \text{if } \{1 + \exp[-2(L_i - 30)/10]\}^{-1} \leq 0.05 \\ 0.95 & \text{if } \{1 + \exp[-2(L_i - 30)/10]\}^{-1} \geq 0.95 \\ \{1 + \exp[-2(L_i - 30)/10]\}^{-1} & \text{otherwise} \end{cases}$$

We generated patients' responses differently for different treatments from

$$Y(1) \sim \frac{1}{2}N(3 - 0.2L + \sqrt{L} - 0.1W, \sigma) + \frac{1}{2}N(2 - 0.2L + \sqrt{L} - 0.1W, \sigma)$$

and

$$Y(0) \sim N(-0.2L + \sqrt{L} - 0.1W, \sigma)$$

with $\sigma = 0.4$. Under the simulation truth the treatment effect, $E[Y(1) - Y(0)|x = (L, W)] = 2.5$, is constant across L, W .

We implemented inference for $E(Y_i|x_i, Z_i)$ using the DDP-GP model. Let $\hat{Y}_i(z) = E(Y_{n+1}|L_{n+1} = L_i, W_{n+1} = W_i, Z_{n+1} = z, \text{data})$ denote the posterior expected response for a future

patient. The estimated average treatment effect (ATE) is defined as $\frac{1}{n} \sum_{i=1}^n [\hat{Y}_i(1) - \hat{Y}_i(0)]$.

We then applied linear regression method to the simulated data to estimate the average treatment effect. We used data of the patients who received treatment 1 to fit the linear model of treatment 1: $Y_i(1) = \beta_{10} + \beta_{11}L_i + \beta_{12}W_i + \epsilon_i$. We used data of patients with treatment 0 to fit: $Y_i(0) = \beta_{00} + \beta_{01}L_i + \beta_{02}W_i + \epsilon_i$. Denoting the least-square estimates by $\hat{\beta}_{zj}$ for $z = 0, 1$ and $j = 0, 1, 2$. We applied two fitted regression model to estimate mean potential response of each patient given treatment 1 and treatment 0: $\hat{E}\{Y_i(z)\} = \hat{\beta}_{z0}L_i + \hat{\beta}_{z2}W_i$ for $z = 0, 1$, and $i = 1, \dots, n$. Then we can estimate average treatment effect as

$$\frac{1}{n} \sum_i [\hat{E}\{Y_i(1)\} - \hat{E}\{Y_i(0)\}].$$

Then we used IPTW to estimate ATE. The IPTW method assigns each patients i a weight b_i equal to the inverse of an estimate of $p(Z_i=1|x_i)$. Let $\pi_i = p(Z_i=1|x_i)$. An estimate of π_i is obtained by fitting a logistic regression model. When $Z_i = 1$, $b_i = 1/\pi_i$; when $Z_i = 0$, $b_i = 1/(1 - \pi_i)$. We define the IPTW mean outcome estimator

$$\text{IPTW}(Z=z) = \frac{\sum_i I(Z_i=z) b_i Y_i}{\sum_i I(Z_i=z) b_i}$$

and corresponding average treatment effect estimate $ATE_{IPTW} = IPTW(Z=1) - IPTW(Z=0)$.

Finally, We used AIPW, which is a doubly robust generalization of the IPTW, to estimate ATE.

$$ATE_{AIPW} = \frac{1}{n} \sum_{i=1}^n \left\{ \left[\frac{I(Z_i=1)Y_i}{\hat{\pi}_i} - \frac{I(Z_i=0)Y_i}{1-\hat{\pi}_i} \right] - \frac{I(Z_i=1) - \hat{\pi}_i}{\hat{\pi}_i(1-\hat{\pi}_i)} \left[(1-\hat{\pi}_i)\hat{E}(Y_i|Z_i=1, x_i) + \hat{\pi}_i\hat{E}(Y_i|Z_i=0, x_i) \right] \right\}.$$

We simulated 100 trials and draw the density plot of estimated regime effects by DDP-GP linear regression, IPTW and AIPW in 100 trials. The truth is indicated by a black vertical line.

Under this simulation setting, it seems linear regression worked as well as DDP-GP model, which are less variable and close to true effects. However, DDP-GP model took much longer time to fit by MCMC sampling.

5.2. Multistage Regimes

Also, we simulated the data under the same setting as indicated in simulation 4 in the article. We simulated samples of size $n = 200$ and patients' blood glucose values $L_j \sim N(100, 10^2)$.

Patients initially were randomized equally between two induction therapies $Z^1 \in \{a_1, a_2\}$. Patients who were resistant (R) to the assigned induction therapies were then assigned salvage treatment $Z^{2,1} \in \{b_{11}, b_{12}\}$. Salvage treatments were randomized using the rule $p(Z^{2,1} = b_{11}|L_i) = 0.8(L_i < 100) + 0.2\mathbb{I}(L_i \geq 100)$. Patients who achieved C and subsequently suffered disease progression (P), were given salvage treatment $Z^{2,2} \in \{b_{21}, b_{22}\}$, using $p(Z^{2,2} = b_{21}|L_i) = 0.2(L_i < 100) + 0.85\mathbb{I}(L_i \geq 100)$. The survival time for each patient was evaluated as

$$T_i = \begin{cases} T_i^{(0,R)} + T_i^{(R,D)} & \text{if patient } i \text{ had sequence } (L, Z^1, T^{(0,R)}, Z^{2,1}) \\ T_i^{(0,C)} + T_i^{(C,P)} + T_i^{(P,D)} & \text{if patient } i \text{ had sequence } (L, Z^1, T^{(0,C)}, T^{(C,P)}, Z^{2,2}). \end{cases}$$

We simulated $T_i^{(0,R)} \sim LN(\beta^{(0,R)} x_i^{(0,R)}, \sigma^{(0,R)})$ and $T_i^{(0,C)} \sim LN(\beta^{(0,C)} x_i^{(0,C)}, \sigma^{(0,C)})$, where $\beta^{(0,R)} = (2, 0.02, 0)$, $\beta^{(0,C)} = (1.5, 0.03, -0.8)$, with $x_i^{(0,R)} = (1, L_i, Z_i^1)$ for $k \in \{(0, R), (0, C)\}$. For the second stage transitions $k \in \{(R, D), (C, P), (P, D)\}$, we generated $T_i^k \sim LN(\beta^k x_i^k, \sigma^k)$, where $\beta^{(R,D)} = (-0.5, 0.03, 0.2, 0.5, 0.3)$, $x_i^{(R,D)} = (1, L_i, Z_i^1, \log(T_i^{(0,R)}), Z_i^{2,1})$, $\beta^{(C,P)} = (1, 0.05, 1, -0.6)$, $x_i^{(C,P)} = (1, L_i, Z_i^1, \log(T_i^{(0,C)}))$, $\beta^{(C,D)} = (0.8, 0.04, 1.5, -1, 0.5, 0.5)$, $x_i^{(P,D)} = (1, L_i, Z_i^1, \log(T_i^{(0,C)}), \log(T_i^{(C,P)}), Z_i^{2,2})$. We simulated 100 trials.

We used several different models to fit the simulated data and estimated mean survival time for each DTR ($Z^1, Z^{2,1}, Z^{2,2}$) using fitted models. The models we have implemented including DDP-GP model proposed in the article, IPTW and AFT regression models with different specification of error distribution. The AFT model is in the form

$$\log(T_i) = \beta_0 + \beta_1 z_{i1} + \cdots + \beta_p z_{ip} + \sigma \varepsilon_i.$$

We assume two parametric error(ε) distributions for AFT models: normal distribution and extreme value distribution with two parameters. Accordingly, the distribution of T is log-normal distribution and weibull distribution, respectively. We also implemented a semiparametric AFT model with smooth error distribution, which is expressed as a mixture of G-splines. We compared the estimated treatment regime effect with the truth. The differences are summarized in the boxplot, which is arranged by different methods and by eight possible DTRs.

In our simulation scenario, the true distribution for the error term is normal distribution. We can see from the boxplot, when the error distribution is correctly specified (normal) in the AFT model, the estimated mean survival time is very close to the truth; when the error distribution is not correctly specified (extreme value) in the AFT model, the estimated mean survival time is a little bit biased. AFT model with error distribution expressed as a mixture of G-splines is more flexible. We don't need to specify any parametric error distribution for the AFT model so as to reduce bias caused by misspecification.

Compared with DDP-GP model proposed in the article, semiparametric AFT model takes much shorter time to fit and performs even a little bit better in this simulation scenario. Maybe it is because survival time between states is linearly generated in the simulation setting. DDP-GP model might be able to accommodate to more complicated data structure better.

6. Personalized Treatment Regime

The proposed approach in the article gave one general choice of treatment regimes for all the patients since they evaluated mean overall survival time by averaging over the empirical covariate distribution. We think we can apply the DDP-GP approach in the context of policy search algorithm by evaluating individual-specific treatment regime and gave treatment recommendations for patients based on their baseline covariates and disease progression.

References

- Chakraborty B, Laber EB, Zhao Y-Q. Inference About the Expected Performance of a Data-Driven Dynamic Treatment Regime. *Clinical Trials*. 2014; 11:408–417. [PubMed: 24925083]
- Doove L, Dusseldorp E, Van Deun K, Van Mechelen I. A Novel Method for Estimating Optimal Tree-Based Treatment Regimes in Randomized Clinical Trials. 2015 Abstract, available at <https://lirias.kuleuven.be/handle/123456789/493191>.
- Hirano K, Porter JR. Asymptotics for Statistical Treatment Rules. *Econometrica*. 2009; 77:1683–1701.
- Kang C, Janes H, Huang Y. Combining Biomarkers to Optimize Patient Treatment Recommendations. *Biometrics*. 2014; 70:695–707. [PubMed: 24889663]
- Laber EB, Linn KA, Stefanski LA. Interactive Model Building for Q-Learning. *Biometrika*. 2014; 101:831–847. [PubMed: 25541562]
- Laber EB, Lizotte DJ, Qian M, Pelham WE, Murphy SA. Dynamic Treatment Regimes: Technical Challenges and Applications. *Electronic Journal of Statistics*. 2014; 8:1225. [PubMed: 25356091]
- Laber EB, Zhao YQ. Tree-Based Methods for Individualized Treatment Regimes. *Biometrika*. 2015; 102:501–514. [PubMed: 26893526]
- Moodie EEM, Dean N, Sun YR. Q-Learning: Flexible Learning About Useful Utilities. *Statistics in Biosciences*. 2013; 6:1–21.
- Orellana L, Rotnitzky A, Robins JM. Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes, Part I: Main content. *The International Journal of Biostatistics*. 2010; 6
- Robins JM. Optimal Structural Nested Models for Optimal Sequential Decisions. In: Lin, DY.; Heagerty, PJ., editors. *Proceedings of the Second Seattle Symposium in Biostatistics (Vol. 179 of Lecture Notes in Statistics)*. New York: Springer; 2004. p. 189-326.
- Robins JM, Orellana L, Rotnitzky A. Estimation and Extrapolation of Optimal Treatment and Testing Strategies. *Statistics in Medicine*. 2008; 27:4678–4721. [PubMed: 18646286]
- Rubin DB. Bayesian Inference for Causal Effects: The Role of Randomization. *Annals of Statistics*. 1978; 6:34–58.
- Schulte PJ, Tsiatis AA, Laber EB, Davidian M. Q- and A-Learning Methods for Estimating Optimal Dynamic Treatment Regimes. *Statistical Science*. 2014; 29:640–661. [PubMed: 25620840]
- Splawa-Neyman J, Dabrowska DM, Speed TP. On the Application of Probability Theory to Agricultural Experiments. *Essay on Principles*. Section 9. *Statistical Science*. 1990; 5:465–472.
- Tanner MA, Wong WH. The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*. 1987; 82:528–540.
- Taylor JMG, Cheng W, Foster JC. Reader Reaction to “A Robust Method for Estimating Optimal Treatment Regimes” by Zhang et al. (2012). *Biometrics*. 2015; 71:267–273. [PubMed: 25228049]
- Van Der Vaart A. On Differentiable Functionals. *The Annals of Statistics*. 1991:178–204.
- Wahed AS, Thall PF. Evaluating Joint Effects of Induction– Salvage Treatment Regimes on Overall Survival in Acute Leukaemia. *Journal of the Royal Statistical Society, Series C*. 2013; 62:67–83.
- Zhang B, Tsiatis AA, Laber EB, Davidian M. A Robust Method for Estimating Optimal Treatment Regimes. *Biometrics*. 2012; 68:1010–1018. [PubMed: 22550953]
- Zhang B, Tsiatis AA, Laber EB, Davidian M. Robust Estimation Of Optimal Dynamic Treatment Regimes for Sequential Treatment Decisions. *Biometrika*. 2013; 100:681–694.
- Zhang Y, Laber EB, Tsiatis A, Davidian M. Using Decision Lists to Construct Interpretable and Parsimonious Treatment Regimes. *Biometrics*. 2015; 71:895–904. [PubMed: 26193819]
- Zhao Y, Kosorok MR, Zeng D. Reinforcement Learning Design for Cancer Clinical Trials. *Statistics in Medicine*. 2009; 28:3294–3315. [PubMed: 19750510]
- Zhao Y, Zeng D, Laber EB, Kosorok MR. New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes. *Journal of the American Statistical Association*. 2015; 110:583–598. [PubMed: 26236062]
- Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association*. 2012; 107:1106–1118. [PubMed: 23630406]

- Zhao YQ, Zeng D, Laber EB, Song R, Yuan M, Kosorok MR. Doubly Robust Learning for Estimating Individualized Treatment With Censored Data. *Biometrika*. 2014; 102:151–168.
- Zhou X, Kosorok MR. Nearest Neighbor Rules for Optimal Treatment Regimes. Under Review. 2016; 1:1–15.
- Zhou X, Mayer-Hamblett N, Khan U, Kosorok MR. Residual Weighted Learning for Estimating Individualized Treatment Rules. *Journal of the American Statistical Association*. 2015 forthcoming.

Figure 1.

Simulation results for single-stage trials under Settings 1 (left) and 2 (right). The boxplots show the differences between estimated and true causal effects over the 100 simulated datasets using the dependent Dirichlet process model (“DDP-GP”) and the semiparametric models with linear (“Ln”) and nonlinear (“NLn”) means and Gaussian (“Gs”) and mixture of Gaussians (“MGs”) residuals.

Figure 2.

Simulation results for three-stage trials under Setting 3. The boxplots show the differences between estimated and true causal effects over the 100 simulated datasets using the Dependent Dirichlet Process model (“DDP-GP”) and the semiparametric models with linear (“Ln”) and nonlinear (“NLn”) means and Gaussian (“Gs”) and mixture of Gaussians (“MGs”) residuals.

Figure 3.

Simulation results for three-stage trials under Setting 4. The boxplots show the differences between estimated and true causal effects over the 100 simulated datasets using the Dependent Dirichlet Process model (“DDP-GP”) and the semiparametric models with linear (“Ln”) and nonlinear (“NLn”) means and Gaussian (“Gs”) and mixture of Gaussians (“MGs”) residuals.