Uncertainty Quantification in Heterogeneous Treatment Effect Estimation with Gaussian-Process-Based Partially Linear Model

Shunsuke Horii[1], **Yoichi Chikahara**[2]

[1]Center for Data Science, Waseda University, Tokyo, Japan
[2]NTT Communication Science Laboratories, Kyoto, Japan
s.horii@waseda.jp, chikahara.yoichi@gmail.com

Abstract

Estimating heterogeneous treatment effects across individuals has attracted growing attention as a statistical tool for performing critical decision-making. We propose a Bayesian inference framework that quantifies the uncertainty in treatment effect estimation to support decision-making in a relatively small sample size setting. Our proposed model places Gaussian process priors on the nonparametric components of a semiparametric model called a partially linear model. This model formulation has three advantages. First, we can analytically compute the posterior distribution of a treatment effect without relying on the computationally demanding posterior approximation. Second, we can guarantee that the posterior distribution concentrates around the true one as the sample size goes to infinity. Third, we can incorporate prior knowledge about a treatment effect into the prior distribution, improving the estimation efficiency. Our experimental results show that even in the small sample size setting, our method can accurately estimate the heterogeneous treatment effects and effectively quantify its estimation uncertainty.

1 Introduction

Assessing heterogeneous treatment effects across individuals provides a key foundation for making critical decisions. For instance, understanding how greatly medical treatment effects are different across patients is helpful for precision medicine, and evaluating the impact of education programs on learning outcomes is essential for personalized learning.

A widely used treatment effect measure is a conditional average treatment effect (CATE), which is an average treatment effect across individuals with identical feature attributes. CATE estimation is challenging when the number of features of an individual is large. Many methods aim to express the complex nonlinearity between treatment effects and features, using a nonparametric regression model, such as tree-based models (Hahn, Murray, and Carvalho 2020; Hill 2011; Wager and Athey 2018) and neural networks (Johansson, Shalit, and Sontag 2016; Shalit, Johansson, and Sontag 2017; Yoon, Jordon, and Van Der Schaar 2018; Wu et al. 2023).

However, most of these methods only output a single-point estimate and cannot consider the uncertainty in CATE estimation. This drawback is fatal because decision-making under uncertainty is usual in many applications, especially when we can only access a small amount of observational data. An obvious example would be medical treatment planning. For example, in the US, more than half of hospitals have fewer beds than 100 (Wiens, Guttag, and Horvitz 2014), illustrating the difficulty of obtaining large-scale data and the importance of considering the uncertainty.

To support decision-making in such critical applications, we propose a Bayesian framework for quantifying the CATE estimation uncertainty. To deal with a small sample size setting, we focus on a semi-parametric model called a partially linear model (Engle et al. 1986), which is linear with respect to the treatment but is nonlinear with respect to features, thanks to the nonparametric components.

Our key idea is to place Gaussian process priors on the nonparametric components in a partially linear model. This idea has three advantages. First, we can analytically compute the posterior distribution of CATE. Such analytical computation is much more computationally efficient than the approximate Bayesian inference, which is required with the complex tree-based models (Hahn, Murray, and Carvalho 2020). Second, we can theoretically guarantee the asymptotic consistency of the posterior distribution of CATE under some mild conditions. This theoretical guarantee also makes a striking contrast with the tree-based models, which have no consistency guarantee due to their model complexity. Third, we can incorporate the prior knowledge about the CATE to improve the estimation accuracy. To take prior knowledge example, consider the education program evaluation, where it is known that past academic performance is an important feature called a *treatment effect modifier* (Hernán and Robins 2020, Chapter 4), which affects the effect of an education program (Yeager et al. 2019). We can utilize this prior knowledge by designing the kernel functions used in the Gaussian process priors. By contrast, integrating such prior knowledge is impossible with the existing Gaussian-process-based model (Alaa and van der Schaar 2018), which puts the priors not on the CATE but on the outcomes. Furthermore, our method can address not only binary treatment but also continuous-valued treatment (Hirano and Imbens 2004), which widens the scope of applications and is helpful, for instance, in determining the appropriate drug dosage for precision medicine (Bica, Jordon, and van der Schaar 2020).

Method	AP	CG	PK
(Alaa and van der Schaar 2018)	✓	✓	
(Hahn, Murray, and Carvalho 2020)			✓
Proposed method	✓	✓	✓

Table 1: Comparison with existing Bayesian methods: AP, CG, and PK are acronyms for Analytic form of Posterior, Consistency Guarantee, and Prior Knowledge incorporation.

Our contributions are summarized as follows:

- We establish a Bayesian framework that can effectively quantify the CATE estimation uncertainty in a relatively small sample size setting (Table 1). To achieve this, we put Gaussian process priors on the nonparametric components of a partially linear model.
- We theoretically prove that the posterior distribution of CATE concentrates around the true one, as the sample size goes to infinity (Section 4).
- We experimentally show that the proposed method can accurately estimate the CATE and effectively quantify its estimation uncertainty, especially when given small and high-dimensional observational data.

The full version of this paper, including the technical appendices, is available at https://arxiv.org/abs/2312.10435. Our code is publicly available at https://github.com/holyshun/GP-PLM.

2 Preliminaries

2.1 Problem Setup

Our target estimand, CATE, is the average effect of treatment T on outcome Y in a subgroup of individuals with identical feature attributes $\boldsymbol{X} = \boldsymbol{x}$. Here we consider a binary or continuous-valued treatment ($T \in \{0, 1\}$ or $T \in \mathbb{R}$), a continuous-valued outcome ($Y \in \mathbb{R}$), and a d-dimensional continuous-valued feature vector ($\boldsymbol{X} \in \mathbb{R}^d$).

In case of binary treatment $T \in \{0, 1\}$, a treatment effect for an individual is measured as the difference between random variables called *potential outcomes*, $Y^{(1)} - Y^{(0)}$, where $Y^{(0)}$ and $Y^{(1)}$ represents outcome Y when an individual is untreated ($T = 0$) and treated ($T = 1$), respectively (Rubin 1974). Unfortunately, we can never observe treatment effect $Y^{(1)} - Y^{(0)}$ because we only observe outcome $Y = TY^{(1)} + (1 - T)Y^{(0)}$ and can never jointly observe two potential outcomes $Y^{(0)}$ and $Y^{(1)}$. For this reason, we focus on the average, CATE, which can be estimated from the data and is defined as the conditional expected value:

$$\begin{aligned}\mathrm{CATE}(\boldsymbol{x}) &= \mathbb{E}[Y^{(1)} - Y^{(0)} | \boldsymbol{X} = \boldsymbol{x}] \\ &= \mathbb{E}[Y^{(1)} | \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y^{(0)} | \boldsymbol{X} = \boldsymbol{x}]. \end{aligned} \quad (1)$$

We can similarly define the CATE for continuous-valued treatment $T \in \mathbb{R}$, which represents the degree of treatment (e.g., the amount of chemotherapy). In this case, potential outcome $Y^{(t)}$ expresses the outcome when $T = t$ ($t \in \mathbb{R}$). The mean potential outcome across individuals with $\boldsymbol{X} = \boldsymbol{x}$, i.e., $\mathbb{E}[Y^{(t)} | \boldsymbol{X} = \boldsymbol{x}]$, can be regarded as a function of t, which is called a *dose response function*. By taking the value difference of this function between treatment values $T = t$ and $T = t'$ ($t, t' \in \mathbb{R}$), we can measure the CATE as

$$\mathrm{CATE}(\boldsymbol{x}, t, t') = \mathbb{E}[Y^{(t')} | \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y^{(t)} | \boldsymbol{X} = \boldsymbol{x}]. \quad (2)$$

Hence, in both treatment setups, we need to estimate mean potential outcome $\mathbb{E}[Y^{(t)} | \boldsymbol{X} = \boldsymbol{x}]$ for treatment $t \in \mathcal{T}$, where $\mathcal{T} = \{0, 1\}$ or $\mathcal{T} \subseteq \mathbb{R}$. To achieve this, we make two standard assumptions. One is the overlap condition (a.k.a., *positivity*), i.e., $0 < p(t|\boldsymbol{x}) < 1$ for all $t \in \mathcal{T}$ and for all $\boldsymbol{x} \in \mathcal{X}$ such that $p(\boldsymbol{x}) > 0$. The other is the *strongly ignorability* condition (Imbens and Rubin 2015), $\{Y^{(t)} : t \in \mathcal{T}\} \perp\!\!\!\perp T | \boldsymbol{X}$; this conditional independence relation is satisfied if features $\boldsymbol{X}$ include all confounders and contain only *pretreatment variables*, which are not affected by treatment T unlike mediators and colliders (Elwert and Winship 2014).[1] Under these two assumptions, the mean potential outcome can be reformulated as [2]

$$\mathbb{E}[Y^{(t)} | \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y | \boldsymbol{X} = \boldsymbol{x}, T = t].$$

That is, the mean potential outcome is reduced to the conditional expectation $\mathbb{E}[Y | \boldsymbol{X} = \boldsymbol{x}, T = t]$. To represent this conditional expectation, we employ a partially linear model, which is described below.

2.2 Partially Linear Model

A partially linear model is a semi-parametric regression model introduced by Engle et al. (1986). A widely used formulation of this model is

$$Y = \theta T + f(\boldsymbol{X}) + \varepsilon, \quad (3)$$

where $\theta \in \mathbb{R}$ is an unknown parameter representing a treatment effect, $f \colon \mathbb{R}^d \to \mathbb{R}$ is an unknown nonlinear function that expresses how greatly outcome Y differ depending on the values of features $\boldsymbol{X}$, and $\varepsilon \overset{i.i.d.}{\sim} \mathcal{N}(0, s_\varepsilon^{-1})$ is a noise that follows a zero-mean Gaussian with precision $s_\varepsilon > 0$.

The model formulation (3) has two advantages. First, compared with nonparametric regression models such as tree-based models and neural networks, the estimation requires a much smaller sample size even when feature vector $\boldsymbol{X}$ is high-dimensional. Such a sample-size efficiency is essential to achieve practical applications with high data acquisition costs. Second, despite this efficiency, the model (3) can represent the complex nonlinearity between outcome Y and $\boldsymbol{X}$, using nonlinear function $f(\boldsymbol{X})$.

By contrast, a drawback of the model (3) is that it cannot capture the treatment effect heterogeneity. This is because the treatment effect parameter θ is a constant with respect to features $\boldsymbol{X}$; hence, it cannot express how greatly the treatment effect changes depending on $\boldsymbol{X}$'s values.

To overcome this drawback, we focus on the following variant of the partially linear model:

$$Y = \theta(\boldsymbol{X})T + f(\boldsymbol{X}) + \varepsilon, \quad (4)$$

[1] In addition, we assume that features $\boldsymbol{X}$ do not include pretreatment colliders, as with the standard CATE estimation methods.

[2] For the derivation, see, e.g., the survey by Yao et al. (2021).

where $\theta\colon \mathbb{R}^d \to \mathbb{R}$ is an unknown nonlinear function. A significant difference from (3) is that $\theta(\cdot)$ in (4) is a function and can represent how strongly a treatment effect varies with $\boldsymbol{X}$'s values, thus addressing treatment effect heterogeneity.

With the model (4), the CATEs for binary and continuous-valued treatments in (1) and (2) are given by

$$\mathrm{CATE}(\boldsymbol{x}) = \theta(\boldsymbol{x}); \quad \mathrm{CATE}(\boldsymbol{x}, t, t') = (t' - t)\theta(\boldsymbol{x}). \quad (5)$$

Hence, the CATE estimation reduces to the problem of estimating function $\theta(\cdot)$. In fact, other treatment effect measure called a *conditional derivative effect*, $\lim_{\xi \to 0} \xi^{-1} \mathbb{E}[Y^{(t+\xi)} - Y^{(t)} | \boldsymbol{X} = \boldsymbol{x}]$ ($t \in \mathbb{R}$), can also be inferred by estimating $\theta(\cdot)$, which we detail in Appendix A.

Estimator $(t' - t)\theta(\boldsymbol{x})$ in (5) assumes that the CATE is linear with respect to treatment $T \in \mathbb{R}$; this assumption might be restrictive in some applications. However, if we have prior knowledge about the functional relationship between outcome Y and treatment T, it is straightforward to use a pre-specified nonlinear function, $h\colon \mathbb{R} \to \mathbb{R}$ (e.g., $h(T) = \sqrt{T}$), to reformulate $\theta(\boldsymbol{X})T$ in (4) as $\theta(\boldsymbol{X})h(T)$ and the CATE as $\mathrm{CATE}(\boldsymbol{x}, t, t') = (h(t') - h(t))\theta(\boldsymbol{x})$. One idea for how to formulate function h is to follow the functional form between Y and T of the parametric models that are commonly used in the field. For instance, in medical treatment planning, a sigmoid function is widely used in the dose-response curve models (Hill 1910; Hamilton, Russo, and Thurston 1977). Developing a data-driven way to infer function h is left as our future work.

Indeed, the partially linear model formulation (4) has also been studied in the treatment effect estimation framework called *Double/Debiased Machine Learning* (DML) (Chernozhukov et al. 2018). This framework is founded on the Frequentist approach and quantifies the uncertainty in estimating function θ with the confidence interval. However, as shown by Van der Vaart (2000), in a finite sample size setting, there is no theoretical guarantee about the uncertainty estimation with a confidence interval, and hence, the uncertainty estimation can be inaccurate. To resolve this issue, we develop a Bayesian approach that infers the posterior distribution of θ.

3 Proposed Model

This section presents the derivation of the posterior distribution of function θ in the partially linear model in (4).

Our posterior distribution can be formulated as follows. Suppose that each observation is obtained as $(t_i, \boldsymbol{x}_i, y_i) \overset{i.i.d.}{\sim} p(t, \boldsymbol{x}, y)$ for $i = 1, \ldots, n$ and that we have n observations $\mathcal{D}_n = (\boldsymbol{t}_n, \boldsymbol{X}_n, \boldsymbol{y}_n)$, where $\boldsymbol{t}_n = (t_1, \ldots, t_n)$, $\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, and $\boldsymbol{y}_n = (y_1, \ldots, y_n)$. Our goal is to estimate the CATE values for a pre-specified set of m feature vector values $\tilde{\boldsymbol{X}}_m = (\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_m)$. Hence, our target posterior can be formulated as posterior predictive distribution $p(\tilde{\boldsymbol{\theta}}_m | \mathcal{D}_n, \tilde{\boldsymbol{X}}_m)$, where $\tilde{\boldsymbol{\theta}}_m = (\theta(\tilde{\boldsymbol{x}}_1), \ldots, \theta(\tilde{\boldsymbol{x}}_m))$. Note that this posterior predictive is different from that of the standard Gaussian process regression: it is the distribution of a function representing the CATE, not outcome Y.

3.1 Priors

To formulate the posterior predictive distribution, we place the Gaussian process prior distributions on functions θ and f in the partially linear model (4) as

$$\theta(\cdot) \sim \mathcal{GP}(0, C(\cdot, \cdot; \boldsymbol{\omega}_\theta)), \quad (6)$$

$$f(\cdot) \sim \mathcal{GP}(0, C(\cdot, \cdot; \boldsymbol{\omega}_f)), \quad (7)$$

where $C(\cdot, \cdot; \boldsymbol{\omega})$ is the covariance function with parameter $\boldsymbol{\omega}$. Here, for notation simplicity, we set the mean functions of the Gaussian processes to zero. Note that such zero-mean priors do not lose generality because they never restrict the posterior means, which are updated with the observed data (See, e.g., Williams and Rasmussen (2006)).

A common formulation of covariance function $C(\cdot, \cdot; \boldsymbol{\omega})$ in (6) and (7) is the radial basis function (RBF) kernel:

$$C(\boldsymbol{x}, \boldsymbol{x}'; \omega) = \exp\left\{-\omega \|\boldsymbol{x} - \boldsymbol{x}'\|^2\right\} \quad (\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d), \quad (8)$$

where $\omega > 0$ is a hyperparameter. We can also design the covariance function by utilizing our prior knowledge. For example, if some features in $\boldsymbol{X}$ are known to be treatment effect modifiers, i.e., important features that explain treatment effect heterogeneity, we can formulate $C(\cdot, \cdot; \boldsymbol{\omega}_\theta)$ as

$$C(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\omega}_\theta) = \exp\left\{-\sum_{k=1}^{d} \omega_{\theta,k}(x_k - x'_k)^2\right\}, \quad (9)$$

where $\boldsymbol{\omega}_\theta = (w_{\theta,1}, \ldots, w_{\theta,d})$ is a vector of hyperparameters, whose k-th element $\omega_{\theta,k} \in \mathbb{R}^{\geq 0}$ represents the k-th feature's importance, which is given based on prior knowledge. In Appendix E, we show that using covariance function (9) leads to better estimation performance.

3.2 Derivation of Posterior Predictive

We show that we can analytically compute posterior predictive $p(\tilde{\boldsymbol{\theta}}_m | \mathcal{D}_n, \tilde{\boldsymbol{X}}_m)$ if the prior hyperparameters, $\boldsymbol{\omega}_\theta$ and $\boldsymbol{\omega}_f$, and noise distribution parameter s_ε are given.

To derive this analytic form, we take three steps. First, we derive the joint distribution $p(\Theta, \boldsymbol{y}_n | \boldsymbol{t}_n, \boldsymbol{X}_n, \tilde{\boldsymbol{X}}_m)$, where $\Theta = (\boldsymbol{\theta}_n, \tilde{\boldsymbol{\theta}}_m, \boldsymbol{f}_n)$ denotes a set of function values, including $\boldsymbol{\theta}_n = (\theta(\boldsymbol{x}_1), \ldots, \theta(\boldsymbol{x}_n))$ and $\boldsymbol{f}_n = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n))$. Second, by conditioning $\boldsymbol{y}_n$, we obtain joint posterior $p(\Theta | \mathcal{D}_n, \tilde{\boldsymbol{X}}_m)$. Finally, we marginalize out $\boldsymbol{\theta}_n$ and $\boldsymbol{f}_n$ in Θ to derive posterior predictive $p(\tilde{\boldsymbol{\theta}}_m | \mathcal{D}_n, \tilde{\boldsymbol{X}}_m)$.

Joint Distribution $p(\Theta, \boldsymbol{y}_n | \boldsymbol{t}_n, \boldsymbol{X}_n, \tilde{\boldsymbol{X}}_m)$ is given as a product of the likelihood and the joint prior:

$$p(\Theta, \boldsymbol{y}_n | \boldsymbol{t}_n, \boldsymbol{X}_n, \tilde{\boldsymbol{X}}_m) = p(\boldsymbol{y}_n | \boldsymbol{\theta}_n, \boldsymbol{f}_n, \boldsymbol{t}_n) p(\Theta | \boldsymbol{X}_n, \tilde{\boldsymbol{X}}_m). \quad (10)$$

Joint prior $p(\Theta | \boldsymbol{X}_n, \tilde{\boldsymbol{X}}_m)$ in (10) is given by the Gaussian process priors in (6) and (7) and hence is formulated as the multivariate Gaussian:

$$p(\Theta | \boldsymbol{X}_n, \tilde{\boldsymbol{X}}_m) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\Theta\Theta}), \quad (11)$$

where $\boldsymbol{\Sigma}_{\Theta\Theta}$ denotes the following covariance matrix:[3]

$$\boldsymbol{\Sigma}_{\Theta\Theta} = \begin{pmatrix} \boldsymbol{\Phi}_{nn} & \boldsymbol{\Phi}_{nm} & \boldsymbol{O} \\ \boldsymbol{\Phi}_{nm}^T & \boldsymbol{\Phi}_{mm} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{\Psi}_{nn} \end{pmatrix},$$

whose elements are given by

$$\begin{aligned} \boldsymbol{\Phi}_{nn} &= \left(C(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\omega}_\theta)\right)_{1\le i,j\le n}, \\ \boldsymbol{\Phi}_{nm} &= \left(C(\boldsymbol{x}_i, \tilde{\boldsymbol{x}}_j; \boldsymbol{\omega}_\theta)\right)_{1\le i\le n, 1\le j\le m}, \\ \boldsymbol{\Phi}_{mm} &= \left(C(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{x}}_j; \boldsymbol{\omega}_\theta)\right)_{1\le i,j\le m}, \\ \boldsymbol{\Psi}_{nn} &= \left(C(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\omega}_f)\right)_{1\le i,j\le n}. \end{aligned} \tag{12}$$

By contrast, likelihood $p(\boldsymbol{y}_n|\boldsymbol{\theta}_n, \boldsymbol{f}_n, \boldsymbol{t}_n)$ in (10) is given by the partially linear model with a Gaussian noise in (4). Hence, it is formulated as the multivariate Gaussian:

$$p(\boldsymbol{y}_n|\boldsymbol{\theta}_n, \boldsymbol{f}_n, \boldsymbol{t}_n) = \mathcal{N}(\boldsymbol{T}_n\boldsymbol{\theta}_n + \boldsymbol{f}_n, s_\varepsilon^{-1}\boldsymbol{I}), \tag{13}$$

where $\boldsymbol{T}_n = \mathrm{diag}(\boldsymbol{t}_n)$ is a diagonal matrix, whose diagonal component is $\boldsymbol{t}_n$.

Thus, both $p(\Theta|\boldsymbol{X}_n, \tilde{\boldsymbol{X}}_m)$ and $p(\boldsymbol{y}_n|\boldsymbol{\theta}_n, \boldsymbol{f}_n, \boldsymbol{t}_n)$ are given as multivariate Gaussians. Therefore, joint distribution $p(\Theta, \boldsymbol{y}_n|\boldsymbol{t}_n, \boldsymbol{X}_n, \tilde{\boldsymbol{X}}_m)$ is also a multivariate Gaussian. This Gaussian has mean $\boldsymbol{0}$ because both the joint prior in (11) and the likelihood in (13) are zero-mean; the mean in (13), $\boldsymbol{T}_n\boldsymbol{\theta}_n + \boldsymbol{f}_n$, is zero because $p(\boldsymbol{\theta}_n, \boldsymbol{f}_n|\boldsymbol{X}_n)$ is zero-mean. As regards covariance matrix $\boldsymbol{\Sigma}$, we can explicitly express precision matrix $\boldsymbol{S} = \boldsymbol{\Sigma}^{-1}$ as

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{\Phi}^{-1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Psi}_{nn}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & s_\epsilon\boldsymbol{I} \end{pmatrix} + \begin{pmatrix} s_\epsilon\boldsymbol{T}_n^2 & \boldsymbol{O} & s_\epsilon\boldsymbol{T}_n & -s_\epsilon\boldsymbol{T}_n \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ s_\epsilon\boldsymbol{T}_n & \boldsymbol{O} & s_\epsilon\boldsymbol{I} & -s_\epsilon\boldsymbol{I} \\ -s_\epsilon\boldsymbol{T}_n & \boldsymbol{O} & -s_\epsilon\boldsymbol{I} & \boldsymbol{O} \end{pmatrix}, \tag{14}$$

where $\boldsymbol{\Phi}$ is the block matrix with the elements in (12):

$$\boldsymbol{\Phi} = \begin{pmatrix} \boldsymbol{\Phi}_{nn} & \boldsymbol{\Phi}_{nm} \\ \boldsymbol{\Phi}_{nm}^T & \boldsymbol{\Phi}_{mm} \end{pmatrix}.$$

Joint Posterior $p(\Theta|\mathcal{D}_n, \tilde{\boldsymbol{X}}_m)$ is obtained from joint distribution $p(\Theta, \boldsymbol{y}_n|\boldsymbol{t}_n, \boldsymbol{X}_n, \tilde{\boldsymbol{X}}_m)$ in (10) by conditioning $\boldsymbol{y}_n$.

To confirm this, consider the covariance of the joint distribution in (10), i.e., $\boldsymbol{\Sigma}$, whose precision matrix $\boldsymbol{S} = \boldsymbol{\Sigma}^{-1}$ is given by (14). Let us rephrase this covariance matrix as [4]

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\Theta\Theta} & \boldsymbol{\Sigma}_{\Theta\boldsymbol{y}_n} \\ \boldsymbol{\Sigma}_{\boldsymbol{y}_n\Theta} & \boldsymbol{\Sigma}_{\boldsymbol{y}_n\boldsymbol{y}_n} \end{pmatrix}.$$

Then, using the formula of conditional multivariate Gaussians (see e.g., Bishop (2006); Williams and Rasmussen (2006)), we can condition $\boldsymbol{y}_n$ and show that joint posterior $p(\Theta|\mathcal{D}_n, \tilde{\boldsymbol{X}}_m)$ is the following multivariate Gaussian:

$$p(\Theta|\mathcal{D}_n, \tilde{\boldsymbol{X}}_m) = \mathcal{N}(\boldsymbol{\mu}_{\Theta|\boldsymbol{y}_n}, \boldsymbol{\Sigma}_{\Theta|\boldsymbol{y}_n}). \tag{15}$$

whose mean $\boldsymbol{\mu}_{\Theta|\boldsymbol{y}_n}$ and covariance $\boldsymbol{\Sigma}_{\Theta|\boldsymbol{y}_n}$ are given by

$$\begin{aligned} \boldsymbol{\mu}_{\Theta|\boldsymbol{y}_n} &= \boldsymbol{M}\boldsymbol{y}_n, \\ \boldsymbol{\Sigma}_{\Theta|\boldsymbol{y}_n} &= \boldsymbol{\Sigma}_{\Theta\Theta} - \boldsymbol{M}\boldsymbol{\Sigma}_{\boldsymbol{y}_n\Theta}, \end{aligned}$$

where $\boldsymbol{M} = \boldsymbol{\Sigma}_{\Theta\boldsymbol{y}}\boldsymbol{\Sigma}_{\boldsymbol{y}\boldsymbol{y}}^{-1}$.

Posterior Predictive $p(\tilde{\boldsymbol{\theta}}_m|\mathcal{D}_n, \tilde{\boldsymbol{X}}_m)$ is obtained by marginalizing out $\boldsymbol{\theta}_n$ and $\boldsymbol{f}_n$ in Θ from joint posterior $p(\Theta|\mathcal{D}_n, \tilde{\boldsymbol{X}}_m)$ in (15).

To see this, consider the submatrices in $\boldsymbol{M}$ and $\boldsymbol{\Sigma}_{\Theta|\boldsymbol{y}_n}$:

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{M}_{\boldsymbol{\theta}} \\ \boldsymbol{M}_{\tilde{\boldsymbol{\theta}}} \\ \boldsymbol{M}_{\boldsymbol{y}} \end{pmatrix},$$

$$\boldsymbol{\Sigma}_{\Theta|\boldsymbol{y}_n} = \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{\theta}|\boldsymbol{y}_n} & \boldsymbol{\Sigma}_{\boldsymbol{\theta}\tilde{\boldsymbol{\theta}}|\boldsymbol{y}_n} & \boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{f}|\boldsymbol{y}_n} \\ \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}\boldsymbol{\theta}|\boldsymbol{y}_n} & \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}|\boldsymbol{y}_n} & \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}\boldsymbol{f}|\boldsymbol{y}_n} \\ \boldsymbol{\Sigma}_{\boldsymbol{f}\boldsymbol{\theta}|\boldsymbol{y}_n} & \boldsymbol{\Sigma}_{\boldsymbol{f}\tilde{\boldsymbol{\theta}}|\boldsymbol{y}_n} & \boldsymbol{\Sigma}_{\boldsymbol{f}\boldsymbol{f}|\boldsymbol{y}_n}, \end{pmatrix}.$$

With these notations, by marginalizing out $\boldsymbol{\theta}_n$ and $\boldsymbol{f}_n$, we can formulate posterior predictive $p(\tilde{\boldsymbol{\theta}}_m|\mathcal{D}_n, \tilde{\boldsymbol{X}}_m)$ as the following multivariate Gaussian:

$$p(\tilde{\boldsymbol{\theta}}_m|\mathcal{D}_n, \tilde{\boldsymbol{X}}_m) = \mathcal{N}(\boldsymbol{M}_{\tilde{\boldsymbol{\theta}}}\boldsymbol{y}_n, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\theta}}\tilde{\boldsymbol{\theta}}|\boldsymbol{y}_n}). \tag{16}$$

By marginalizing $\boldsymbol{f}_n$ and formulating the posterior in this way, we remove the estimation bias arising from nuisance parameter f, which corresponds to the *confounding bias* arising from the confounders in features $\boldsymbol{X}$. With such Bayesian inference, we have no need to estimate the *propensity score* model, unlike the Frequentist approach. We detail the reason for this in Section 5.2.

The posterior predictive in (16) requires computation time $O(n^3)$ for sample size n, which might be problematic when n is large. However, as with the standard Gaussian process regression, we can apply various approximation techniques, such as the sparse GP (Quinonero-Candela and Rasmussen 2005), to deal with a large-scale dataset.

3.3 Addressing Unknown Hyperparameters

So far, we have derived the posterior predictive under the assumption that the values of hyperparameters $\boldsymbol{\omega}_\theta$, $\boldsymbol{\omega}_f$, and s_ε are given. Since their true values are unknown in practice, we must determine them using the observed data. Below, we present the two data-driven approaches.

One is to put the priors on these hyperparameters. In this case, the posterior predictive differs from (16), and the analytic form is no longer available. Hence, we will need to approximate the posterior predictive, using Markov chain Monte Carlo (MCMC) methods, such as the Metropolis-Hastings (MH) algorithm. Unfortunately, this approximation requires much computation time and might be impractical.

For this reason, in our experiments, we took the other approach, which estimates the values of hyperparameters $\boldsymbol{\omega}_\theta$,

[3] $\boldsymbol{O}$ and $\boldsymbol{I}$ are the zero and identity matrices, respectively. In this paper, all zero and identity matrices are denoted by $\boldsymbol{O}$ and $\boldsymbol{I}$, regardless of their matrix sizes.

[4] For example, $\boldsymbol{\Sigma}_{\Theta\Theta}$ is the submatrix of $\boldsymbol{S}^{-1}$, consisting of rows to 1 to $2n$ and columns 1 to $2n$.

$\boldsymbol{\omega}_f$, and s_ε by maximizing the marginal likelihood:

$$p(\boldsymbol{y}_n|\boldsymbol{t}_n, \boldsymbol{X}_n; \boldsymbol{\omega}_\theta, \boldsymbol{\omega}_f, s_\varepsilon) = \int\int p(\boldsymbol{y}_n, \boldsymbol{\theta}_n, \boldsymbol{f}_n|\boldsymbol{t}_n, \boldsymbol{X}_n; \boldsymbol{\omega}_\theta, \boldsymbol{\omega}_f, s_\varepsilon)\mathrm{d}\boldsymbol{\theta}_n\mathrm{d}\boldsymbol{f}_n,$$

where we marginalize out $\boldsymbol{\theta}_n$ and $\boldsymbol{f}_n$ from joint distribution:

$$\begin{aligned}&p(\boldsymbol{y}_n, \boldsymbol{\theta}_n, \boldsymbol{f}_n|\boldsymbol{t}_n, \boldsymbol{X}_n; \boldsymbol{\omega}_\theta, \boldsymbol{\omega}_f, s_\varepsilon)\\ =&p(\boldsymbol{y}_n|\boldsymbol{\theta}_n, \boldsymbol{f}_n, \boldsymbol{t}_n; s_\varepsilon)p(\boldsymbol{\theta}_n|\boldsymbol{X}_n; \boldsymbol{\omega}_\theta)p(\boldsymbol{f}_n|\boldsymbol{X}_n; \boldsymbol{\omega}_f). \quad (17)\end{aligned}$$

Here the three distributions in the r.h.s. of (17) are given as

$$\begin{aligned}&p(\boldsymbol{y}_n|\boldsymbol{\theta}_n, \boldsymbol{f}_n, \boldsymbol{t}_n; s_\varepsilon) = \mathcal{N}(\boldsymbol{T}_n\boldsymbol{\theta}_n + \boldsymbol{f}_n, s_\varepsilon^{-1}\boldsymbol{I}),\\ &p(\boldsymbol{\theta}_n|\boldsymbol{X}_n; \boldsymbol{\omega}_\theta) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Phi}_{nn}),\\ &p(\boldsymbol{f}_n|\boldsymbol{X}_n; \boldsymbol{\omega}_f) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}_{nn}),\end{aligned}$$

respectively. Hence, the joint distribution in (17) is also a multivariate Gaussian with mean $\boldsymbol{0}$ and covariance matrix

$$\begin{pmatrix} s_\epsilon\boldsymbol{I} & -s_\epsilon\boldsymbol{T}_n & -s_\epsilon\boldsymbol{I} \\ -s_\epsilon\boldsymbol{T}_n & s_\epsilon\boldsymbol{T}_n^2 + \boldsymbol{\Phi}_{nn}^{-1} & s_\epsilon\boldsymbol{T}_n \\ -s_\epsilon\boldsymbol{I} & s_\epsilon\boldsymbol{T}_n & s_\epsilon\boldsymbol{I} + \boldsymbol{\Psi}_{nn}^{-1} \end{pmatrix}^{-1}.$$

Using the matrix formula for the Schur complement, we can marginalize $\boldsymbol{\theta}_n$ and $\boldsymbol{f}_n$ from this multivariate Gaussian and obtain the marginal likelihood $p(\boldsymbol{y}_n|\boldsymbol{t}_n, \boldsymbol{X}_n; \boldsymbol{\omega}_\theta, \boldsymbol{\omega}_f, s_\varepsilon)$ as a multivariate Gaussian with mean $\boldsymbol{0}$ and covariance matrix

$$s_\epsilon^{-1}\boldsymbol{I} + \boldsymbol{T}_n\boldsymbol{\Phi}_{nn}\boldsymbol{T}_n + \boldsymbol{\Psi}_{nn}.$$

This marginal likelihood is not necessarily convex. For this reason, in our experiments, we maximize it with respect to $\boldsymbol{\omega}_\theta$, $\boldsymbol{\omega}_f$, and s_ε by combining the grid search and the gradient descent method. See Appendix D.2 for the details.

4 Theoretical Analysis

This section aims to guarantee the asymptotic convergence of our posterior distribution. In particular, we prove the *strong posterior consistency* (Ghosal, Ghosh, and Ramamoorthi 1999), whose definition can be used for nonparametric Bayesian models, including Gaussian processes.

This consistency notion is determined whether the random measure representing the posterior converges to the true data-generating distribution, as sample size $n \to \infty$. With the partially linear model, the data-generating distribution is defined with functions θ and f. Since we put the Gaussian process priors, these functions themselves have randomness. For this reason, we consider the posterior of their joint density, $p_{\theta,f}(\boldsymbol{x}, t, y) = p(y|\boldsymbol{x}, t, \theta, f)p(t|\boldsymbol{x})p(\boldsymbol{x})$. Our goal is to show that as $n \to \infty$, this posterior concentrates around its true joint density, p_{θ_0,f_0} with high probability, where θ_0 and f_0 denote the true data-generating functions.

To achieve this goal, we extend the results by Ghosal, Ghosh, and Ramamoorthi (1999), which proves the posterior consistency for the standard Gaussian process regression problems. As with these results, we make Assumptions (**P**; Smoothness of priors), (**F**; Bounded feature space), (**T**; True functions), and (**E**; Exponential decay of priors). Among them, Assumption (**F**) requires feature vector values $\boldsymbol{x}$ to belong to a bounded subset of $\mathbb{R}^d$, and Assumption (**T**) imposes the condition that true functions θ_0 and f_0 belong to the reproducing kernel Hilbert space (RKHS) of a kernel function used in the covariance function in Gaussian process priors. Regarding the technical assumptions on priors (Assumptions (**P**) and (**E**)), we detail them in Appendix B.

To extend the results by Ghosal, Ghosh, and Ramamoorthi (1999) to the CATE estimation problem, we make an additional assumption on the boundedness of the conditional moments of treatment T given features $\boldsymbol{X}$:

(B) (**Boundedness of conditional moments**) There exist constants $C_1 > 0$ and $C_2 > 0$ such that

$$\mathbb{E}[T|\boldsymbol{X}] < C_1 \quad \mathbb{P}_{\boldsymbol{X}}\text{-}a.s.; \quad \mathbb{E}\left[T^2|\boldsymbol{X}\right] < C_2 \quad \mathbb{P}_{\boldsymbol{X}}\text{-}a.s.$$

This assumption imposes the conditional mean and variance of T given $\boldsymbol{X}$ to be at most C_1 and C_2, respectively.

Using Assumptions (**P**), (**F**), (**T**), (**E**), and (**B**), we prove the consistency of the posterior of θ and f. For this purpose, we use the L_1 metric between $p_{\theta,f}$ and p_{θ_0,f_0}:

$$\|p_{\theta,f} - p_{\theta_0,f_0}\|_{L_1} = \sum_t \iint |p_{\theta,f}(\boldsymbol{x}, t, y) - p_{\theta_0,f_0}(\boldsymbol{x}, t, y)|\mathrm{d}y\mathrm{d}\boldsymbol{x}.$$

Let $\mathbb{P}_0^n$ be the true distribution of sample $\mathcal{D}_n = (\boldsymbol{t}_n, \boldsymbol{X}_n, \boldsymbol{y}_n)$, and Π be the prior distribution of $p_{\theta,f}$ when the parameters of θ and f are distributed according to priors $\Pi_{\tau_\theta}, \Pi_{\tau_f}, \Pi_{\lambda_\theta}$, and Π_{λ_f}. Then the following theorem holds.

Theorem 1. *Suppose that Assumptions (**P**), (**F**), (**T**), (**E**), and (**B**) hold. Then for any $\epsilon > 0$,*

$$\Pi\left((\theta, f) : \|p_{\theta,f} - p_{\theta_0,f_0}\|_{L_1} > \epsilon \mid \mathcal{D}_n\right) \to 0 \qquad (18)$$

with $\mathbb{P}_0^n$-probability 1.

Proof. See Appendix C for the formal proof. Here, we provide a proof sketch. From Assumptions (**P**), (**F**), (**T**), and (**E**), we can show that the probability that joint density $p_{\theta,f}$ is not a smooth function is exponentially small. Hence, we only have to consider a set of smooth joint density functions. With the above four assumptions, we can bound the *metric entropy* of this smooth function set. This allows us to bound the covering number and to upper bound the probability of the event that $p_{\theta,f}$ does not enter the neighborhood of p_{θ_0,f_0}, using a union bound. Finally, using Assumption (**B**), we can bound the Kullback-Leibler divergence between $p_{\theta,f}$ and p_{θ_0,f_0}, which enables us to bound the L_1 metric, using Pinsker's inequality. □

5 Related Work

5.1 Bayesian Approaches to CATE Estimation

The key idea of our method is to put a Gaussian process prior on function $\theta(\boldsymbol{X})$ in (4), which represents how the CATE varies with the values of features $\boldsymbol{X}$. While Alaa and van der Schaar (2018) also employ the Gaussian process priors, they propose to place them on each mean potential outcome function, which is denoted by $f_t(\cdot)$ in their model formulation:

$$Y^{(t)} = f_t(\boldsymbol{x}) + \varepsilon, \quad t \in \{0, 1\}.$$

Compared with this model, ours has two advantages. First, as described in Section 1, it can incorporate prior knowledge about the CATE by designing covariance function $C(\cdot, \cdot, \boldsymbol{\omega}_\theta)$ in (6); we provide a formulation example in (9), which can utilize the prior knowledge about treatment effect modifiers. Second, it can deal with the continuous-valued treatment setup; hence, the scope of applications is wider.

To develop a Bayesian approach under the continuous-valued treatment setup, several methods use a tree-based model called Bayesian additive regression trees (BART) (Hill 2011; Woody et al. 2020; Hahn, Murray, and Carvalho 2020). As reported by Dorie et al. (2017), these methods empirically work well on many synthetic benchmark datasets. However, due to the complexity of tree-based models, their performance is not theoretically guaranteed, demonstrating that it is uncertain whether they can be used for the crucial applications that involve critical decision-making. By contrast, using the Gaussian process priors, we have derived the asymptotic consistency of posterior distribution, thus yielding more reliable CATE estimation results. Furthermore, as illustrated in (16), we can analytically compute the posterior when given the hyperparameter values. Thus, our method requires a much smaller computation time, compared with the tree-based methods, which rely on the computationally demanding approximate Bayesian inference.

5.2 Non-Bayesian Approaches with Partially Linear Model

As described in Section 2.2, the partially linear model in (4) has been studied in the DML-based methods (Chernozhukov et al. 2018), which are founded on the Frequentist approach. Closest to our work is the R Learner (Nie and Wager 2021), which also makes the assumption that function θ in (4) is an element of RKHS (i.e., Assumption (**T**) in Section 4). For this reason, our method can be regarded as the Bayesian counterpart of the R Learner.

A large difference between the DML-based methods and ours is how to remove the confounding bias arising from the confounders in features $\boldsymbol{X}$. To achieve this, the DML-based methods use conditional distribution $p(t|\boldsymbol{x})$ (a.k.a., propensity score) [5], which is expressed with the following model:

$$T = \rho(\boldsymbol{X}) + \eta, \quad \mathbb{E}[\eta|\boldsymbol{X}] = 0. \tag{19}$$

They estimate function ρ in (19) and use the estimated function to eliminate the estimation bias due to nuisance parameter f in the partially linear model (4). By contrast, with our Bayesian approach, we do not need to estimate the propensity score from the data because we can remove the confounding bias, simply by marginalizing out function f. As explained by Li, Ding, and Mealli (2023), the main reason is that it is common in Bayesian inference to model the priors such that their parameters are mutually independent. This independence relation implies that the parameters of propensity score ρ are conditionally independent of functions (θ, f) conditioned on sample $\mathcal{D}_n = (\boldsymbol{t}_n, \boldsymbol{X}_n, \boldsymbol{y}_n)$. Thus, the propensity score model does not affect the inference, and we do not need to estimate the propensity score.

A serious disadvantage of the DML-based methods is that their uncertainty estimate is founded on a confidence interval, which has no theoretical guarantee in a small sample size setting (Van der Vaart 2000). As claimed in Section 1, this disadvantage is fatal if we focus on the applications related to decision-making under uncertainty. To overcome this disadvantage, we have established a Bayesian inference framework that can effectively quantify the CATE estimation uncertainty in the finite sample size regime.

6 Experiments

Since we have no access to the true CATE values with real-world data, we evaluated the performance of our method with synthetic and semi-synthetic data.

Synthetic data: As with Nie and Wager (2021), we prepared synthetic data that follow the partially linear model (4).

We generated four datasets with binary treatment using Setup A, B, C, and D in Nie and Wager (2021). In all setups, the number of features $\boldsymbol{X}$ is $d = 6$. Each setup provides different formulations of distributions $p(\boldsymbol{x})$ and $p(t|\boldsymbol{x})$ and functions $\theta(\boldsymbol{X})$ and $f(\boldsymbol{X})$ in (4). The main difference lies in functions $\theta(\boldsymbol{X})$ and $f(\boldsymbol{X})$: smooth θ and f (Setup A); smooth θ and non-differentiable f (Setup B); constant θ and smooth f (Setup C); and non-differentiable θ and f (Setup D). We detail these setups in Appendix D.1.

As regards continuous-valued treatment setup, we modified the above four setups in Nie and Wager (2021) to generate the values of treatment $T \in \mathbb{R}$. In particular, we considered its data-generating process $T = \rho(\boldsymbol{X}) + \eta$ (i.e., (19)) and formulated function ρ in a different way: the linear function (Setup A), the constant function (Setup B), and the non-linear and non-differentiable function (Setups C and D).

Semi-synthetic data: For binary treatment setup, we used the Atlantic Causal Inference Conference (ACIC) dataset (Shimoni et al. 2018), including 1000 observations (514 of whom are treated and 486 are untreated). The data of $d = 177$ features come from the Linked Birth and Infant Death Data (LBIDD) (MacDorman and Atkinson 1998), while those of treatment and outcome are simulated.

Unfortunately, there is no well-established benchmark dataset for the continuous-valued treatment setup, unlike the binary treatment setup. For this reason, we focus only on synthetic data experiments for continuous-valued cases.

Baselines: We compared our method with the three baselines: the Bayesian causal forest (BCF) method (Hahn, Murray, and Carvalho 2020),[6] the R Learner (Nie and Wager 2021),[7] which employs a kernel function to infer function θ in the partially linear model (4), and the Bayesian linear regression model. In Appendix F, we present the comparison with the additional baseline, the stableCFR method (Wu

[5] For continuous-valued treatment $t \in \mathbb{R}$, it is called a *generalized propensity score* (Imbens 2000).

[6] https://github.com/jaredsmurray/bcf

[7] We used the original implementation in https://github.com/xnie/rlearner for the binary treatment setup. As regards continuous-valued treatment setup, since there is no original implementation, we employed the KernelDML class in the EconML package downloaded from https://econml.azurewebsites.net/index.html.

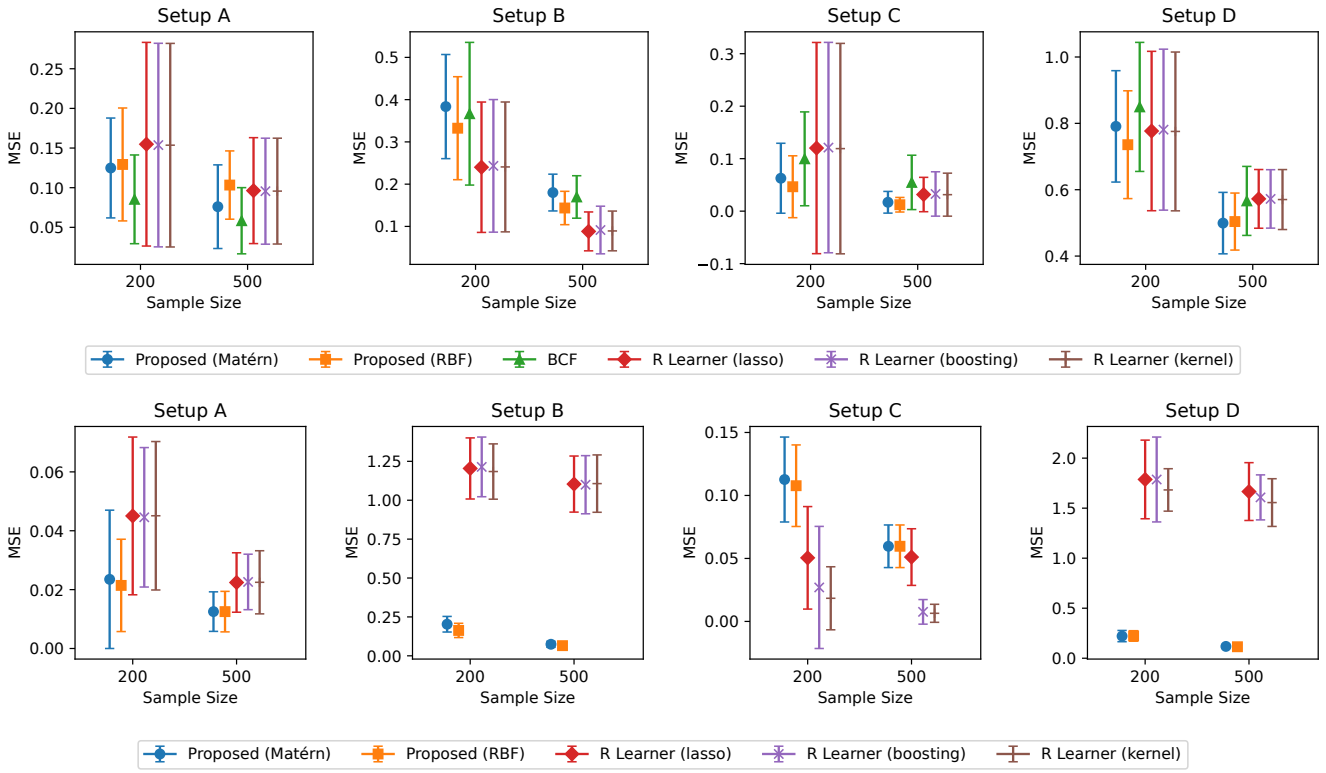

Figure 1: MSEs on synthetic datasets. (Top): binary treatment setup; (Bottom): continuous-values treatment setup. Lower is better.

et al. 2023), which is a recent neural-network-based method.

As regards the continuous-valued treatment setup, we compared our method only with the R Learner because the original implementation of BCF is unavailable.

To examine the advantages of using a partially linear model, we also compared with the Bayes optimal estimator when employing a linear model defined as $Y = (\boldsymbol{\beta}_\theta^\top \boldsymbol{X})T + \boldsymbol{\beta}_f^\top \boldsymbol{X} + \varepsilon$. In this case, the Bayes optimal estimator of the CATE is given by the posterior mean of $\boldsymbol{\beta}_\theta^\top \boldsymbol{X}$.

Evaluation measures: To measure the CATE estimation performance, we conducted 100 experiments and computed the average and standard deviation of the mean squared error (MSE): $\frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}(\tilde{\boldsymbol{x}}_i) - \theta(\tilde{\boldsymbol{x}}_i))^2$, where $\hat{\theta}(\cdot)$ denotes the CATE estimated with training data, and $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_m \overset{i.i.d.}{\sim} p(\boldsymbol{x})$ are the feature values in test data. In all experiments, we used $n = 200$ or $n = 500$ observations as training data and $m = 100$ observations as test data. With the ACIC dataset, the true CATE value, $\theta(\boldsymbol{x}_i)$, is given by a difference between the simulated potential outcomes, and we evaluated the MSE by randomly selecting training and test data from 1000 observations. Note that under the continuous-valued treatment setup, CATE $\theta(\boldsymbol{x})$ in the MSE corresponds to $(t' - t)\theta(\boldsymbol{x})$ in (5) when $t' = t + 1$.

To examine the performance of uncertainty estimation, we computed the 95% credible interval for the Bayesian methods (i.e., the proposed method and BCF) and the 95% confidence interval for the Frequentist-based method (i.e., the R Learner). To measure the quality of these intervals, we computed the coverage ratio (i.e., the ratio in which the true CATE value is included) and the length of the interval.

Results: Figure 1 presents the MSEs of each method on synthetic datasets. Here, we omitted the results with the Bayesian linear regression model due to its extremely large MSE values compared with other methods. See Appendix F for the results.

Our method achieved better or more competitive performance in binary and continuous-valued treatment setups than the BCF and the R Learner. Our method worked well when the data were generated from the partially linear model with non-differentiable functions (i.e., Setups B and D). In particular, our method outperformed the other two baselines under Setup D. These results demonstrate that even if function $\theta(\cdot)$, which represents the CATE, is non-differentiable, our method can approximate it with the smooth functions induced by the kernel function. One of the reasons why our method can perform such effective inference is that it does not rely on any approximate posterior computation, unlike the BCF method. The proposed method performs worse than other methods in Setup C of the continuous-values treatment setup. Our method failed to approximate the CATE function θ, which was given as a constant function. This is because it can be difficult to approximate such a too-simple function with smooth functions in a small sample size setting.

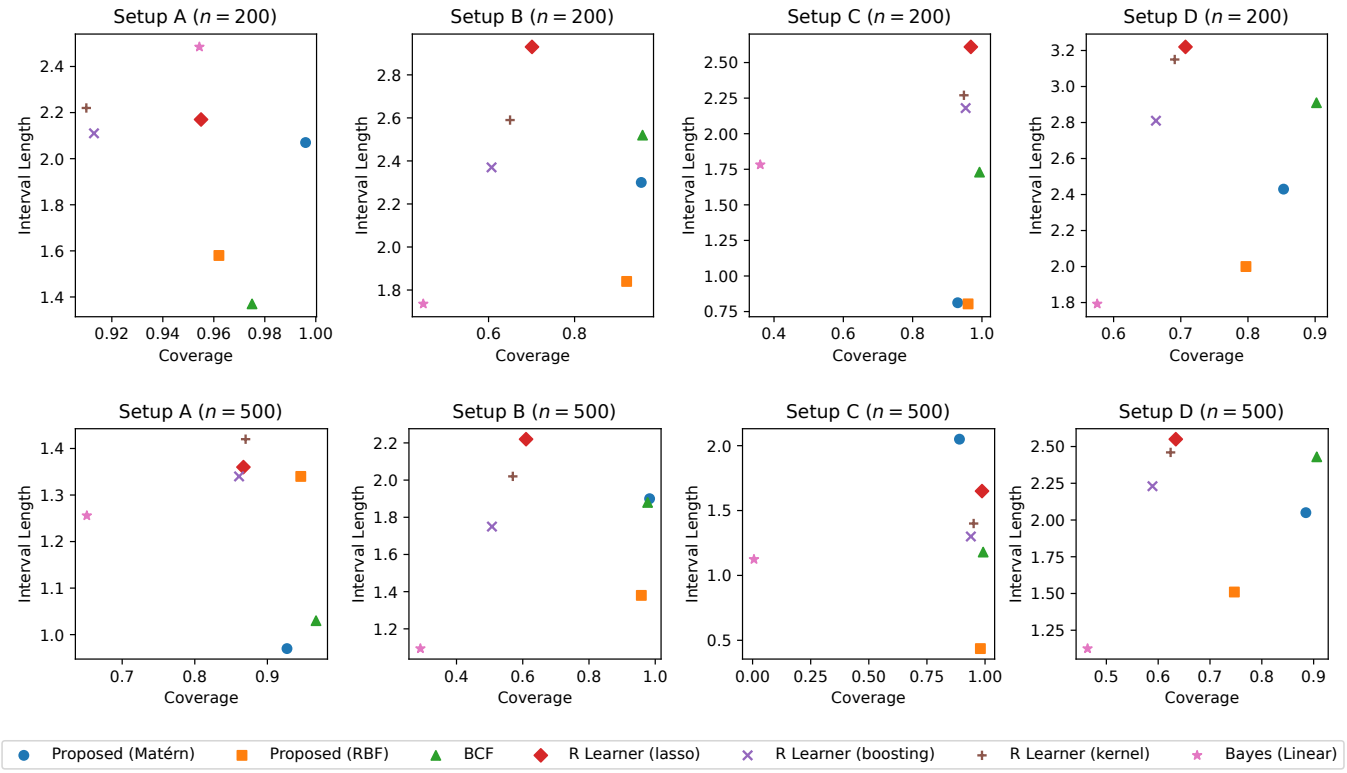

Figure 2: Coverage ratio and interval length of credible/confidence intervals on synthetic data under binary treatment setup. (Top): the sample size $n = 200$; (Bottom): the sample size $n = 500$. Methods closer to the bottom right corner are better.

With the ACIC dataset, we obtained similar results (Figure 3). We observed that the performance of the R Learner strongly depended on the choice of regression models for estimating the conditional expectations, $\mathbb{E}[T|\boldsymbol{X} = \boldsymbol{x}]$ and $\mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}]$. When using kernel regression (i.e., R Learner (kernel); omitted in Figure 3), the MSEs were extremely large: 71.8 ± 11.0 ($n = 200$) and 52.3 ± 4.3 ($n = 500$). Such unstable performance may raise serious doubt for practitioners because appropriately selecting the regression model requires deep understanding of the data analysis. By contrast, our method worked well, regardless of the choice of kernel functions (i.e., Matérn and RBF kernels).

Figure 2 shows the performance of uncertainty quantification on synthetic data. A higher coverage ratio and a shorter interval length are better: the methods closer to the bottom right corner exhibit better performance. Especially when n is small, our method and BCF achieved larger coverage ratios than the R Learner while keeping the length small, thus demonstrating the effectiveness of the Bayesian approaches.

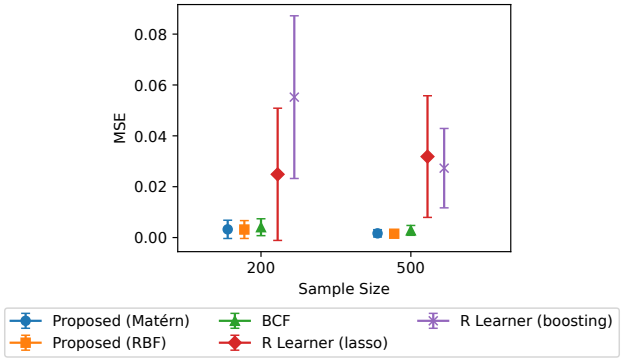

Figure 3: MSEs on ACIC dataset. MSEs of R Learner (kernel) are omitted due to their extremely large values.

7 Conclusion

We proposed a Bayesian framework that quantifies the CATE estimation uncertainty with the posterior distribution. The key idea is to put Gaussian process priors on the nonparametric components in a partially linear model. This idea offers a computationally efficient Bayesian inference with a closed-form posterior of the CATE. Moreover, it enables us to incorporate prior knowledge about the CATE, leading to an effective posterior inference, as empirically demonstrated in our experimental results (Appendix E).

Theoretically, we prove that the posterior has asymptotic consistency under some mild conditions. Our future work constitutes further investigation. In particular, we will investigate the relationship between the minimax information rate and the assumed class of nonlinear functions, as with the results of the Gaussian-process-based model (Alaa and van der Schaar 2018).

Acknowledgments

This research is partially supported by the Telecommunications Advancement Foundation, and No. 22K12156 of Grant-in-Aid for Scientific Research Category (C), Japan Society for the Promotion of Science.

References

Alaa, A. M.; and van der Schaar, M. 2018. Bayesian nonparametric causal inference: Information rates and learning algorithms. *IEEE Journal of Selected Topics in Signal Processing*, 12(5): 1031–1046.

Bica, I.; Jordon, J.; and van der Schaar, M. 2020. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33: 16434–16445.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer.

Chernozhukov, V.; Chetverikov, D.; Demirer, M.; Duflo, E.; Hansen, C.; Newey, W.; and Robins, J. 2018. Double/debiased machine learning for treatment and structural parameters.

Dorie, V.; Hill, J.; Shalit, U.; Scott, M.; and Cervone, D. 2017. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition.

Elwert, F.; and Winship, C. 2014. Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology*, 40: 31–53.

Engle, R. F.; Granger, C. W.; Rice, J.; and Weiss, A. 1986. Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American statistical Association*, 81(394): 310–320.

Ghosal, S.; Ghosh, J. K.; and Ramamoorthi, R. 1999. Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1): 143–158.

Hahn, P. R.; Murray, J. S.; and Carvalho, C. M. 2020. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3): 965–1056.

Hamilton, M. A.; Russo, R. C.; and Thurston, R. V. 1977. Trimmed Spearman-Karber method for estimating median lethal concentrations in toxicity bioassays. *Environmental science & technology*, 11(7): 714–719.

Hernán, M. A.; and Robins, J. M. 2020. *Causal Inference: What if. Boca Raton: Chapman & Hill/CRC*.

Hill, A. V. 1910. The possible effects of the aggregation of the molecules of hemoglobin on its dissociation curves. *Journal of Physiology*, 40: iv–vii.

Hill, J. L. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1): 217–240.

Hirano, K.; and Imbens, G. W. 2004. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164: 73–84.

Imbens, G. W. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3): 706–710.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Johansson, F.; Shalit, U.; and Sontag, D. 2016. Learning representations for counterfactual inference. In *ICML*, 3020–3029.

Li, F.; Ding, P.; and Mealli, F. 2023. Bayesian causal inference: a critical review. *Philosophical Transactions of the Royal Society A*, 381(2247): 20220153.

MacDorman, M. F.; and Atkinson, J. O. 1998. Infant mortality statistics from the linked birth/infant death data set–1995 period data. *Mon Vital Stat Rep, 46(suppl 2):1-22*.

Nie, X.; and Wager, S. 2021. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2): 299–319.

Quinonero-Candela, J.; and Rasmussen, C. E. 2005. A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6: 1939–1959.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5): 688.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: Generalization bounds and algorithms. In *ICML*, 3076–3085.

Shimoni, Y.; Yanover, C.; Karavani, E.; and Goldschmnidt, Y. 2018. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046*.

Van der Vaart, A. W. 2000. *Asymptotic statistics*, volume 3. Cambridge university press.

Wager, S.; and Athey, S. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523): 1228–1242.

Wiens, J.; Guttag, J.; and Horvitz, E. 2014. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 21(4): 699–706.

Williams, C. K.; and Rasmussen, C. E. 2006. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Woody, S.; Carvalho, C. M.; Hahn, P. R.; and Murray, J. S. 2020. Estimating heterogeneous effects of continuous exposures using bayesian tree ensembles: revisiting the impact of abortion rates on crime. *arXiv preprint arXiv:2007.09845*.

Wu, A.; Kuang, K.; Xiong, R.; Li, B.; and Wu, F. 2023. Stable estimation of heterogeneous treatment effects. In *International Conference on Machine Learning*, 37496–37510. PMLR.

Yao, L.; Chu, Z.; Li, S.; Li, Y.; Gao, J.; and Zhang, A. 2021. A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5): 1–46.

Yeager, D. S.; Hanselman, P.; Walton, G. M.; Murray, J. S.; Crosnoe, R.; Muller, C.; Tipton, E.; Schneider, B.; Hulleman, C. S.; Hinojosa, C. P.; et al. 2019. A national experiment reveals where a growth mindset improves achievement. *Nature*, 573(7774): 364–369.

Yoon, J.; Jordon, J.; and Van Der Schaar, M. 2018. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *ICLR*.