Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes

Ahmed M. Alaa[1] and Mihaela van der Schaar[1,2]
[1]Electrical Engineering Department
University of California, Los Angeles
ahmedmalaa@ucla.edu

[2]Department of Engineering Science
University of Oxford
mihaela.vanderschaar@eng.ox.ac.uk

May 4, 2017

Abstract

Predicated on the increasing abundance of electronic health records, we investigate the problem of inferring *individualized* treatment effects using observational data. Stemming from the potential outcomes model, we propose a novel *multi-task* learning framework in which factual and counterfactual outcomes are modeled as the outputs of a function in a vector-valued reproducing kernel Hilbert space (vvRKHS). We develop a nonparametric Bayesian method for learning the treatment effects using a multi-task Gaussian process (GP) with a linear coregionalization kernel as a prior over the vvRKHS. The Bayesian approach allows us to compute individualized measures of confidence in our estimates via pointwise credible intervals, which are crucial for realizing the full potential of precision medicine. The impact of selection bias is alleviated via a *risk-based empirical Bayes* method for adapting the multi-task GP prior, which jointly minimizes the empirical error in factual outcomes and the uncertainty in (unobserved) counterfactual outcomes. We conduct experiments on a dataset for a randomized trial of an interventional social program for improving cognitive skills of premature infants, and an observational dataset for the survival benefit of left ventricular assist devices in cardiac patients wait-listed for a heart transplant. In both experiments, we show that our method significantly outperforms the state-of-the-art.

1 Introduction

Clinical trials entail enormous costs: the average costs of multi-phase trials in vital therapeutic areas such as the respiratory system, anesthesia and oncology are \$115.3 million, \$105.4 million, and \$78.6 million, respectively [1]. Moreover, due to the high costs and difficulty of patient recruitment, randomized controlled trials often exhibit small sample sizes, which hinders the discovery of heterogeneous therapeutic effects across different patient subgroups [2]. With the advent of electronic health records (EHRs), currently deployed in more than 75% of hospitals in the U.S. according to the latest ONC data brief[1], there has been a growing interest in using machine learning algorithms to infer heterogeneous treatment effects from the readily available observational EHR data. This interest glints in recent initiatives, such as STRATOS [3], which focus on developing new methods for conducting retrospective cohort studies, in addition to the various recent works on causal inference from observational data developed by the machine learning community [4-11].

2 Problem Setup

We consider the setting in which a specific treatment is applied to a population of subjects (patients), where each subject i possesses a d-dimensional *feature* $X_i \in \mathcal{X}$, and two *potential outcomes* $Y_i^{(1)}, Y_i^{(0)} \in \mathbb{R}$ that correspond to the subject's response with and without the treatment, respectively. The potential outcomes $Y_i^{(0)}$ and $Y_i^{(1)}$ are random variables that are drawn from a conditional distribution $\mathbb{P}(Y_i^{(0)}, Y_i^{(1)} \,|\, X_i = x)$. The causal effect of the treatment on every *individual* subject manifests through the random variable $(Y_i^{(1)} - Y_i^{(0)}) \,|\, X_i = x$. Hence, we define the *individualized treatment effect* (ITE) for subject i as the expected effect of the treatment on that subject, i.e.

$$T(x) = \mathbb{E}\left[Y_i^{(1)} - Y_i^{(0)} \,\middle|\, X_i = x \right]. \tag{1}$$

Our goal is to efficiently estimate the function $T(x)$ from an *observational* dataset, e.g. a retrospective cohort study [26, 34] or a hospital's electronic health record [15]. A typical observational dataset $\mathcal{D}$ comprises n independent and identically distributed samples of the random tuple $\{X_i, W_i, W_i\, Y_i^{(W_i)} + (1 - W_i)\, Y_i^{(1-W_i)}\}$, where $W_i \in \{0, 1\}$ is a treatment assignment indicator that indicates whether or not subject i has received the treatment under consideration. The outcomes $Y_i^{(W_i)}$ and $Y_i^{(1-W_i)}$ are known as the *factual* and the *counterfactual* outcomes, respectively [6, 11]. Treatment assignments are generally dependent on features, i.e. $W_i \not\perp\!\!\!\perp X_i$. The conditional distribution $\mathbb{P}(W_i = 1 | X_i = x)$, also known as the *propensity score* of subject i [1, 7, 21], reflects the underlying (unknown) policy for assigning the treatment to subjects. Throughout this paper, we respect the standard assumptions of *unconfoundedness* (or *ignorability*) and *overlap* [1, 3, 8, 12, 15, 32-34]. The former posits that treatment assignments are independent of the outcomes conditional on features, i.e. $Y_i^{(0)}, Y_i^{(1)} \perp\!\!\!\perp W_i \,|\, X_i$, whereas the latter requires that

[1] `https://www.healthit.gov/sites/default/files/briefs/`

$0 < \mathbb{P}(W_i = 1|X_i = x) < 1, \forall x \in \mathcal{X}$. The setting described above is known as the Rubin-Neyman causal model [22, 23, 36].

Individual-based causal inference using observational data is challenging. Since we only observe one of the potential outcomes for every subject i, we never observe the treatment effect $Y_i^{(1)} - Y_i^{(0)}$ for any of the subjects, and hence we cannot resort to standard supervised learning to estimate $T(x)$. Moreover, the dataset $\mathcal{D}$ exhibits *selection bias*, which may render the estimates of $T(x)$ inaccurate if the treatment assignment for individuals with $X_i = x$ is strongly biased (i.e. $\mathbb{P}(W_i = 1|X_i = x)$ is close to 0 or 1). Since our primary motivation for addressing this problem comes from its application potential in precision medicine, it is important to associate our estimate of $T(.)$ with a pointwise measure of confidence in order to properly guide therapeutic decisions for individual patients.

3 Multi-task Learning for Causal Inference

Vector-valued Potential Outcomes Function We adopt the following *signal-in-white-noise* model for the potential outcomes:

$$Y_i^{(w)} = f_w(X_i) + \epsilon_{i,w},\ w \in \{0, 1\} \tag{2}$$

where $\epsilon_{i,w} \sim \mathcal{N}(0, \sigma_w^2)$ is a Gaussian noise variable. It follows from (2) that $\mathbb{E}[Y_i^{(w)} \,|\, X_i = x] = f_w(x)$, and hence the ITE can be estimated as $\hat{T}(x) = f_1(x) - f_0(x)$. Previous works adopted two different approaches for estimating the ITE function $T(x)$ using $\mathcal{D}$. The first approach learns two separate regression models $f_w(.) : \mathcal{X} \to \mathbb{R}, w \in \{0, 1\}$, using data from treated and control groups, and estimates the ITE as $\hat{T}(x) = f_1(x) - f_0(x)$ [13, 25]. The second approach learns one regression model that treats the treatment assignment as an input feature, i.e. $f_w(x) = f(x, w), f(., .) : \mathcal{X} \times \{0, 1\} \to \mathbb{R}$, and estimates the ITE as $\hat{T}(x) = f(x, 1) - f(x, 0)$ [11, 18, 23]. We depart from those approaches by introducing a new regression model that learns a vector-valued *potential outcomes* (PO) function $\mathbf{f}(.) : \mathcal{X} \to \mathbb{R}^2$, with d inputs (features) and 2 outputs (potential outcomes); the ITE estimate is the projection of the PO function on the vector $\mathbf{e} = [-1\ \ 1]^T$, i.e. $\hat{T}(x) = \mathbf{f}^T(x)\,\mathbf{e}$.

Consistent pointwise estimation of the ITE function $T(x)$ generally requires restricting the PO function $\mathbf{f}(x)$ to some regularity class [23]. To this end, we model the PO function $\mathbf{f}(x)$ as belonging to a *vector-valued Reproducing Kernel Hilbert Space* (vvRKHS) $\mathcal{H}_{\mathbf{K}}$ equipped with an inner product $\langle ., . \rangle_{\mathcal{H}_{\mathbf{K}}}$, and with a *reproducing kernel* $\mathbf{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{2\times 2}$, where $\mathbf{K}$ is a symmetric matrix-valued function such that for any $x, x' \in \mathcal{X}, \mathbf{K}(x, x')$ is a positive semi-definite matrix [2]. Our choice for the vvRKHS as a regularity class for $\mathbf{f}(x)$ is motivated by its algorithmic advantages[2]; by virtue of the *representer theorem*, we know that

[2] Consistency of learning functions in the RKHS was studied in [Sec. 7.2.1, 15]. Selection of the RKHS as a regularity class for the PO function $\mathbf{f}(x)$ is not limiting since different selections for the kernel $\mathbf{K}$ correspond to various familiar function spaces, including the Sobolev space [19].

learning the infinite-dimensional PO function entails estimating a finite number of coefficients evaluated at the input points $\{X_i\}_{i=1}^n$ in $\mathcal{D}$ [22].

Multi-task Learning The vector-valued model for the PO function allows conceptualizing causal inference as a multi-task learning problem. That is, the observational dataset $\mathcal{D} = \{X_i, W_i, Y_i^{(W_i)}\}_{i=1}^n$ can be though of as comprising training data for two (related) learning tasks with target functions $f_0(.)$ and $f_1(.)$, and with W_i acting as the "task index" for the i^{th} training point [2, 5]. For an estimate $\mathbf{f}(x)$ of the PO function, the corresponding true loss functional is given by

$$\mathcal{L}(\mathbf{f}) = \int_{x \in \mathcal{X}} \left(\mathbf{f}^T(x)\,\mathbf{e} - T(x)\right)^2 \cdot \mathbb{P}(X = x)\,dx. \tag{3}$$

The loss functional in (3), originally introduced in [9], is known as the *precision in estimating heterogeneous effects* (PEHE), and is used to quantify the "goodness" of $\hat{T}(x)$ in capturing the heterogeneity of the true ITE function $T(x)$ [3, 9, 11, 18]. A conspicuous challenge that arises when learning the "PEHE-optimal" PO function $\mathbf{f}$ is that we cannot compute the empirical PEHE for a particular $\mathbf{f} \in \mathcal{H}_{\mathbf{K}}$ since the treatment effect samples $\{Y_i^{(1)} - Y_i^{(0)}\}_{i=1}^n$ are not available in the observational data. On the other hand, using a loss function that evaluates the squared losses of $f_0(x)$ and $f_1(x)$ separately (as in conventional multi-task learning [Sec. 3.2, 2]) can be highly problematic: in the presence of a strong selection bias, the empirical losses for $\mathbf{f}(.)$ with respect to factual outcomes may not generalize to the counterfactual outcomes, leading to a large PEHE loss.

In order to establish a proxy for the empirical PEHE, we first consider an "oracle" that has access to counterfactual outcomes. For such an oracle, the finite-sample empirical PEHE is

$$\hat{\mathcal{L}}(\mathbf{f}; \mathbf{K}, \mathbf{Y}^{(\mathbf{W})}, \mathbf{Y}^{(\mathbf{1}-\mathbf{W})}) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{f}^T(X_i)\,\mathbf{e} - (1 - 2W_i)\left(Y_i^{(1-W_i)} - Y_i^{(W_i)}\right)\right)^2, \tag{4}$$

where $\mathbf{Y}^{(\mathbf{W})} = [Y_1^{(W_1)}, \ldots, Y_n^{(W_n)}]^T$ and $\mathbf{Y}^{(\mathbf{1}-\mathbf{W})} = [Y_1^{(1-W_1)}, \ldots, Y_n^{(1-W_n)}]^T$. When $\mathbf{Y}^{(\mathbf{1}-\mathbf{W})}$ is accessible, estimating the PEHE-optimal PO function $\mathbf{f}(.)$ becomes an ordinary supervised learning problem, the solution to which is given by the following representer Theorem.

Theorem 1 (Representer Theorem for Oracle Causal Inference). *For any* $\mathbf{f}^* \in \mathcal{H}_{\mathbf{K}}$ *satisfying*

$$\mathbf{f}^* = \arg\min_{\mathbf{f} \in \mathcal{H}_{\mathbf{K}}} \hat{\mathcal{L}}(\mathbf{f}; \mathbf{K}, \mathbf{Y}^{(\mathbf{W})}, \mathbf{Y}^{(\mathbf{1}-\mathbf{W})}) + \lambda\,||\mathbf{f}||^2_{\mathcal{H}_{\mathbf{K}}},\ \lambda \in \mathbb{R}_+, \tag{5}$$

we have that $\hat{T}^*(.) = (\mathbf{f}^*(.))^T\,\mathbf{e} \in$ *span*$\{\tilde{\mathbf{K}}(., X_1), \ldots, \tilde{\mathbf{K}}(., X_n)\}$, *where* $\tilde{\mathbf{K}}(.,.) = \mathbf{e}^T\,\mathbf{K}(.,.)\,\mathbf{e}$. *That is,* $\hat{T}^*(.)$ *admits a representation* $\hat{T}^*(.) = \sum_{i=1}^n \alpha_i\,\tilde{\mathbf{K}}(., X_i)$, $\alpha = [\alpha_1, \ldots, \alpha_n]^T$, *where*

$$\alpha = (\tilde{\mathbf{K}}(\mathbf{X}, \mathbf{X}) + n\,\lambda\,\mathbf{I})^{-1}((\mathbf{1} - 2\mathbf{W}) \odot (\mathbf{Y}^{(\mathbf{1}-\mathbf{W})} - \mathbf{Y}^{(\mathbf{W})})), \tag{6}$$

where $\odot$ denotes component-wise product, $\tilde{\mathbf{K}}(\mathbf{X},\mathbf{X}) = (\tilde{\mathbf{K}}(X_i,X_j))_{i,j}$, $\mathbf{W} = [W_1,\ldots,W_n]^T$. □

Theorem 1 follows from the generalized representer Theorem [22] (The proof is provided in Appendix A), and it characterizes the PEHE-optimal interpolant $\mathbf{f}^*(.)$ obtained via regularized empirical PEHE minimization by an oracle learner that knows the counterfactual outcomes.

A Bayesian Perspective on Causal Inference Regularized empirical risk minimization in vvRKHS is equivalent to Bayesian inference with a Gaussian process (GP) prior [Sec. 2.2, 2]. Therefore, we can interpret $\hat{T}^*(.)$ as the posterior mean of $T(.)$ given a GP prior with a covariance kernel $\tilde{\mathbf{K}}$. Since we know from Theorem 1 that $\tilde{\mathbf{K}} = \mathbf{e}^T\mathbf{K}\mathbf{e}$, the GP prior on $T(.)$ is equivalent to a *multi-task* GP prior on the PO function $\mathbf{f}(.)$ with a kernel $\mathbf{K}$. The Bayesian view of the problem is advantageous for two reasons. First, it allows computing individualized (pointwise) measures of uncertainty in $\hat{T}(.)$ via posterior credible intervals. Second, it allows reasoning about the unobserved counterfactual outcomes in a Bayesian fashion, and hence provides a natural proxy for the oracle learner's empirical PEHE in (4). In particular, we define the Bayesian PEHE risk $R(\theta,\mathbf{f};\mathcal{D})$ for a PO function $\mathbf{f}(.)$ that belongs to a vvRKHS with kernel $\mathbf{K}_\theta$, where $\theta \in \Theta$ is a kernel *hyper-parameter*, as follows

$$R(\theta,\mathbf{f};\mathcal{D}) = \mathbb{E}_\theta\left[\hat{\mathcal{L}}(\mathbf{f};\mathbf{K}_\theta,\mathbf{Y}^{(\mathbf{W})},\mathbf{Y}^{(\mathbf{1-W})})\,\Big|\,\mathcal{D}\right], \tag{7}$$

The expectation in (7) is taken with respect to $\mathbf{Y}^{(\mathbf{1-W})}|\mathcal{D}$. The Bayesian PEHE risk $R(\theta,\mathbf{f};\mathcal{D})$ is simply the oracle learner's empirical loss in (4) marginalized over the posterior distribution of the unobserved counterfactuals $\mathbf{Y}^{(\mathbf{1-W})}$, and hence it incorporates the posterior uncertainty in counterfactual outcomes without explicit propensity modeling. The optimal interpolant $\mathbf{f}^*(.)$ that minimizes the Bayesian PEHE risk is given in the following Theorem.

Theorem 2 (Risk-based Empirical Bayes). *The minimizer $(\mathbf{f}^*,\theta^*)$ of $R(\theta,\mathbf{f};\mathcal{D})$ is given by*

$$\mathbf{f}^* = \mathbb{E}_{\theta^*}[\,\mathbf{f}\,|\,\mathcal{D}\,],\;\; \theta^* = \arg\min_{\theta\in\Theta}\left[\underbrace{\left\|\mathbf{Y}^{(\mathbf{W})} - \mathbb{E}_\theta[\,\mathbf{f}\,|\,\mathcal{D}\,]\right\|_2^2}_{\text{Empirical factual error}} + \underbrace{\left\|\mathrm{Var}_\theta[\,\mathbf{Y}^{(\mathbf{1-W})}\,|\,\mathcal{D}\,]\right\|_1}_{\text{Posterior counterfactual variance}}\right],$$

where $Var_\theta[.|.]$ is the posterior variance and $\|.\|_p$ is the p-norm. □

The proof is provided in Appendix B. Theorem 2 shows that model selection (i.e. selecting the hyper-parameter θ) is instrumental in alleviating the impact of selection bias. This is because, as the Theorem states, the optimal hyper-parameter θ^* minimizes the empirical squared loss of $\mathbf{f}^*$ with respect to the factual outcomes $\mathbf{Y}^{(\mathbf{W})}$ with the posterior variance of the counterfactual outcomes as a regularizer. Hence, when the observational data exhibits a significant selection bias, θ^* will carve a kernel $\mathbf{K}_{\theta^*}$ that not only fits the factual outcomes, but also generalizes well to the unobserved counterfactuals. It comes as no surprise that $\mathbf{f}^* = \mathbb{E}_{\theta^*}[\,\mathbf{f}\,|\,\mathcal{D}\,]$; that is, $\mathbb{E}_{\theta^*}[\,\mathbf{f}\,|\,\mathcal{D}, \mathbf{Y}^{(\mathbf{1-W})}\,]$ is equivalent to the Oracle's solution in Theorem 1, and thus by the law of iterated expectations, $\mathbb{E}_{\theta^*}[\,\mathbf{f}\,|\,\mathcal{D}\,] = \mathbb{E}_{\theta^*}[\,\mathbb{E}_{\theta^*}[\,\mathbf{f}\,|\,\mathcal{D}, \mathbf{Y}^{(\mathbf{1-W})}\,]\,|\,\mathcal{D}\,]$ is the oracle's solution marginalized over the posterior distribution of counterfactual outcomes. The ITE estimate is given by $\hat{T}^*(x) = \mathbb{E}_{\theta^*}[\,\mathbf{f}^T\,|\,\mathcal{D}\,]\mathbf{e}$.

The model selection approach suggested by Theorem 2 is known as the *risk-based empirical Bayes* method [Sec. 2, 19], and it departs from *likelihood-based empirical Bayes* (i.e. evidence maximization [5]) in that it calibrates the GP prior so as to minimize the risk of the posterior mean $\mathbb{E}_{\theta^*}[\,\mathbf{f}\,|\,\mathcal{D}\,]$ (see [Eq. (1.5), 19]) rather than maximizing the likelihood of the observations. While none of the two methods display a conclusive superiority to the other in ordinary non-parametric regression, the risk-based method is clearly a more sensible approach in our setting. This is because evidence maximization selects a kernel $\mathbf{K}_\theta$ that only fits the factual outcomes, and hence may not necessarily lead to a good estimate of the treatment effect. Contrarily, the risk-based method penalizes the factual empirical error $(Y_i^{(W_i)} - \mathbb{E}_\theta[\,\mathbf{f}(X_i)\,|\,\mathcal{D}\,])^2$ for every subject i with the posterior variance of her counterfactual outcome $\mathrm{Var}_\theta[\,Y_i^{(1-W_i)}\,|\,\mathcal{D}\,]$; we might think of this procedure as being a Bayesian analog for propensity score re-weighting [1, 4, 8], with the propensity score indirectly manifesting in the posterior variance of the counterfactual outcome.

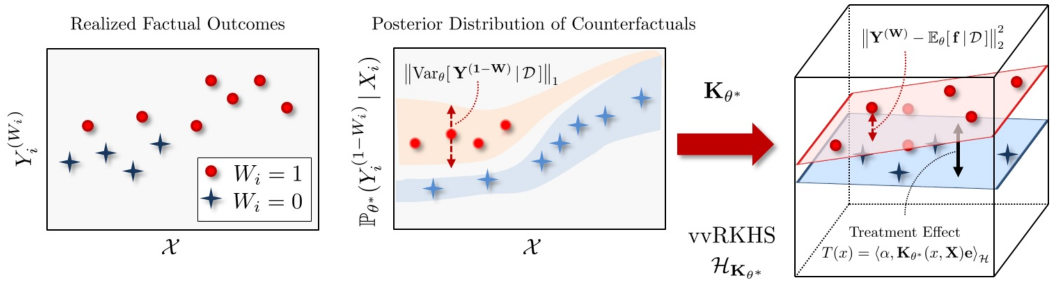
Realized Factual Outcomes
$Y_i^{(W_i)}$
$W_i = 1$
$W_i = 0$
$\mathcal{X}$
Posterior Distribution of Counterfactuals
$\mathbb{P}_{\theta^*}(Y_i^{(1-W_i)}\,|\,X_i)$
$\|\mathrm{Var}_\theta[\,\mathbf{Y}^{(\mathbf{1-W})}\,|\,\mathcal{D}\,]\|_1$
$\mathcal{X}$
$\mathbf{K}_{\theta^*}$
vvRKHS
$\mathcal{H}_{\mathbf{K}_{\theta^*}}$
$\|\mathbf{Y}^{(\mathbf{W})} - \mathbb{E}_\theta[\,\mathbf{f}\,|\,\mathcal{D}\,]\|_2^2$
Treatment Effect
$T(x) = \langle \alpha, \mathbf{K}_{\theta^*}(x, \mathbf{X})\mathbf{e} \rangle_{\mathcal{H}}$

Figure 1: Pictorial depiction for model selection via risk-based empirical Bayes.

A Feature Space Interpretation

4 Causal Multi-task Gaussian Processes (CMGPs)

In this Section, we provide a recipe for causal inference using multi-task GPs. Following the discussion in Section 2, we model the PO function $\mathbf{f} \sim \mathcal{GP}(0, \mathbf{K}_\theta)$ as a random function drawn from a GP prior with d inputs and 2 outputs; hence $T(x) \sim \mathcal{GP}(0, \tilde{\mathbf{K}}_\theta)$ is drawn from a single-output GP with $\tilde{\mathbf{K}}_\theta = \mathbf{e}^T \mathbf{K}_\theta\,\mathbf{e}$. We call this model a *Causal Multi-task Gaussian Process* (CMGP).

Constructing the CMGP Kernel The two response surfaces $f_0(.)$ and $f_1(.)$ may display different levels of heterogeneity (smoothness), and may have different relevant features. This is often the case in medical settings where, depending on the nature of the intervention, treated patient groups are generally more likely to exhibit a more heterogeneous response surface as compared to the control groups. Standard intrinsic coregionalization models for constructing vector-valued kernels impose the same covariance parameters for all outputs [5], which indeed limits the interaction between the treatment assignments and the patients' features. To that end, we construct a *linear model of coregionalization* (LMC) [2, 15], which mixes two intrinsic coregionalization models as follows

$$\mathbf{K}_\theta(x, x') = \mathbf{A}_0\, k_0(x, x') + \mathbf{A}_1\, k_1(x, x'), \tag{8}$$

where $k_w(x, x'), w \in \{0, 1\}$, is the *radial basis function* (RBF) with automatic relevance determination, i.e. $k_w(x, x') = \exp\left(-\frac{1}{2}(x - x')^T\, \mathbf{R}_w^{-1}\, (x - x')\right), \mathbf{R}_w = \text{diag}(\ell_{1,w}^2, \ell_{2,w}^2, \ldots, \ell_{d,w}^2)$, with $\ell_{d,w}$ being the *length scale* parameter of the d^{th} feature in $k_w(.,.)$, whereas $\mathbf{A}_0$ and $\mathbf{A}_1$ are given by

$$\mathbf{A}_0 = \begin{bmatrix} \beta_{00}^2 & \rho_0 \\ \rho_0 & \beta_{01}^2 \end{bmatrix}, \; \mathbf{A}_1 = \begin{bmatrix} \beta_{10}^2 & \rho_1 \\ \rho_1 & \beta_{11}^2 \end{bmatrix}. \tag{9}$$

The parameters $(\beta_{ij}^2)_{ij}$ and $(\rho_i)_i$ determine the variances and correlations of the two response surfaces $f_0(x)$ and $f_1(x)$. The LMC kernel introduces degrees of freedom that allow the two response surfaces to have different covariance functions and relevant features. When $\beta_{00} >> \beta_{01}$ and $\beta_{11} >> \beta_{10}$, the length scale parameter $\ell_{d,w}$ can be interpreted as the relevance of the d^{th} feature to the response surface $f_w(.)$. The set of all hyper-parameters is $\theta = (\sigma_0, \sigma_1, \mathbf{R}_0, \mathbf{R}_1, \mathbf{A}_0, \mathbf{A}_1)$.

Adapting the Prior via Risk-based Empirical Bayes Following Theorem 2, we adapt the CMGP prior to the observations in $\mathcal{D}$ via risk-based empirical Bayes. In order to avoid overfitting to the factual outcomes $\mathbf{Y}^{(\mathbf{W})}$, we evaluate the empirical error in factual outcomes via leave-one-out cross validation (LOO-CV) with Bayesian regularization [16]; the regularized objective function is thus given by $\hat{R}(\theta; \mathcal{D}) = \eta_0\, Q(\theta) + \eta_1\, \|\theta\|_2^2$, where

$$Q(\theta) = \left\| \text{Var}_\theta[\, \mathbf{Y}^{(\mathbf{1-W})}\, |\, \mathcal{D}] \right\|_1 + \sum_{i=1}^{n} \left(Y_i^{(W_i)} - \mathbb{E}_\theta[\mathbf{f}(X_i)\, |\, \mathcal{D}_{-i}] \right)^2, \tag{10}$$

and $\mathcal{D}_{-i}$ is the dataset $\mathcal{D}$ with subject i removed, whereas η_0 and η_1 are the Bayesian regularization parameters. For the second level of inference, we use the improper Jeffrey's prior as an ignorance prior for the regularization parameters, i.e. $\mathbb{P}(\eta_0) \propto \frac{1}{\eta_0}$ and $\mathbb{P}(\eta_1) \propto \frac{1}{\eta_1}$. This allows us to integrate out the regularization parameters [Sec. 2.1, 16], leading to a revised objective function $\hat{R}(\theta; \mathcal{D}) = n\, \log(Q(\theta)) + (10 + 2\, d)\, \log(\|\theta\|_2^2)$ (See Appendix C for a detailed analysis). It is important to note that LOO-CV with squared loss has often been considered to be unfavorable in ordinary GP regression as it leaves one degree of freedom undetermined [Sec. 5.4.2, 5]; this problem does not arise in our setting since the term $\left\|\text{Var}_\theta[\,\mathbf{Y^{(1-W)}}\,|\,\mathcal{D}\,]\right\|_1$ involves all the variance parameters, and hence the objective function $\hat{R}(\theta;\mathcal{D})$ does not depend solely on the posterior mean.

Causal Inference via CMGPs Algorithm 1 sums up the entire causal inference procedure. It first invokes the routine `Initialize-hyperparameters`, which uses the sample variance and up-crossing rate of $\mathbf{Y^{(W)}}$ to initialize θ (see Appendix D). Such an automated initialization procedure allows running our method without any user-defined inputs, which facilitates its usage by researchers conducting observational studies. (The only inputs to the algorithm are the observational dataset $\mathcal{D}$, and the desired Bayesian coverage rate γ.) Having initialized θ (line 3), the algorithm finds a locally optimal θ^* using gradient descent (lines 5-12), and then estimates the ITE function and the associated credible intervals (lines 13-17). In Algorithm 1, $\mathbf{X} = [\{X_i\}_{W_i=0}, \{X_i\}_{W_i=1}]^T$, $\mathbf{Y} = [\{Y_i^{(W_i)}\}_{W_i=0}, \{Y_i^{(W_i)}\}_{W_i=1}]^T$, $\mathbf{\Sigma} = \text{diag}(\sigma_0^2\,\mathbf{I}_{n-n_1}, \sigma_1^2\,\mathbf{I}_{n_1})$, $n_1 = \sum_i W_i$, $\text{erf}(x) = \frac{1}{\sqrt{\pi}}\int_{-x}^{x} e^{-y^2}dy$, and $\mathbf{K}_\theta(x) = (\mathbf{K}_\theta(x, X_i))_i$.

We use a re-parametrized version of the Adaptive Moment Estimation (ADAM) gradient descent algorithm for optimizing θ [12]; we first apply a transformation $\phi = \exp(\theta)$ to ensure that all covariance parameters remain positive, and then run ADAM to minimize $\hat{R}(\log(\phi_t); \mathcal{D})$. (Analytic expressions for the gradient $\nabla_\phi \hat{R}(\log(\phi_t); \mathcal{D})$ are provided in Appendix D.) The ITE function is estimated as the posterior mean of the CMGP (line 14). The credible interval $\mathcal{C}_\gamma(x)$ with a Bayesian coverage of γ for a subject with feature x is defined as $\mathbb{P}_\theta(T(x) \in C_\gamma(x)) = \gamma$, and is computed straightforwardly using the error function of the normal distribution (lines 15-17). The computational burden of Algorithm 1 is dominated by the $O(n^3)$ matrix inversion in line 13; for large observational studies, this can be ameliorated using conventional sparse approximations [Sec. 8.4, 5].

Algorithm 1 Causal Inference via CMGPs

```
1: **Input:** Observational dataset $\mathcal{D}$, Bayesian coverage $\gamma$
2: **Output:** ITE function $\hat{T}(x)$, credible intervals $\mathcal{C}_\gamma(x)$
3: $\theta \leftarrow$ Initialize-hyperparameters($\mathcal{D}$)
4: $\phi^0 \leftarrow \exp(\theta)$, $t \leftarrow 0$, $m_t \leftarrow 0$, $v_t \leftarrow 0$,
5: **repeat**
6:     $m_{t+1} \leftarrow \beta_1 m_t + (1-\beta_1) \cdot \phi_t \odot \nabla_\phi \hat{R}(\log(\phi_t); \mathcal{D})$
7:     $v_{t+1} \leftarrow \beta_2 v_t + (1-\beta_2) \cdot (\phi_t \odot \nabla_\phi \hat{R}(\log(\phi_t); \mathcal{D}))^2$
8:     $\hat{m}_{t+1} \leftarrow m_t/(1-\beta_1^t)$, $\hat{v}_{t+1} \leftarrow v_t/(1-\beta_2^t)$
9:     $\phi_{t+1} \leftarrow \phi_t \odot \exp\left(-\eta \cdot \hat{m}_{t+1}/(\sqrt{\hat{v}_{t+1}} + \epsilon)\right)$
10:    $t \leftarrow t+1$
11: **until convergence**
12: $\theta^* \leftarrow \log(\phi_{t-1})$
13: $\mathbf{\Lambda}_{\theta^*} \leftarrow (\mathbf{K}_{\theta^*}(\mathbf{X}, \mathbf{X}) + \mathbf{\Sigma})^{-1}$
14: $\hat{T}(x) \leftarrow (\mathbf{K}_{\theta^*}^T(x)\, \mathbf{\Lambda}_{\theta^*}\, \mathbf{Y})^T \mathbf{e}$
15: $\mathbf{V}(x) \leftarrow \mathbf{K}_{\theta^*}(x,x) - \mathbf{K}_{\theta^*}(x)\, \mathbf{\Lambda}_{\theta^*}\, \mathbf{K}_{\theta^*}^T(x)$
16: $\hat{I}(x) \leftarrow \mathrm{erf}^{-1}(\gamma)\, (2\mathbf{e}^T \mathbf{V}(x) \mathbf{e})^{\frac{1}{2}}$
17: $\mathcal{C}_\gamma(x) \leftarrow [\hat{T}(x) - \hat{I}(x), \hat{T}(x) + \hat{I}(x)]$
```

5 Experiments

5.1 The Dataset

We evaluated the performance of our algorithm through the semi-simulated dataset based on the Infant Health and Development Program (IHDP) introduced in [12]. The IHDP is intended to enhance the cognitive and health status of low birth weight, premature infants through pediatric follow-ups and parent support groups. The semi-simulated dataset in [12] is based on covariates from a real randomized experiment that evaluated the impact of the IHDP on the subjects' IQ test scores at the age of three: selection bias is introduced by removing a subset of the treated population. All outcomes (response surfaces) are simulated. The response surface data generation process was not designed to favor our method: we used the standard non-linear "Response Surface B" setting in [12] (also used in [15]) to generate the response surfaces. The dataset

Method	CMGP	BART	RF	CF	k-NN	BNN-2-2
PEHE	0.43 ± 0.1	1.7 ± 0.2	1.6 ± 0.1	2.3 ± 0.1	2.9 ± 0.2	1.9 ± 0.1

Table 1: Performance comparisons with standard causal inference benchmarks.

comprises 747 subjects (608 control and 139 treated), and there are 25 covariates associated with each subject.

5.2 Benchmarks

We compared our algorithm to various state-of-the-art methods including BART [12] (the winner of the Causal Inference Data Analysis Challenge at the 2016 Atlantic Causal Inference Conference), in addition to two recently developed algorithms for estimating individualized treatment effects: Causal Forests (CF) [24, 33] and Balancing Neural Networks (BNN) with the BNN-2-2 configuration (i.e. 2 output layers and 2 representation layers) [15]. We also compare our method with a standard matching approach, k-nearest neighbor (k-NN) [26], and classical direct modeling approaches that fit separate regression models for the two potential outcomes using Random Forests (RF) [17] and Gaussian Processes (GP) [19].

5.3 Evaluation Methodology and Criteria

We performed 10 held-out experiments to select the hyper-parameters of all the algorithms under consideration, and 1000 experiments to evaluate the performance of the algorithms. In each experiment, we draw new values for the two potential outcomes of all subjects according to the “Response Surface B" model in [12]. (The same evaluation setup was used in [12] and [15].) For BART, we use the default prior as in [12]. We evaluated the performance of every algorithm by measuring its *Precision in Estimating Heterogeneous Effects* (PEHE) metric introduced in [12]. This metric reflects the accuracy of an algorithm in estimating the “heterogeneity" of a treatment's effect; it measures the accuracies of the estimates for both the factual and the counter-factual outcomes. The PEHE is computed as the root-mean-square error of the estimates for the treatment effect as follows PEHE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}((\hat{f}_1(x_i) - \hat{f}_1(x_i)) - T(x_i))^2}$. The PEHE is computable since we have simulated outcomes in our experiments, and hence we have access to all the counter-factual outcomes.

5.4 Results

The results in Table 5.4 clearly demonstrate the significant gains achieved by CMGPs in terms of the accuracy in estimating the individualized treatment effects. As expected, the k-NN algorithm displays the worst performance since it relies on a fixed, non-adaptive distance metric that fails to cope with the selection bias. The gain achieved by GPs with respect to the tree-based algorithms

(RF, BART and CF) results from the fact that GPs assign a prior distribution on a space of smooth functions (RHKS), whereas tree-based algorithms compute their estimates by averaging many non-smooth functions. That is, trees average many discontinuous functions, and these functions are all very coarse zero-order-hold approximations for the true function, so they need a large number of samples to converge to the true function since the true response function in any given practical setting is indeed smooth. In the Bayesian context, this translates in the GP's posterior contraction rate being faster than that for tree-based algorithms [25, 28, 29], and hence the estimated treatment effect function $\hat{T}(x)$ converges more quickly to the true function $T(x)$ for a given x. Unlike tree based methods, CMGPs estimate the kernel parameters first using empirical Bayes and then "adapts" its prior to the data. (The kernel parameters (length scale) determine the level of smoothness of the functions over which the GP prior is placed.)

References

[1] A. Abadie and G. W. Imbens. Matching on the Estimated Propensity Score. *Econometrica*, 84(2):781-807, 2016.

[2] M. A. Alvarez, L. Rosasco, N. D. Lawrence. Kernels for Vector-valued Functions: A Review. *Foundations and Trends ®in Machine Learning*, 4(3):195-266, 2012.

[3] S. Athey and G. Imbens. Recursive Partitioning for Heterogeneous Causal Effects. *Proceedings of the National Academy of Sciences*, 113(27):7353-7360, 2016.

[4] H. Bang and J. M. Robins. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61(4):962-973, 2005.

[5] E. V. Bonilla, K. M. Chai, and C. Williams. Multi-task Gaussian Process Prediction. In *NIPS*, 2007.

[6] L. Bottou, J. Peters, J. Candela, J. Quinonero, D. Charles, M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson,. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *JMLR*, 14(1):3207-3260, 2013.

[7] M. Caliendo and S. Kopeinig. Some Practical Guidance for the Implementation of Propensity Score Matching. *Journal of Economic Surveys*, 22(1):31-72, 2008.

[8] H. A. Chipman, E. I. George, R. E. McCulloch, et al. BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 4(1):266-298, 2010.

[9] M. Dudk, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *ICML*, 2011.

[10] M. J. Funk, D. Westreich, C. Wiesen, T. Strmer, M. A. Brookhart, and M. Davidian. Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology*, 173(7):761-767, 2011.

[11] J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Counterfactual Prediction with Deep Instrumental Variables Networks. arXiv preprint arXiv:1612.09596, 2016.

[12] J. L. Hill. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 2012.

[13] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schlkopf. Nonlinear Causal Discovery with Additive Noise Models. In *NIPS*, 2009.

[14] S. M. Iacus, G. King, and G. Porro. Causal Inference without Balance Checking: Coarsened Exact Matching. *Political Analysis*, 1-24, 2011.

[15] F. D. Johansson, U. Shalit, and D. Sontag. Learning Representations for Counter-factual Inference. In *ICML*, 2016.

[16] D. Kingma and J. Ba. ADAM: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980, 2014.

[17] M. Lu, S. Sadiq, D. J. Feaster, and H. Ishwaran. Estimating Individual Treatment Effect in Observational Data using Random Forest Methods. arXiv preprint arXiv:1701.05306, 2017.

[18] K. E. Porter, S. Gruber, M. J. Van Der Laan, and J. S. Sekhon. The Relative Performance of Targeted Maximum Likelihood Estimators. *The International Journal of Biostatistics*, 7(1):1-34, 2011.

[19] Carl Edward Rasmussen. Gaussian Processes for Machine Learning. *Citeseer*, 2006.

[20] P. R. Rosenbaum. Optimal Matching for Observational Studies. *Journal of the American Statistical Association*, 84(408): 1024-1032, 1989.

[21] P. R. Rosenbaum and D. B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41-55, 1983.

[22] D. B Rubin. Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of statistics*, 34-58, 1978.

[23] D. B Rubin. Causal Inference using Potential Outcomes. *Journal of the American Statistical Association*, 2011.

[24] U. Shalit, F. Johansson, and D. Sontag. Estimating Individual Treatment Effect: Generalization Bounds and Algorithms. arXiv preprint arXiv:1606.03976, 2016.

[25] S. Sniekers, A. van der Vaart. Adaptive Bayesian Credible Sets in Regression with a Gaussian Process Prior. *Electronic Journal of Statistics*, 9(2):2475-2527, 2015.

[26] E. A. Stuart. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25(1):1, 2010.

[27] A. Swaminathan and T. Joachims. Batch Learning from Logged Bandit Feedback Through Counter-factual Risk Minimization. *JMLR*, 16(1): 1731-1755, 2015.

[28] B. Szabo, A. van der Vaart, and H. van Zanten. Honest Bayesian Confidence Sets for the ℓ_2-norm. *Journal of Statistical Planning and Inference*, 166:36-51, 2015.

[29] B. Szabo, A. van der Vaart, J. van Zanten. Frequentist Coverage of Adaptive Nonparametric Bayesian Credible Sets. *The Annals of Statistics*, 43(4): 1391-1428, 2015.

[30] M. Taddy, M. Gardner, L. Chen, and D. Draper. A Nonparametric Bayesian Analysis of Heterogeneous Treatment Effects in Digital Experimentation. *Journal of Business and Economic Statistics*, 2016.

[31] L. Tian, A. A. Alizadeh, A. J. Gentles, and R. 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.

[32] D. Tran, F. J. Ruiz, S. Athey, and D. M. Blei. Model Criticism for Bayesian Causal Inference. arXiv preprint arXiv:1610.09037, 2016.

[33] S. Wager and S. Athey. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. arXiv preprint arXiv:1510.04342, 2015.

[34] Y. Xie, J. E. Brand, and B. Jann. Estimating Heterogeneous Treatment Effects with Observational Data. *Sociological Methodology*, 42(1):314-347, 2012.

[35] H. Wackernagel. Multivariate Geostatistics: an Introduction with Applications. *Springer Science and Business Media*, 2013.

[36] J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the Application of Probability Theory to Agricultural Experiments. *Statistical Science*, 5(4): 465-472, 1990.

[37] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, and N. de Freitas. Learning to Learn by Gradient Descent by Gradient Descent. NIPS 2016.