Weighted Gaussian Process for Estimating Treatment Effect

Junfeng Wen Negar Hassanpour Russell Greiner
Department of Computing Science
University of Alberta, Edmonton, AB, Canada
junfengwen@gmail.com, {hassanpo, rgreiner}@ualberta.ca

Abstract

Estimating treatment effect is crucial in many fields, including but not limited to medicine, psychology and economics. Accurate estimation of treatment effect is difficult in most observational studies, as those collected examples are inevitably biased: distributions of sample covariates between treatment and control groups are misaligned due to experimental conditions or constraints. To address this issue, we borrow covariate shift correction techniques from the transfer machine learning community and incorporate them into weighted Gaussian process for effective bias correction. Our method can (1) correct sample bias, (2) predict both population and individual treatment effects, and (3) provide corresponding confidence intervals.

1 Introduction

A patient's clinical treatment should be based on his/her individual conditions. In order to identify whether and to what degree a treatment is helpful, researchers often have to estimate its effect from observational data. This is because running randomized control trials are at best expensive, or in most cases, infeasible. Therefore, accurate estimation of treatment effect from *observational studies* is of crucial importance. We can estimate the treatment effect by looking at the different outcomes of treatment and control groups. However, naïve estimation methods may suffer from sample bias because, realistically, the assignment of treatment can depend on characteristics of the patient: for example, clinicians are more likely to provide treatments to sicker patients. Moreover, the treatment effect may be different for different individuals, and practitioners would prefer to know the confidence of the estimation.

The issues of sample bias and the correction techniques involved are not limited to medical domain. Another plausible application is A/B testing in the web/app industry (Crook et al., 2009). In A/B testing, two groups of users are respectively presented with two variants of a product (website/app layout), and then the test designer will examine the difference in outcome (website visits/revenue). Similar to medical domain, here the outcome difference (in analogy to treatment effect) could be inaccurate if the user groups are biased. As such, the approach described in this paper remains applicable for A/B testing.

In this work, we borrow certain covariate shift techniques (Sugiyama and Kawanabe, 2012) from the transfer learning community to handle sample bias of treatment effect estimation, and more importantly, provide confidence on the estimation using weighted Gaussian process (GP). We will first provide the problem specification and related work (Section 2), then we will describe the sample bias issue in treatment effect estimation and how to solve it with importance reweighting in Section 3, followed by our Bayesian interpretation of weighted learning (Section 4). We will show that the resultant weighted Gaussian process can (1) correct the sample bias, (2) predict population/individual treatment effect, and (3) provide respective confidence intervals. Finally, in Section 5, we will demonstrate the effectiveness of our method in both synthetic and real-world RCT datasets.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

2 Problem Specification and Related Work

Throughout the paper, we use capital letters (e.g., X, K) for matrices, bold letters (e.g., $\mathbf{x}, \mathbf{y}$) for vectors, and non-bold letters (e.g., x, y) for scalars. We will also abuse these notations to represent both random variables and their realizations. We are given a dataset $(X, \mathbf{y}_0, \mathbf{y}_1, \mathbf{t})$, where $X \in \mathbb{R}^{n\times d}$ is the covariate data matrix, $\mathbf{t} \in \{0,1\}^n$ is the vector of indicator variables whose i^{th} entry represents whether the i^{th} patient belongs to control or treatment group, and $\mathbf{y}_0, \mathbf{y}_1 \in \mathbb{R}^n$ represent the outcomes after being controlled or treated respectively. Note that for the i^{th} patient, only one of y_{i0}, y_{i1} will be observed and the other will be missing. For example, when $t_i = 1$, we observed the treatment outcome y_{i1} but not the control outcome y_{i0}. The goal is to estimate the sample average treatment effect on the treated (SATT): $\mathbb{E}_{\mathbf{x}\sim\widehat{p}(\mathbf{x}|t=1)}[y_1 - y_0]$, where the expectation is taken over the empirical distribution $\widehat{p}(\mathbf{x}|t=1)$. Because of sample bias (i.e., $p(\mathbf{x}|t=1) \neq p(\mathbf{x}|t=0)$), a major difficulty lies in effectively estimating $\mathbb{E}_{\mathbf{x}\sim\widehat{p}(\mathbf{x}|t=1)}[y_0]$. Of course, it would be advantageous to have models that can predict treatment effect on unseen new patients.

Many existing methods focus on estimating the *propensity score* $e(\mathbf{x}) \stackrel{\text{def}}{=} p(t=1|\mathbf{x})$ (Rosenbaum and Rubin, 1983; Rubin, 2006) by employing, for instance, logistic regression, and then correcting sample bias with it. However, estimating $e(\mathbf{x})$ can be as difficult as finding $y_1(\mathbf{x}), y_0(\mathbf{x})$ for patient $\mathbf{x}$. Instead, our proposed method models $y_1(\mathbf{x}), y_0(\mathbf{x})$ directly with (weighted) Gaussian process, which is also beneficial for extrapolation to unseen new patients. Matching and inverse probability weighting (IPW) are two types of methods among others (Austin, 2011) that are most relevant to our method.

Matching methods (Stuart, 2010) couple each treated individual with another "similar" individual in the control group (or a combination of several controlled individuals). The similarity is usually measured by some metrics, such as the distance in the covariate space $\mathcal{X}$ or their propensity score difference. After matching, hopefully the matched sample would have reduced bias so that the treatment effect can be estimated as usual. A major drawback of matching is that the sample is not utilized effectively: some data in the control group could be left out and completely ignored. Our method does intend to match the treatment and control groups, but we achieve it by aligning the whole groups instead, which is closely related to IPW.

Inverse probability weighting (Lunceford and Davidian, 2004) assigns all individuals a weight: $w(\mathbf{x}) = 1$ for treated individual and $w(\mathbf{x}) = \widehat{e}(\mathbf{x})/(1-\widehat{e}(\mathbf{x}))$ for controlled individual, where the estimated propensity score $\widehat{e}$ is computed from sample. SATT is then estimated from this weighted sample. The performance of IPW is highly dependent on the accuracy of propensity score estimation, which might be a complicated task especially in high dimensional $\mathcal{X}$ space. To overcome this obstacle, we aim to estimate the ratio $\widehat{e}(\mathbf{x})/(1-\widehat{e}(\mathbf{x}))$ directly, using techniques from the transfer learning community (Sugiyama et al., 2007; Sugiyama, Suzuki, and Kanamori, 2012). Combining these well-developed techniques with weighted Gaussian process, we are able to effectively estimate SATT together with confidence intervals.

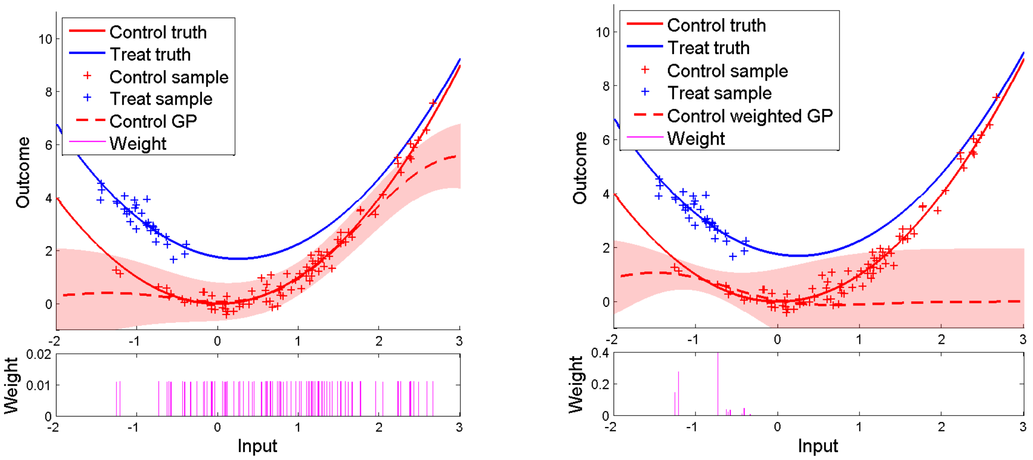
3 Treatment Effect Estimation

In this section, we will discuss the sample bias issue in treatment effect estimation, and we will address it using weighted Gaussian process in the next section. To begin, we can decompose the joint distribution $p(\mathbf{x}, y, t)$ as

$$p(\mathbf{x}, y, t) = p(t) \cdot p(\mathbf{x}|t) \cdot p(y|\mathbf{x}, t).$$

Even when there is no prior preference (i.e., $p(t=0) = p(t=1)$), $p(\mathbf{x}|t=0)$ and $p(\mathbf{x}|t=1)$ are generally different, so are $p(y|\mathbf{x}, t=0)$ and $p(y|\mathbf{x}, t=1)$. Since the difference $y_{i1} - y_{i0}$ is never observed, we could instead model regressors $f_0(\mathbf{x}_i), f_1(\mathbf{x}_i) : \mathcal{X} \mapsto \mathbb{R}$ for $p(y|\mathbf{x}, t=0)$ and $p(y|\mathbf{x}, t=1)$ respectively, and then use their difference $f_1 - f_0$ to estimate the treatment effect. Such modelling is appealing as it allows us to predict outcome and possibly its associated confidence interval for any individual or group.

Gaussian process (Rasmussen and Williams, 2005) is a reasonable choice for the task. In order to estimate SATT, the key problem is measuring the hypothetical control outcome of a treated patient $(y_0|\mathbf{x}, t=1)$. Applying direct Gaussian process modelling on $p(y|\mathbf{x}, t=0)$ could be problematic due to sample bias (i.e., $p(\mathbf{x}|t=0) \neq p(\mathbf{x}|t=1)$). As shown in Figure 1a, modelling with equal weights on control group may result in inaccurate estimation of $y_0|\mathbf{x}, t=1$ due to sample bias.



(a) Gaussian process

(b) Weighted Gaussian process

Figure 1: The blue/red curves are the true models for treatment/control outcomes respectively. The blue/red crosses are examples from treatment/control groups respectively. Note that the treatment group concentrates on $[-1.5, 0]$ while the control group spreads all over the x-axis. The dashed red curves have different meanings in the figures: Figure 1a shows the mean of GP learned from control group with equal weights $w_i = 1$, while Figure 1b shows the mean of weighted GP learned from control group with adjusted weights (patients similar to treatment group have larger weights). The weights are shown at the bottom graphs as purple bars for each control patient (note that their y-scales are different). It is clear that such adjustment on weights is beneficial since the prediction of $y_0|\mathbf{x}, t = 1$ (hypothetical control outcome of treated patient in $[-1.5, 0]$) is more accurate. Moreover, the confidence interval (red shades) shrinks significantly on the treatment region, meaning that we are more confidence about the prediction.

Such sample bias is very plausible in practice because, for instance, doctors are more inclined to offer treatment to sicker patients (or based on some other internal criteria of theirs). With proper importance weights w_i on the control data points, we could achieve better estimation on $y_0|\mathbf{x}, t = 1$ (Figure 1b). Therefore, our method can be summarized as follows:

1. Use GP to model the treatment outcome $p(y|\mathbf{x}, t = 1)$ from treatment group with equal weights.
2. Use weighted GP to model the control outcome $p(y|\mathbf{x}, t = 0)$ from weighted control group.
3. Apply both GPs to the treatment group and calculate the model differences as SATT. Both individual and sample treatment effects can be estimated. It is also convenient to provide relevant confidence intervals from Eq.(2) and Eq.(3) in the next section.

The only problem left is how to compute importance weights for control group patients. In the statistics community, the weight is usually computed by $\widehat{w}(\mathbf{x}) = \widehat{e}(\mathbf{x})/(1 - \widehat{e}(\mathbf{x}))$, where $\widehat{e}(\mathbf{x})$ is estimated propensity score. A closer look at the propensity score reveals its connection to importance weight in the transfer learning community. In transfer learning, or specifically in the covariate shift scenario (Shimodaira, 2000; Sugiyama and Kawanabe, 2012), there are two marginal distributions: the source and target distributions of inputs $p_S(\mathbf{x})$, $p_T(\mathbf{x})$. If we look $e(\mathbf{x})$ as the target marginal $p_T(\mathbf{x})$ and $1 - e(\mathbf{x})$ as the source marginal $p_S(\mathbf{x})$, then finding $w(\mathbf{x}) = e(\mathbf{x})/(1 - e(\mathbf{x})) = p_T(\mathbf{x})/p_S(\mathbf{x})$ is just usual weight function estimation, which has been extensively studied in transfer learning community (Huang et al., 2007; Sugiyama, Suzuki, and Kanamori, 2012; Wen, Yu, and Greiner, 2014; Cortes, Mohri, and Muñoz Medina, 2015). In fact, in our context, $p_T(\mathbf{x}) = p(\mathbf{x}|t = 1)$ and $p_S(\mathbf{x}) = p(\mathbf{x}|t = 0)$. With equal prior $p(t = 1) = p(t = 0)$, we have

$$w(\mathbf{x}) = \frac{e(\mathbf{x})}{1 - e(\mathbf{x})} = \frac{p(t = 1|\mathbf{x})}{p(t = 0|\mathbf{x})} = \frac{p(\mathbf{x}|t = 1)}{p(\mathbf{x}|t = 0)} = \frac{p_T(\mathbf{x})}{p_S(\mathbf{x})}.$$

In the following, we will use Kullback-Leibler Importance Estimation Procedure (KLIEP) (Sugiyama et al., 2007) to compute this weight, which has been well recognized for its good performance in the transfer learning community (Kanamori, Hido, and Sugiyama, 2009). KLIEP minimizes $\mathrm{KL}[p_T(\mathbf{x}) \| \widehat{p}_T(\mathbf{x})]$, the KL divergence between true target marginal $p_T(\mathbf{x})$ and estimated target marginal $\widehat{p}_T(\mathbf{x}) = \widehat{w}(\mathbf{x}) p_S(\mathbf{x})$. Since neither $p_T(\mathbf{x})$ nor $p_S(\mathbf{x})$ is known in practice, empirical substitutions are used and after basic manipulations, the objective of KLIEP becomes:

$$\max_{\widehat{w}} \quad \sum_{j=1}^{n_T} \log\left(\widehat{w}\left(\mathbf{x}_j^{(T)}\right)\right) \quad \text{s.t.} \quad \frac{1}{n_S} \sum_{i=1}^{n_S} \widehat{w}\left(\mathbf{x}_i^{(S)}\right) = 1 \quad \text{and} \quad \widehat{w}(\mathbf{x}) \geqslant 0, \forall \mathbf{x} \in \mathcal{X}$$

where n_T, n_S are number of treated (target) patients and number of controlled (source) patients respectively. The non-negativity constraint is not hard to understand since $\widehat{w}(\mathbf{x})$ is modelling density ratio $p_T(\mathbf{x})/p_S(\mathbf{x})$. The normalization constraint comes from the fact that the estimated target marginal $\widehat{p}_T(\mathbf{x})$ should be a proper density function:

$$1 = \int \widehat{p}_T(\mathbf{x}) d\mathbf{x} = \int \widehat{w}(\mathbf{x}) p_S(\mathbf{x}) d\mathbf{x} \approx \frac{1}{n_S} \sum_{i=1}^{n_S} \widehat{w}\left(\mathbf{x}_i^{(S)}\right).$$

Eventually, the weight function is parametrized as a mixture of Gaussians: $\widehat{w}(\mathbf{x}) = \sum_l \alpha_l \varphi_l(\mathbf{x})$, where $\alpha_l \geqslant 0$ are the mixing coefficients and $\varphi_l(\mathbf{x})$ are basis Gaussians. KLIEP is then optimized over α_l. It is a convex problem and thus can be solved efficiently.

4 Weighted Gaussian Process

This section describes how to incorporate importance weights into Gaussian process (GP) and derive relevant posterior distributions. We will briefly review GP before going to its weighted version. There are two equivalent perspectives of GP regression: parameter-space and function-space view.

4.1 Parameter-space View

First of all, let us investigate the standard linear generative model:

$$f(\mathbf{x}_i) = \boldsymbol{\theta}^\top \mathbf{x}_i, \quad y_i = f(\mathbf{x}_i) + \epsilon_i,$$

where $\boldsymbol{\theta}$ is model parameters with Gaussian prior[1] $\mathcal{N}(\mathbf{0}, I)$ and ϵ_i are i.i.d. additive noise drawn from $\mathcal{N}(0, \sigma^2)$. Given a dataset $(X, \mathbf{y})$, it is easy to see that the posterior distribution of $\boldsymbol{\theta}$ satisfies

$$\begin{aligned} p(\boldsymbol{\theta}|X, \mathbf{y}) \propto p(\mathbf{y}|X, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}) &\propto \exp\left(-\frac{1}{2\sigma^2}(X\boldsymbol{\theta} - \mathbf{y})^\top (X\boldsymbol{\theta} - \mathbf{y})\right) \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\theta}\right) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^\top \left(\sigma^{-2} X^\top X + I\right)(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})\right), \end{aligned} \tag{1}$$

where $\bar{\boldsymbol{\theta}} = (X^\top X + \sigma^2 I)^{-1} X^\top \mathbf{y}$. That is, the posterior is also Gaussian with mean $\bar{\boldsymbol{\theta}}$ and covariance $\Sigma = (\sigma^{-2} X^\top X + I)^{-1}$. Now consider the distribution of the predictive value $f_\star \overset{\text{def}}{=} f(\mathbf{x}_\star)$ at a new point $\mathbf{x}_\star$. By marginalizing $\boldsymbol{\theta}$, we have

$$p(f_\star|\mathbf{x}_\star, X, \mathbf{y}) = \int p(f_\star|\mathbf{x}_\star, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}|X, \mathbf{y})\, d\boldsymbol{\theta} = \mathcal{N}(\mathbf{x}_\star^\top \bar{\boldsymbol{\theta}},\ \mathbf{x}_\star^\top \Sigma \mathbf{x}_\star).$$

Because

$$\begin{aligned} (X^\top X + \sigma^2 I)^{-1} X^\top &= (X^\top X + \sigma^2 I)^{-1} X^\top (XX^\top + \sigma^2 I)(XX^\top + \sigma^2 I)^{-1} \\ &= (X^\top X + \sigma^2 I)^{-1} (X^\top XX^\top + \sigma^2 X^\top)(XX^\top + \sigma^2 I)^{-1} \\ &= (X^\top X + \sigma^2 I)^{-1} (X^\top X + \sigma^2 I) X^\top (XX^\top + \sigma^2 I)^{-1} \\ &= X^\top (XX^\top + \sigma^2 I)^{-1}, \end{aligned}$$

we can rewrite the mean as

$$\mathbf{x}_\star^\top \bar{\boldsymbol{\theta}} = \mathbf{x}_\star^\top (X^\top X + \sigma^2 I)^{-1} X^\top \mathbf{y} = \mathbf{x}_\star^\top X^\top (XX^\top + \sigma^2 I)^{-1} \mathbf{y}.$$

[1] Note that the prior can be specified in other forms. We use zero mean and identity covariance here for notation simplicity.

We can use the matrix inversion lemma (Boyd and Vandenberghe, 2004) to rewrite the covariance Σ:

$$\Sigma = (\sigma^{-2}X^\top X + I)^{-1} = I - X^\top (XX^\top + \sigma^2 I)^{-1} X.$$

These variational representations allow us to use the "kernel trick" and replace all inner products of $\mathbf{x}$ by a kernel $\kappa(\cdot,\cdot)$, so that

$$f_\star | \mathbf{x}_\star, X, \mathbf{y} \ \sim \ \mathcal{N}\left(\mu_\star, \sigma_\star^2\right) \quad \text{with} \quad \begin{aligned} \mu_\star &\overset{\text{def}}{=} \mathbf{k}_\star^\top (K + \sigma^2 I)^{-1}\mathbf{y} \\ \sigma_\star^2 &\overset{\text{def}}{=} \kappa(\mathbf{x}_\star, \mathbf{x}_\star) - \mathbf{k}_\star^\top (K + \sigma^2 I)^{-1}\mathbf{k}_\star, \end{aligned} \tag{2}$$

where $\mathbf{k}_\star = [\kappa(\mathbf{x}_\star, \mathbf{x}_1), \cdots, \kappa(\mathbf{x}_\star, \mathbf{x}_n)]^\top$ and $K_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ with $i, j = 1, \cdots, n$. Therefore, we can both predict $f_\star$ by the mean and provide its associated confidence interval because we have its distribution.

4.2 Function-space View

Gaussian process is defined as a (possibly infinite) collection of random variables, any finite subset of which have a joint Gaussian distribution. It is completely specified by its mean function and covariance function. For instance, a real process $f(\mathbf{x})$ indexed by $\mathbf{x}$ can be written as $f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$.

Suppose we use zero mean function $\mu(\mathbf{x}) = 0$ and assume $y = f(\mathbf{x}) + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Given a dataset $\{X, \mathbf{y}\}$ and a new point $\mathbf{x}_\star$, the joint distribution of the outcomes is

$$\begin{bmatrix} \mathbf{y} \\ f_\star \end{bmatrix} \ \sim \ \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K + \sigma^2 I & \mathbf{k}_\star \\ \mathbf{k}_\star^\top & \kappa(\mathbf{x}_\star, \mathbf{x}_\star) \end{bmatrix}\right).$$

Then we can find the marginal distribution of $f_\star$:

$$f_\star | \mathbf{x}_\star, X, \mathbf{y} \ \sim \ \mathcal{N}\left(\mu_\star, \sigma_\star^2\right),$$

which is exactly the same as Eq.(2). The chosen kernel is now interpreted as covariance function.

4.3 Reweighting Data Instances

To understand the Bayesian interpretation of weighted GP, consider maximizing the posterior (Eq.(1)), which is equivalent to minimizing its negative in log scale:

$$\min_{\boldsymbol{\theta}} \ -\log p(\boldsymbol{\theta}|X, \mathbf{y}) \Longleftrightarrow \min_{\boldsymbol{\theta}} \ \frac{1}{2}\|X\boldsymbol{\theta} - \mathbf{y}\|_2^2 + \frac{\sigma^2}{2}\|\boldsymbol{\theta}\|_2^2 \Longleftrightarrow \min_{\boldsymbol{\theta}} \ \frac{1}{2}\sum_{i=1}^n (\mathbf{x}_i^\top \boldsymbol{\theta} - y_i)^2 + \frac{\sigma^2}{2}\|\boldsymbol{\theta}\|_2^2.$$

It is clear that this is ridge regression (Hastie, Tibshirani, and Friedman, 2009), treating every data point with equal importance (here, with weight one). In order to correct sample bias (or specifically covariate shift), one may impose different importance weights w_i and optimize the following instead:

$$\min_{\boldsymbol{\theta}} \quad \frac{1}{2}\sum_{i=1}^n w_i \cdot (\mathbf{x}_i^\top \boldsymbol{\theta} - y_i)^2 + \frac{\sigma^2}{2}\|\boldsymbol{\theta}\|_2^2 \iff \min_{\boldsymbol{\theta}} \quad \frac{1}{2}(X\boldsymbol{\theta} - \mathbf{y})^\top W (X\boldsymbol{\theta} - \mathbf{y}) + \frac{\sigma^2}{2}\|\boldsymbol{\theta}\|_2^2$$

with $w_i \geqslant 0, \sum_i w_i = n$ and W being a diagonal matrix with w_i on the diagonal. Equivalently, we are now maximizing $p(\boldsymbol{\theta}) \prod_i p^{w_i}(y_i|\mathbf{x}_i, \boldsymbol{\theta})$. Beyond this point, it is straightforward to see that by replacing X with $\widetilde{X} = W^{\frac{1}{2}}X$ and $\mathbf{y}$ with $\tilde{\mathbf{y}} = W^{\frac{1}{2}}\mathbf{y}$, we can obtain the distribution of $f_\star$:

$$f_\star | \mathbf{x}_\star, \widetilde{X}, \tilde{\mathbf{y}} \sim \mathcal{N}\left(\tilde{\mu}_\star, \tilde{\sigma}_\star^2\right) \quad \text{with} \quad \begin{aligned} \tilde{\mu}_\star &\overset{\text{def}}{=} \tilde{\mathbf{k}}_\star^\top (\widetilde{K} + \sigma^2 I)^{-1}\tilde{\mathbf{y}} \\ \tilde{\sigma}_\star^2 &\overset{\text{def}}{=} \kappa(\mathbf{x}_\star, \mathbf{x}_\star) - \tilde{\mathbf{k}}_\star^\top (\widetilde{K} + \sigma^2 I)^{-1}\tilde{\mathbf{k}}_\star, \end{aligned} \tag{3}$$

where $\tilde{\mathbf{k}}_\star = W^{\frac{1}{2}}\mathbf{k}_\star$ and $\widetilde{K} = W^{\frac{1}{2}}KW^{\frac{1}{2}}$. The importance of understanding the distribution of $f_\star$ in weighted GP is that we can now provide its confidence interval when instances are weighted. The adjustments with W in Eq.(3) are similar to that of normalized kernel (Weiss, 1999; Xu, White, and Schuurmans, 2009). The difference is that, instead of computed from the kernel K, the weights here are pre-computed for sample bias correction.

5 Experiments

In this section, we evaluate the proposed method, denoted as weighted Gaussian process (WGP), based on how accurately it can calculate sample average treatment effect on the treated (SATT) on both synthetic datasets and a real-world study. In order to acquire true treatment effect, we investigate randomised control trial (RCT) data, in which the true effect can be calculated as $\bar{y}_1 - \bar{y}_0$, the difference in means of treated and controlled outcomes.

5.1 Synthetic Data

We generated a synthetic dataset to compare the effectiveness of the proposed method against the other methods in the literature.

Data Generation. Two quadratic functions were constructed to represent the true underlying outcome for control group and another for treatment group:

$$f_0(x) = x^2, \quad f_1(x) = x^2 + |x - 3.5|/2$$

We generated X_0 and X_1 (each representing samples from control and treatment group respectively) with 250 random one-dimensional examples from $\mathcal{N}(0, 1)$. The outcome vectors $\mathbf{y}_0$ and $\mathbf{y}_1$ were generated from $f_0(X_0)$ and $f_1(X_1)$ plus a small Gaussian noise from $\mathcal{N}(0, 0.3^2)$. This is basically how we generated data in Figure 1 earlier. Figure 2 (top) shows the Gaussian distributions fitted by the control and treatment samples. These distributions are almost identical, suggesting that this synthetic data is indeed RCT.

To mimic observational data, we skewed the dataset such that examples with larger x-values have a higher chance to be assigned to the control group. Likewise, examples with smaller x-values have a higher chance to be assigned to the treatment group. In other words, the skewed dataset is a subset of the original RCT data, but it is clearly biased in its treatment assignment procedure. The true SATT can be computed from each skewed treated sample based on the shift in the second term of f_1. This sampling procedure is repeated 50 times to imitate 50 different observational studies. Figure 2 (middle) shows the 50 skewed datasets fitted by Gaussian distributions.

Methods and Results. We are going to compare our method with Baseline, Bayesian additive regression trees (BART) (Chipman, George, and McCulloch, 2010) and targeted maximum likelihood learning (TMLE) (van der Laan and Rubin, 2006). The Baseline method is simply modelling the skewed dataset with Gaussian process on the control group, and then estimate SATT using the difference in true treated outcome and estimated controlled outcome of the treatment group patients. On such observational dataset, training a naïve model like this without adjustment for discrepancy between distributions would usually result in an inaccurate estimation of SATT. BART is an ensemble method that can be viewed as a representative of matching approaches because of its tree nature. TMLE is an efficient doubly robust approach for estimating any target value of data distribution (in our case, the SATT). For our method, we used Gaussian kernel as covariance function with kernel width chosen by the median pair-wise distance in the dataset, and σ^2 was selected from a wide range of candidates by cross-validation.

The results are shown in the Table 1. Entries marked as Truth reflect the true SATT calculated based on $f_1 - f_0$ on the treated set. The Baseline method performs poorly on estimating SATT since it does not consider sample bias correction. Both BART and WGP have effectively corrected sample bias and produced reasonable SATT estimates. Their standard deviations are also very close. The correction of WGP can be directly seen in Figure 2 (bottom), where the control group is obviously shifted towards treatment group. TMLE did correct the bias to some extend compared to Baseline, but not as significant as BART or WGP. Moreover, its standard deviation is higher than the others.

5.2 CO-MED Dataset

Dataset Overview. The real-world RCT dataset that we used is *Combining Medications to Enhance Depression Outcomes* (CO-MED) (Rush et al., 2011). This study was designed and conducted by the U.S. National Institute of Mental Health (NIMH), to compare the effectiveness of administering a combination of antidepressant medications in the first treatment step instead of one antidepressant medication alone, for people with chronic or recurrent major depressive disorder. The control group had received *Escitalopram + Placebo* , while the treatment group had received *Venlafaxine XR*

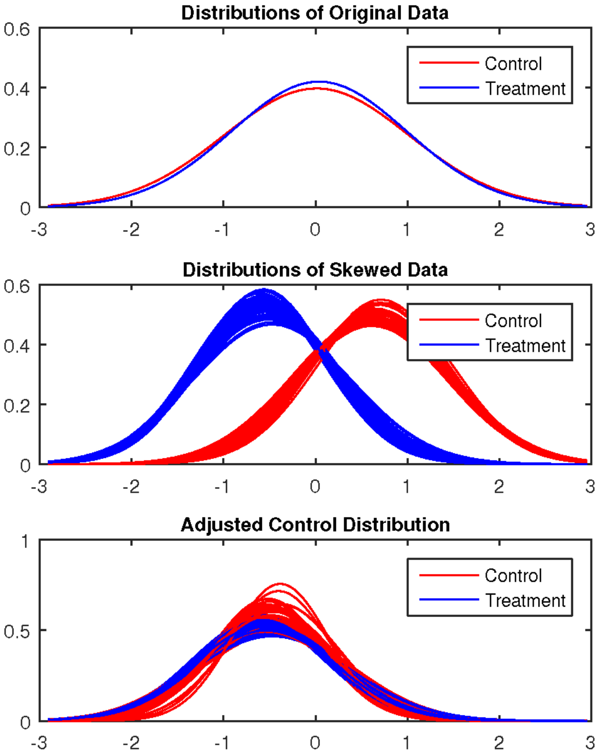


Figure 2: Gaussian distributions fitted to control and treatment groups for the synthetic RCT dataset (top); 50 observational (i.e., skewed) datasets (middle); and adjusted control distributions by KLIEP that attempts to shift the control distribution to match the treatment distribution (bottom).

Method	Synthetic	CO-MED
Truth	2.02 (± 0.02)	-1.43
Baseline	2.18 (± 0.10)	-0.69 (± 0.18)
BART	1.87 (± 0.09)	-1.19 (± 0.32)
TMLE	2.21 (± 0.34)	-1.28 (± 0.65)
WGP	2.13 (± 0.10)	-1.48 (± 0.32)

Table 1: SATT estimation results. Mean and standard deviation (in parenthesis) over 50 runs.

#	Description
1	Gender
2	Feel more self confident than usual
3	Feel heart racing
4	Been having lots of great ideas
5	Plan to commit suicide
6	Ability to focus / sustain attention
7	Somatic anxiety
8	Work and interests
9	Retardation
10	Mood (anxious)
11	Mood (irritable)
12	Sympathetic arousal
13	Worry daily
14	Probability of fall asleep b/c worry
15	Tension in muscles b/c anxiety
16	Trouble concentrating b/c worry
17	Snappy/irritable b/c worry
18	Sleep onset insomnia
19	Suicidal ideation
20	Money satisfaction
21	Friends satisfaction
22	Back pain
23	Severity of nasal congestion
24	Severity of muscular cramps
25	Menstrual irregularities
26	Work hours missed due to reasons other than health problems

Table 2: Most predictive features by LARS

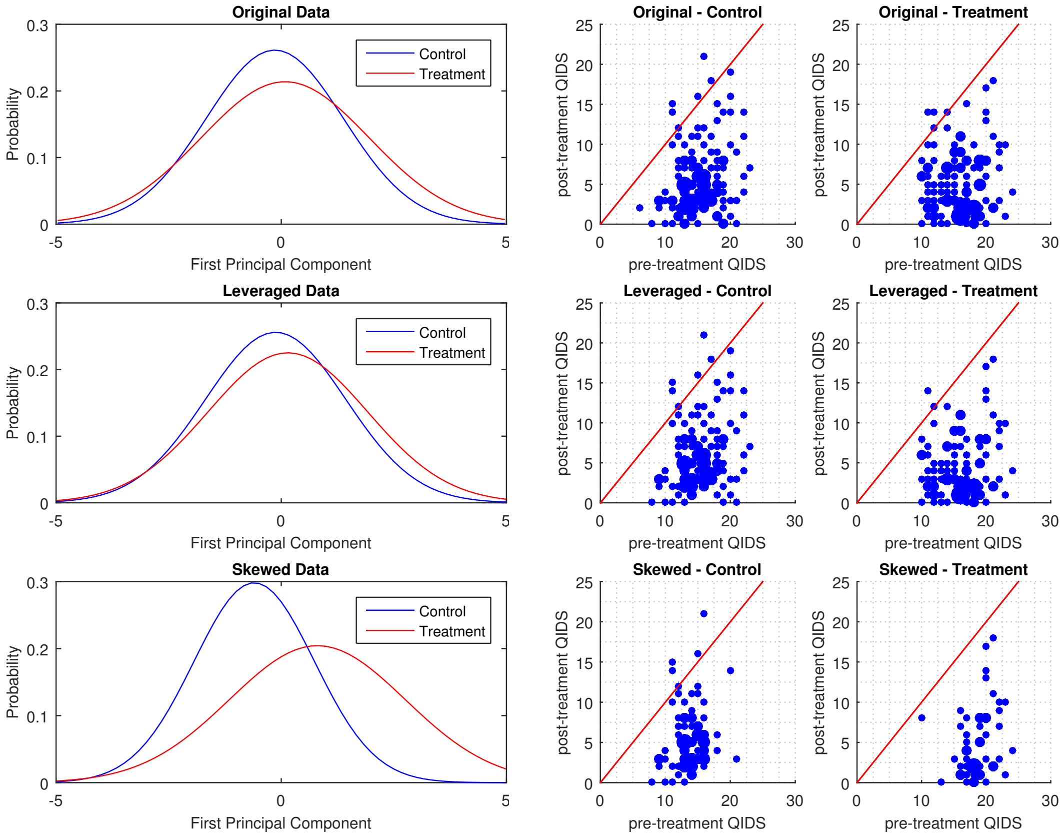
+ *Mirtazapine*. The eligible participants were randomly allocated to one of the groups. Several assessment forms were completed throughout the study by both the patients and their respective psychiatrists including demographic information, as well as psychiatric and medical measurements.

The primary outcome measure of the study is QIDS-SR16 score[2] at the end-point (week 28, post-treatment). The study had recruited 444 patients, 261 of whom completed the 28 weeks duration of the study (136 in the control group and 125 in the treatment group). 362 features were compiled from the assessment forms. We then used least-angle regression (LARS) (Efron et al., 2004) and found 26 key features (cross-validated) that were most predictive for estimating outcome. Table 2 lists the description of these key features.

Preprocessing. Existing analysis of the CO-MED study showed no significant difference in outcome between prescribing a combination of medications over mono-therapy with Escitalopram in patients with chronic and/or recurrent major depressive disorder (Rush et al., 2011). This is further confirmed by calculating the true effect from the RCT dataset using $\bar{y}_1 - \bar{y}_0$. The effect is -0.457, which is a small remission for a score that ranges from 0 to 27.

Such small effect makes subsequent analysis overwhelmingly difficult for all methods due to the presence of noise in the dataset. Therefore, we leverage (increase) the effect by removing some controlled patients with huge remission as well as some treated patients with insignificant remission at the end-point of the study. It should be noted that, this leverage process maintained the RCT nature of the data. This is confirmed in Figure 3a, which shows the distributions fitted to control and treatment

[2] Quick Inventory of Depressive Symptomatology - Self-Report. It ranges from 0 (none) to 27 (severe).



(a) (b)

Figure 3: CO-MED dataset. (a) Distributions of control versus treatment groups over the first principal component direction. The control and treatment groups are well-aligned before (top) and after (middle) leveraging the treatment effect, which suggests that the leveraged data remains RCT. After skewing the dataset to mimic observational study, the distributions are no longer matched (bottom), indicating that the skewed data is not RCT anymore. (b) Patients' QIDS-SR16 scores measured at end-point versus start-point. Note that the size of the circles represent the number of patients with same [start-point, end-point] QIDS-SR16 score tuple. The 45° line indicates patients with no effect after the study. We can see that the leveraged dataset closely resembles the original dataset, while in the skewed dataset, patients with severe depressive disorder are more likely to receive treatment.

groups are still closely aligned after leverage process. The treatment effect on this leveraged dataset is now -1.43, which would be considered as the true effect for estimation algorithms.

Finally, we generate observational data by assigning higher probability of being treated for sicker patients (based on pre-treatment QIDS-SR16 score) in the treatment group, and higher probability of being controlled for healthier patients in the control group. This sampling procedure is repeated 50 times to generate 50 observational studies for further analysis. Figure 3b is a visualization that summarizes the effect of treatment and control on each participant in this study (original data, leveraged data and one particular skewed data).

Results. We again compare WGP with BART and TMLE. For this dataset, we use linear kernel since it achieves better performance than Gaussian kernel, in terms of cross-validated prediction error. The experiment results are shown in Table 1. The Truth value in CO-MED is a single number because it is directly computed based on the whole RCT dataset with fixed treated and controlled patients. Again, we can see that without proper sample bias correction, the Baseline method performs very poorly on estimating SATT. Our method, although slightly over-claims the remission effect, produces the closest value to the truth compared to BART and TMLE. Moreover, WGP maintains a reasonable deviation on the prediction compared to the competitors.

References

Austin, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research* 46(3):399–424.

Boyd, S., and Vandenberghe, L. 2004. *Convex optimization*. Cambridge university press.

Chipman, H. A.; George, E. I.; and McCulloch, R. E. 2010. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics* 266–298.

Cortes, C.; Mohri, M.; and Muñoz Medina, A. 2015. Adaptation algorithm and theory based on generalized discrepancy. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, 169–178. New York, NY, USA: ACM.

Crook, T.; Frasca, B.; Kohavi, R.; and Longbotham, R. 2009. Seven pitfalls to avoid when running controlled experiments on the web. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1105–1114. ACM.

Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R.; et al. 2004. Least angle regression. *The Annals of statistics* 32(2):407–499.

Hastie, T.; Tibshirani, R.; and Friedman, J. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York.

Huang, J.; Smola, A. J.; Gretton, A.; Borgwardt, K. M.; and Schölkopf, B. 2007. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, volume 19, 601.

Kanamori, T.; Hido, S.; and Sugiyama, M. 2009. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research* 10(Jul):1391–1445.

Lunceford, J. K., and Davidian, M. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine* 23(19):2937–2960.

Rasmussen, C. E., and Williams, C. K. I. 2005. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Rosenbaum, P. R., and Rubin, D. B. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Rubin, D. B. 2006. *Matched sampling for causal effects*. Cambridge University Press.

Rush, A. J.; Trivedi, M. H.; Stewart, J. W.; Nierenberg, A. A.; Fava, M.; Kurian, B. T.; Warden, D.; Morris, D. W.; Luther, J. F.; Husain, M. M.; et al. 2011. Combining medications to enhance depression outcomes (co-med): acute and long-term outcomes of a single-blind randomized study. *American Journal of Psychiatry*.

Shimodaira, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* 90(2):227–244.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25(1):1.

Sugiyama, M., and Kawanabe, M. 2012. *Machine learning in non-stationary environments: introduction to covariate shift adaptation*. The MIT Press.

Sugiyama, M.; Nakajima, S.; Kashima, H.; Von Buenau, P.; and Kawanabe, M. 2007. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*.

Sugiyama, M.; Suzuki, T.; and Kanamori, T. 2012. *Density ratio estimation in machine learning*. Cambridge University Press.

van der Laan, M. J., and Rubin, D. 2006. Targeted maximum likelihood learning. *The International Journal of Biostatistics* 2(1).

Weiss, Y. 1999. Segmentation using eigenvectors: a unifying view. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, 975–982. IEEE.

Wen, J.; Yu, C.-N.; and Greiner, R. 2014. Robust learning under uncertain test distributions: Relating covariate shift to model misspecification. In *International Conference on Machine Learning*, 631–639.

Xu, L.; White, M.; and Schuurmans, D. 2009. Optimal reverse prediction: a unified perspective on supervised, unsupervised and semi-supervised learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 1137–1144. ACM.