Title

Optimally Balanced Gaussian Process Propensity Scores for Estimating Treatment Effects

Permalink

https://escholarship.org/uc/item/9056c7j0

Journal

Journal of the Royal Statistical Society Series A (Statistics in Society), 183(1)

ISSN

0964-1998

Authors

Vegetabile, Brian G
Gillen, Daniel L
Stern, Hal S

Publication Date

2020

DOI

10.1111/rssa.12502

Peer reviewed

Published in final edited form as:
J R Stat Soc Ser A Stat Soc. 2020 January ; 183(1): 355–377. doi:10.1111/rssa.12502.

Optimally Balanced Gaussian Process Propensity Scores for Estimating Treatment Effects

Brian G. Vegetabile[1], **Daniel L. Gillen**[2], **Hal S. Stern**[2]

[1]RAND Corporation, Santa Monica, CA, 90401, USA

[2]Department of Statistics, Donald Bren School of Information & Computer Sciences, University of California, Irvine, CA, 92697-3425, USA

Abstract

Propensity scores are commonly employed in observational study settings where the goal is to estimate average treatment effects. This paper introduces a flexible propensity score modeling approach, where the probability of treatment is modeled through a Gaussian process framework. To evaluate the effectiveness of the estimated propensity score, a metric of covariate imbalance is developed that quantifies the discrepancy between the distributions of covariates in the treated and control groups. It is demonstrated that this metric is ultimately a function of the hyperparameters of the covariance matrix of the Gaussian process and therefore it is possible to select the hyperparameters to optimize the metric and minimize overall covariate imbalance. The effectiveness of the GP method is compared in a simulation against other methods of estimating the propensity score and the method is applied to data from Dehejia and Wahba (1999) to demonstrate benchmark performance within a relevant policy application.

Keywords

Causal Inference; Covariate Balance; Gaussian Process; Nonparametric Estimation

1 | INTRODUCTION

The propensity score was introduced in Rosenbaum and Rubin (1983) in the context of study designs with binary treatment regimes and has become a common tool for conducting causal inference in observational studies. The propensity score is loosely defined as the probability of an individual or unit in a study being in the treated group given a set of pretreatment covariates. Rosenbaum and Rubin (1983) demonstrated that under certain assumptions, adjustment on the propensity score enables unbiased estimation of the average treatment effect or the average treatment effect among the treated.

There is a large literature on estimating the propensity score for binary treatment regimes. A common approach is to utilize a statistical model building procedure (e.g. logistic regression) that is iterated until a functional form of the propensity score is arrived at that

Correspondence: Brian Vegetabile, RAND Corporation, 1776 Main St, Santa Monica, CA 90401, USA, bvegetab@rand.org.

sufficiently “balances” covariates (Imbens and Rubin, 2015; Imai and Ratkovic, 2014). Procedures of this type are computationally fast to perform since models are often built in a step-wise fashion. They typically achieve the goal of approximately balancing covariates but, because of their restricted parametric form, may not accurately estimate treatment assignment. Alternative approaches rely on nonparametrically estimating the propensity score through modern statistical learning methods without a need to specify a functional form (Woo et al., 2008; Lee et al., 2010; McCaffrey et al., 2004). These methods are attractive because they can provide accurate estimation of the treatment assignment mechanism while requiring fewer modeling decisions to be made. However, they are often computationally demanding due to their flexibility. This paper focuses on a new procedure for nonparametric estimation of the propensity score using Gaussian processes. In contrast to other nonparametric estimation procedures which focus on hyperparameter selection to provide unbiased estimation of the propensity score, the hyperparameters of the model are selected to optimally balance the marginal distributions of pretreatment covariates in the treated and control groups.

In a study comparing two groups, say a treated and a control group, estimation of the propensity score can be addressed as a binary regression problem. A Bayesian latent variable approach to binary regression is to assume that the probability of success (in this example the probability of being treated) given a set of covariates is related to a random (e.g. Gaussian) latent “score” for each observation. An extension of this approach is to assume that these latent scores arise from a Gaussian process. A Gaussian process (GP) is a collection of random variables, for which any finite subset of the collection has a joint Gaussian distribution. The GP is described by specifying a mean function and a covariance function, where the mean and covariance may depend on the set of covariates and unknown parameters. Rasmussen and Williams (2006) provide an overview of GP modeling in the context of classification problems.

In the context of propensity score estimation, the covariance function of the GP plays a key role in modeling the latent scores which determine the probability of treatment. Generally, covariance functions are constructed by using a kernel function that relates the covariance of the latent scores for two individuals to the distance between the covariates for those individuals. The idea being that if two individuals’ observed covariates are similar, then they should have a similar probability of being assigned to the treatment (control) group. In this way, estimation of the propensity score using a GP is analogous to methods in observational studies that make use of the distance between sets of covariates to create matched pairs (Stuart, 2010; Rosenbaum, 2010). The covariance function of the GP can also be parameterized by hyperparameters that allow for heterogeneity in how each dimension of the covariate space affects the covariances of the latent scores. Methods that utilize GPs (e.g. Rasmussen and Williams (2006) for binary classification) typically optimize the hyperparameters with respect to the marginal likelihood function (where the latent score variables have been marginalized over). This does not address one of the key assumptions required for causal inference in observational studies, that of covariate balance, thus we propose a method to optimize hyperparameters with respect to a metric of covariate imbalance.

This paper introduces a new approach to estimating the propensity score using Gaussian processes and optimizing hyperparameters with respect to covariate balance. Section 2 provides an overview of causal inference, defines a function for measuring covariate imbalance, and introduces our method for estimating the propensity score. Section 3 provides a simulation study comparing the optimally balanced Gaussian process approach against other methods currently in use for estimating the propensity score in binary settings. Section 4 applies the method to data from LaLonde (1986) in order to compare with results from Dehejia and Wahba (1999) and to provide benchmark performance against this common data set. Section 5 provides discussion and conclusions.

2 | METHODOLOGY

2.1 | Overview of Causal Inference

We utilize the potential outcomes framework of Neyman and Rubin (Splawa-Neyman et al., 1990; Rubin, 1974) for estimating causal effects. For each sampled unit i, let $T_i \in \{0, 1\}$ represent a binary treatment assignment, where $T_i = 1$ and $T_i = 0$ represent membership in a treated group and a control group, respectively. Let Y_i^t be defined as the potential response for unit i under the treatment exposure $T = t$, e.g., Y_i^0 would represent the potential response for unit i under the control exposure. One possible measure of the effect of treatment for unit i is $\tau_i \equiv Y_i^1 - Y_i^0$, which can be summarized across individuals by assessing the average treatment effect (ATE), $\tau_{ATE} = E(Y^1 - Y^0)$. Alternatively it may make more sense to measure the average treatment effect in the treated group (ATT), which is the average of the individual treatment effects conditioned upon the treatment exposure, $\tau_{ATT} = E(Y^1 - Y^0 \mid T = 1)$. An example of when this estimand may be useful is in situations where a treated group is observed and a control group is constructed from a large data source for which relevant data is available and therefore an estimate of a treatment effect is only suitable for those who have been treated. We cannot observe both Y_i^1 and Y_i^0 for a given unit under the *exact* same conditions (i.e., time, environment, etc), and can therefore never measure a true causal effect for a unit in the counterfactual sense (see Holland, 1986, the *Fundamental Problem of Causal Inference*). What we do get to observe is the response for that unit arising from the treatment actually received, that is $Y_i^{obs} = I(T_i = 1)Y_i^1 + I(T_i = 0)Y_i^0$, complicating estimation and inference for τ_{ATE} and τ_{ATT}.

Inference in the potential outcome framework within observational settings can be made possible through the use of two assumptions: strong ignorability given a set of pretreatment covariates and the stable unit treatment value assumption (see Imbens and Rubin, 2015, Chapter 1 for an overview). An assumption of strong ignorability implies that a unit's potential outcomes are conditionally independent of the treatment assignment given covariates and that there is a positive probability of being assigned to either treatment group, mathematically that is $\left(Y_i^1, Y_i^0\right) \coprod T_i | X_i$ and $0 < P(T_i = 1 | X_i = x_i) < 1$ for all i and where x_i is a D-dimensional vector of pretreatment covariates; for example x_i may be a vector embedded in $\mathscr{R}^D$. The stable unit treatment value assumption further implies that there is no interference between units and that there is no hidden treatment variability at different treatment levels. Under these assumptions it is possible to obtain unbiased estimates of the ATE, or the ATT, using observational data by conditioning on the observed covariate vector X.

Conditioning on a D-dimensional vector of covariates may be difficult and thus it may be desirable to condition upon a transformation of X that is of lower dimension. Rosenbaum and Rubin (1983) defined one such transformation as a balancing score, where a balancing score is any function $b(\cdot)$ such that $X \amalg T \mid b(X)$. If treatment assignment is strongly ignorable given X, then it is strongly ignorable given a balancing score $b(X)$ (see Rosenbaum and Rubin, 1983, Theorem 3). Thus, under strong ignorability, conditioning on a balancing score enables unbiased estimation for the ATT or ATE. Rosenbaum and Rubin (1983) demonstrated that the propensity score, defined as $e(x) = Pr(T = 1|X = x)$, is a balancing score. This implies that conditioning on the propensity score is adequate for estimating treatment effects under strong ignorability. In this paper we focus on weighting estimators of the ATE and ATT based on inverse probability weighting similar to those devised in Horvitz and Thompson (1952). Specifically, given a sample $i = 1, \ldots, N$ and estimates of the propensity score, $\hat{e}(x_i)$, the ATE or the ATT may be estimated as follows,

$$\hat{\tau}^* = \frac{\sum_{i=1}^{N} w_i^* I(T_i = 1)Y^{obs}}{\sum_{i=1}^{N} w_i^* I(T_i = 1)} - \frac{\sum_{i=1}^{N} w_i^* I(T_i = 0)Y^{obs}}{\sum_{i=1}^{N} w_i^* I(T_i = 0)} \tag{1}$$

The weights w^* are chosen for the appropriate estimand (defined later in Section 2.3 for the ATE and ATT) and the expression $I(\cdot)$ is an indicator function that evaluates to one if the statement is true. Weighting estimators can be shown to provide unbiased estimates for the ATE and the ATT (Hirano et al., 2003; Stuart, 2010; Imbens and Rubin, 2015).

The assertion of the assumption of strong ignorability requires further discussion. In practice, the assumption that $\left(Y_i^1, Y_i^0\right) \amalg T_i | X_i$, will require that all confounding covariates have been identified; an untestable assumption in observation settings (see Imbens and Rubin (2015)). While this is a clear limitation of observational studies, there have been many proposed methods to address the sensitivity of estimates of treatment effects to the presence of potentially unobserved confounding variables (e.g., the methods outlined in Rosenbaum (2010)). The assumption that $0 < P(T_i = 1|X_i = x_i) < 1$ is often referred to as the positivity assumption, or the overlap assumption, and is also an important assumption that must be well understood when performing an analysis. This positivity assumption, in large samples, provides a region of the covariate space where we would expect to observe outcomes under both treatments, i.e., overlap. Therefore estimates of causal effects within the potential outcomes framework can only be obtained for this region where we observe both outcomes. To enforce this assumption in practice many choose some threshold $\eta \in (0, 0.5)$ and restrict the analysis to individuals whose estimated propensity score satisfies $\hat{e}(x_i) > \eta$ and $\hat{e}(x_i) < 1 - \eta$. We note that this changes the applicable population of individuals that are available to study and therefore changes the causal estimand (see Li et al. (2017)). While it is useful to understand these assumptions in practice, we will assume they hold throughout the remainder of this manuscript.

2.2 | Estimation of Propensity Scores using Gaussian Processes

To estimate the propensity score, we assume that it can be modeled through a probit transformation of a latent variable f_i, i.e., $e(x_i) = P(T_i = 1|X_i = x_i) = \Phi(f_i)$, where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and the dependence of f_i on X_i is left implicit to keep notation simple. For $i = 1, \ldots, N$, let,

$$T_i|X_i = x_i \sim Bernoulli(\Phi(f_i)), \tag{2}$$

and make the assumption that the collection of latent scores f_i arise from a Gaussian process. Let $\mathbf{X}$ represent the $N \times D$ matrix of pretreatment covariates for all units and let θ be a vector of hyperparameters. Then

$$\mathbf{f}|\mathbf{X}, \theta \sim GP(0, K(\mathbf{X}, \theta)), \tag{3}$$

where $K(\mathbf{X}, \theta)$ models the covariance of the observed set of the process and we make the common a priori assumption that the mean of the process is zero.

The covariance matrix $K(\mathbf{X}, \theta)$ describes the 'similarity' between the latent scores for units within the study. The matrix is constructed using a kernel function, $k(x_i, x_j; \theta)$, to compute the covariance between the latent score for unit i and the latent score for unit j given hyperparameters θ. There are many specifications for kernel functions that may be chosen (see Rasmussen and Williams, 2006, Chapter 4 for designing kernel functions). Two common kernel functions for binary regression models are the squared-exponential kernel, $k_{se}(\mathrm{x}_i,\mathrm{x}_j; \rho) = \exp\left\{-\frac{\rho^2}{2}\sum_{d=1}^{D}\left(\mathrm{x}_{i,d} - \mathrm{x}_{j,d}\right)^2\right\}$, and the normalized polynomial kernel, $K_{np}(\mathrm{x}_i,\mathrm{x}_j; \sigma_0, \rho) = \left(\frac{\mathrm{x}_i^T\mathrm{x}_j + \sigma_0^2}{\sqrt{\mathrm{x}_i^T\mathrm{x}_i + \sigma_0^2}\sqrt{\mathrm{x}_j^T\mathrm{x}_j + \sigma_0^2}}\right)^p$. To estimate the propensity score, our model uses an additive function of the squared exponential and the normalized polynomial kernel which is a valid kernel (Rasmussen and Williams, 2006, see Section 4.2.4). That is, we consider a kernel of the form

$$k\left(\mathrm{x}_i, \mathrm{x}_j; \theta\right) = k_{se}\left(\mathrm{x}_i, \mathrm{x}_j; \rho\right) + K_{n\rho}\left(\mathrm{x}_i, \mathrm{x}_j; \sigma_0, p = 1\right), \tag{4}$$

where we fix $p = 1$ to only consider first-order polynomial terms and therefore $\theta = (\rho, \sigma_0)$. The first-order normalized polynomial kernel allows for long range dependencies of the latent scores within the data, while the squared exponential kernel captures local variability in the latent scores.

Within this model both $\mathbf{f}$ and θ are unknown. One approach to inference is to specify a fully Bayesian model by providing a prior distribution for θ and then sample from the posterior distribution of $\mathbf{f}, \theta|\mathbf{T}, \mathbf{X}$ using Markov Chain Monte Carlo (MCMC). What complicates this approach is that causal applications require a specification of the propensity score that is also a balancing score, and therefore only vectors of $\mathbf{f}$ that balance covariates are helpful. Therefore in practice it is adequate to find θ such that a functional of $\mathbf{f}|\mathbf{T}, \mathbf{X}, \theta$ balances covariates (e.g., a function of $E(\mathbf{f}|\mathbf{T}, \mathbf{X}, \theta)$). One way to choose such a θ would be to maximize a form of the marginal likelihood with respect to θ where the latent scores have been integrated out of the joint distribution of $\mathbf{f}$, $\mathbf{T}|\mathbf{X}$, θ. This approach is complicated by the non-Gaussian likelihood (though it is not impossible) and it is an approach that indirectly attempts to balance covariates by using an unbiased estimate of $E(\mathbf{f}|\mathbf{T}, \mathbf{X}, \theta)$, relying on the fact that an unbiased estimate of the propensity score is a balancing score. We propose an approach that is more direct in that we model $\mathbf{f}|\mathbf{T}, \mathbf{X}, \theta$ and select the value of θ that yields propensity score estimates, i.e. $\Phi(E(\mathbf{f}|\mathbf{T}, \mathbf{X}, \theta))$, that provide an optimal level of covariate balance.

Given a value of θ, the conditional posterior distribution $\mathbf{f}|\mathbf{T}, \mathbf{X}, \theta$ is as follows,

$$\begin{aligned} & p(\mathbf{f}|\mathbf{T}, \mathbf{X}, \theta) \propto p(\mathbf{T}|\mathbf{X},\mathbf{f}\,)p(\mathbf{f}|\mathbf{X}, \theta) \\ & \propto |K(\mathbf{X}, \theta)|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{f}^T K(\mathbf{X}, \theta)^{-1}\mathbf{f}\right\} \prod_{i=1}^{N} \Phi(f_i)^{T_i}(1-\Phi(f_i))^{1-T_i}. \end{aligned} \tag{5}$$

This conditional posterior distribution is not tractable and must be approximated. Three common methods exist for approximation (Rasmussen and Williams, 2006): (1) Laplace Approximation; (2) MCMC sampling; and (3) Expectation Propagation. We utilize expectation propagation (EP) for its computational efficiency relative to MCMC and the fact that, in classification problems, the EP approximation has been shown to provide more comparable results to those obtained through MCMC sampling of the conditional posterior distribution than the Laplace Approximation (Kuss and Rasmussen, 2005).

In this setting, expectation propagation replaces each individual treatment assignment likelihood component, i.e. $g_i = \Phi(f_i)^{T_i}(1-\Phi(f_i))^{1-T_i}$, with an approximation that is a scaled Gaussian distribution, $q_i = \frac{Z_i{*}}{\sqrt{2\pi\sigma_i^2{*}}} \exp\left(-\frac{(f_i-\mu_i{*})^2}{2\sigma_i^2{*}}\right)$. The terms μ_{i*}, σ_{i*}^2 are the center and scale of the Gaussian approximation, respectively. The factor Z_{i*}, ensures that the constant of integration between the true posterior distribution and the approximation agree. Using this approximation it follows that

$$p(\mathbf{f}|\mathbf{T}, \mathbf{X}, \theta) \propto |K(\mathbf{X}, \theta)|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{f}^T K(\mathbf{X}, \theta)^{-1}\mathbf{f}\right\} \prod_{i=1}^{N} \Phi(f_i)^{T_i}(1-\Phi(f_i))^{1-T_i} \tag{6}$$

$$\approx |K(\mathbf{X}, \theta)|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{f}^T K(\mathbf{X}, \theta)^{-1}\mathbf{f}\right\} \prod_{i=1}^{N} \frac{Z_{i*}}{\sqrt{2\pi\sigma_{i*}^2}} \exp\left(-\frac{(f_i-\mu_{i*})^2}{2\sigma_{i*}^2}\right) \tag{7}$$

$$\propto \exp\left\{-\frac{1}{2}\mathbf{f}^T K(\mathbf{X}, \theta)^{-1}\mathbf{f}\right\} \exp\left\{-\frac{1}{2}(\mathbf{f}-\boldsymbol{\mu}_*)^T \sum\nolimits_*^{-1} (\mathbf{f}-\boldsymbol{\mu}_*)\right\} \tag{8}$$

where $\boldsymbol{\mu}_*$ is the vector of μ_{i*} and $\sum_* = diag(\sigma_{i*}^2)$. Note that in Equation (8) it appears that the Z_{i*} are irrelevant, but they play an important role in the EP algorithm to ensure that the appropriate μ_{i*}, σ_{i*}^2 are selected.

The EP approximation to the distribution of $\mathbf{f}|\mathbf{T}, \mathbf{X}, \theta$ can then be found using properties of normal distributions. The utility of the approach depends on finding acceptable Z_{i*}, μ_{i*}, and σ_{i*}^2 for all i. The EP algorithm as specified in Rasmussen and Williams (2006) and Kuss and Rasmussen (2005) is performed in an iterative fashion until the approximating parameters converge to stable values. The EP algorithm of Rasmussen and Williams (2006) can be significantly improved with respect to computational runtime if done in parallel (Van Gerven et al., 2010; Tolvanen et al., 2014). The implementation that we utilize is a parallel version of the EP algorithm (See Van Gerven et al. (2010) and Tolvanen et al. (2014) for overviews). Using the approximation of the distribution $\mathbf{f}|\mathbf{T}, \mathbf{X}, \theta$ resulting from the EP algorithm, the propensity score can be estimated from its mean, $\hat{e}(\mathrm{x}_i) = \Phi(E(\mathrm{f}_i | \mathrm{T}, \mathrm{X}, \theta))$. The results in this section assume that θ is already known, Section 2.4 describes our approach for selecting θ in order to minimize covariate imbalance.

2.3 | Measuring Covariate Imbalance

One method of assessing the adequacy of a propensity score model is to measure the degree to which it balances the distributions of the pretreatment covariates in the treated and control groups. Here we develop a measure of covariate imbalance as a function of the moments of the covariate distributions conditioned upon treatment type. Let $X_{i,d}$ be the d^{th} dimension of the covariate vector for individual i. Then for each dimension of the covariate space corresponding to a continuous covariate, we define

$$\bar{X}_d(t) = \frac{\sum_{i=1}^{N} w_i^* I(T_i = t) X_{i,d}}{\sum_{i=1}^{N} w_i^* I(T_i = t)} \quad \text{and} \quad s_d^2(t) = \frac{\sum_{i=1}^{N} w_i^* I(T_i = t)\left(X_{i,d} - \bar{X}_d(t)\right)^2}{\sum_{i=1}^{N} w_i^* I(T_i = t)} \tag{9}$$

as the weighted mean and weighted sample variance for each covariate in each group, respectively. For binary covariates let $\bar{X}_d(t)$ be defined the same way but let $s_{G,d}^2 = \bar{X}_d(t)\left(1 - \bar{X}_d(t)\right)$. In this work, ordinal covariates are treated as continuous covariates and categorical covariates are transformed to binary covariates using dummy variables. The weights, w_i^*, are defined in two different ways, depending on the estimand of interest. They are defined for the ATE and the ATT as follows (Hirano et al., 2003; Stuart, 2010):

$$w_i^{ATE} = \frac{I(T_i = 1)}{\hat{e}(\mathrm{x}_i)} + \frac{I(T_i = 0)}{1 - \hat{e}(\mathrm{x}_i)} \quad \text{and} \quad w_i^{ATT} = I(T_i = 1) + I(T_i = 0)\frac{\hat{e}(\mathrm{x}_i)}{1 - \hat{e}(\mathrm{x}_i)} \tag{10}$$

The weighted means and variances defined in Equation (9) can be used to assess covariate imbalance. Imbens and Rubin (2015) define the standardized difference in the means and the logarithm of the ratio of the sample standard deviations as

$$\Delta_d = \frac{\bar{X}_d(1) - \bar{X}_d(0)}{\sqrt{\left(s_d^2(1) + s_d^2(0)\right)/2}} \quad \text{and} \quad \Gamma_d = \log(s_d(1)) - \log(s_d(0)), \tag{11}$$

respectively. When the propensity score adequately balances the conditional distributions of the pretreatment covariates, both $|\Delta_d|$ and $|\Gamma_d|$ should be small. We operationalize this notion by defining $\mathscr{B}_d$ as the covariate imbalance in dimension d, where

$$\mathscr{B}_d = \begin{cases} |\Delta_d|^2 + |\Gamma_d|^2 & \text{if } \mathrm{X}_d \text{ is a continuous covariate} \\ |\Delta_d|^2 & \text{if } \mathrm{X}_d \text{ is a binary covariate} \end{cases} \tag{12}$$

and use a measure of overall covariate imbalance that is a sum of the covariate imbalance in each dimension,

$$\mathscr{B} = \sum_{d=1}^{D} \mathscr{B}_d . \tag{13}$$

Two choices made in the proposed definition of $\mathscr{B}$ deserve further explanation. First, the balance metrics used for continuous and binary distributions differ. The purpose of the measure of covariate imbalance is to quantify the total difference between the moments of the distributions of the pretreatment covariates between the treated and control groups. The distribution of a binary covariate is completely defined by the first moment and therefore Γ_d for such a covariate is a function of the first moment. This implies that including Γ_d in the imbalance measure for binary covariates would, in effect, more heavily weight the first moments of binary covariates than those of continuous covariates. In simulation studies this negatively impacted performance. The variance terms, Γ_d, are needed for continuous covariates to reduce the bias that would arise if the distributions were not balanced with respect to the second moment. A second feature of our measure is that each term in $\mathscr{B}_d$ is squared. The typical advice is to evaluate covariate balance by considering the absolute value of each term (Imbens and Rubin, 2015). Squaring each term in $\mathscr{B}_d$ penalizes solutions that allow some dimensions of imbalance to remain large while others approach zero. In simulations this was found to improve performance by ensuring that no dimensions have substantial imbalance. This is analogous to the difference between minimizing absolute error loss and minimizing squared error loss in multivariate estimation problems.

2.4 | Minimizing Covariate Imbalance

Section 2.2 described our approach for estimating propensity scores given hyperparameter θ. Section 2.3 defined a measure of the overall covariate imbalance, $\mathscr{B}$, which is a function of the estimated propensity scores (and hence a function of θ). We propose to find a value of θ which minimizes total covariate imbalance,

$$\theta_{opt} = \arg\min_{\theta} \mathscr{B}(\theta) . \tag{14}$$

The function $\mathscr{B}(\theta)$ relies on the EP approximation to find $E(\mathbf{f}|\mathbf{T}, \mathbf{X}, \theta)$, and therefore the derivatives of $\mathscr{B}(\theta)$ are not easily obtained. To optimize covariate imbalance we utilize a derivative-free optimization routine called “Bounded Optimization BY Quadratic Approximation (BOBYQA)” defined in Powell (2009) and implemented in the R package minqa. The algorithm is a method that optimizes a function when first derivatives are not available, as in our application. Additionally, the boundedness of the optimization routine allows for specifying regions of the parameter space that are valid, such as specifying that the parameter θ must be positive. The algorithm requires starting estimates for θ. If all covariates have been standardized (i.e., mean zero and variance of one for continuous covariates and binary covariates transformed to $X_i \in \{1, -1\}$), then the initial value of $\theta_{init} = (1, 1)$ has provided satisfactory performance in our simulations. Once a vector of parameter values, θ_{opt}, has been found that minimizes covariate balance, the propensity score is estimated as $\hat{e}(\mathrm{x}_i) = \Phi\left(E\left(f_i|\mathbf{T}, \mathbf{X}, \theta_{opt}\right)\right)$ and an estimate of a treatment effect can be found using weighted least squares.

2.5 | Software Implementation

The methods developed here have been implemented within the R programming language in a package called gpbalancer. The primary function gpbal takes as inputs a set of observed covariates and treatment assignments and a covariance function and finds the set of hyperparameters of the defined covariance function that minimize Equation 13. The package contains a few covariance functions that are useful for estimation, but these can be extended if necessary. At the time of publication the development version of this package can be found at https://github.com/bvegetabile/gpbalancer where examples of usage can be found.

3 | SIMULATION STUDY

We provide a simulation study to investigate the performance of our method and compare its effectiveness in estimating treatment effects against other propensity score estimation methods. Section 3.1 focuses on estimating the ATE where there is a true propensity score model used to generate the treatment assignments. In this setting we can compare results with the true propensity score, as well as against other methods, to assess the relative performance of our approach. Section 3.2 focuses on simulation results for estimating the ATT. These sections consider estimating treatment effects under two potential outcome models: (1) potential outcomes are linearly related to a covariate with a constant treatment effect and (2) there is a non-constant treatment effect (effect modification by a covariate).

3.1 | ATE Simulation Study

3.1.1 | Simulation Setting—

To compare methods for estimating the ATE, 1000 data sets, each consisting of 500 observations, were simulated with treatment assignment determined using a true propensity score model that is a function of two covariates X_1 and X_2, such that $X_1 \sim \mathcal{N}(0, 1)$ and $X_2 \sim \mathit{Bernoulli}(0.4)$. The potential outcome models for this section are defined to be functions of the continuous covariate $X0_1$ and are described in Table 1. The simple potential outcome models of this section demonstrate settings where the bias of the estimated treatment effect is a function of differences in the distributions of the covariate X_1 conditioned upon treatment type. The first potential outcome model only requires balancing the mean of the covariate as this is the only characteristic of the distribution that will affect the treatment effect estimates. Because the second potential outcome model includes an exponential term the bias is a function of more moments than we include in our imbalance metric; this case demonstrates the performance where we only consider the first two moments of X_1.

In this section we consider two different models for the true propensity score,

$$P\left(T_i = 1 \middle| X_{1,i}, X_{2,i}\right) = \alpha_1 \times \Phi\left(g_j(X_i, \beta)\right) + \alpha_2,$$

for $j = 1, 2$. The function $g_j(X_i, \beta)$ takes one of two forms: a polynomial with linear and interaction terms, $g_1(X_i, \beta) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$, or a second order polynomial with no interaction terms, $g_2(X_i, \beta) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2$. The values α_1 and α_2 were chosen to restrict the values of the propensity score to be in the interval $(\alpha_2, \alpha_1 + \alpha_2)$ and thereby ensure that the positivity assumption is valid. Additionally, by strategically choosing α and β values, it is possible to construct functions that are difficult to model using logistic or probit regression. Settings of the parameters for the two propensity score models are defined in Table 2. Figure 1 visualizes these functions for 500 sample draws as functions of X_1 and X_2.

Within each simulation, eight approaches for estimating the ATE are compared. They are defined in Table 3. First the ME is estimated without adjustment, i.e. $\hat{\tau} = \frac{\sum_i T_i Y^{obs}}{\sum_i T_i} - \frac{\sum_i (1 - T_i) Y^{obs}}{\sum_i (1 - T_i)}$, to provide baseline measures of performance. The average treatment effect is then estimated by weighting using the true propensity score to construct the weights and provides performance measures that could be obtained if the true propensity score were known. Next, weighting adjustment is performed using four nonparametric propensity score estimation methods. The first two methods are our optimally balanced Gaussian process model using the ATE weights previously defined. The first specification of our model utilizes the additive kernel defined in Equation (4) and the second model has a kernel such that $\kappa_{se}\left(X_i, X_j; \rho\right) = \exp\left\{-\sum_{d=1}^{D} \rho_d^2 \left(X_{i,d} - X_{j,d}\right)^2 / 2\right\}$ a version of the squared exponential kernel where each dimension has its own inverse-length scale parameter ρ_d. Additionally, we add a small value (i.e., $1e{-}6$) to the diagonal of each kernel for numerical stability. The other two nonparametric approaches are a method utilizing gradient boosting machines (GBM) available in the twang package in R (McCaffrey et al., 2004) and Bayesian additive regression trees (Chipman et al., 2010; Hill, 2011) available in the BART package in R. Finally we compare the nonparametric methods against methods where the propensity score is estimated parametrically. We consider two methods: the Covariate Balancing Propensity Score, CBPS (Imai and Ratkovic, 2014) available in the CBPS package in R and a generalized linear model (GLM) using logistic regression. For each of these parametric models the model is a misspecified model as defined in Table 3. We note that any parametric model would be “misspecified” if it is missing parameters to capture the effect of α_1 and α_2.

3.1.2 | ATE Simulation Results—Each table of results provides simulation summaries for the eight different adjustment methods outlined in Table 3 under the two different potential outcome models defined in Table 1. The columns of the tables are grouped together into three sections. The first three columns provide the proportion of the 1000 simulations that were declared to be mean balanced, i.e., $|\Delta_d| < \delta$ for all d and for various thresholds of δ. The next four columns are results for the potential outcomes that were linear and related to X_1 and the last four columns are results when the treatment effect varies as a function of X_1. Each group of four columns contains the following simulation summaries without conditioning on balance: the mean bias and mean absolute bias, the mean reduction in bias as compared with no adjustment by the propensity score, the empirical standard error of the simulation ATE estimates, and finally the empirical mean squared error for the ATE. The rows of the table are in the same order as Table 3.

Tables 4 and 5 provide results for the cases where nonparametric modeling should outperform models utilizing parametric assumptions. Table 4 contains results for when the true data generating propensity score was a linear polynomial with interaction terms and Table 5 contains results for when it was a second-order polynomial. The first significant result is that the optimally balanced Gaussian process propensity score methods balance covariates more effectively than other methods, across different thresholds of δ and under both propensity score settings. When the true propensity score was a linear function with interaction terms (Table 4) the CBPS methods also performed well for balancing covariates, while the other nonparametric methods performed well when the true propensity score was a second-order polynomial (Table 5). Neither performed well under both propensity score functions. This suggests that our method is applicable in a wider range of data-generating settings than competing methods for estimating the ATE.

While balancing covariates is an important step in performing causal inference, the primary goal is to minimize the bias of the estimated average treatment effect through adjustment on an estimated propensity score. Tables 4 and 5 demonstrate that the optimally balanced Gaussian process propensity score provides unbiased estimation of the ATE. This demonstrates that the balance achieved through the optimization procedure is not at the expense of other properties of the estimator. Additionally, we see that the mean squared error is often the smallest value in each table and correspondingly provides low empirical standard errors of the estimator.

The alternative nonparametric methods perform similarly to the optimally balanced Gaussian process propensity score in estimating the ATE. These methods generally obtain performance in estimating average treatment effects that are similar to the true data generating propensity score. Across both tables it appears that nonparametric models of the propensity score provide more consistent performance, for both balancing covariates and estimating the ATE, than parametric models. Specifically consider row 7 (GLM-Logistic Regression) of Table 5, the method provides very good performance for estimating the ATE under the the “Linear Related to X_1” setting and in almost all simulations the propensity score estimates balances the covariates, but the weights did not perform well for estimating the ATE when the outcome model contained effect modification. Alternatively consider row 8 (CBPS) of Table 4, this demonstrates a setting where there is good covariate balance, yet the performance in estimating the ATE under the effect modification case is worse than for all nonparametric methods. These examples demonstrate that while a method may be able to balance covariates, it does not guarantee optimal performance in estimating treatment effects.

Comparing across all ATE simulations, the optimally balanced Gaussian process propensity score provides the best performance in aggregate for both balancing the covariates and removing bias in estimating the ATE.

3.2 | ATT Simulation Study

3.2.1 | Simulation Setting—The results of the previous section demonstrated the performance of the method in estimating the ATE when the outcome models were only related to one covariate. In this section, we further explore the performance of the optimally balanced Gaussian process methodology, but in this case for estimating the ATT and when there is a multivariate covariate distribution consisting of both continuous and binary random variables that are related to the potential outcome models. Similar to the last section, we consider 1000 simulated data sets, each of 500 observations each. In this simulation, each $X = (X_1, X_2, X_3, X_4, X_5)^T$ is generated such that, $X_1 \sim \mathcal{N}(0, 1)$, $X_2 \sim \mathcal{N}(0, 1)$, $X_3 \sim Bernoulli(p_3 = 0.3)$, $X_4 \sim Bernoulli(p_4 = \Phi(x_1))$ and $X_5 \sim Bernoulli(p_5 = \Phi(x_2))$. Further, the true propensity score is defined as, $P(T = 1|X = x) = \Phi(f(x))$ where, $f(x) = 0.5x_1 + 0.25x_2 + 0.1x_1x_2x_3 + 0.05x_2x_5 + 0.025x_4$. Under this data generation setting, we then simulate $T_i | X_i = x_i \sim Bernoulli(\Phi(f(x_i)))$. In this section, we again consider two potential outcome settings, but they are now functions of more than one covariate as demonstrated in Table 6.

Similar to the ATE Setting discussed in Section 3.1, we compare the optimally balanced Gaussian process propensity score estimation method against other methods of estimating the propensity score. For our method, we utilize the kernels defined in Section 3.1.1, but now we use the ATT weighting as defined in Section 2.3 when measuring covariate imbalance. We consider seven methods in total for estimating the propensity score in the ATT case as outlined in Table 7 and described previously. In contrast to the previous section the “No Adjustment” estimator does not make sense for estimating the ATT and therefore we do not use it.

3.2.2 | ATT Simulation Results

Similar to Section 3.1.2 we review the results for estimating the ATT using the optimally balanced GP method and compare it against other methods of estimating the propensity score; Table 8 provides simulation results. The table is grouped by potential outcome setting and each group contains performance metrics for the simulations as in Section 3.1 without conditioning on covariate balance.

Table 8 demonstrates that the optimally balanced Gaussian process propensity score again performs well for balancing covariates, as it was intended to do. The results also demonstrate that our method is comparable with CBPS, a method which also optimizes with respect to covariate balance though with parametric assumptions. Finally, we see that all methods provide bias reduction for estimating the ATT, but methods that optimize on covariate imbalance often provide lower levels of MSE and absolute bias.

4 | APPLICATION

This section focuses on comparing propensity score models constructed in Dehejia and Wahba (1999) to propensity score estimates from our developed method. The analysis of Dehejia and Wahba (1999) was itself a replication of earlier work by LaLonde (1986), and the data set of Dehejia and Wahba (1999) is often considered a benchmark data set to demonstrate the performance of propensity score estimation methods. The research goal of LaLonde (1986) was focused on the extent to which observational data can be used to replicate results obtained through controlled experiments. The original analysis assessed the effect of a job training program, the National Supported Work (NSW) Demonstration, on post-exposure earnings. Within the original study there was randomization to either a control group that did not get training or a treatment group that did receive training. LaLonde (1986) then collected six observational data sets and tried to use these data as pools of control individuals with which to replicate the experimental results; three from the Panel Study on Income Dynamics (PSID-1, PSID-2, PSID-3) and three from the Current Population Survey-Social Security File (CPS-1, CPS-2, CPS-3). The data sets PSID-2, PSID-3 and CPS-2, CPS-3 are subsets from PSID-1 and CPS-1 data sets, respectively, and were chosen because the author believed that the subsetted individuals were more similar to the NSW experimental treatment group. Dehejia and Wahba (1999) then extended these results using propensity score estimation based upon the work of Rosenbaum and Rubin (1983) to estimate the ATT. In this section, we provide similar analyses using the propensity score models of Dehejia and Wahba (1999) and our optimally balanced Gaussian process propensity score using inverse-probability weighted estimation for the ATT.

The data we will analyze are the subset of individuals that were utilized in Dehejia and Wahba (1999). This subset of data focused on men who were assigned to treatment after 1975 and the outcome measure was the post-intervention earnings in 1978. The data set contains the following variables for each individual: age, education in years, indicators of whether the individual was black or Hispanic, an indicator of whether the individual was married, an indicator of "no degree", and retrospective earnings in 1974 and 1975. Table 9 provides summaries of the data. It is clear that the observational data sets are different from the experimental data indicating that adjustment is necessary to remove the effects of potential confounding variables. In particular individuals within the experimental data were often younger, were less often married, and had significantly less earnings in 1974 and 1975; additionally there was a higher representation of black individuals within the experimental data.

Table 10 provides two sections of results for each data set (i.e., a row in this table): 1) the first two columns contain summaries of unweighted covariate imbalance, as defined by Equation (13), comparing the control data set to the NSW treatment group, and estimates of the treatment effect without weighting adjustment; 2) the remaining columns are results for covariate imbalance measures and estimates of the ATT using propensity score models from Table 3 of Dehejia and Wahba (1999)[1] and propensity scores estimated using our optimally balanced Gaussian process approach. We note that the original Dehejia and Wahba (1999) analysis did not contain a similar weighting analysis for the ATT, but we provide one here using their prescribed propensity score functions. The first point of comparison is that using the experimental data (first row of Table 10), the estimated difference in post-intervention earnings between the treated and control group was $1794, suggesting that the NSW job training program may have provided a benefit for individuals who were similar to those who were enrolled within the study. Now, as compared to the experimental results, the unweighted estimates of the effect of treatment using the observational controls would suggest that there was no effect of the training program, but this assertion can be doubted as the value of the covariate imbalance metrics indicates large differences in the covariate distributions between the experimental and observational control sources. The next result is that weighting by the estimated propensity score clearly provides benefits for reducing covariate imbalance to levels that are more comparable with the experimental control group and that our optimally balanced Gaussian process propensity score estimation method obtains imbalance metrics that are smaller in magnitude than the experimental covariate imbalance metric. We also see that the models of Dehejia and Wahba (1999) do not provide a level of covariate imbalance that is comparable to our method across the PSID data sets, but the methods are similar across CPS data sets (for a full comparison of covariate imbalance measures across all covariate dimensions see the supplemental information). In particular, many of the covariate dimensions within the PSID data sets contain covariate imbalance measures that are approximately 0.2 or larger, often considered an indicator that propensity score estimates are inadequate. Next, the results in columns 5 and 7 demonstrate that estimates similar to those from the experimental results can be obtained using observational control groups by ATT weighting using the estimated propensity score (though the results are no longer statistically significant). Both methods provide estimated treatment effects that are now consistent with those obtained using the experimental data, as compared with no weighting adjustment, i.e., they are now of a similar magnitude and in the correct direction. We see that when there is a large remaining imbalance, as demonstrated in the PSID data sets using the models of Dehejia and Wahba (1999), that the estimated treatment effect is larger than the estimated treatment effect using our methodology. When the methods provide similar levels of covariate imbalance, as is the case in the CPS data sets, the estimated treatment effects agree more. We note that while the results are similar, the optimally balanced Gaussian process methodology required many fewer modeling decisions in specifying afunctional form for the propensity score in each observational data set.

5 | DISCUSSION

Estimation of the propensity score is an often-used tool in causal analyses of observational data. Estimation of average treatment effects, either the ATE or the ATT, through propensity score weighting provides a flexible method of allowing researchers to control for pretreatment covariates. Often though, researchers make parametric modeling assumptions when estimating the propensity score and these assumptions may not be adequate to remove bias in the estimated treatment effects due to covariate imbalance. This paper describes a nonparametric estimation strategy that utilizes a Gaussian process model to estimate the

[1] All models were logistic regression. For PSID-1, covariates included were: *Age*, Age^2, *Education*, $Education^2$, *I (Married)*, *I (NoDegree)*, *I (Black)*, *I (Hispanic)*, $RE74$, $RE75$, $RE74^2$, $RE75^2$, $I(RE74 = 0) \times I(Black)$. For PSID-2 & PSID-3, covariates included were: *Age*, Age^2, *Education*, $Education^2$, *I (Married)*, *I (NoDegree)*, *I (Black)*, *I (Hispanic)*, $RE74$, $RE75$, $RE74^2$, $RE75^2$, $I(RE74 = 0)$, $I(RE75 = 0)$. For CPS-1, CPS-2 & CPS-3, covariates included were: *Age*, Age^2, Age^3, *Education*, $Education^2$, *I (Married)*, *I (NoDegree)*, *I (Black)*, *I (Hispanic)*, $RE74$, $RE75$, $I(RE74 = 0)$, $I(RE75 = 0)$, $Education \times RE74$

probability of treatment given pretreatment covariates. The hyperparameters of the Gaussian process are chosen to minimize an overall covariate imbalance metric. The potential of the proposed method was highlighted using a series of simulations and an application that replicated findings from Dehejia and Wahba (1999).

Our Gaussian process propensity score method is advantageous due to the flexibility in modeling the propensity score and the fact that what are truly needed in a causal analysis are estimates from a balancing score. Our paper demonstrated that optimizing propensity score estimates to minimize a metric of covariate imbalance can provide better performance. This advantage was demonstrated in Section 3, where the methods that consistently performed best in estimating either the ATE or the ATT were those that were selected to minimize the covariate imbalance metric. While the CBPS is also optimized towards this pursuit, it was demonstrated in Table 4 that nonparametric methods can provide comparable or better performance as it relates to the treatment effect estimation stage of a causal analysis. The Gaussian process approach provides propensity score estimates that balance covariates, and provide the required bias reduction in estimating treatment effects and lower mean squared error. A secondary advantage relates to the positivity assumption in causal inference. Logisitic regression and other parametric models make an assumption that as we reach more and more extreme values of the covariate space the probability of treatment goes to either zero or one. This may not be a valid assumption and the Gaussian process propensity score method (and other nonparametric methods) allows flexibility for modeling more extreme values of the covariate space.

There are limitations associated with our approach though. The first limitation of the method is that of computational runtime. Gaussian processes are computationally challenged by the need to invert a dense covariance matrix at each step of the algorithm, a process which scales at $O(N^3)$. These computational restrictions limit the maximum size of a data set that can be used to estimate the propensity score. For example consider Table 11, which provides timing comparisons across the various methods under the setting of a propensity score model that is defined by a linear polynomial with interaction terms similar to the one used in Section 3.1. We see that as the size of the data set increases, the computational burden of the GP approach increases exponentially. These simulations were run on a 2013 iMac with a quad-core 3.4 GHz Intel Core i5 processor and 16 GB of memory.

We have already made attempts to reduce this computational burden by implementing a version of the expectation propagation algorithm that provides an approximation of the posterior distribution of the latent scores and is run in parallel across multiple cores. There are also other considerations to further reduce this computational burden, such as those that rely on geometric assumptions to create a sparse covariance matrix (e.g., assuming the correlation between points can be set to zero past a certain distance), reduced rank approximations to the covariance matrix, or alternative approximations of the posterior distribution to reduce runtime. Further research will focus on reducing the computational burden associated with this approach.

Another limitation of our Gaussian process propensity score method is that the optimization routine could be considered a black-box procedure that does not allow the user to apply important domain-specific knowledge. That is, the algorithm provides an optimal solution with respect to the defined loss function, but that does not mean it provides hyperparameters such that the imbalance of all, let alone specific, covariate dimensions go to zero. In our application setting, for example, the mean imbalance of previous earnings covariates ($RE74$ and $RE75$) were still found to be somewhat high ($|\Delta| \sim 0.2$) after propensity score adjustment in the PSID-1 data set (see the supplemental information for detailed covariate imbalance metrics). Such considerations are important because previous earnings may be correlated with both an individual's inclusion into training programs, and with theirfuture earnings. Therefore residual imbalance within important dimensions, such as previous earnings, may imply a certain level of bias cannot be removed when estimating treatment effects. The implication is that our method does not relieve the analyst from thinking deeply about the causal mechanisms within the data and potential sources of bias when there is difficulty in achieving a minimal level of covariate imbalance.

While not demonstrated here, one potential adjustment to our procedure to address this would be to augment the loss function, prior to estimating treatment effects, so that imbalance in certain dimensions receive more weight in our metric if they are believed to be more important in the final analysis. Alternatively, as was demonstrated in Section 3, different kernel functions have slightly different properties. Choosing, or constructing a new kernel function for the application at hand may be an alternative way to incorporate domain-specific structure. This may be necessary to correctly model the propensity score in difficult situations. Finally, if neither of these solutions provide adequate covariate balance, it may be a case in which there is no function which can be used to balance covariates for the estimand of interest.

It is worth noting that convergence of our proposed propensity weighting method is likely to be slower than root-N. van der Vaart and van Zanten (2011) derive an upper bound for the quadratic risk of the nonparametricGP model under Matern and squared exponential kernels, which in turn yields an upper bound on the Kullback-Leibler information between the predictive and true data distribution. This provides some insight into the rate of convergence of the GP weights in our propensity model. More specifically, they show that the quadratic error rate associated with a GP fit is bounded by the smoothness of the underlying function to be approximated, the smoothness of the specified kernel, and the number of covariates used in the prediction model. This impact on the propensity weights will then carry through to the convergence of the estimated treatment effect since consistency of the propensity weights is necessary to ensure consistency of the estimate of treatment effect. Despite this, we note the relatively strong performance of our estimator and other flexible nonparametric estimators that we have compared to in finite sample simulation studies with reasonable sample sizes.

Our simulation results and application show that by using the optimally balanced Gaussian process approach to propensity score modeling, we are able to balance covariates in many settings and enable estimation of the ATE or the ATT. It is clear from this study that under the stable unit treatment value assumption and strong ignorability given the covariates, estimating the propensity score using Gaussian processes and optimizing parameters of the model to minimize metrics of covariate imbalance is an effective nonparametric modeling strategy which provides unbiased estimation of treatment effects.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

Research reported in this article was supported by NIMH Silvio O. Conte Centers of the National Institutes of Health under award number P50 MH 096889 (Brian Vegetabile and Hal Stern) and NIH Award P50 AG16573 (Daniel L. Gillen). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Funding information

National Institute of Health (NIH), National Institute of Mental Health, Grant/Award Number: P50 MH 096889 (Vegetabile, Stern); NIH, National Institute on Drug Abuse: Grant/Award Number R01 DA 045049 (Vegetabile); and NIH, National Institute on Aging: Grant/Award Number R01 AG 053555 (Gillen);

REFERENCES

Chipman HA, George EI, McCulloch REet al. (2010) BART: Bayesian additive regression trees. The Annals of Applied Statistics, 4, 266–298.

Dehejia RH and Wahba S (1999) Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American Statistical Association, 94, 1053–1062.

Hill JL (2011) Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics, 20, 217–240. URL:10.1198/jcgs.2010.08162.

Hirano K, Imbens GW and Ridder G (2003) Efficient estimation of average treatment effects using the estimated propensity score. Econometrica, 71, 1161–1189.

Holland PW (1986) Statistics and causal inference. Journal of the American Statistical Association, 81, 945–960. URL:http://www.jstor.org/stable/2289064.

Horvitz DG and Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47, 663–685. URL:http://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483446.

Imai K and Ratkovic M (2014) Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76, 243–263. URL:10.1111/rssb.12027.

Imbens GW and Rubin DB (2015) Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press.

Kuss M and Rasmussen CE (2005) Assessing approximate inference for binary Gaussian process classification. Journal of Machine Learning Research, 6, 1679–1704. URL:http://dl.acm.org/citation.cfm?id=1046920.1194901.

LaLonde RJ (1986) Evaluating the econometric evaluations of training programs with experimental data. The American Economic Review, 604–620.

Lee BK, Lessler J and Stuart EA (2010) Improving propensity score weighting using machine learning. Statistics in Medicine, 29, 337–346. [PubMed: 19960510]

Li F, Morgan KL and Zaslavsky AM (2017) Balancing covariates via propensity score weighting. Journal of the American Statistical Association, 1–11.

McCaffrey DF, Ridgeway G and Morral AR (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological Methods, 9, 403–425. [PubMed: 15598095]

Powell MJ (2009) The bobyqa algorithm for bound constrained optimization without derivatives. Cambridge NA Report NA2009/06, University of Cambridge, Cambridge.

Rasmussen CE and Williams CK (2006) Gaussian Processes for Machine Learning. Cambridge, Massachusetts: The MIT Press.

Rosenbaum PR (2010) Design of Observational Studies. Springer Series in Statistics. New York NY: Springer Verlag.

Rosenbaum PR and Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. Biometrika, 70, 41–55. URL:http://biomet.oxfordjournals.org/content/70/1/41.abstract.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66, 688.

Splawa-Neyman J, Dabrowska DM and Speed TP (1990) On the application of probability theory to agricultural experiments. Essay on Principles. Section 9. Statistical Science, 5, 465–472. URL:http://www.jstor.org/stable/2245382.

Stuart EA (2010) Matching methods for causal inference: A review and a look forward. Statistical Science, 25, 1. [PubMed: 20871802]

Tolvanen V, Jylänki P and Vehtari A (2014) Expectation propagation for nonstationary heteroscedastic Gaussian process regression. In 2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 1–6. IEEE.

van der Vaart A and van Zanten H (2011) Information rates of nonparametric gaussian process methods. Journal of Machine Learning Research, 12, 2095–2119.

Van Gerven MA, Cseke B, De Lange FP and Heskes T (2010) Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. NeuroImage, 50, 150–161. [PubMed: 19958837]

Woo M-J, Reiter JP and Karr AF (2008) Estimation of propensity scores using generalized additive models. Statistics in Medicine, 27, 3805–3816. [PubMed: 18366144]

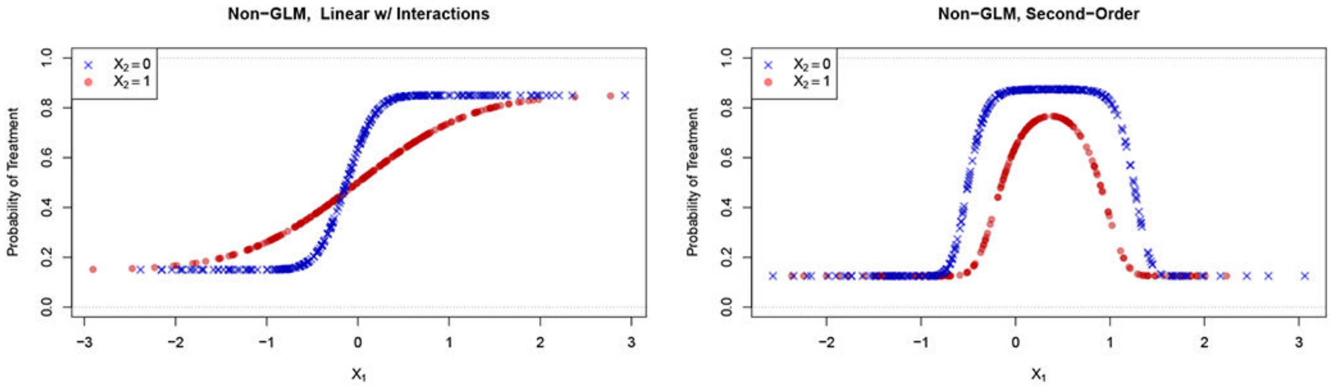

FIGURE 1.

Visualization of Propensity Score Simulation Settings. Parameter values are described in Table 2.

TABLE 1

Models for Potential Outcomes for Section 3.1. The error terms were simulated such that $\epsilon_{T=t,i} \sim N(0,0.5^2)$ for $t \in \{0,1\}$ for each unit i.

Potential Outcome Setting	Treatment Response	Control Response
1) Linear Related to X_1	$Y^1 = X_1 + 3 + \epsilon_{T=1}$	$Y^0 = X_1 + \epsilon_{T=0}$
2) Effect Modification in X_1	$Y^1 = \exp(X_1) + 4X_1 + 3 + \epsilon_{T=1}$	$Y^0 = -X^2_1 - \exp(X_1) + \epsilon_{T=0}$

TABLE 2

Parameter settings used to create propensity score functions

Setting	**Parameters, $\gamma = (\beta_0, \beta_1, \beta_2, \beta_3, \alpha_1, \alpha_2)$**
Non-GLM, Linear w/ Interactions	(0.5, 4, −0.5, −3, 0.7, 0.15)
Non-GLM, Second Order	(2.5, 3, −4, −2, 0.75, 0.125)

TABLE 3

Models to be compared for estimating the ATE under various simulation settings and the R package used to estimate them. The first two methods (no adjustment and adjustment by the true propensity score) provide baseline performance measures. The next four methods are nonparametric methods of estimating the propensity score. The package gpbalancer can be found at https://github.com/bvegetabile/gpbalancer. The last two rows are parametric methods of estimating the propensity score. Note that the parametric models are misspecified in the sense that they are missing terms to handle the fact that $\alpha_1 \neq 1$ and $\alpha_2 \neq 0$ in the data-generating propensity score model.

No.	Adjustment Weighting Method	R package
1	No Adjustment	-
2	True Propensity Score	-
3	Optimally Balanced Gaussian Process Propensity Score - (Normalized Polynomial + Squared Exponential, Common ρ)	gpbalancer
4	Optimally Balanced Gaussian Process Propensity Score - (Squared Exponential, Covariate specific ρ_d)	gpbalancer
5	Gradient Boosted Machine	twang
6	Bayesian Additive Regression Trees	BART
7	Generalized Linear Model - $g(X, \beta) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$	glm
8	Covariate Balancing Propensity Score - $g(X, \beta) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$	CBPS

TABLE 4

Simulation results for estimating the ATE in the non-GLM setting where the true propensity score was linear and included interaction terms. The first column describes the adjustment methods. The next three columns provide the proportion of simulated data sets where the estimated propensity score balanced covariates. The next columns demonstrate the mean bias, the mean reduction in bias as compared with no adjustment by the propensity score, the empirical standard error of the simulation ATE estimates, and finally the empirical mean squared error for the ATE.

	Prop. Bal. **(|Δ_d| < δ *for all d*)**			***Linear Related to X*** $_1$					***Effect Modification Related to X*** $_1$				
Adjustment Method	**δ = 0.1**	**δ = 0.15**	**δ = 0.2**	**Bias**	**Abs. Bias**	**%ABR**	**Emp. S.E.**	**MSE**	**Bias**	**Abs. Bias**	**%ABR**	**Emp. S.E.**	**MSE**
No Adjustment	0.000	0.000	0.000	0.959	0.959	-	0.082	0.927	1.803	1.803	-	0.383	3.399
True Propensity Score	0.380	0.661	0.843	0.002	0.099	89.544	0.125	0.016	−0.031	0.440	73.432	0.552	0.305
Opt. Bal. GP PS (NPSE)	**0.998**	**1.000**	**1.000**	**0.001**	**0.025**	**97.352**	**0.032**	**0.001**	−0.077	0.318	80.457	0.386	0.154
Opt. Bal. GP PS (SE)	0.994	0.999	**1.000**	0.008	0.027	97.193	0.034	**0.001**	**−0.020**	**0.309**	**81.407**	**0.382**	**0.146**
GBM (twang)	0.071	0.449	0.811	0.156	0.156	83.880	0.050	0.027	0.247	0.398	78.439	0.436	0.251
BART	0.000	0.191	0.818	0.169	0.169	82.452	0.038	0.030	0.249	0.376	79.810	0.402	0.223
GLM - Logistic Regression	0.090	0.228	0.393	−0.306	0.307	68.066	0.227	0.145	0.628	1.365	24.630	2.258	5.486
CBPS	0.837	0.981	**1.000**	0.020	0.062	93.471	0.075	0.006	0.532	0.809	56.118	0.943	1.171

TABLE 5

Simulation results for estimating the ATE in the non-GLM setting where the true propensity score was a second-order polynomial. The column headings are discussed in Table 4.

	Prop. Bal. **($\|\Delta_d\| < \delta$ *for all d*)**			***Linear Related to*** X_1					***Effect Modification Related to*** X_1				
Adjustment Method	**δ = 0.1**	**δ = 0.15**	**δ = 0.2**	**Bias**	**Abs. Bias**	**%ABR**	**Emp. S.E.**	**MSE**	**Bias**	**Abs. Bias**	**%ABR**	**Emp. S.E.**	**MSE**
No Adjustment	0.000	0.000	0.000	0.426	0.426	-	0.086	0.189	1.332	1.332	-	0.343	1.892
True Propensity Score	0.338	0.584	0.804	−0.002	0.103	74.306	0.129	0.017	−0.028	0.566	51.304	0.701	0.492
Opt. Bal. GP PS (NPSE)	0.969	**0.996**	**1.000**	−0.027	0.041	89.804	0.044	0.003	−0.185	0.387	64.966	0.446	0.232
Opt. Bal. GP PS (SE)	**0.973**	**0.996**	0.999	**−0.019**	**0.037**	**90.793**	0.044	**0.002**	−0.123	0.365	67.624	0.437	0.206
GBM (twang)	0.665	0.939	0.995	0.055	0.064	85.130	0.053	0.006	**0.079**	0.382	69.033	0.471	0.228
BART	0.894	0.991	**1.000**	0.056	0.060	86.152	**0.043**	0.005	0.106	**0.357**	**71.378**	0.429	**0.195**
GLM - Logistic Regression	0.895	0.981	0.997	−0.052	0.057	86.788	0.049	0.005	1.813	1.813	−36.127	0.564	3.606
CBPS	0.004	0.053	0.273	0.224	0.224	47.054	0.048	0.052	1.476	1.476	−10.464	**0.413**	2.349

TABLE 6

Models for Potential Outcomes for simulations used for demonstrating performance in estimating the ATT

Potential Outcome Setting	Potential Response Functions	Difference: $E(Y^1 - Y^0 \mid X)$
1) Constant Treatment Effect	$Y^1 = 5X^2{}_1 + X_1X_3 - 4X_2 + 50X_5 + 10 + \epsilon_{T=1}$	10
	$Y^0 = 5X^2{}_1 + X_1X_3 - 4X_2 + 50X_5 + \epsilon_{T=0}$	
2) Effect Modification	$Y^1 = 5X_1{}^2 + X_1X_3 - 4X_2 + 50X_5 + 10X_1 - 3X_2{}^3 + \epsilon_{T=1}$	$10X_1 - 3X_2{}^3$
	$Y^0 = 5X^2{}_1 + X_1X_3 - 4X_2 + 50X_5 + \epsilon_{T=0}$	

TABLE 7

Models to be compared for estimating the ATT under various simulation settings and the R package used to estimate them.

No.	Adjustment Weighting Method	R package
1	True Propensity Score	-
2	Optimally Balanced Gaussian Process Propensity Score - (Normalized Polynomial + Squared Exponential, Common ρ)	gpbalancer
3	Optimally Balanced Gaussian Process Propensity Score - (Squared Exponential, Covariate specific ρ_d)	gpbalancer
4	Gradient Boosted Machine	twang
5	Bayesian Additive Regression Trees	BART
6	Generalized Linear Model: Logistic Regression - $X^{T}\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_1{}^2 + \beta_3 X_2 + \beta_4 X^{2}{}_2 + +\beta_5 X_3 + \beta_6 X_4 + \beta_7 X_5$	glm
7	Covariate Balancing Propensity Score - $\beta_0 + \beta_1 X_1 + \beta_2 X^{2}{}_1 + \beta_3 X_2 + \beta_4 X^{2}{}_2 + +\beta_5 X_3 + \beta_6 X_4 + \beta_7 X_5$	CBPS

TABLE 8

Results for estimating the ATT. The first column lists the adjustment method and the next three columns demonstrate the ability of the method to provide mean covariate balance at various thresholds. The next eight columns are grouped together into two groups of four columns and provide the mean bias for the ATT, the mean absolute bias of the ATT, the empirical standard error of the ATT estimates, and the empirical mean squared error across the 1000 simulations.

	Prop. Bal. **(**$\|\Delta_d\| < \delta$ ***for all d*****)**			***Linear Related to*** X_1				***Effect Modification Related to*** X_1			
Adjustment Method	$\delta = 0.1$	$\delta = 0.15$	$\delta = 0.2$	**Bias**	**Abs. Bias**	**Emp. S.E.**	**MSE**	**Bias**	**Abs. Bias**	**Emp. S.E.**	**MSE**
True Propensity Score	0.817	0.976	0.998	0.059	2.249	2.812	7.904	0.054	2.299	2.916	8.497
Opt. Bal. GP PS (NPSE)	**0.793**	0.968	0.996	0.648	0.840	0.883	1.199	0.642	1.158	1.345	2.219
Opt. Bal. GP PS (SE)	0.171	0.577	0.895	**0.499**	**0.727**	0.841	**0.955**	**0.493**	**1.089**	1.321	**1.985**
GBM (twang)	0.770	**0.976**	**0.998**	1.388	1.756	1.631	4.584	1.382	1.901	1.923	5.604
BART	0.138	0.492	0.815	0.825	0.996	0.931	1.545	0.819	1.258	1.398	2.622
GLM - Logistic Regression	0.605	0.908	0.980	0.570	1.252	1.632	2.986	0.563	1.476	1.922	4.007
CBPS	0.706	0.946	0.984	0.778	0.801	**0.644**	1.020	0.772	1.120	**1.202**	2.041

TABLE 9

Summaries of experimental and observational data from Dehejia and Wahba (1999). The first row is the number of observations from that data source and the next rows provide summaries of the mean and standard deviations for that specific category. Clearly the observational data sources differ in many ways from the experimental data set.

	Experimental Data		**Observational Control Data**					
	NSW - Treated	**NSW - Control**	**PSID-1**	**PSID-2**	**PSID-3**	**CPS-1**	**CPS-2**	**CPS-3**
N_{obs}	185	260	2490	253	128	15992	2369	429
Age	25.82 (7.14)	25.05 (7.04)	34.85 (10.44)	36.09 (12.06)	38.26 (12.84)	33.23 (11.04)	28.25 (11.69)	28.03 (10.77)
Education	10.35 (2.01)	10.09 (1.61)	12.12 (3.08)	10.77 (3.17)	10.3 (3.16)	12.03 (2.87)	11.24 (2.58)	10.24 (2.85)
I (Black)	0.84 (0.36)	0.83 (0.38)	0.25 (0.43)	0.39 (0.49)	0.45 (0.5)	0.07 (0.26)	0.11 (0.32)	0.2 (0.4)
I (Hispanic)	0.06 (0.24)	0.11 (0.31)	0.03 (0.18)	0.07 (0.25)	0.12 (0.32)	0.07 (0.26)	0.08 (0.28)	0.14 (0.35)
I (Married)	0.19 (0.39)	0.15 (0.36)	0.87 (0.34)	0.74 (0.44)	0.7 (0.46)	0.71 (0.45)	0.46 (0.5)	0.51 (0.5)
I (No degree)	0.71 (0.45)	0.83 (0.37)	0.31 (0.46)	0.49 (0.5)	0.51 (0.5)	0.3 (0.46)	0.45 (0.5)	0.6 (0.49)
RE74	2095.57 (4873.4)	2107.03 (5676.96)	19428.75 (13404.18)	11027.3 (10793.28)	5566.87 (7226.76)	14016.8 (9569.5)	8727.96 (8965.95)	5619.24 (6780.83)
RE75	1532.06 (3210.54)	1266.91 (3097.01)	19063.34 (13594.22)	7569.22 (9024.06)	2610.7 (5550.69)	13650.8 (9270.11)	7397.23 (8110.5)	2466.48 (3288.16)

TABLE 10

Estimates of the ATT and summaries of covariate imbalance using data from Dehejia and Wahba (1999) before and after adjusting for estimated propensity scores.

	No Weighting Adjustment		Dehejia & Wahba (1999)		GP Propensity Scores	
Dataset	**Balance, $\mathscr{B}$**	**Est. Diff.**	**Balance, $\mathscr{B}$**	**ATT Est.**	**Balance, $\mathscr{B}$**	**ATT Est.**
NSW Control	0.37	1794 (633)	-	-	-	-
PSID-1	17.55	−15205 (1155)	0.70	3028 (942)	0.12	892 (1002)
PSID-2	8.36	−3647 (960)	0.50	2112 (1099)	0.10	1708 (1025)
PSID-3	5.47	1070 (900)	0.92	2284 (1250)	0.35	1803 (1075)
CPS-1	16.91	−8498 (712)	0.07	1782 (771)	0.06	1273 (717)
CPS-2	8.89	−3822 (671)	0.15	1735 (1024)	0.09	1748 (735)
CPS-3	4.57	−635 (657)	0.27	1116 (1029)	0.26	1568 (808)

TABLE 11

Comparisons of computational runtime across the methods considered in Section 3. The runtimes are averaged across 5 simulated data sets using the data-generating procedure of Section 3.1 where the true propensity score was a linear function with interaction terms.

	Runtime in Seconds		
Method	N_{obs} **= 100**	N_{obs} **= 500**	N_{obs} **= 1000**
Generalized Linear Model - Logistic Regression	<0.01	<0.01	<0.01
Optimally Balanced GP Propensity Score (NPSE)	0.18	3.88	20.57
Optimally Balanced GP Propensity Score (SE)	0.11	2.51	16.63
Bayesian Additive Regression Trees	1.72	2.94	5.15
Covariate Balancing Propensity Score	0.25	0.18	0.29
Gradient Boosted Machines (twang)	2.89	6.64	10.94