Journal Title
XX(X):1–29

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Propensity-score based methods for causal inference in observational studies with non-binary treatments

Shandong Zhao[1], David A. van Dyk[2] and Kosuke Imai[3]

Abstract

Propensity score methods are a part of the standard toolkit for applied researchers who wish to ascertain causal effects from observational data. While they were originally developed for binary treatments, several researchers have proposed generalizations of the propensity score methodology for non-binary treatment regimes. Such extensions have widened the applicability of propensity score methods and are indeed becoming increasingly popular themselves. In this article, we closely examine two methods that generalize propensity scores in this direction, namely, the propensity function (PF), and the generalized propensity score (GPS), along with two extensions of the GPS that aim to improve its robustness. We compare the assumptions, theoretical properties, and empirical performance of these methods. On a theoretical level, the GPS and its extensions are advantageous in that they are designed to estimate the full dose response function rather than the average treatment effect that is estimated with the PF. We compare GPS with a new PF method, both of which estimate the dose response function. We illustrate our findings and proposals through simulation studies, including one based on an empirical study about the effect of smoking on healthcare costs. While our proposed PF-based estimator preforms well, we generally advise caution in that all available methods can be biased by model misspecification and extrapolation.

Keywords

covariate adjustment, generalized propensity score, model diagnostics, propensity function, smooth coefficient model, stabilized weights, nonparametric models

1 Introduction

One of the most common strategies used in numerous scientific disciplines to make causal inferences in observational studies is to adjust for observed confounding variables. Researchers find that results based on regression adjustments, however, can be sensitive to model specification, especially when applied to data where the treatment and control groups differ substantially in terms of their pre-treatment covariates. The propensity score methods of Rosenbaum and Rubin [23], hereafter RR, aim to address this fundamental problem by reducing the covariate imbalance between the two groups. RR showed that

[1] Department of Statistics, University of California, Irvine, CA 92697 USA
[2] Department of Mathematics, Imperial College London, SW7 2AZ, United Kingdom
[3] Department of Government and Department of Statistics, Harvard University, Cambridge MA 02138 USA

Corresponding author:
Kosuke Imai, Professor of Government and of Statistics, Institute for Quantitative Social Sciece, Harvard University, 1737 Cambridge Street, MA 02138, USA.
Email: Imai@Harvard.Edu

under the assumption of no unmeasured confounding, adjusting for the propensity score, rather than potentially high-dimensional covariates, is sufficient for unbiased estimation of causal effects and this can be accomplished using simple nonparametric methods such as matching and subclassification.

Despite their popularity, one limitation of the original propensity score methods is that they are only applicable to studies with a treatment and control group, that is, studies with a binary treatment. Some years ago, several researchers proposed generalization of the propensity score methodology for non-binary treatment regimes [11, 14, 15, 21]. Such extensions have widened the applicability of propensity score methods and are indeed becoming increasingly popular themselves.

These methods, however, require users to overcome the challenges of (i) correctly modeling a treatment variable as a function of a possibly large number of pre-treatment covariates and (ii) modeling the response variable. Both represent significant challenges in practice despite recent progress on more robust modeling strategies [e.g., 9, 25]. Standard diagnostics based on the comparison of the covariate distributions between the treatment and control groups are not directly applicable to non-binary treatment regimes and the final inference can be quite sensitive to the choice of response model. This motivated Flores et al. [8], hereafter FFGN, to propose two extensions to the method of Hirano and Imbens [11] that aim to provide more robust estimation through a more flexible response model.

In this article, we closely examine two propensity score-based methods for causal inference with non-binary treatments, namely, the propensity function (PF) of Imai and van Dyk [14], hereafter IvD; and the generalized propensity score (GPS) of Hirano and Imbens [11], hereafter HI, along with the FFGN extensions. We compare the assumptions and theoretical properties of these methodologies when used to estimate a dose response function (DRF) and examine their empirical performance in practice. Our primary message is cautionary: estimating the full DRF with a continuous treatment in an observational study is challenging. Researchers should be cautious when attempting to do so.

The methods of IvD and HI each formalize a framework for inference, but are based on ideas that appear earlier in the literature [15, 16, 19, 22]. In this article we focus on the estimation of the DRF in an observational study in which both the covariates and treatment are fixed, as outlined in HI. This is the simplest setting beyond estimating an average treatment effect and is a well-defined testing ground for the clear comparison of available methods. As we shall see, even in this simple setting standard methods may exhibit unacceptable statistical properties.

The methods of IvD and HI are not the only ones that can be used to estimate the effect of non-binary treatments in observational studies. Inverse probability weighting (IPW) [21, hereafter RHB], for example, is a general method that (like GPS) enjoys wide application beyond the estimation of a DRF, such as estimating the effect of time-varying treatments and with longitudinal data [e.g., 10, 12, 20]. Ertefaie and Stephens [6] provides a comparison of IPW and GPS in this setting. In part due to space constraints, we focus on using GPS and PF to estimate the DRF, and discuss IPW only insofar as it comes into one of the extensions of the GPS proposed by FFGN.

The remainder of this article is organized into five sections. Section 2 reviews the theoretical properties of the original propensity score methodology and its interrelated generalizations. While the GPS is designed to estimate the DRF, the PF estimates the average treatment effect. In Section 3, we compare the methods of IvD, HI, and FFGN both theoretically and empirically. We demonstrate that the response model used by HI is less flexible than those typically used with propensity score methods and that the methods proposed by FFGN to address this problem can exhibit undesirable properties. We also show that one of FFGN's methods can be improved by using the stabilized weights of RHB, effectively implementing IPW to estimate the full DRF. In Section 4, we compare these methods with a new proposal and show how the method of IvD can also be extended for robust estimation of the full DRF. The efficacy of the proposed methodology is illustrated through simulation studies in Section 4 and an empirically-based study in Section 5. Section 6 offers concluding remarks while Appendix A presents additional simulation results and Appendix B introduces a robust variant of HI's method. Table 1 lists and defines the abbreviations used in the text.

Table 1. List of abbreviations used in text.

Abbreviation	Meaning
DRF	dose response function
FFGN	the paper by Flores, Flores-Lagunes, Gonzalez and Neumann [8]
GPS	generalized propensity score
HI	the paper by Hirano and Imbens [11]
IPW	inverse probability weighting
IvD	the paper by Imai and van Dyk [14]
PF	propensity function
RHB	the paper by Robins, Hernán and Brumback [21]
RR	the paper by Rosenbaum and Rubin [23]
SCM	smooth coefficient model

2 Methods

Suppose we have a simple random sample of size n with each unit consisting of a p-dimensional column vector of pretreatment covariates, $\boldsymbol{X}_i$, the observed univariate treatment, T_i, and the outcome variable, Y_i. Although IvD's method can be applied to multivariate treatments, here we assume the treatment is univariate to facilitate comparison with the method of HI. We omit the subscript when referring to generic values of $\boldsymbol{X}_i$, T_i, and Y_i.

We denote the potential outcomes by $\mathcal{Y} = \{Y_i(t), t \in \mathcal{T} \text{ for } i = 1, \ldots, n\}$, where $\mathcal{T}$ is a set of possible treatment values and $Y_i(t)$ is a function that maps a particular treatment level of unit i, to its outcome. This setup implies the *stable unit treatment value assumption* [24] that the potential outcome of each unit is not a function of treatment level of other units and that the same version of treatment is applied to all units. In addition, we assume *strong ignorability of treatment assignment*, i.e., $Y(t) \perp\!\!\!\perp T \mid \boldsymbol{X}$ and $p(T = t \mid \boldsymbol{X}) > 0$ for all $t \in \mathcal{T}$, which implies no unmeasured confounding (RR).

2.1 The propensity score with a binary treatment

RR considered the case of treatment variables that take on only two values, $\mathcal{T} = \{0, 1\}$, where $T_i = 1$ ($T_i = 0$) implies that unit i receives (does not receive) the treatment and defined the *propensity score* to be the conditional probability of assignment to treatment given the observed covariates, i.e., $e(\boldsymbol{X}) = p(T = 1 \mid \boldsymbol{X})$. In practice, $e(\boldsymbol{X})$ is typically estimated using a parametric treatment assignment model $p_\psi(T = 1 \mid \boldsymbol{X})$ where ψ is a vector of unknown parameters. The appropriateness of the fitted model can be assessed via the celebrated balancing property of $e(\boldsymbol{X})$, namely, that covariates should be independent of the treatment conditional on the propensity score, $\boldsymbol{X} \perp\!\!\!\perp T \mid e(\boldsymbol{X})$. In particular, the fitted model, $\hat{e}(\boldsymbol{X}) = p_{\hat{\psi}}(T = 1 \mid \boldsymbol{X})$ should not be accepted unless adjusting for $\hat{e}(\boldsymbol{X})$ results in adequate balance.

In order to estimate causal quantities, we must properly adjust for $\hat{e}(\boldsymbol{X})$. RR propose three techniques: matching, subclassification, and covariance adjustment. We focus on subclassification and covariance adjustment because they are more closely related to the generalizations for non-binary treatments. The key advantage of propensity scores when applying these methods, and the inverse weighting method discussed below, is dimension reduction. They only require adjustment for a scalar variable $\hat{e}(\boldsymbol{X})$ rather than for the entire covariate vector.

With subclassification (RR), we adjust for $\hat{e}(\boldsymbol{X})$ by dividing the observations into several subclasses based on $\hat{e}(\boldsymbol{X})$. Individual response models are then fitted within each subclass, adjusting for $\hat{e}(\boldsymbol{X})$ and sometimes $\boldsymbol{X}$ along with T. The overall causal effect is then computed as the weighted average of the within-class coefficients of T, with weights proportional to the size of subclass. The standard error of the causal effect is typically computed by treating the within-subclass estimates as independent of one another.

With covariance adjustment (RR), we regress the response variable on $\hat{e}(\boldsymbol{X})$ separately for the treatment and control groups. Specifically, we divide the data into the treatment and control groups and fit the regression model, $\mathrm{E}(Y \mid \boldsymbol{X}, T = t) = \alpha_t + \beta_t \cdot \hat{e}(\boldsymbol{X})$, separately for $t = 0, 1$. The average causal effect is

then estimated as

$$\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\alpha}_1 + \hat{\beta}_1 \cdot \hat{e}(\boldsymbol{X}_i) - \hat{\alpha}_0 - \hat{\beta}_0 \cdot \hat{e}(\boldsymbol{X}_i)\right) = (\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0) \cdot \overline{\hat{e}(\boldsymbol{X})} \tag{1}$$

where $\overline{\hat{e}(\boldsymbol{X})}$ is the sample mean of the estimated propensity score.

Subclassification and covariance adjustment (along with methods such as matching and IPW) aim to provide robust flexible adjustment for $\hat{e}(\boldsymbol{X})$ in the response model. As we shall see below, the flexibility of the response model is important especially for non-binary treatment regimes. This is because unlike the treatment assignment model which has an effective diagnostic tool based on the balancing property of propensity scores, the response model lacks such diagnostics.

2.2 Propensity score: methods for non-binary treatments

Suppose now that $\mathcal{T}$ is a more general set of treatment values, perhaps categorical or continuous. It is in this setting that that IvD introduced the PF and that HI introduced the GPS. (IvD also allow for multivariate treatments, which we do not discuss in this paper.) In what follows, we review and compare these generalizations of propensity score methods. In particular, we consider the following aspects of propensity score adjustment with binary treatments that both IvD and HI generalize:

1. Treatment assignment model: Model the distribution of the treatment assignment given covariates to estimate the propensity score, i.e., $\hat{e}(\boldsymbol{X})$
2. Diagnostics: Validate $\hat{e}(\boldsymbol{X})$, by checking for covariate balance, i.e., $T \perp\!\!\!\perp \boldsymbol{X} \mid \hat{e}(\boldsymbol{X})$
3. Response model: Model the distribution of the response given the treatment, adjusting for $\hat{e}(\boldsymbol{X})$ via matching, subclassification, covariance adjustment, or IPW
4. Causal quantities of interest: Estimate the causal quantities of interest and their standard error based on the fitted response model

2.2.1 Treatment assignment model. As in the case of the binary treatment, we begin by modeling the distribution of the observed treatment assignment given the covariates using a parametric model, $p_\psi(T \mid \boldsymbol{X})$, where ψ is a set of parameters. Common choices of $p_\psi(T|\boldsymbol{X})$ include the Gaussian or multinomial regression models when the treatment variable is continuous or categorical, respectively. HI define the GPS as $R = r(T, \boldsymbol{X}) = p_\psi(T \mid \boldsymbol{X})$. That is, the GPS is equal to the treatment assignment model density (or mass) function *evaluated* at the observed treatment variable and covariate for a particular individual. This is analogous to the propensity score for the binary treatment, which can be written as $e(\boldsymbol{X}) = r(1, \boldsymbol{X}) = p_\psi(T = 1 \mid \boldsymbol{X})$.

Before HI coined the term GPS, the same quantity was used by RHB in IPW. In particular, RHB considered using weights equal to $1/r(T, \boldsymbol{X})$ and then, noting the instability of these weights, suggested using the stabilized weights $W = W(T, \boldsymbol{X}) = p_{\psi_0}(T)/p_\psi(T \mid \boldsymbol{X})$ instead. For example, if we use a normal linear model for $p_\psi(T \mid \boldsymbol{X})$, i.e., $T_i \mid \boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{X}_i^T\beta, \sigma^2)$, we might use a normal model for $p_{\psi_0}(T)$, i.e., $T_i \sim \mathcal{N}(\mu, \tau^2)$. Although RHB first proposed the quantity, we use HI's now standard term, GPS, for $r(T, \boldsymbol{X})$.

IvD summarize $p_\psi(T \mid \boldsymbol{X})$ in a manner that is qualitatively different. First, they define the PF to be the entire conditional density (or mass) function of the treatment, namely $e_\psi(\cdot \mid \boldsymbol{X}) = p_\psi(\cdot \mid \boldsymbol{X})$. This is also analogous to the propensity score for the binary treatment case because $e_\psi(\cdot \mid \boldsymbol{X})$ is completely determined by $e(\boldsymbol{X}) = p_\psi(T = 1 \mid \boldsymbol{X})$. In order to summarize the PF, IvD introduce the *uniquely parameterized propensity function* assumption which states that for every value of $\boldsymbol{X}$, there exists a unique finite-dimensional parameter, $\boldsymbol{\theta} \in \Theta$, such that $e_\psi(\cdot \mid \boldsymbol{X})$ depends on $\boldsymbol{X}$ only through $\boldsymbol{\theta}_\psi(\boldsymbol{X})$. In other words, $\boldsymbol{\theta}$ uniquely represents $e\{\cdot \mid \boldsymbol{\theta}_\psi(\boldsymbol{X})\}$, which we may therefore write as $e(\cdot \mid \boldsymbol{\theta})$ or simply $\boldsymbol{\theta} = \boldsymbol{\theta}_\psi(\boldsymbol{X}_i)$. For example, if we model the treatment, $T_i \sim \mathcal{N}(\boldsymbol{X}_i^\top\boldsymbol{\beta}, \sigma^2)$ with $\psi = (\boldsymbol{\beta}, \sigma^2)$, the scalar $\boldsymbol{\theta}_i = \boldsymbol{X}_i^\top\boldsymbol{\beta}$ uniquely represents $e_\psi(\cdot \mid \boldsymbol{X}_i)$. In practice, ψ, ψ_0, $\boldsymbol{\theta}_i$, W_i, R_i, and r_i are estimated from data; we denote their estimates by $\hat{\psi}$, $\hat{\psi}_0$, $\hat{\boldsymbol{\theta}}_i$, $\hat{W}_i$, $\hat{R}_i$, and $\hat{r}_i$, respectively.

2.2.2 Diagnostics. Diagnostics for the treatment assignment model rely on balancing properties of the IPW, PF and GPS. For example, IvD shows that the PF is a balancing score, i.e., $T \perp\!\!\!\perp \boldsymbol{X} \mid e(\cdot \mid \boldsymbol{\theta})$. IvD suggest checking balance by regressing each covariate on T and $\hat{\boldsymbol{\theta}}$, e.g., using Gaussian and/or logistic regression and comparing the distribution of the t-statistics for each of the resulting regression coefficients of T with the standard normal distribution via a normal quantile plot. Improvement in balance can be assessed by constructing the plot again in the same manner except that $\hat{\boldsymbol{\theta}}$ is left out of each regression. Although not typical used, this diagnostic is equally applicable in the binary treatment case.

HI, on the other hand, show that $\mathbf{1}\{T = t\}$ is independent of $\boldsymbol{X}$ given $r(t, \boldsymbol{X})$, where $\mathbf{1}\{\cdot\}$ is an indicator function and the GPS is evaluated at $t \in \mathcal{T}$. Following the covariate balancing property for the binary propensity score, HI construct a series of binary treatments by coarsening the original treatment T in the form $\{t_j < T \leq t_{j+1}\}$ for some $t_1, t_2, \ldots, t_J$. Covariate balance is then checked for these binary treatment variables by first subclassifying units on $\hat{r}(\widetilde{T}_j, \boldsymbol{X})$, where $\widetilde{T}_j$ is the median of the treatment variable among units with $\mathbf{1}\{t_j < T \leq t_{j+1}\} = 1$. Then, two-sample t-tests are performed within each subclass to compare the mean of each covariate among units with $\mathbf{1}\{t_j < T \leq t_{j+1}\} = 0$ against that among units with $\mathbf{1}\{t_j < T \leq t_{j+1}\} = 1$. Finally the within-subclass differences in means and the variances of these differences are combined to compute a single t-statistic for each covariate. HI suggest repeating this diagnostics for several choices of $\{t_1, \ldots, t_J\}$ that cover the range of observed T.

Because the same treatment models are used with both methods the diagnostics for each may be used for the other. In any case, failure to reject the null hypothesis of perfect balance does not imply balance and hence the diagnostics must be interpreted carefully. In fact, a small (within subclass) sample size, may limit the ability to detect a lack of balance [13].

2.2.3 Response model. The response models proposed by IvD and HI are quite different, with HI relying more heavily on parametric assumptions. IvD propose two response models. The first is completely analogous to the subclassification technique proposed by RR. The data are subclassified on $\hat{\boldsymbol{\theta}}$ and individual response models are fitted within each subclass, adjusting for $\hat{\boldsymbol{\theta}}$ and typically $\boldsymbol{X}$ along with T. The second is a smooth coefficient model (SCM), which allows the intercept and slope to vary smoothly as a function of the PF

$$E(Y \mid T, \hat{\boldsymbol{\theta}}) \;=\; f(\hat{\boldsymbol{\theta}}) + g(\hat{\boldsymbol{\theta}}) \cdot T, \tag{2}$$

where $f(\cdot)$ and $g(\cdot)$ are unknown smooth continuous functions. In our numerical illustrations, we fit this model using the R package `mgcv` developed by Simon Wood, in which smooth functions are represented as a weighted sum of known basis functions; and the likelihood is maximized with an added smoothness penalization term. We use penalized cubic regression splines as the basis functions, with dimension equal to five.

In contrast, HI propose to estimate the conditional expectation of the response as a function of the observed treatment, T, and the GPS, $\hat{R}$. They recommend using a flexible parametric function of the two arguments and give the following Gaussian quadratic regression model,

$$E(Y \mid T, \hat{R}) \;=\; \alpha_0 + \alpha_1 \cdot T + \alpha_2 \cdot T^2 + \alpha_3 \cdot \hat{R} + \alpha_4 \cdot \hat{R}^2 + \alpha_5 \cdot T \cdot \hat{R}. \tag{3}$$

This can be viewed as a generalization of RR's covariance adjustment technique, which in the binary treatment case involves regressing Y on $\hat{e}(\boldsymbol{X})$ separately for the treatment and control groups. HI, on the other hand, parametrically estimate the average outcome for all possible treatment levels simultaneously via the quadratic regression on T given in (3). Non-parametric response models for the GPS are suggested by FFGN, see Section 2.2.5.

2.2.4 Estimating causal quantities. The PF was designed to estimate the average causal effect. Under (2) this involves averaging $g(\hat{\boldsymbol{\theta}}_i)$ across all units. Bootstrap standard errors are computed by resampling the data and refitting both the treatment assignment and response models. With subclassification, computing the estimated average causal effect proceeds exactly as in the binary case. Because a response model is fit

conditional on T within each subclass, we can also in principle average these fitted models and estimate the DRF. While we illustrate this possibility in our simulations, we advocate a flexible non-parametric approach in Section 4.1.

In contrast, to estimate the DRF, HI computes the average potential outcome on a grid of treatment values. In particular, at treatment level t, they compute

$$\hat{E}\{Y(t)\} \;=\; \frac{1}{N}\sum_{i=1}^{N}\Big(\hat{\alpha}_0+\hat{\alpha}_1\cdot t+\hat{\alpha}_2\cdot t^2+\hat{\alpha}_3\cdot \hat{r}(t,\boldsymbol{X}_i)+\hat{\alpha}_4\cdot \hat{r}(t,\boldsymbol{X}_i)^2+\hat{\alpha}_5\cdot t\cdot \hat{r}(t,\boldsymbol{X}_i)\Big). \tag{4}$$

In (4) the mean response had all individuals received dose t is estimated by computing a *dose-specific* score: r is evaluated at t, not T. While not unprecedented [15, 22], this distinguishes the method from the PF, for which there is a single score for each individual. This difference has ramifications. When the inverse GPS is used as a weight, it cannot be stabilized in the standard manner because $p_{\hat{\psi}_0}(t)$ does not vary among individuals so that $p_{\hat{\psi}_0}(t)/r(t,\boldsymbol{X}_i)\propto 1/r(t,\boldsymbol{X}_i)$. Standard errors can be calculated using the bootstrap, taking into account the estimation of both the GPS and model parameters.

In practice, we are often interested in the *relative* DRF, $E\{Y(t)-Y(0)\}$, which compares the average outcome under each treatment level with that under the control, i.e., $t=0$. Of course, in some studies there is no control *per se* and we revert to $E\{Y(t)\}$. In our simulation studies we report the relative DRF while in our applied example we report the DRF which is more appropriate in its particular context.

2.2.5 Extensions of the method of HI. Unfortunately, the quadratic regression in (3) is not sufficiently flexible for robust estimation of the DRF, see Section 3. Bia et al. [1] and FFGN point out that misspecification of (3) can result in biased causal quantities and consider similar nonparametric regression methods. In particular, they generalize (3) with,

$$E(Y\mid T,\hat{R}) \;=\; \beta(T,\hat{R}) \tag{5}$$

where $\beta(T,\hat{R})$ is a flexible nonparametric model; in our numerical studies we use a SCM.* The DRF, $\hat{E}\{Y(t)\}$, and its standard errors are computed as in (4), but with

$$\hat{E}\{Y(t)\} \;=\; \frac{1}{N}\sum_{i=1}^{N}\hat{\beta}[t,\hat{r}(t,\boldsymbol{X}_i)] \tag{6}$$

Because the SCM is a function of the GPS, we refer to this method as SCM(GPS).

FFGN also propose a second method which involves a GPS version of inverse probability weighting. We refer to this method as IPW_0 because it uses naive rather than stabilized weights; recall that stabilization is not possible when using $1/\hat{r}(t,\boldsymbol{X}_i)$ as weights, see Section 2.2.4. The IPW_0 estimate of the DRF is

$$\hat{E}\{Y(t)\} \;=\; \frac{\sum_{i=1}^{N}\tilde{K}_{h,X}(T_i-t)\cdot Y_i}{\sum_{i=1}^{N}\tilde{K}_{h,X}(T_i-t)}, \tag{7}$$

where $\tilde{K}_{h,X}(T_i-t)=K_h(T_i-t)/\hat{r}(t,\boldsymbol{X}_i)$, $K(\cdot)$ is a kernel function with the usual properties, h is a bandwidth satisfying $h\to 0$ and $Nh\to\infty$ as $N\to\infty$, and $K_h(\cdot)=h^{-1}K(\cdot/h)$. This is the local constant

*There are different nonparametric methods based on this model. For example, FFGN propose a nonparametric kernel estimator with a polynomial regression of order 1 [7], while Bia et al. [1] propose using radial basis functions to avoid the need for tensor product splines (they also propose a penalized spline method with tensor product) [see also 2]. We use the SCM to facilitate comparisons of the methods. As with (2), we use the `mgcv` package with penalized cubic regression splines as the basis functions with dimension equal to five for both T and $\hat{R}$ along with a tensor product. We leave the comparison of these different nonparametric methods to future research.

Table 2. Estimates of the DRF. Methods differ in their response models, but not their treatment models.

Estimate	Description
IvD	Using $\hat{\boldsymbol{\theta}}$ to form S subclasses, fit a linear model within each subclass and average the S fitted models to estimate DRF.
HI	Estimate the DRF with (3) and (4).[a]
SCM(GPS)	Same as HI, but with the quadratic regressions in (3) and (4) replaced by a SCM as in (5) and (6).
IPW_0	Same as HI, but with the quadratic regressions in (3) and (4) replaced by the kernel smoothing regression in (7) and (8).
IPW_{SW}	Same as IPW_0, but with $\tilde{K}_{h,X}(T_i - t) = K_h(T_i - t)\hat{W}(T_i, \boldsymbol{X}_i)$.
SCM(PF)	Estimate the DRF with (11) and (12), see Section 4.

[a] In the linear fit of Simulation II, the quadratic models in (3) and (4) are replaced by linear models.

regression (Nadaraya-Watson) estimator but now with each individual's kernel weight being divided by its GPS at t. To avoid boundary bias and to simplify derivative estimation, the IPW_0 estimates $E\{Y(t)\}$ using a local linear regression of Y on T with a weighted kernel function $\tilde{K}_{h,X}(T_i - t)$, i.e.,

$$\hat{E}\{Y(t)\} \;=\; \frac{D_0(t)S_2(t) - D_1(t)S_1(t)}{S_0(t)S_2(t) - S_1^2(t)}, \tag{8}$$

where $S_j(t) = \sum_{i=1}^{N} \tilde{K}_{h,X}(T_i - t)(T_i - t)^j$ and $D_j(t) = \sum_{i=1}^{N} \tilde{K}_{h,X}(T_i - t)(T_i - t)^j Y_i$. The global bandwidth can be chosen following the procedure of Fan and Gijbels [7]. We use (8) as the IPW_0 estimator in our numerical studies.

Unfortunately, IPW_0 can be very unstable, owing to the infinite variance of $1/r(t, \boldsymbol{X}_i)$, at least when the treatment is continuous (RHB). To improve IPW_0 we replace $1/\hat{r}(t, \boldsymbol{X}_i)$ with Robins' stabilized weight, $\hat{W}(T, \boldsymbol{X})$. We denote this method IPW_{SW}, where the subscript indicates its stabilized weights. Notice that these weights are evaluated at T rather than t. This is a fundamental difference from the methods laid out by HI. In this regard IPW_{SW} should be viewed as IPW in flexible kernel smoothing regression, namely via (7), but with $\tilde{K}_{h,X}(T_i - t) = K_h(T_i - t)\hat{W}(T_i, \boldsymbol{X}_i)$. Table 2 summarizes the specific estimates of the DRF that we compare in Sections 3-5.

3 Comparing the GPS and the PF

In this section, we examine the differences between the methods of IvD, HI, and FFGN using both simulation studies and theoretical comparisons. The key differences lie in how the method summarizes $p(T \mid \boldsymbol{X})$: the GPS evaluate this density at the observed covariate, whereas the PF uniquely parameterizes it. As we show below, this difference leads to alternative response models and markedly divergent results.

3.1 Simulation study I

In our first simulation study, we generate $2,000$ observations, each of which includes a single continuous covariate, X, a continuous univariate treatment, T, and a response variable, Y. We simulate $X_i \overset{\text{ind}}{\sim} \mathcal{N}(0.5, 0.25)$ and $T_i \mid X_i \overset{\text{ind}}{\sim} \mathcal{N}(X_i, 0.25)$ and assume that the potential outcome is distributed as $Y_i(t) \mid T_i, X_i \overset{\text{ind}}{\sim} \mathcal{N}(10X_i, 1)$ for all $t \in \mathcal{T}$. (Here and elsewhere we parameterize the Normal distribution in terms of its mean and variance.) Under this simulation setup, the true treatment effect is zero and the true DRF is five for all t. We deliberately choose this simple setting where any reasonable method should perform well. Fitting a simple linear regression of Y on T yields a statistically significant treatment effect estimate of roughly five. However, adjusting for X in the regression model is sufficient to yield an estimate that is much closer to and is not statistically different from the true effect of zero.

Using the correctly specified treatment assignment model, $T_i \mid X_i \overset{\text{ind}}{\sim} \mathcal{N}(X_i, 0.25)$, the marginal distribution of the the treatment, $T_i \sim \mathcal{N}(0.5, 0.5)$, and the response models given in Table 2, we implement

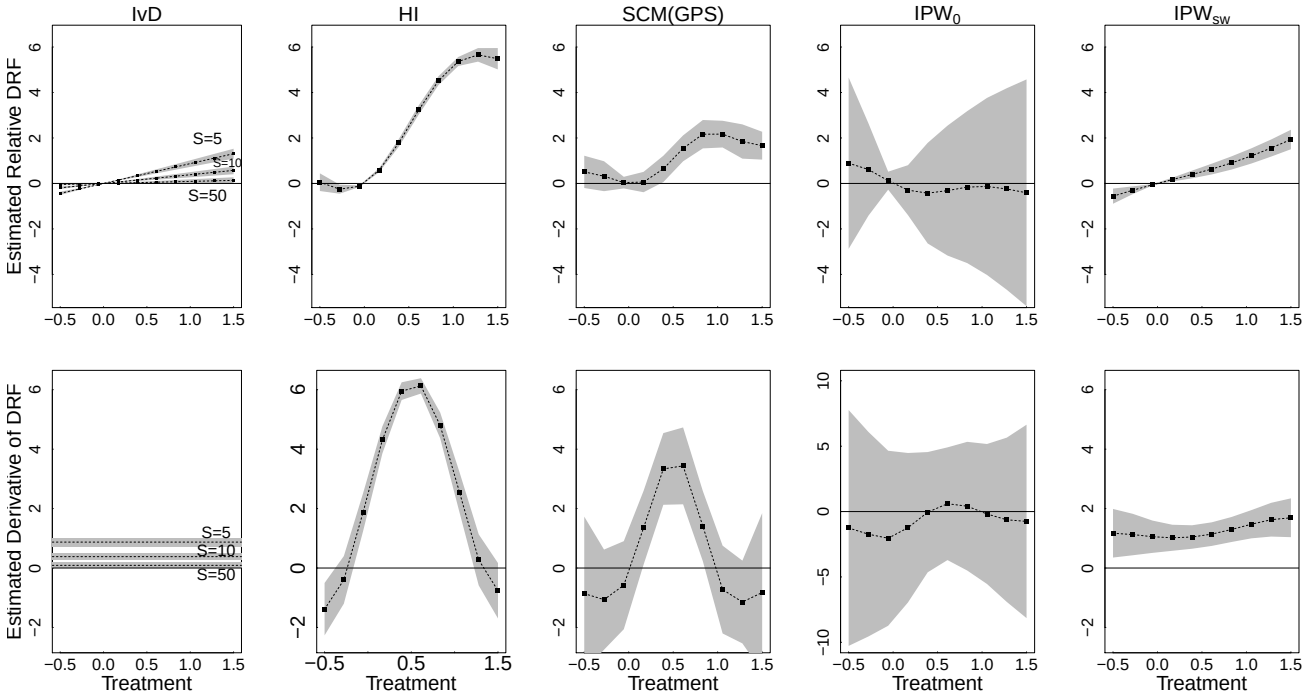

Figure 1. Results of Simulation Study I. The first row plots the estimated relative DRF with the horizontal solid line representing the true relative DRF. For IvD, we use $S = 5, 10, 50$ subclasses. The solid diagonal line for the method of HI is the unadjusted regression of Y on T. The second row plots the estimated derivatives of the DRF with the solid line representing the truth. In both rows, the grey shaded areas represent 95% confidence intervals. The estimated derivative for IPW_0 is plotted on a different scale as its standard error is significantly larger than that of the other methods.

the HI, IvD, SCM(GPS), IPW_0, and IPW_{SW} methods. For the purposes of illustration, we do not adjust for $\hat{\boldsymbol{\theta}}$ within each subclass when using IvD's method. Owing to the linear structure of the generative model, doing so would dramatically reduce bias even with a small number of subclasses. Here we illustrate, instead, how bias can be reduced by increasing the number of subclass; we implement IvD with $S = 5, 10$, and 50 subclasses. For the methods other than IvD, we use a grid of ten equally spaced points between -0.5 and 1.5, $t_1, \ldots, t_D$ with $D = 10$, to compute the relative DRF and its derivative. Standard errors are computed using 1,000 bootstrap replications.

Figure 1 presents the results. In the first row, we plot the estimated relative DRF while the second row plots the estimated derivative of the DRF. For HI, SCM(GPS), IPW_0, and IPW_{SW} the derivative is computed as

$$\frac{1}{2}\left[\frac{\hat{E}\{Y(t_{d+1})\} - \hat{E}\{Y(t_d)\}}{t_{d+1} - t_d} + \frac{\hat{E}\{Y(t_d)\} - \hat{E}\{Y(t_{d-1})\}}{t_d - t_{d-1}}\right] \tag{9}$$

for $d = 2, \ldots D - 1$. For $d = 1$, we simply use the first term in (9) and for $d = 10$ we use the second term in (9). For IvD, the derivative is the weighted average of the within subclass linear regression coefficient; 95% point-wise confidence intervals are shaded gray.

Figure 1 shows that even in this simple simulation, all methods except IPW_0 miss the true relative DRF and its derivative, albeit to differing degrees. The behavior of IvD's estimate improves with more subclasses, a luxury we can afford here because of the large sample size. IvD makes the general recommendation that the within subclass models be adjusted for $\boldsymbol{X}$ or at least for $\hat{\boldsymbol{\theta}}$. Because of the simple structure of this simulation, doing so would result in a correctly specified model even with a single subclass, eliminating

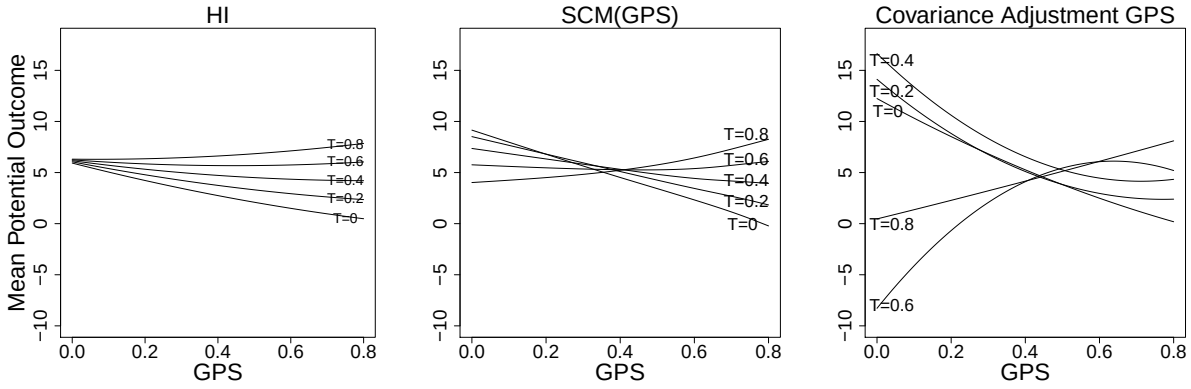

Figure 2. The Varying Flexibility of the Response Models. The plots show the mean potential outcome as a function of the GPS and T under HI's quadratic response model (left panel), fitted SCM(GPS) (middle panel), and covariance adjustment GPS (right panel). Covariance adjustment GPS fits a quadratic regression ($Y \sim R + R^2$) in each of several subclasses based on T, see Appendix B for details. We use 10 subclasses but only plot five. Subclassification is by far the most flexible of the three response models.

bias in the estimated average treatment effect.[†] We do not recommend estimating the DRF by averaging the unadjusted within subclass models, but do so here to facilitate comparisons between the methods. We propose a new estimate of the DRF using the PF in Section 4.1.

The performance of HI's method is particularly poor; it differs only slightly from the unadjusted regression. Although SCM(GPS) offers limited improvement, it also introduces a cyclic artifact into the fit. We see this pattern again and discuss it in Section A.3. IPW_0, on the other hand, results in an unstable fit that is characterized by large standard errors. The performance of these methods are especially troubling both because the GPS was expressly designed to estimate the DRF and because the current simulation setup is so simple. Given their performance here, it is difficult to expect these methods to succeed in more realistic settings. While IPW_{SW} also misses the mark in Simulation I, it is important to emphasize that RHB did not propose to use IPW to estimate the full DRF. The primary goal of this paper is to explain why the GPS-based methods can fail and to provide a more robust estimate of the DRF.

One reason that GPS-based methods can perform poorly is that their response model are based on overly strong parametric assumptions, especially (3). This is illustrated in Figure 2 which compares the fitted mean potential outcome as a function of the GPS and T under the HI model (left panel) and under SCM(GPS) (middle panel) (Appendix A.1 provides additional discussion based on analytical calculations.) The fitted potential outcomes differs substantially and are considerably more constrained under the quadratic model of HI. To fit an even more flexible response model, we subclassified the data into 10 subclasses based on T, and fit a quadratic regression for Y as a function of the GPS separately within each of the subclasses. Five out of the 10 within subclass fit are plotted in the right most plot in Figure 2. The results differs substantially from HI and reveals the considerable constraint of the quadratic response model. Subclassifying on T in this way leads to a new response model and a corresponding new GPS-based estimate of the DRF; this method is discussed in Appendix B.

Given the relatively large sample size used in Simulation I ($n = 2000$), we replicate the study with a reduced sample size ($n = 500$). Although we observe an increase in standard errors for both the estimated relative DRFand its derivative, the qualitative relative performance of the methods remains the same; results appear in Appendix A.2.

[†]It would also complicate estimation of the DRF. Because the treatment assignment mechanism is strongly ignorable given the propensity function (IvD), we aim to adjust for the propensity function in a robust manner in the response model. Thus, adjusting for $\hat{\theta}$ within the subclasses poses no conceptional problem. In practice, however, $\hat{\theta}$ tends to be fairly constant within subclasses and its coefficient tends to be correlated with the intercept. A solution is to recenter $\hat{\theta}$ within each subclass. Because, we propose a more robust strategy for estimating the DRF using the PF in Section 4.1, however, we do not purse such adjustment strategies here.

3.2 Simulation study II

Although IvD's response model in Simulation I is misspecified in terms of its adjustment for X, the method may benefit from its assumption that the DRF is linear in T. We address this in the second simulation that compares the performance of the methods under several alternative generative models. We also explore the frequency properties of the methods.

Suppose we have a simple random sample of $2,000$ observations that includes a trivariate normal covariate, (X_1, X_2, X_3), with mean vector $(1, 1, 4)$, component variances all equal to one, $\text{Corr}(X_1, X_2) = 0.3$, $\text{Corr}(X_1, X_3) = -0.4$, and $\text{Corr}(X_2, X_3) = 0.6$. Suppose further that the treatment is generated according to $T \mid \boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(X_1 - X_2^2 + 0.5X_3, 1)$ and the response is generated according to one of four response models:

Gaussian Linear DRF: $Y(t) \mid t, \boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(X_1 + 2X_2 + t, 9)$,

Gaussian Quadratic DRF: $Y(t) \mid t, \boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}((X_1 + 2X_2 + t)^2, 9)$,

Lognormal DRF: $Y(t) \mid t, \boldsymbol{X} \overset{\text{ind}}{\sim} \exp\left\{\dfrac{\mathcal{N}(X_1 + 2X_2 + t, 9)}{5}\right\}$, or

Gaussian Piecewise-Quadratic DRF: $Y(t) \mid t, \boldsymbol{X} \overset{\text{ind}}{\sim} \begin{cases} \mathcal{N}((X_1 + 2X_2)^2, 9) & t \leq 0 \\ \mathcal{N}((X_1 + 2X_2 + t)^2, 9) & t > 0 \end{cases}$

Unlike the Gaussian response models the lognormal model exhibits significant heteroskedasticity; the central 95% of the conditional variances range from 0.3 to 17.8. On the other hand, the piecewise-quadratic model shows the greatest degree of nonlinearity.

To isolate the difference between the methods, we use the correctly specified treatment model in our analyses. (We use the R function `density` to estimate the marginal distribution of T for use in stabilized weights.) For the fitted response model under HI and IvD's methods, we consider Gaussian regression models that are linear and quadratic in T. In particular for the method of IvD we fit (i) $Y \sim T$ and (ii) $Y \sim T + T^2$ within each of $S = 10$ equally sized subclasses, and for the method of HI we fit (i) $Y \sim T + R + R^2 + R \cdot T$ and (ii) $Y \sim T + T^2 + R + R^2 + R \cdot T$. With IvD, the relative DRF is computed by averaging the coefficients of the within subclass models. The response models given in Table 2 are used for SCM(GPS), IPW_0, and IPW_{SW}. For the GPS-based methods and IPW, the relative DRF is evaluated at ten equally spaced values of t between -1.9 to 3.4. The entire procedure was repeated using all methods on each of 1,000 data sets generated with the same covariate and treatment models and with each of the three generative response models. All of the fitted response models are misspecified in their adjustment for X and/or T, as we expect in practice. Thus, this simulation study investigates the robustness of the methods to typical misspecification of the response model.

Figures 3 and 4 report the average of the estimated relative DRFs across the simulations (dashed lines) along with their two standard deviation intervals (shaded regions). The true relative DRF functions are plotted as solid lines.

The IvD method performs reasonably well when the generative model is linear (row 1). When IvD is fit with a quadratic model (column 3), the DRF is estimated with little bias though the estimate has higher variability. IvD exhibits significant bias only when the fitted model is linear and the true DRF is non-linear (column 1, rows 2 and 3). The HI method, on the other hand, exhibits appreciable bias even when the fitted response model matches the true model in its functional dependence on t. Like the IvD method, the bias is most acute when the fitted model is linear but the true DRF is not. Unlike with IvD, however, the 95% frequency intervals of HI miss the true value appreciably across a wide range of treatment values.

The first three columns of Figure 4 give result for SCM(GPS), IPW_0, and IPW_{SW} and show that IPW_0 and IPW_{SW} improve on HI for both the quadratic and piecewise-quadtratic DRF, although the variances are larger. For the linear DRF, SCM(GPS) is comparable to HI while IPW_0 exhibits enormous variance,

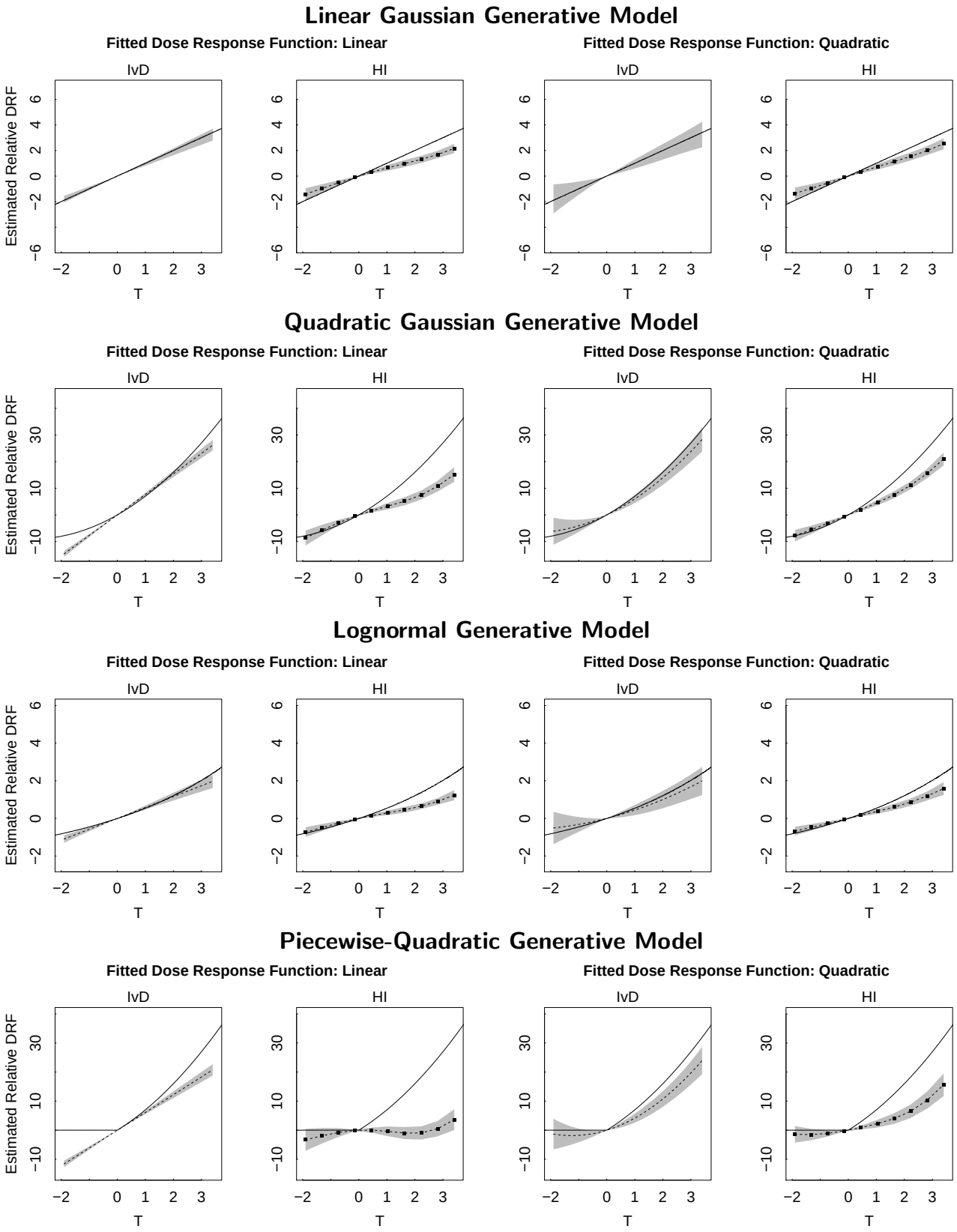

Figure 3. Estimated Relative DRFs in Simulation Study II Using the Methods of IvD and HI. The solid lines plot the true relative DRFs, the dashed lines plot the means of the fitted relative DRFs across 1000 simulations, and the gray shaded regions plot two standard deviation pointwise intervals across the 1000 fits. The evenly-spaced grid of evaluation points used with HI are also plotted as solid circles. The method of HI shows a substantial bias with all six combinations of generative and fitted response models. The method of IvD, on the other hand, exhibits significant bias only when the fitted model is linear and the generative is not.

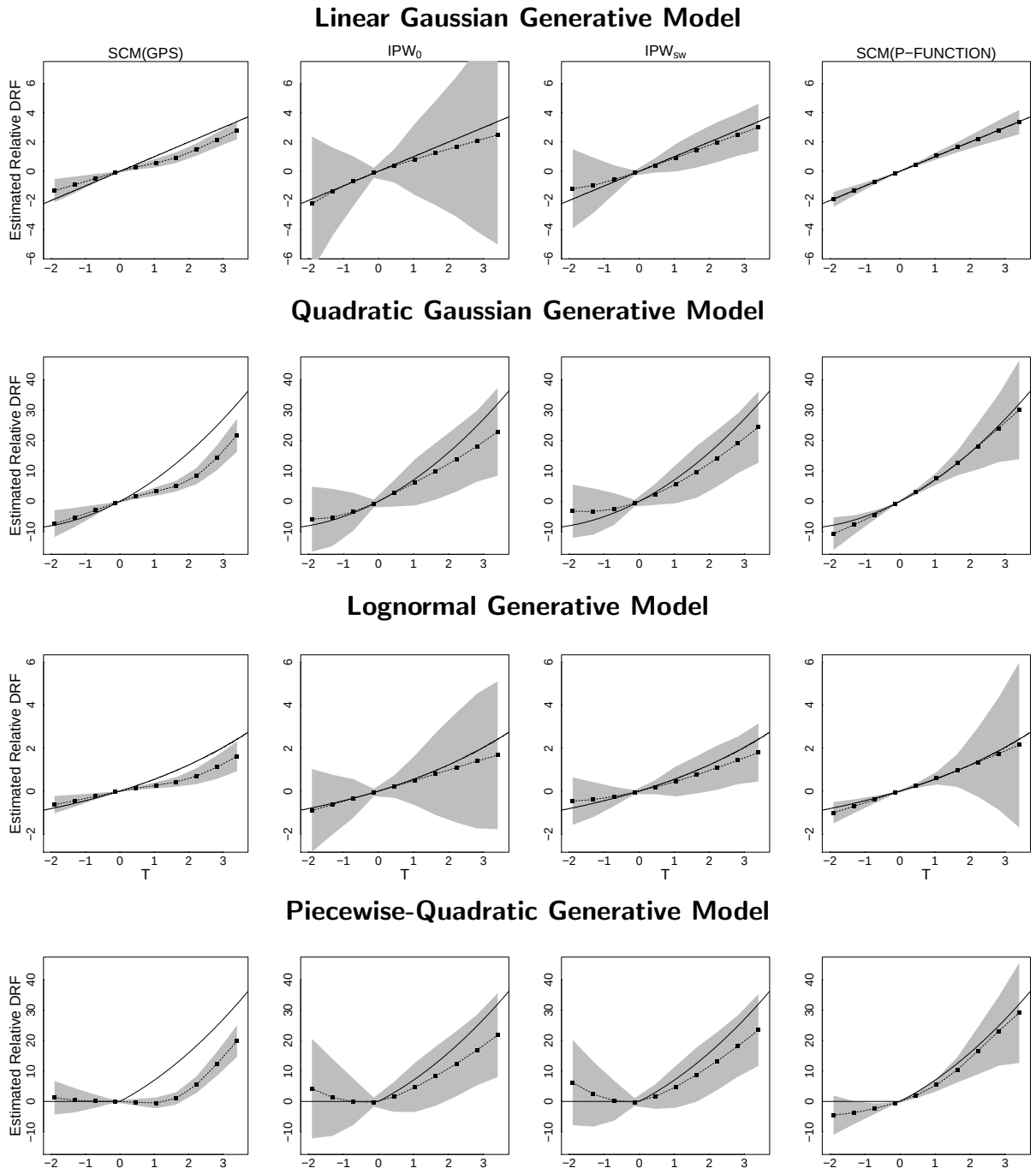

Figure 4. Estimated Relative DRFs in Simulation Study II for the SCM(GPS), IPW_0, IPW_{SW}, and SCM(PF) Methods. Solid lines, dashed lines and gray regions represent the true relative DRFs the means of the 1000 fitted relative DRFs and 95% pointwise intervals. Points represent the evenly-spaced grid points. The SCM(PF) method is discussed in Section 4.1.

and IPW_{SW} exhibits moderate variance. Except for the variance of IPW_0, the heteroscedasticity of the lognormal response model does not significantly effect the three methods.

As in Simulation I, we replicate Simulation II with a reduced sample size of $n = 500$. Although standard errors are larger with the smaller sample size, the relative performance of the methods is qualitatively similar; see Appendix A.2.

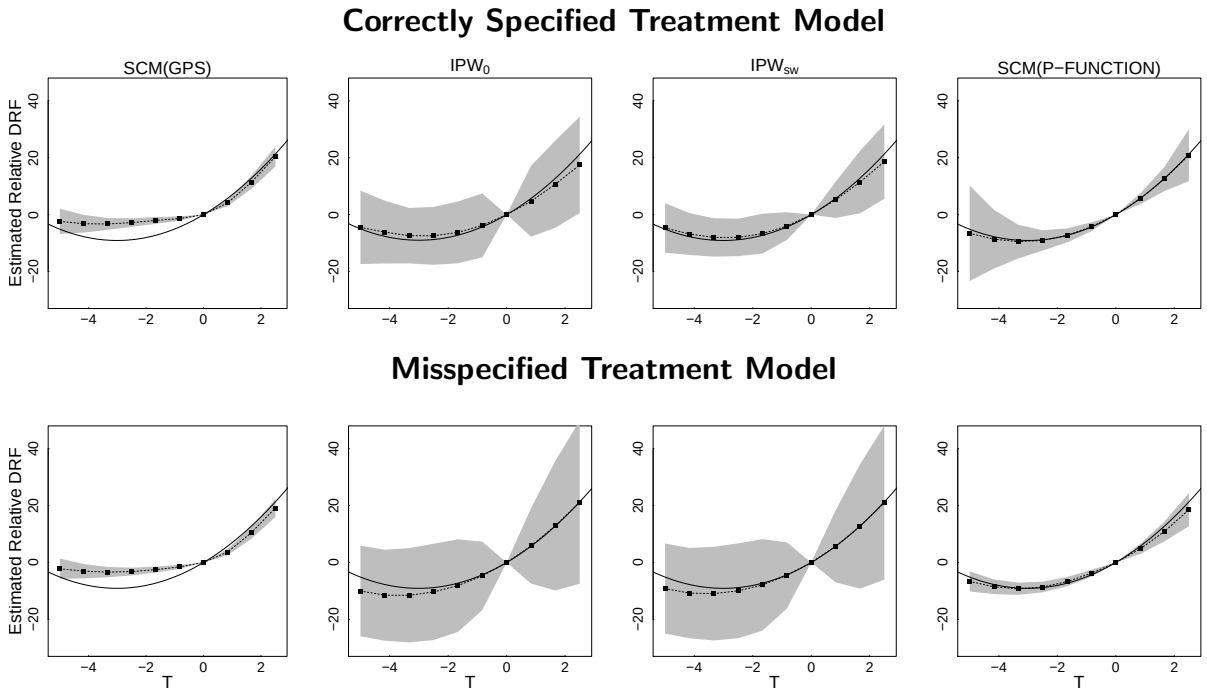

Figure 5. Estimated Relative DRFs under the Heteroscedastic Treatment of Simulation Study III. Solid lines, dashed lines and gray regions represent the true relative DRFs the means of the 1,000 fitted relative DRFs and 95% pointwise intervals.

3.3 Simulation study III

This simulation study aims to investigate the effect of a heteroscedastic treatment and the robustness of the methods to misspecification of the treatment model. The simulation setup is exactly as in Simulation study II (i.e., the same sample size, covariate distribution, and replication) but we consider a heteroscedastic treatment, $T \mid \boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(X_1 - X_2^2,\ 0.25X_3^2)$. The response is generated according to the Gaussian quadratic DRF of Simulation study II. This simulation is repeated using two fitted treatment models:

Correctly Specified Treatment Model: $T \mid \boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 X_1 + \beta_2 X_2^2,\ \sigma^2 X_3^2)$

Misspecified Treatment Model: $T \mid \boldsymbol{X} \overset{\text{ind}}{\sim} \mathcal{N}(\beta_0 + \beta_1 X_1 + \beta_2 X_2^2,\ \sigma^2)$.

(We consider misspecification of both the mean and variance of the treatment model in Section 5.2.) The parameters of both treatment models are fit via maximum likelihood and the marginal treatment model is fit using the R function `density`. The estimated DRFs fit with SCM(GPS), IPW_0[‡], and IPW_{SW} using the response models in Table 2 appear in the first three columns of Figure 5. The specification of the treatment model has little effect on SCM(GPS). For both of the IPW methods, the choice of treatment model effects the variance more than the bias of the fitted DRF.

Taking the results of Simulations I–III together, IPW_{SW} preforms better than the other methods. The GPS methods (HI, SCM(GPS), and IPW_0) are explicitly designed to estimate the DRF but seem ill-suited to the task.

[‡]Using the bandwidth of Fan and Gijbels [7] as suggested by FFGN leads to numerical instability in 708 of the 1000 datasets fit with IPW_0 under the correctly specified treatment model. The bandwidth can be tuned to improve stability; results in Figure 5 are based on a single bandwidth that gave stable results in 793 of the 1,000 datasets. Although this procedure would be difficult to implement with a single dataset, it gives the best possible representation to IPW_0. The standard bandwidth of Fan and Gijbels [7] was used with the misspecified treatment model.

3.4 Theoretical considerations and methodological implications

Both the GPS and PF generalize RR's propensity score. In the binary case, $p(T|\boldsymbol{X})$ is uniquely determined by $e(\boldsymbol{X}) = p(T = 1|\boldsymbol{X})$. IvD focus on *uniquely* determining the full conditional distribution of T given $\boldsymbol{X}$, and assume this conditional distribution is parameterized and can be uniquely represented by $\boldsymbol{\theta}$. HI, on the other hand, does not constrain the treatment assignment model in this way and instead, following the binary propensity score, evaluates $p(T|\boldsymbol{X})$ to compute propensity weights or GPS. In this way the GPS does *not* uniquely determine $p(T|\boldsymbol{X})$. There may be multiple distributions that are equal when evaluated at a particular t. The assumption of a uniquely parameterized propensity function constrains the choice of treatment model that can be used for a PF. In practice, however, the same treatment models are typically used by both methods.

Comparing IvD and HI, the assumption of IvD allows a stronger form of *strong ignorability of the treatment assignment given the propensity function*. In particular, Result 2 of IvD states

Ignorability of IvD: $p\{Y(t) \mid T, e(\cdot \mid \boldsymbol{\theta})\} = p\{Y(t) \mid e(\cdot \mid \boldsymbol{\theta})\}$ for every t,

Whereas, in their Theorem 2.1., HI show

Ignorability of HI: $p_T\{t \mid r(t, \boldsymbol{X}), Y(t)\} = p_T\{t \mid r(t, \boldsymbol{X})\}$ for every t.

In the case where T is categorical, HI's ignorability implies that $\mathbf{1}\{T = t\}$ and $Y(t)$ are independent given $r(t, \boldsymbol{X})$, where the GPS is evaluated at the particular value of t in the indicator function. Although achieving conditional independence of $Y(t)$ and T would require conditioning on a family of GPS, HI provide an insightful moment calculation to show how the response model described in Section 2.2 can be used to compute the DRF. Nonetheless conditioning on either R or $r(t, \boldsymbol{X})$, for any particular value of t does not guarantee that T is uncorrelated with the potential outcomes. The fact that the GPS does not constitute a single score for each individual restricts the response models that can be used. Subclassification, for example, is not feasible unless the classifying variable is low dimensional. The advantage of IvD over HI is that $Y(t)$ and T are conditionally independent given the low-dimensional score, $\boldsymbol{\theta}$, enabling the use of a wide-range of response models.

4 Estimating the DRF Using the P-FUNCTION

In this section, we propose a new method for robust estimation of the DRF using the PF. Appendix B discusses another new robust GPS-based method.

4.1 Using the P-FUNCTION in a SCM to estimate the DRF

IvD developed the PF to estimate the average treatment effect, rather than the full DRF. Nonetheless we use the framework of IvD to compute the DRF in Simulation studies I and II (see Figures 1 and 3). The method we employ, however, is constrained by its dependence on the parametric form of the within subclass model. Practitioners would generally prefer a robust and flexible DRF, and here we propose a procedure that allows such estimation. We view this estimate as the best available for a DRF in an observational study.

We begin by writing the DRF as

$$E[Y(t)] = E\Big[\ E[Y(t) \mid \boldsymbol{\theta}]\ \Big] = E\Big[\ E[Y(T) \mid \boldsymbol{\theta}, T = t]\ \Big] \tag{10}$$

where the first equality follows from the law of iterated expectation and the second from the strong ignorability of the treatment assignment given the PF. We estimate the DRF using the right-most expression in (10) which we flexibly model using a SCM,

$$E[Y(T) \mid \boldsymbol{\theta}, T = t] \ = \ f(\boldsymbol{\theta}, T), \tag{11}$$

where $f(\cdot)$ is a smooth function of $\boldsymbol{\theta}$ and T. In practice we replace $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}$ from the fitted treatment model. We approximate the outer expectation in (10) by averaging over the empirical distribution of $\hat{\boldsymbol{\theta}}$, to obtain an estimate of the DRF using a SCM of the PF,

$$\hat{E}[Y(t)] \;=\; \frac{1}{n}\sum_{i=1}^{n} \hat{f}(\hat{\boldsymbol{\theta}}_i, t), \tag{12}$$

where $\hat{f}(\cdot)$ is the fitted SCM. We refer to this method of estimating the DRF as the SCM(PF) method and typically evaluate (12) on a grid of values of $t_1, \ldots, t_D$ evenly spaced in range of the observed treatments. Bootstrap standard errors are computed on the same grid.

Comparing (5)–(6) with (11)–(12), SCM(GPS) and SCM(PF) are algorithmically very similar. The primary difference is the choice between the PF and GPS in the response model. As we shall see, this change has a significant effect on the statistical properties of the estimates. Simply put, $\boldsymbol{\theta}$ is a much better behaved predictor variable than is R. When using Gaussian linear regression for the treatment model, for example, $\boldsymbol{\theta} = \boldsymbol{X}_i^\top \beta$, whereas R is the Gaussian density evaluated at T. As illustrated in Section A.3, the dependence of the GPS on t and the non-monotonicity of this dependence both complicate the response model and pose challenges to robust estimation.

Computing $\hat{E}[Y(t_0)]$ with (12) for some particular t_0 involves evaluating $\hat{f}(\cdot, t_0)$ at every observed value of $\hat{\theta}_i$. Invariably, the range of $\hat{\theta}$ observed among units with T near t_0 is smaller than the total range of $\hat{\theta}$, at least for some values of t_0. Thus, evaluating (12) involves some degree of extrapolation, at least for some values of t. Luckily, this problem is relatively easy to diagnose with a scatter plot of the observed values of $(T_i, \hat{\theta}_i)$. The estimate in (12) may be biased for values of the treatment where the range of observed $\hat{\theta}_i$ is relatively small. As we illustrate in our simulation studies, however, (12) appears quite robust and this bias is small relative to the biases of other available methods.

4.2 Simulation studies I–III revisited

We now revisit the simulation studies from Section 3, which illustrate the potentially misleading results and/or high variance of existing estimates of the DRF. Here, we compare these results with those of SCM(PF). In all cases, SCM(PF) was fitted using the same (correctly specified or misspecified) treatment assignment model and with the same equally-spaced grid points. When fitting the SCM, we continue to use the penalized cubic regression spline basis for both parameters (R and T) and a tensor product to construct a smooth fit of the continuous function $f(\theta, T)$ (see `mgcv` R-package documentation). Figure 6 shows the fitted (relative) DRF for SCM(PF) in Simulation I. The performance of SCM(PF) is a dramatic improvement over all other methods in Simulation I; compare Figures 1 and 6.

The rightmost column of Figure 4 presents the results of the SCM(PF) method in Simulation study II. Comparing Figures 3 and 4 again illustrates the advantages of the proposed method. The fits in Figure 3 are quite dependent on the parametric choice of the response model, whereas the non-parametric fits illustrated in Figure 4 do not require a parametric form. Among the non-parametric methods, the advantage of SCM(PF) over SCM(GPS) and IPW_0 is clear. It essentially eliminates bias with only a small increase in variance. Only IPW_{SW} has comparable statistical properties in this simulation.

In Simulation III, the correctly specified heteroscedastic treatment model is parameterized by a bivariate function; $\boldsymbol{\theta}$ has two components which are equal to the mean and log-variance of the fitted treatment model. Both components of $\boldsymbol{\theta}$ are used as predictor variables in the SCM used to model the response. The rightmost column of Figure 5 shows that SCM(PF) performs very well in Simulation III, especially with the simpler misspecified treatment model.

Overall, SCM(PF) consistently provides better estimates than HI, SCM(GPS), and IPW_0, both in terms of bias and variance. IPW_{SW} also preforms better than these method and is comparable to SCM(PF) in Simulation II (and Simulation IV, see Appendix A.3). In the bulk of our numerical studies (Simulations I and III, as well as in Simulation V in Appendix A.4 and the applied example in Section 5), however, SCM(PF) outperforms even IPW_{SW}.

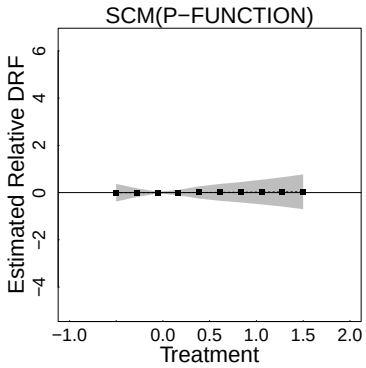

Figure 6. Estimated Relative DRF Using SCM(PF) in Simulation Study I. The solid (dashed) lines represent the true (fitted) relative DRF, the 95% confidence bands are plotted in grey, and the grid points are identical to those in Figure 1. The fitted relative DRF is much improved compared with those of HI, SCM(GPS), IPW0, and IPW_{SW} but without the linear assumptions of IvD (see Figure 1).

5 Example: The effect of smoking on medical expenditures

5.1 Background

We now illustrate the available methods by estimating the DRF of smoking on annual medical expenditures. The data we use were extracted from the 1987 National Medical Expenditure Survey (NMES) by Johnson et al. [17]. Its detailed information about frequency and duration of smoking allows us to continuously distinguish among smokers and estimate the effects of smoking as a function of how much they smoke. The response variable, medical costs, is verified by multiple interviews and additional data from clinicians and hospitals. IvD used the propensity function to estimate the average effect of smoking on medical expenditures. We extend their analysis and study estimation of the full DRF. Like IvD, we adjust for the following subject-level covariates: age at the times of the survey, age when the individual started smoking, gender, race (white, black, other), marriage status (married, widowed, divorced, separated, never married), educational level (college graduate, some college, high school graduate, other), census region (Northeast, Midwest, South, West), poverty status (poor, near poor, low income, middle income, high income), and seat belt usage (rarely, sometimes, always/almost always).

To measure the cumulative exposure to smoking based on the self-reported smoking frequency and duration, Johnson et al. [17] proposed the variable of `packyear`, defined as

$$\texttt{packyear} = \frac{\text{number of cigarettes per day}}{20} \times \text{number of years smoked.} \tag{13}$$

We use log(`packyear`) as our treatment variable. We follow Johnson et al. [17] and IvD and discard all individuals with missing values and conduct a complete-case analysis, yielding a sample of 9,073 smokers. Although in general complete-case propensity-score-based analyses produce biased causal inference unless the data are missing completely at random [4], Johnson et al. [17] showed that accounting for the missing data using multiple imputation did not significantly affect their results.

Because the observed response variable, self-reported medical expenditure, denoted Y, is semicontinuous, we use the two-part model of Duan et al. [5]. This involves first modeling the probability of spending some money on medical care, $\Pr(Y > 0 \mid T, \boldsymbol{X})$, where $T = \log(\texttt{packyear})$, and $\boldsymbol{X}$ represents the covariates; and then modeling the conditional distribution of Y given T and $\boldsymbol{X}$ for those who reported positive medical expenditure. To illustrate and compare methods for computing the DRF, we concentrate on the second part of this model. Because the distribution of Y is skewed, we consider the model $p(\log(Y) \mid Y > 0, T, \boldsymbol{X})$.

For our treatment assignment model, we use a Gaussian linear regression adjusted for all available covariates and the second order terms of two age covariates. The model was fitted using sampling weights

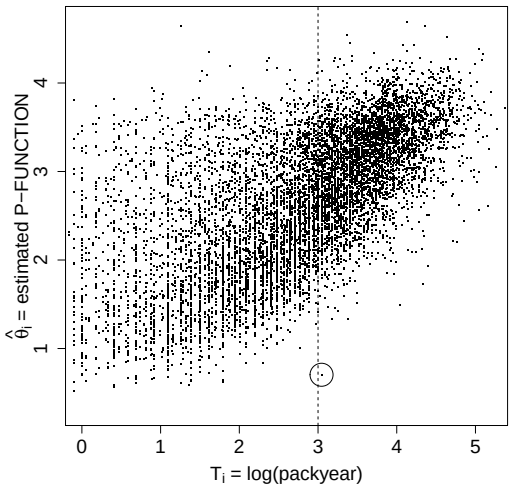

Figure 7. A Diagnostic for SCM(PF). Because the range of the $\hat{\theta}_i$ when $T_i > 3$ is less than the overall range of $\hat{\theta}_i$ estimating the DRF for $t > 3$ involves extrapolation under the SCM and thus possible bias. The single individual with T_i slightly larger than three and $\hat{\theta}_i$ less than one is circled. Although this datapoint may mitigate bias for t near three, the fitted DRF for $t > 3$ may be seriously biased.

provided with the original data. This is the same treatment assignment model used by IvD who demonstrate that it achieves adequate balance.

5.2 Simulation study based on the smoking data

This simulation study aims to mimic the characteristics of the actual data with the goal of comparing the statistical properties of the proposed methods in as realistic a setting as possible. In particular, we do not alter the observed covariates or treatment and use the same fitted treatment model used by IvD. Figure 7 presents a scatter plot of the observed treatment variable, $T_i = \log(\texttt{packyear})$, and the values of the PF from the fitted treatment assignment model, $\hat{\theta}_i$. As discussed in Section 4.1, this plot can be used as a diagnostic for SCM(PF). Recall that this estimate requires that we fit a SCM to predict the response variable as a function of T_i and $\hat{\theta}_i$. To estimate the DRF at t we must evaluate the fitted SCM at $(t, \hat{\theta}_i)$ using each observed $\hat{\theta}_i$ in the data set. This involves extrapolation and thus possible bias if the range of θ_i at a particular value of t is less than the overall range of θ_i. Judging from Figure 7, this is a concern for t greater than about three. There is a solitary individual with T_i slightly above three and $\hat{\theta}_1 < 1$ that is circled in Figure 7. Even this single point can guard against significant extrapolation bias for t less than three, but the concern remains for larger t. We emphasize that this diagnostic is performed before the response model is fit.

To explore the robustness of the methods to different DRFs, we simulate the response variable under three known DRFs and attempt to reconstruct them using HI, SCM(GPS), IPW_0, IPW_{SW}, and SCM(PF). In particular, we assume $\log(Y_i(t)) \sim \mathcal{N}(E[\log(Y_i(t))], 0.5^2)$ with $t = \log(\texttt{packyear})$ and consider three functional forms for $E[\log(Y_i(t))]$:

$$\text{Quadratic DRF:} \quad E[\log(Y_i(t))] = \frac{4}{25} \cdot t^2 + [\log(\texttt{age}_i)]^2$$

$$\text{Piecewise-Linear DRF:} \quad E[\log(Y_i(t))] = \begin{cases} -4 - 0.5 \cdot t + [\log(\texttt{age}_i)]^2, & t \leq 2 \\ -5 - 2.3 \cdot (t-2) + [\log(\texttt{age}_i)]^2, & t > 2, \end{cases}$$

$$\text{Hockey-Stick DRF:} \quad E[\log(Y_i(t))] = \begin{cases} -8.1 + [\log(\texttt{age}_i)]^2, & t \leq 3 \\ -8.1 + 1.5 \cdot (t-3)^2 + [\log(\texttt{age}_i)]^2, & t > 3, \end{cases}$$

where `age` is the age at the time of the survey. We include `age` because it is the covariate most correlated with log(`packyear`) and thus most able to bias a naive analysis. Each of the response models was fitted using the sampling weights.[§]

Each of the five methods was fitted to one data set generated under each of the three DRFs. We evaluate each DRF at ten points equally spaced between the 5% and 95% quantiles of log(`packyear`). The results appear in Figure 8 in which rows correspond to the three generative models and columns represent the method used to fit the DRF. In all plots, the true DRF is plotted as a solid line and a directly fitted SCM of $\log(Y)$ on T as a dashed line. This SCM fit is a simple bench mark; it does not account for covariates in any way, in particular it does not adjust for any summary of the treatment assignment model. The fitted DRFs are plotted with dotted lines and bullets indicate the grid where the estimates are evaluated. The shaded regions represent 95% point-wise bootstrap confidence intervals. The diagnostic described in Figure 7 indicates possible bias in the SCM(PF) method for $t > 3$. Thus, we plot the fit in this region in light grey to emphasize its potential unreliability.

The HI fit misses the true DRF under all three generative models, even the quadratic DRF which coincides with the parametric dependence of $\log(Y)$ on T under HI's response model. Instead, HI's fitted DRF tends to follow the unadjusted SCM fit of $\log(Y)$ on T. Although SCM(GPS) improve somewhat on HI, it still exhibits a cyclic pattern; notice its cubic-like fits in the first and third rows. Unfortunately, IPW_0 again exhibits instability, although in this case it takes the form of bias rather than variance. Although IPW_{SW} performs significantly better than IPW_0 in our simulation studies, here the methods are essentially indistinguishable. Finally, SCM(PF) closely matches the true DRF under all three generative models, at least for $t < 3$. As discussed above, we suspect bias for $t > 3$ and see that the fitted DRF reverts to the unadjusted SCM in this range. The quality of the fit can be improved still further by increasing the dimension of the basis used in the SCM. We do not purse this strategy, however, for fear of over fitting. Overall, SCM(PF) appears to be the most reliable, especially considering the diagnostic that alerts us the ranges of t where there is the potential for bias.

6 Concluding Remarks

Propensity score methods have gained wide popularity among applied researchers in a number of disciplines. Although they were originally designed exclusively for binary treatment regimes, the fact that treatment variables of interest are not binary in many research settings has led to proposals for generalized propensity score methods. These methods are applicable to a variety of non-binary treatment regimes, and their applications are becoming increasingly common.

In this article, we compare two frequently used generalized propensity score methods, the GPS of Hirano and Imbens [11] and the PF of Imai and van Dyk [14], as well as the GPS extensions of Flores et al. [8]. First, we show that the suggested implementation of the HI method is sensitive to misspecification of the response model. Second, we show that while SCM(GPS) exhibits substantial improvement over HI's method, it remains biased and/or can exhibit a cyclic artifact in some situations. Third, we demonstrate that while Flores *et al.*'s IPW_0 can be highly unstable, using RHB's stabilized weights can significantly improve its performance. Finally, while IvD provides a relatively robust estimate of the average causal effect, its main limitation is its inability to estimate the DRF. We show how to obtain an estimate of the DRF based on the PF and empirically compare its performance to that of the other estimates. We also give an explanation as to why the SCM(PF) method outperforms the SCM(GPS) method. While SCM(PF) preforms well in comparison to other methods, it remains biased in realistic settings. *We emphasize that researchers should be cautious when using any method to estimate the full* DRF *with a continuous treatment in an observational study.*

[§]When using IPW_0 or IPW_{SW}, we construct new weights by multipling the weights required by IPW and the sampling weights. We also take the sampling weights into account when estimating the marginal distribution of the treatment with IPW_{SW} (using the `density` function in R). Ignoring the sampling weights leads to similar results.

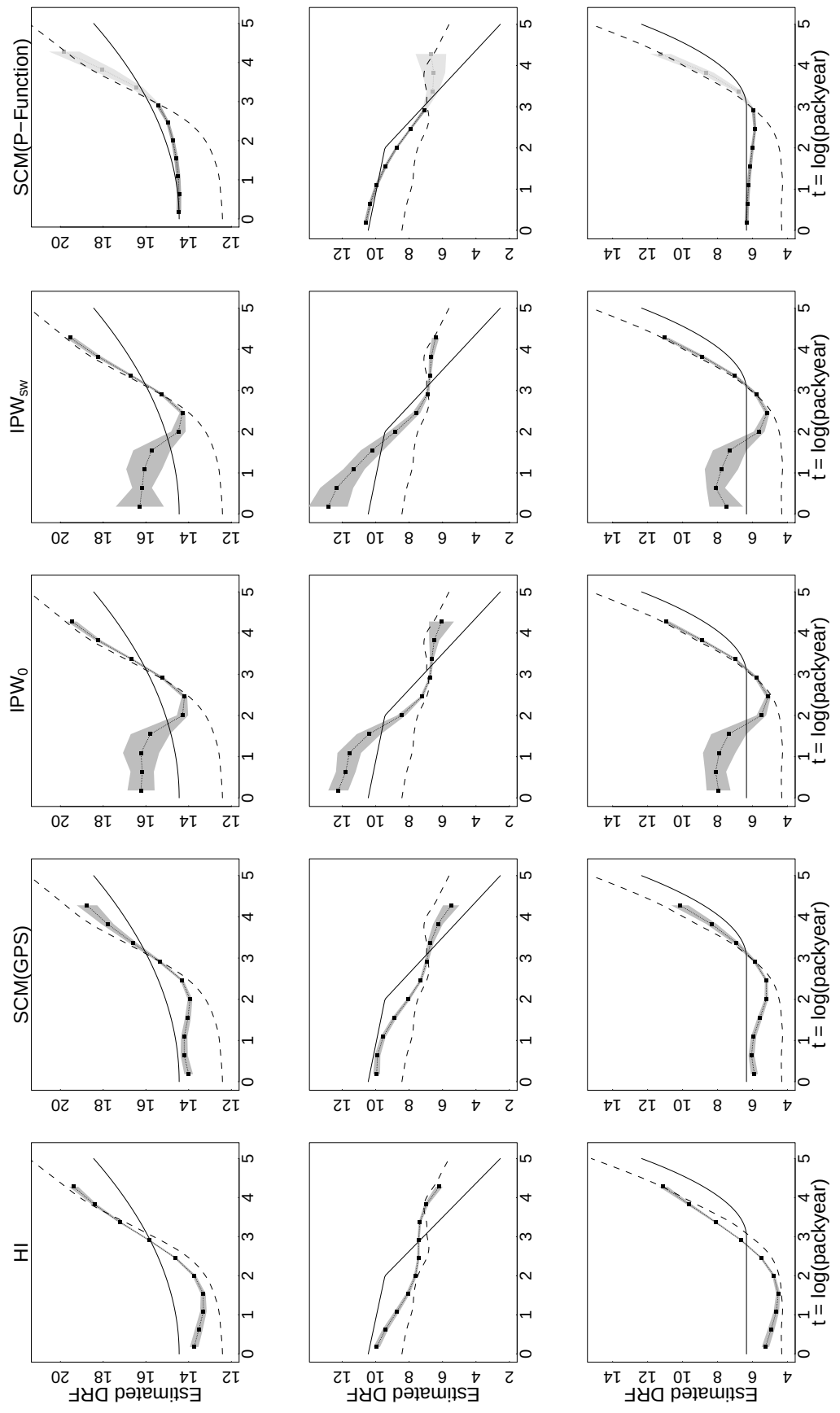

Figure 8. Estimated DRFs for the Simulation Based on Smoking Data. The five columns correspond to the method of HI, $\mathrm{SCM(GPS)}$, $\mathrm{IPW_{SW}}$, $\mathrm{IPW_0}$ and $\mathrm{SCM(PF)}$ respectively. In all plots the dotted lines with bullets correspond to the fitted DRF. The true DRF is plotted as solid lines while dashed lines represent the fitted SCM of $\log(Y)$ on T, unadjusted for the covariates. The evaluation points are evenly-spaced in t. The 95% asymptotic confidence bands plotted in grey are based on 1000 bootstrap replications. A lighter shade of grey is used in the right-most column for $t > 3$ because the estimate is less reliable in this region. The performance of the $\mathrm{SCM(PF)}$ clearly dominates the other methods, especially for $t < 3$.

There are several important challenges that must still be addressed. We have largely assumed that the propensity weights, GPS, and PF can be correctly estimated. This is an optimistic assumption given that modeling a multi-valued or continuous treatment in a high-dimensional covariate space is much more difficult than doing so for a binary treatment. Diagnostic tools developed for the binary treatment case are also not directly applicable to general treatment regimes. Even more challenging is diagnosing misspecification in the response model. As we have illustrated, this can lead to significant bias in the estimated DRF. Our proposals rely on implementing more flexible response models in more natural spaces, but principled diagnostics for the response model remain elusive. Diagnosing and correcting for imbalance in either the PF or the GPS is another difficulty. Since the subpopulation that has propensity for treatment varies with the dose, the estimated dose response function is in effect the treatment effect on a varying subpopulation. Recently, researchers have started to develop new methods for estimating the propensity weights, GPS, and PF in the presence of possible misspecification of the treatment assignment model [e.g., 9, 25]. Future research should also address the estimation of the DRF in the presence of possible misspecification of the response model, as well as diagnostics tools.

References

[1] Bia M, Flores AC and Mattei A (2011) Nonparametric estimators of dose-response functions. *CEPS/INSTEAD* .

[2] Bia M, Flores C, Flores-Lagunes A and Mattei A (2014) A Stata package for the application of semiparametric estimators of dose-response functions. *Stata Journal* 3: 580–604.

[3] Cochran WG (1968) The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 24: 295–313.

[4] D'Agostino RB Jr and Rubin DB (2000) Estimating and using propensity scores with partially missing data. *Journal of the American Statistical Association* 95(451): 749–759.

[5] Duan N, Manning WGJ, Morris CN and Newhouse JP (1983) A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics* 1: 115–126.

[6] Ertefaie A and Stephens DA (2010) Comparing approaches to causal inference for longitudinal data: Inverse probability weighting versus propensity scores. *The International Journal of Biostatistics* 6: 1–24.

[7] Fan J and Gijbels I (1996) *Local Polynomial Modelling and its Applications*. Chapman and Hall, London.

[8] Flores CA, Flores-Lagunes A, Gonzalez A and Neumann TC (2012) Estimating the effects of length of exposure to instruction in a training program: The case of job corps. *The Review of Economics and Statistics* 94(1): 153–171.

[9] Fong C, Hazlett C and Imai K (2018) Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *Annals of Applied Statistics* 12(1): 156–177.

[10] Hernán MA, Brumback B and Robins JM (2002) Estimating the causal effect of zidovudine on cd4 count with a marginal structural model for repeated measures. *Statistics in Medicine* 21: 1689–1709.

[11] Hirano K and Imbens GW (2004) The propensity score with continuous treatments. In: *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, chapter 7. Wiley, pp. 73–84.

[12] Hogan JW and Lancaster T (2004) Instrumental variables and inverse probability weighting for causal inference from longitudinal observational studies. *Statistical Methods in Medical Research* 13: 17–48.

[13] Imai K, King G and Stuart EA (2008) Misunderstandings among experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* 171(2): 481–502.

[14] Imai K and van Dyk DA (2004) Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* 99(467): 854–866.

[15] Imbens GW (2000) The role of the propensity score in estimating dose-response functions. *Biometrika* 87(3): 706–710.

[16] Joffe MM and Rosenbaum PR (1999) Propensity scores. *American Journal of Epidemiology* 150: 327–333.

[17] Johnson E, Dominici F, Griswold M and Zeger SL (2003) Disease cases and their medical costs attributable to smoking: An analysis of the National Medical Expenditure Survey. *Journal of Econometrics* 112: 135–151.

[18] Kang JD and Schafer JL (2007) Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussions). *Statistical Science* 22(4): 523–539.

[19] Lu B, Zanutto E, Hornik R and Rosenbaum PR (2001) Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association* 96(456): 1245–1253.

[20] Moodie EE and Stephens DA (2012) Estimation of dose-response functions for longitudinal data using the generalised propensity score. *Statistical Methods in Medical Research* 21(2): 149–166.

[21] Robins JM, Hernán MA and Brumback B (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology* 11(5): 550–560.

[22] Rosenbaum PR (1987) Model-based direct adjustment. *Journal of the American Statistical Association* 82: 387–394.

[23] Rosenbaum PR and Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1): 41–55.

[24] Rubin DB (1990) Comments on "On the application of probability theory to agricultural experiments. Essay on principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed. *Statistical Science* 5: 472–480.

[25] Zhu Y, Coffman D and Ghosh D (2015) A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of Causal Inference* 3(1): 25–40.

Online Appendix

A Additional Simulation Studies

A.1 Simulation study I: Analytical calculations

Based on a suggestion by an anonymous reviewer, this appendix presents an analytical calculation for Simulation I introduced in Section 3.1. Under the true data generating process, the conditional expectation of the response can be written as a mixture,

$$\mathrm{E}(Y(t) \mid R = r, T = t) \;=\; 10w(r,t)u_1(r,t) + 10(1 - w(r,t))u_2(r,t) \tag{14}$$

for $-\infty < t < \infty$, $0 < r < m = (2\pi\sigma_T^2)^{-1/2}$ where

$$u_1(r,t) \;=\; t - \sqrt{2\sigma_T^2 \log(m/r)}, \quad \text{and} \quad u_2(r,t) \;=\; t + \sqrt{2\sigma_T^2 \log(m/r)} \tag{15}$$

and

$$w(r,t) \;=\; \frac{\phi((u_1(r,t) - \mu_X)/\sigma_X)}{\phi((u_1(r,t) - \mu_X)/\sigma_X) + \phi((u_2(r,t) - \mu_X)/\sigma_X)} \tag{16}$$

with $\phi(\cdot)$ is the standard normal density function. The dose-response function is obtained by integrating equation (14) over the distribution $R_t = r(t, X)$ with $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$. A Monte Carlo analysis reveals that the resulting DRF is a complex function of t, which is difficult to approximate by a quadratic function.

A.2 Simulation study I and II: Smaller sample size

Based on a suggestion of another anonymous reviewer, we replicate Simulations I and II with a reduced sample size of $n = 500$ (instead of $n = 2,000$ as reported in Sections 3.1 and 3.2). Updated versions of Figures 1, 3, and 4 can be found in Figures 9, 10, and 11, respectively. In all cases the standard errors grow as expected, but the relative performance of the algorithms remains qualitatively similar.

A.3 Simulation study IV: The potential cyclic bias of SCM(GPS)

The DRF fitted with SCM(GPS) in Simulation I exhibits a cyclic artifact that does not exist in the underlying DRF; Appendix A.4 provides another simulation study in which this effect is even more pronounced. Here we present a simulation study that investigates the origin of this cyclic bias. In particular, we independently generate $Z_i \sim \text{Bernoulli}(0.5)$, $X_i \sim \mathcal{N}(Z_i, 0.01)$, $T_i \sim \mathcal{N}(X_i, 1)$, and $Y_i \sim \mathcal{N}(4Z_i, 1)$, for $i = 1, \ldots, 2000$. Using the correct treatment model, we estimate the DRF using SCM(GPS) at ten evenly spaced theoretical percentiles of T. We repeat the entire fitting procedure on each of 1,000 replicated data sets and plot the average of the estimated DRF and their pointwise two standard deviation intervals in Figure 12. For comparison, we also present results for IPW_{SW} and SCM(PF). The cyclic bias of the SCM(GPS) fit is evident, whereas both IPW_{SW} and SCM(PF) yield reasonable and similar fits.

To find the source of the cyclic bias, we plot the fitted response model, the SCM given in (5), as a heat map along with a scatter plot of the observed $(T_i, \hat{R}_i)$ in the leftmost panel of Figure 13. The two bell-shaped curves that appear in the plotted values of $(T_i, \hat{R}_i)$ stem from the definition of the GPS; $\hat{R}_i$ is the value of the fitted density function of T. By design, X clusters around the two values of Z; these two clusters correspond to the two bell-shaped curves.

The overlapping bell-shaped curves in the observed $(T_i, \hat{R}_i)$ induce a cyclic pattern in the fitted SCM-response model. To estimate the DRF at t, the fitted response model is evaluated at and averaged over each $\hat{r}(t, \boldsymbol{X}_i)$. This is illustrated in the second panel of Figure 13 which plots $(t, \hat{r}(t, \boldsymbol{X}_i))$ with $t = 0$ on top of the fitted SCM. The cluster of points at the top of the panel land in a local minimum resulting in the dip of the fitted DRF in Figure 12. The third and fourth panels show that as t increases to 0.5 and 1.0,

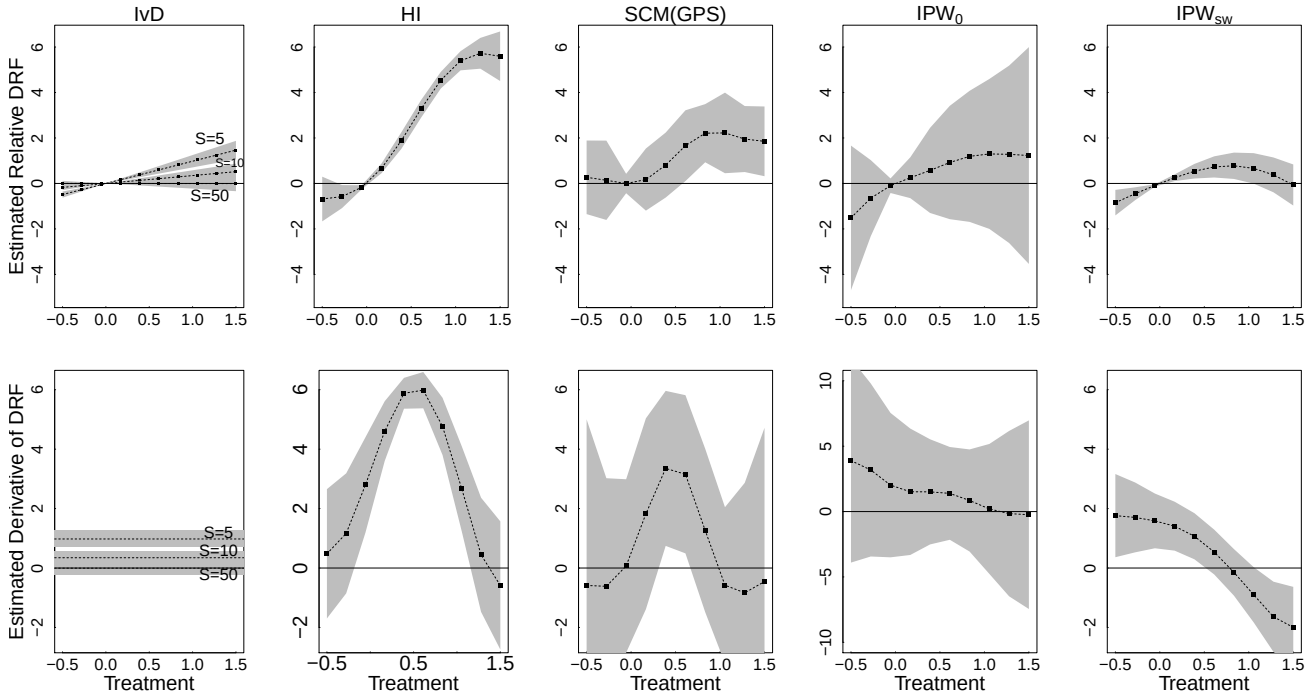

Figure 9. Simulation I Replicated with a Reduced Sample Size of $n = 500$. Results are qualitatively similar to those presented in Figure 1, but with larger standard errors.

the values of $(t, \hat{r}(t, \boldsymbol{X}_i))$ continue to cluster, but the clusters shift from minima to maxima of the fitted SCM, leading to the cyclic pattern in the fitted DRF.

The patterned behavior of the GPS (illustrated in the first panel of Figure 12) means that the response model is particularly difficult to accurately represent, even with a flexible non-parametric model, and that extrapolation is especially likely. Unfortunately, this is inevitable: when estimating the DRF we must evaluate the fitted response model at each value of $\hat{r}(t, \boldsymbol{X}_i)$, including at unobserved combinations of t and $\hat{r}(t, \boldsymbol{X}_i)$, see (5) and (6). This is a difficulty with the underlying response model, regardless of the choice of fitted response model. Although Simulation IV uses a simple setting to clearly explain the cyclic bias of SCM(GPS), the bias persist in more complex settings (see Figures 1, 4, 8, and 14).

A.4 Simulation study V: Further illustration of cyclic bias of SCM(GPS)

Although the SCM(GPS) estimate of the DRF shows some bias in Simulation studies II and III, the cyclic bias that it exhibits in Simulation I is much less pronounced in Figures 4 and 5. To see if the cyclic bias exists in more complex settings, we extend the well-known simulation study of Kang and Schafer [18] to a continuous treatment. In particular, we independently simulate $Z_{ij} \sim \mathcal{N}(0, 1)$ for $i = 1, \ldots, 2000$ and $j = 1, \ldots, 4$ and generate

$$T_i = -Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.1Z_{i4} + \sigma_i,$$

where $\sigma_i \sim \mathcal{N}(0, 1)$, and

$$Y_i = 210 + 27.4Z_{i1} + 13.7Z_{i2} + 13.7Z_{i3} + 13.7Z_{i4} + 10T_i + \epsilon_i$$

where $\epsilon_i \sim N(0, 1)$. We estimate the relative DRF by applying the methods of HI, SCM(GPS), IPW_0, IPW_{SW}, and SCM(PF) to each of 1000 replicated data sets. We use the correctly specified treatment model and the fitted response models given in Table 2. Figure 14 shows that the cyclic bias remains a problem for SCM(GPS) and that large variances continue to plague IPW_0. The stabilized weights of IPW_{SW} clearly improve its performance, relative to IPW_0. Nonetheless, SCM(PF) dominated the other methods.

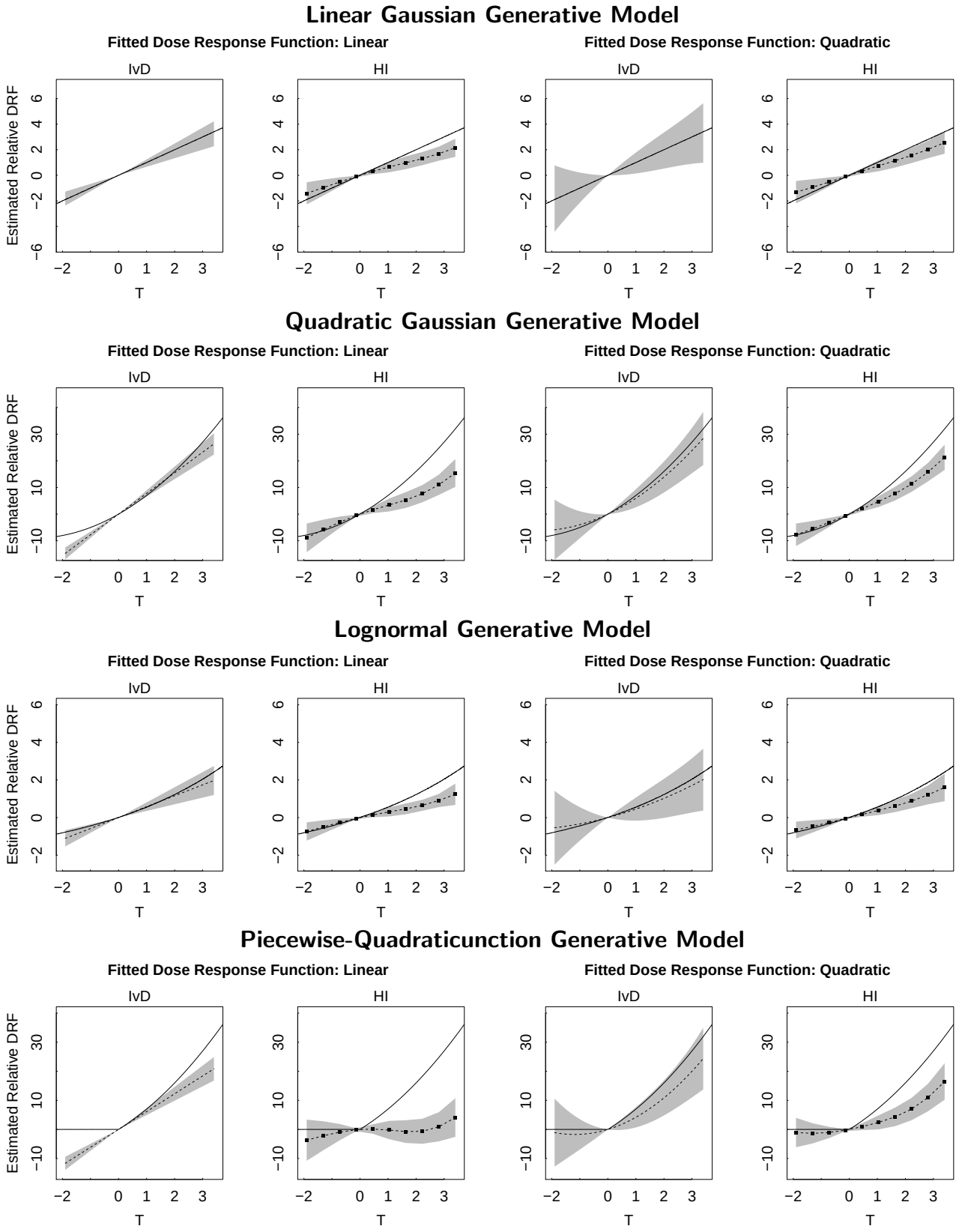

Figure 10. Estimated Relative DRFs Using the Methods of IvD and HI in Simulation Study II, but with a Reduced Sample Size of $n = 500$. Results are qualitatively similar to those presented in Figure 3, but with larger standard errors.

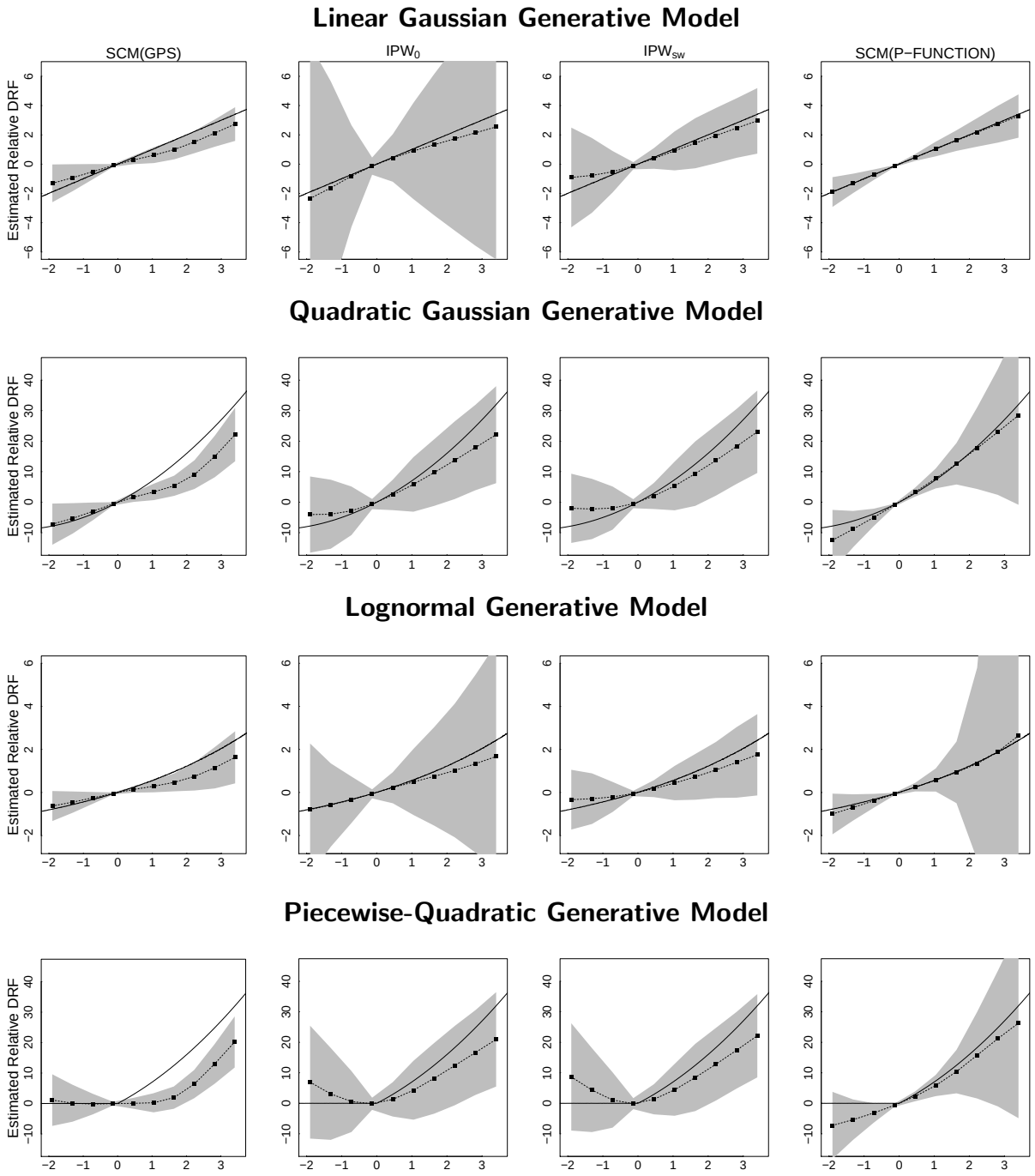

Figure 11. Estimated Relative DRFs for $\mathrm{SCM(GPS)}$, IPW_0, $\mathrm{IPW}_{\mathrm{SW}}$, and $\mathrm{SCM(PF)}$ Methods in Simulation Study II, but with a Reduced Sample Size of $n = 500$. Results are qualitatively similar to those presented in Figure 4, but with larger standard errors.

B Covariance Adjustment GPS

B.1 Covariance adjustment for categorical treatments

One of the response models suggested by RR for a binary treatment in an observational study involves *covariance adjustment*. With this method, the response variable is regressed on the fitted propensity score separately for the treatment and control groups. Suppose we use the GPS in place of the propensity score in the context of a binary treatment. Specifically, for units in the treatment group, we use the ordinary propensity score, $R_i = r(1, \mathbf{X}_i) = p_\psi(T = 1 \mid \mathbf{X}_i)$, but for units assigned to the control group, we use the

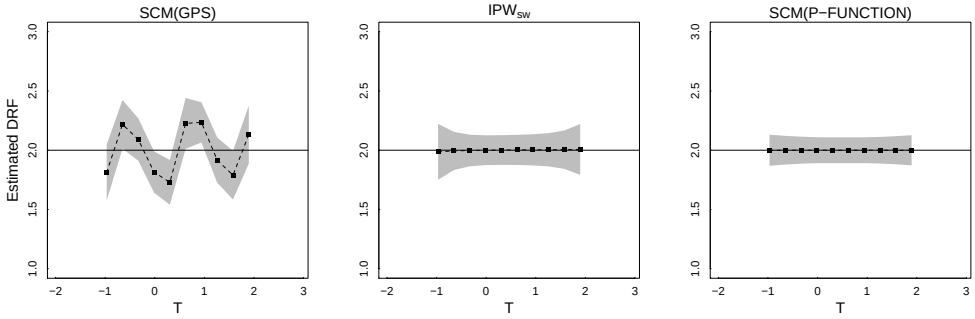

Figure 12. Estimated DRF for Simulation IV. Solid lines, dashed lines and gray regions represent the true DRFs the means of the 1,000 fitted DRFs and 95% pointwise intervals. Only SCM(GPS) exhibits the cyclic bias. The SCM(PF) is introduced in Section 4.

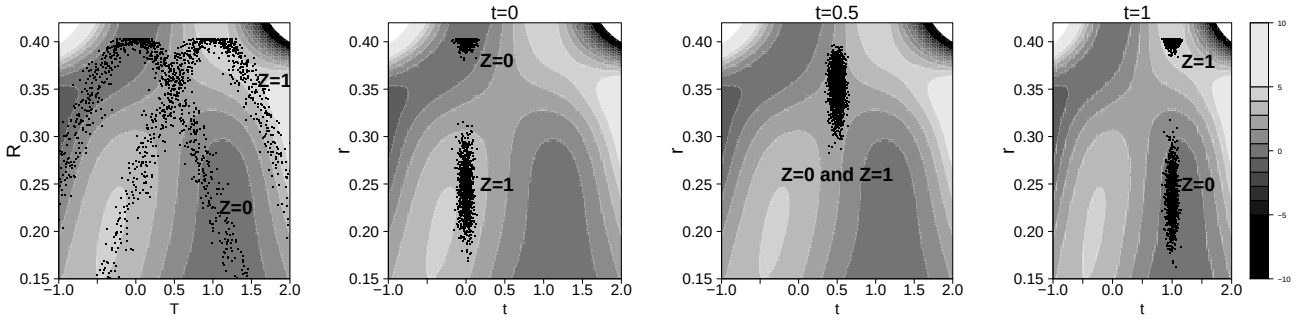

Figure 13. How the SCM(GPS) fit can lead to a cyclic artifact in the the fitted DRF. The leftmost panel overlays a scatter plot of T and the GPS, $\hat{R}$, on a heat map of the fitted SCM(GPS) response model in Simulation IV. The other three panels overlay scatterplots of $(t, \hat{r}(t, \boldsymbol{X}_i))$, with t equal to 0, 0.5, and 1. (We jitter in the T direction to improve visualization.) The panels show that as t increases the $(t, \hat{r}(t, \boldsymbol{X}_i))$ clusters move from local minima to local maxima and back, resulting in a cyclic pattern in the fitted DRF.

probability of control rather than the probability of treatment, $R_i = r(0, \mathbf{X}_i) = p_\psi(T = 0 \mid \mathbf{X}_i)$. Because the GPS is equal to the propensity score for treatment units and is equal to one minus the propensity score for control units [15], it is easy to see that the usual covariance adjustment is equivalent to fitting the following regression model,

$$Y_i \ \sim \ \alpha_t + \beta_t \hat{R}_i, \tag{17}$$

separately for the treatment and control units, i.e., $t = 0$ and 1. The linear transformation of the predictor variable does not effect the predicted value of the response for the control group.

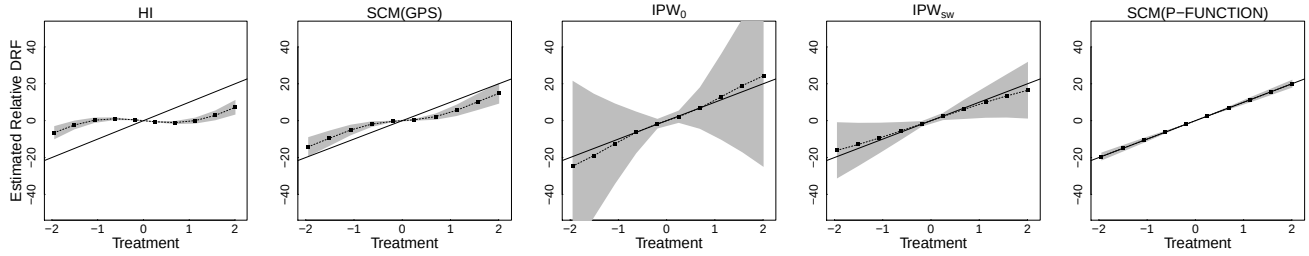

Figure 14. Estimated Relative DRF in Simulation V. Solid lines, dashed lines and gray regions represent the true relative DRFs the means of the 1000 fitted relative DRFs and 95% pointwise intervals. The evaluation points are identical for all plots. SCM(GPS) exhibits a cyclic artifact and IPW_0 is quite unstable. The SCM(PF) method proposed in Section 4.1 again outperforms the other methods.

After fitting the model given in equation (17), the average of the two potential outcomes can be estimated by averaging the fitted values over all units in the sample. That is, we compute

$$\hat{E}\{Y(t)\} \;=\; \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\alpha}_t + \hat{\beta}_t\; \hat{r}(t, \mathbf{X}_i)\right\}, \tag{18}$$

for $t = 0, 1$. The estimated average causal effect is simply the difference $\hat{E}\{Y(1)\} - \hat{E}\{Y(0)\}$, which is equivalent to the estimate reported in equation (1). Thus, with a binary treatment, the method of HI is equivalent to RR's covariate adjustment, except that HI propose a quadratic rather than a linear response model.

Suppose now that the treatment variable is categorical with more than two levels. In principle, exactly the same procedure can be applied. Namely, the regression model given in equation (17) can be fitted separately for units in each treatment group and the average potential outcome can be computed using the formula of equation (18) for each level of the treatment. We refer to this procedure as *covariate adjustment* GPS *for categorical treatments*. The relative DRF can be estimated as $\hat{E}\{Y(t)\} - \hat{E}\{Y(0)\}$ for each t. This procedure's validity follows directly from the theory of RR because it only considers two treatments at a time.

If the categorical treatment variable is ordinal with a meaningful numerical scale, we can use the quadratic regression model of equation (3) suggested by HI. However, such a model is restrictive because the slope for GPS in the model changes in a particular way across the treatment levels. Figure 2 shows that this assumption may be too strong to justify in practice.

The usefulness of the covariance adjustment GPS for categorical variables is limited by our ability to fit multiple regression models with limited data. When the treatment takes a large number of values, the method may be infeasible. This problem is even more acute for continuous treatments where it is simply impossible to fit a separate regression model for each observed treatment level. We now discuss the covariate adjustment for continuous treatments.

B.2 Covariance adjustment GPS for continuous treatments

To use covariance adjustment with a continuous treatment variable, we propose to subclassify the data on the treatment variable rather than on the GPS or the PF. To facilitate the computation of standard errors via bootstrap (see below), we form the subclasses using the theoretical quantiles of the fitted treatment assignment model. This is typically easy to accomplish via Monte Carlo. We draw a large sample from the fitted treatment assignment model with parameters fixed at their fitted values and covariates sampled from their observed values and estimate the theoretical quantiles based on this sample. We also compute the theoretical median of the treatment, or its Monte Carlo approximation, within each subclass and denote it as t_s for $s = 1, \ldots, S$ with S the number of subclasses.

With the subclassifed data in hand, we fit the model defined in equation (17) separately for each subclass. Alternatively, we can use a more flexible model. Here, we consider both quadratic regression, i.e., $Y_i \sim \alpha_t + \beta_t \hat{R}_i + \gamma_t \hat{R}_i^2$, and the SCM given in equation (2) with T replaced by $\hat{R}$. We then compute the GPS for each unit at the median treatment value within each subclass, i.e., $\hat{r}(t_s, \mathbf{X}_i)$ for $i = 1, \ldots, n$ and $s = 1, \ldots, S$. Finally, we estimate the DRF by computing $\hat{E}\{Y(t_s)\}$ for each t_s using equation (18) or an appropriate generalization of it if a different response model is used. The derivative of the DRF at t_s can be estimated as in equation (9). Notice that the grid values at which we compute the DRF are different than those advocated by HI. Ours are based on percentiles of the fitted treatment assignment model, whereas theirs are equally spaced in the range of observed treatments.

The standard bias-variance tradeoff arises when selecting the number of subclasses, S. We generally defer to Cochran's advice and use about five [3]. Sensitivity to the choice of S can be quantified by repeating the entire procedure with S equal to approximately three and ten. One source of bias in this procedure results from using units with a range of treatment values to fit the model given in equation (17) (or a more flexible

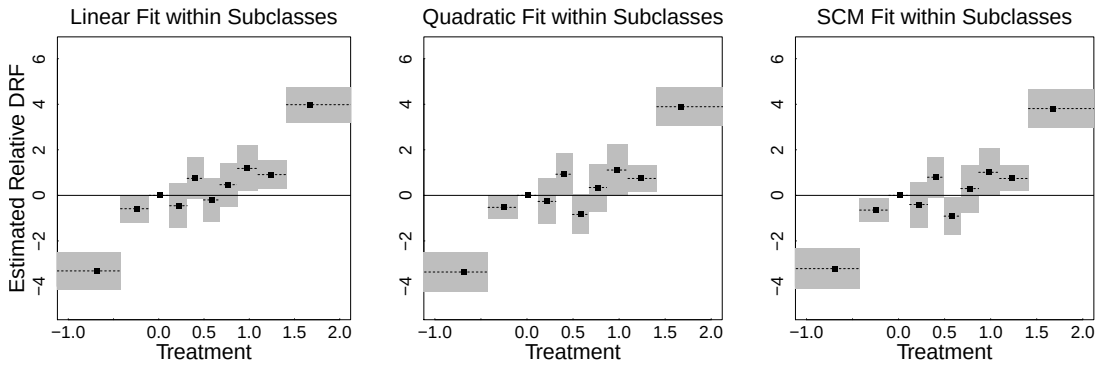

Figure 15. Estimated Relative DRFs in Simulation Study I for the Covariance Adjustment GPS Method. The three plots correspond to the three within subclass models. In all plots the solid (dashed) lines represent the true (fitted) relative DRF and 95% confidence bands based on 1000 bootstrap replications are plotted in grey.

version of it). This bias is especially acute in subclasses with a relatively wide range of the treatment value. If the distribution of the treatment has tails in either direction this correspond to extreme evaluation points of the DRF, t_1 and t_S. Thus, in some cases, we might want to increase the number of subclasses, especially when the extremities of the DRF are of interest. This point is illustrated in Sections B.3.

We approximate the standard errors of the estimated DRF and its estimated derivative via bootstrap resampling. We resample the data, fit the treatment model, subclassify, and compute the DRF and its derivative for each resampled data as described above. We use the same evaluation points, $t_1, \ldots, t_S$ for each resampled data set. Because both the treatment assignment model and the response model are fitted to each bootstrap sample, this procedure accounts for both sources of uncertainty.

B.3 The numerical performance of Covariance Adjustment GPS

We now examine the performance of covariance adjustment GPS in Simulations I, II, and III, as well as in the estimation of the DRF of smoking on annual medical expenditures. In Simulation I, we again use the the correct treatment assignment model and $S = 10$ subclasses with grid points at the $5\%, 15\%, \ldots,$ and 95% quantiles of T. (Using $S = 5$ or 15 gives similar results.) We compare three within subclass response models: (i) $Y \sim R$, (ii) $Y \sim R + R^2$, and (iii) $Y \sim f(R)$, where $f(\cdot)$ is a SCM. The results are shown in Figure 15. The three response models are labelled linear, quadratic, and SCM fit within subclasses, respectively. The response models are conditional on R, rather than on T as in Section 3.1 because covariance adjustment GPS subclassifies on T. As mentioned in Section B.2, the fitted relative DRF exhibit bias in extreme subclasses owing to the relatively large range of treatment levels in these classes. Because the three within subclass models used with covariance adjustment GPS lead to very similar fits, we only present results for the quadratic model in the rest of this section.

Figures 16 and 17 show the estimated relative DRFs in Simulation II and III, respectively. In both simulations, we use a quadratic response model within each of $S = 10$ subclasses. (Using $S = 7$ or 13 and/or the other two within subclass models yields similar results.) The correct treatment assignment model is used in Simulation II. Except in the two most extreme subclasses, the estimated DRF appears to be essentially unbiased. As in Simulation study I, the fitted relative DRF deteriorates in the extreme, more heterogeneous treatment subclasses. Because the distribution of treatment is left skewed, this is less of a problem for the right-most than for the left-most subclass. This along with the blocky nature of the fitted DRF may lead many users to prefer the smooth fitted relative DRF obtained with SCM(PF).

Figure 18 shows the estimated DRF for the simulation based on the applied-example in Section 5. We used $S = 7, 10$, and 13 subclasses along with a linear, quadratic, or SCM fit of $\log(Y)$ on $\hat{R}$ within each subclass. The results are all similar and we only present those with $S = 10$ using a quadratic model within subclass fit. For this fit, the DRF is evaluated at the midpoint of each subclass, corresponding to the theoretical

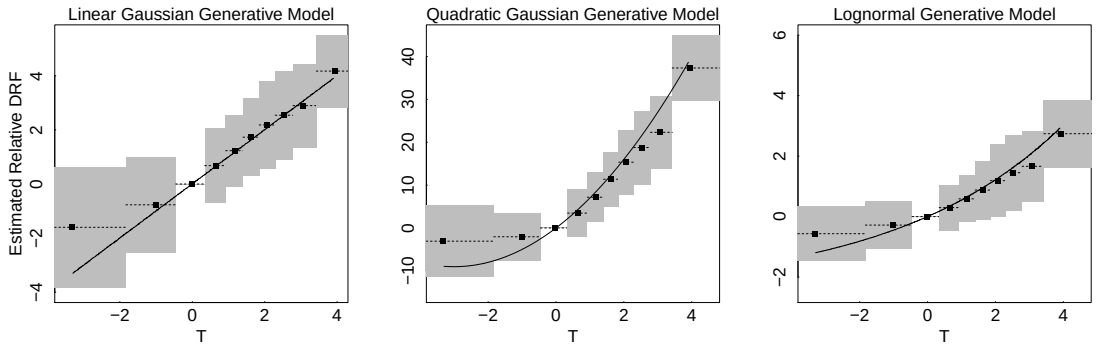

Figure 16. Estimated Relative DRFs in Simulation Study II for the Covariance Adjustment GPS Method. Solid lines represent the true relative DRF and dashed lines the average of the fitted relative DRFs across 1000 simulations. Points represent the theoretical quantiles of T used to construct the subclasses. The grey shaded regions represent pointwise intervals containing 95% of the 1000 fitted relative DRFs. Except in the extreme subclasses, the estimated DRF appears to be essentially unbiased.

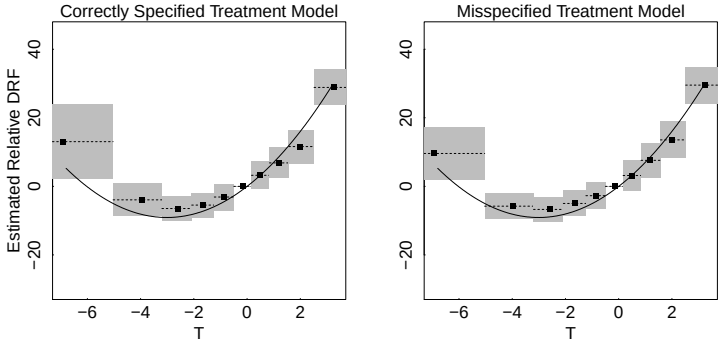

Figure 17. Estimated Relative DRFs in Simulation Study III for the Covariance Adjustment GPS Method. Solid lines represent the true relative DRF and dashed lines the average of the fitted relative DRFs across 1000 simulations. Points represent the theoretical quantiles of T used to construct the subclasses. The grey shaded regions represent pointwise intervals containing 95% of the 1000 fitted relative DRFs.

5%, 15%, ..., 95% quantiles of log(`packyear`). The general shape of the estimated DRFs estimated with covariance adjustment GPS is very similar to those estimated with SCM(GPS), see Figure 8. While both methods show some improvement over HI they are both dominated by SCM(PF).

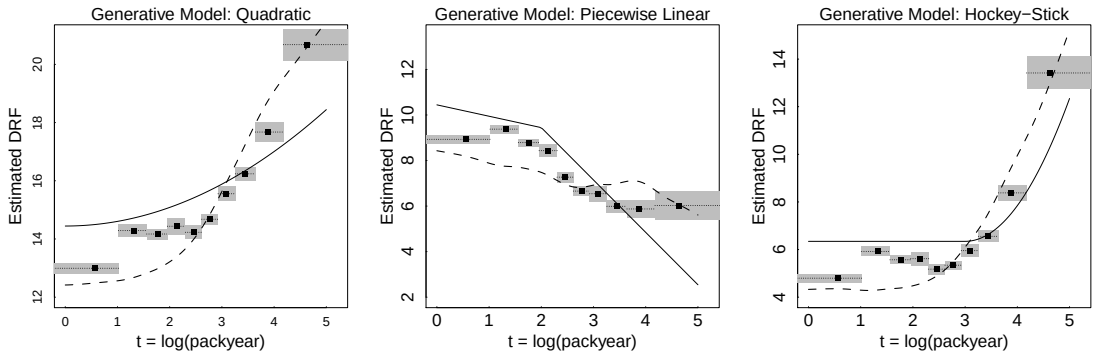

Figure 18. Estimated DRF for the Simulation Based on Smoking Data Using the Covariance Adjustment GPS Method. In all plots the dotted lines with bullets correspond to the fitted DRF. The true DRF is plotted as solid lines while dashed lines represent the fitted SCM of $\log(Y)$ on T, unadjusted for the covariates. Evaluation points are based on the theoretical quantiles of $\log(\texttt{packyear})$. The 95% asymptotic confidence bands plotted in grey are based on 1000 bootstrap replications.