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

Shandong Zhao* David A. van Dyk† Kosuke Imai‡

October 27, 2019

1 Additional Simulation Studies

1.1 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,

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

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

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

and

$$w(r,t) \;=\; \frac{\phi((u_1(r,t) - \mu_X)/\sigma_X)}{\phi((u_1(r,t) - \mu_X)/\sigma_X) + \phi((u_2(r,t) - \mu_X)/\sigma_X)} \tag{3}$$

with $\phi(\cdot)$ is the standard normal density function. The dose-response function is obtained by integrating equation (1) 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.

1.2 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 1, 2, and 3, respectively. In all cases the standard errors grow as expected, but the relative performance of the algorithms remains qualitatively similar.

1.3 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 1.4 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 $Z_i \sim$ 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 $i = 1, \ldots, 2000$.

*Department of Statistics, University of California, Irvine, CA 92697 USA
†Department of Mathematics, Imperial College London, SW7 2AZ, United Kingdom
‡Department of Government and Department of Statistics, Harvard University, Cambridge MA 02138 USA

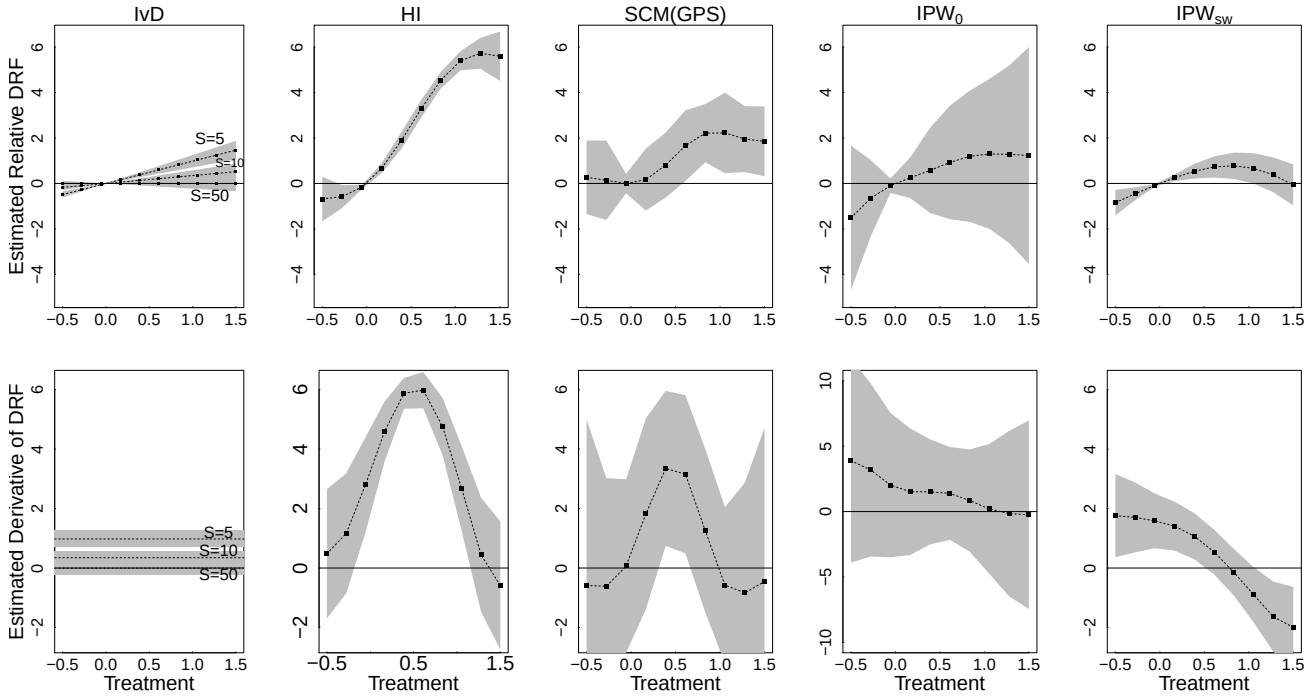

Figure 1: Simulation I Replicated with a Reduced Sample Size of $n = 500$. Results are qualitatively similar to those presented in Figure 1, but with larger standard errors.

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 4. 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 5. 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 5 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 4. The third and fourth panels show that as t increases to 0.5 and 1.0, 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 4) 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 6).

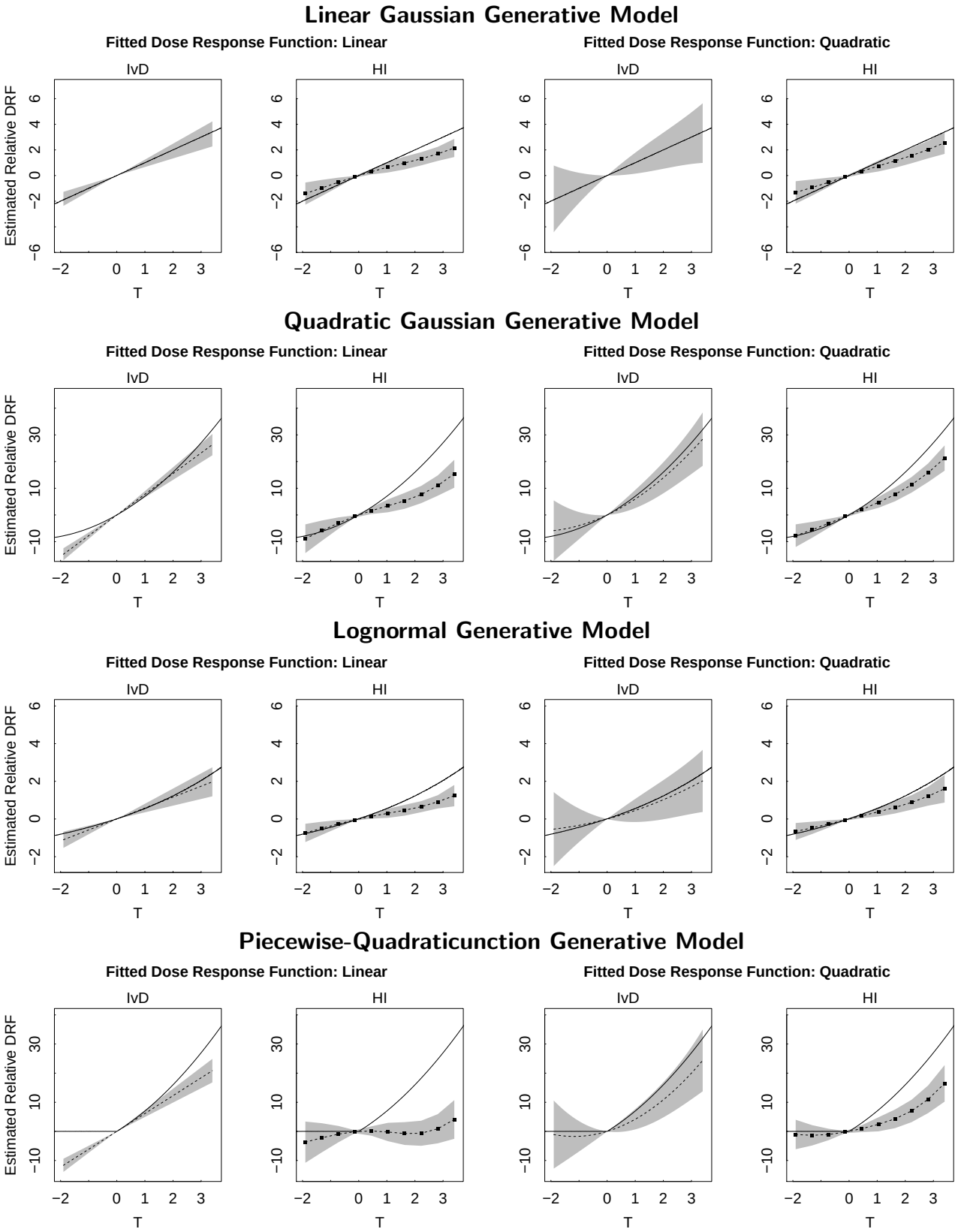

Figure 2: Estimated Relative DRFs Using the Methods of IvD and HI in Simulation Study II, but with a Reduced Sample Size of $n = 500$. Results are qualitatively similar to those presented in Figure 3, but with larger standard errors.

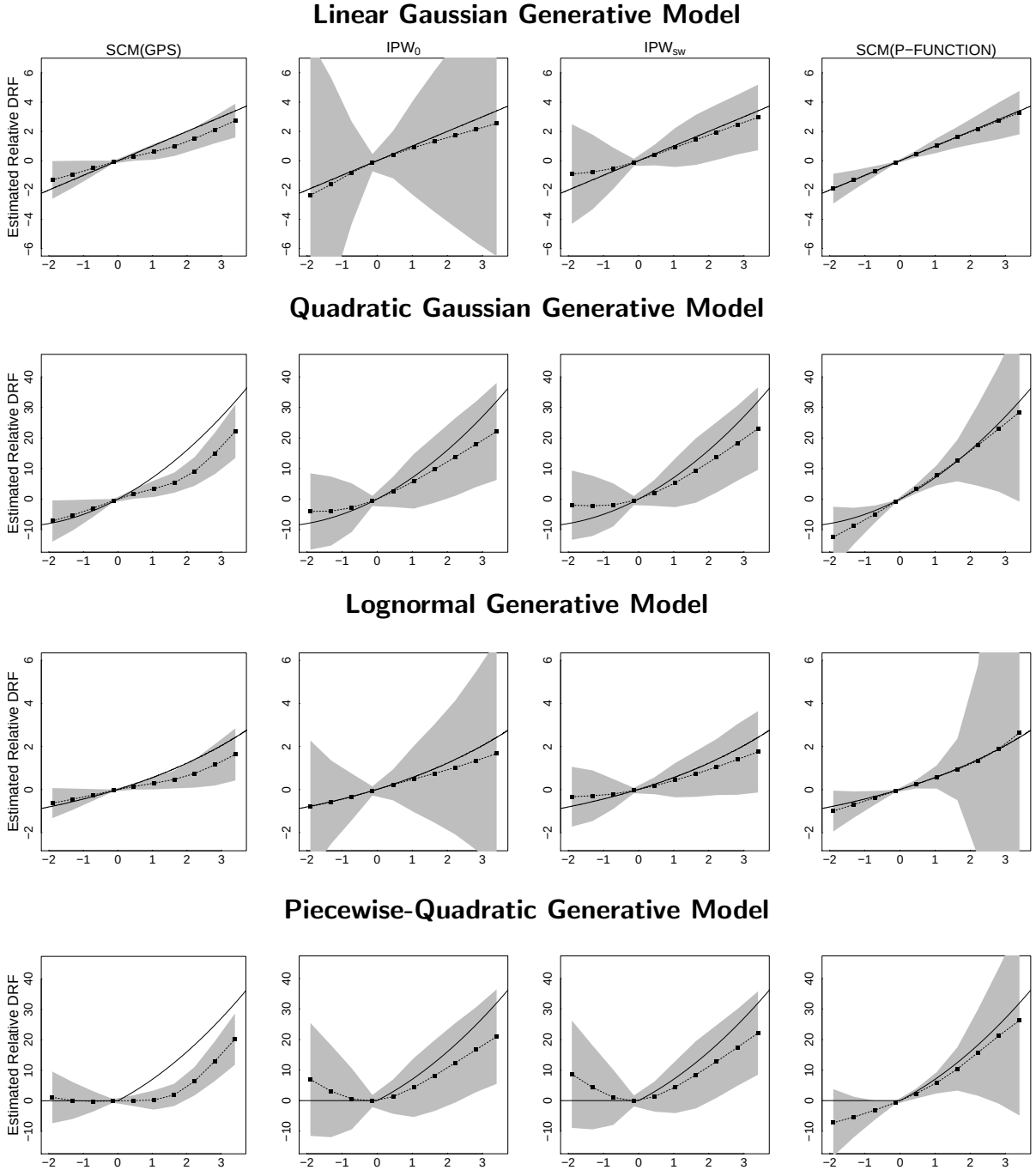

Figure 3: Estimated Relative DRFs for SCM(GPS), IPW_0, IPW_{SW}, and SCM(PF) Methods in Simulation Study II, but with a Reduced Sample Size of $n = 500$. Results are qualitatively similar to those presented in Figure 4, but with larger standard errors.

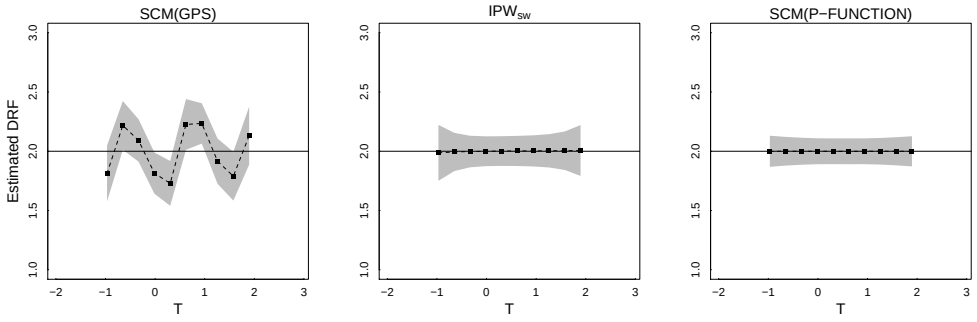

Figure 4: Estimated DRF for Simulation IV. Solid lines, dashed lines and gray regions represent the true DRFs the means of the 1,000 fitted DRFs and 95% pointwise intervals. Only SCM(GPS) exhibits the cyclic bias. The SCM(PF) is introduced in Section 4.

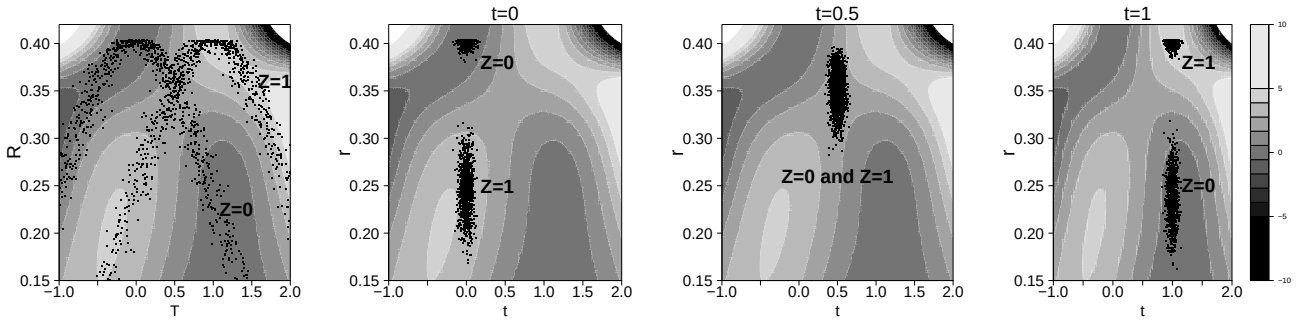

Figure 5: How the SCM(GPS) fit can lead to a cyclic artifact in the the fitted DRF. The leftmost panel overlays a scatter plot of T and the GPS, $\hat{R}$, on a heat map of the fitted SCM(GPS) response model in Simulation IV. The other three panels overlay scatterplots of $(t, \hat{r}(t, \boldsymbol{X}_i))$, 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.

1.4 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 (2007) 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 6 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.

2 Covariance Adjustment GPS

2.1 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 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 (Imbens, 2000), it is easy to see that the usual covariance adjustment is equivalent to fitting the following regression model,

$$Y_i \ \sim \ \alpha_t + \beta_t \hat{R}_i, \tag{4}$$

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.

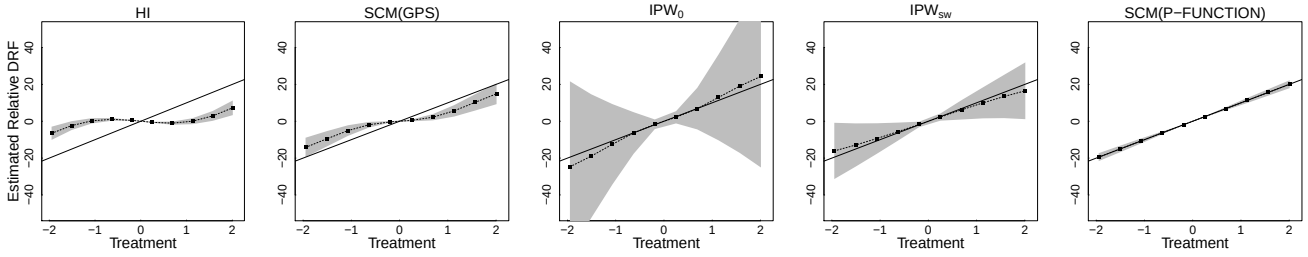

Figure 6: Estimated Relative DRF in Simulation V. Solid lines, dashed lines and gray regions represent the true relative DRFs the means of the 1000 fitted relative DRFs and 95% pointwise intervals. The evaluation points are identical for all plots. SCM(GPS) exhibits a cyclic artifact and IPW_0 is quite unstable. The SCM(PF) method proposed in Section 4.1 again outperforms the other methods.

After fitting the model given in equation (4), 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

$$\hat{E}\{Y(t)\} \;=\; \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\alpha}_t + \hat{\beta}_t\ \hat{r}(t,\mathbf{X}_i)\right\}, \tag{5}$$

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 (4) can be fitted separately for units in each treatment group and the average potential outcome can be computed using the formula of equation (5) 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.

2.2 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 (4) 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 (5) 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 (Cochran, 1968). 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

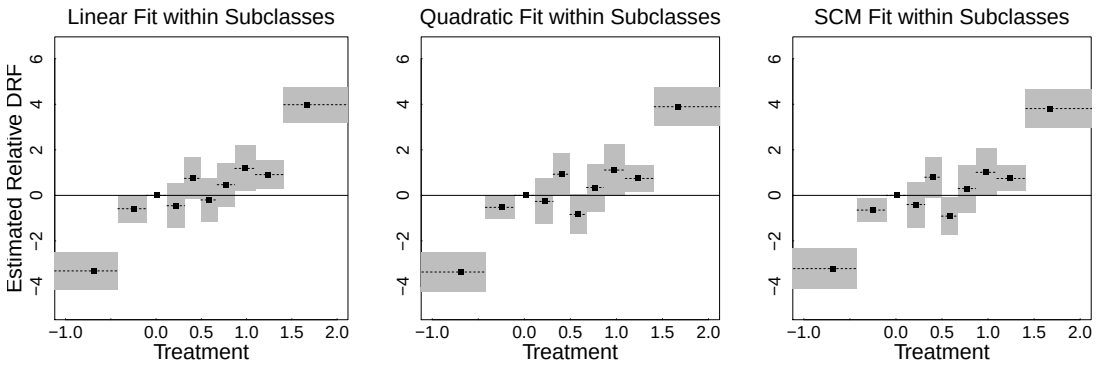

Figure 7: Estimated Relative DRFs in Simulation Study I for the Covariance Adjustment GPS Method. The three plots correspond to the three within subclass models. In all plots the solid (dashed) lines represent the true (fitted) relative DRF and 95% confidence bands based on 1000 bootstrap replications are plotted in grey.

equation (4) (or a more flexible 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 2.3.

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.

2.3 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 7. 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 2.2, 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 8 and 9 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 10 shows the estimated DRF for the simulation based on the applied-example in Section 5. We

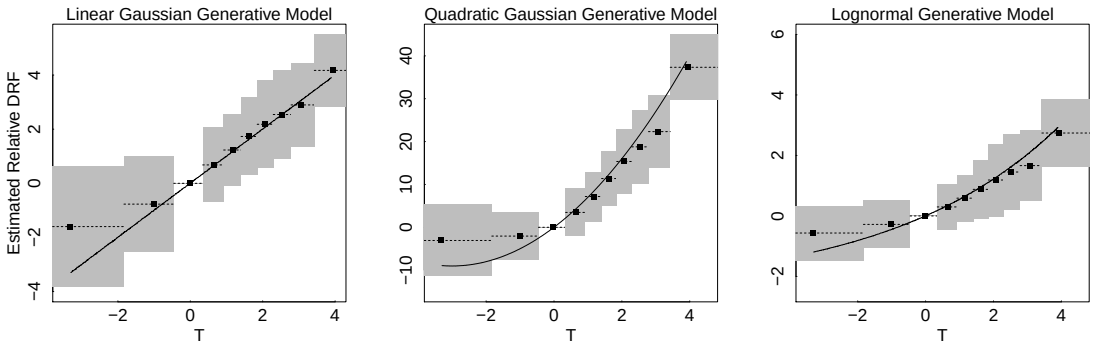

Figure 8: Estimated Relative DRFs in Simulation Study II for the Covariance Adjustment GPS Method. Solid lines represent the true relative DRF and dashed lines the average of the fitted relative DRFs across 1000 simulations. Points represent the theoretical quantiles of T used to construct the subclasses. 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.

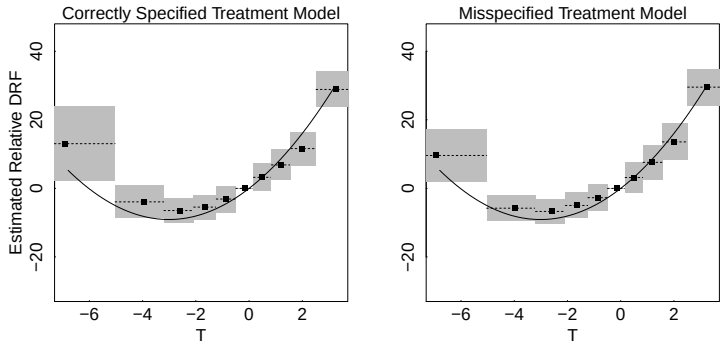

Figure 9: Estimated Relative DRFs in Simulation Study III for the Covariance Adjustment GPS Method. Solid lines represent the true relative DRF and dashed lines the average of the fitted relative DRFs across 1000 simulations. Points represent the theoretical quantiles of T used to construct the subclasses. The grey shaded regions represent pointwise intervals containing 95% of the 1000 fitted relative DRFs.

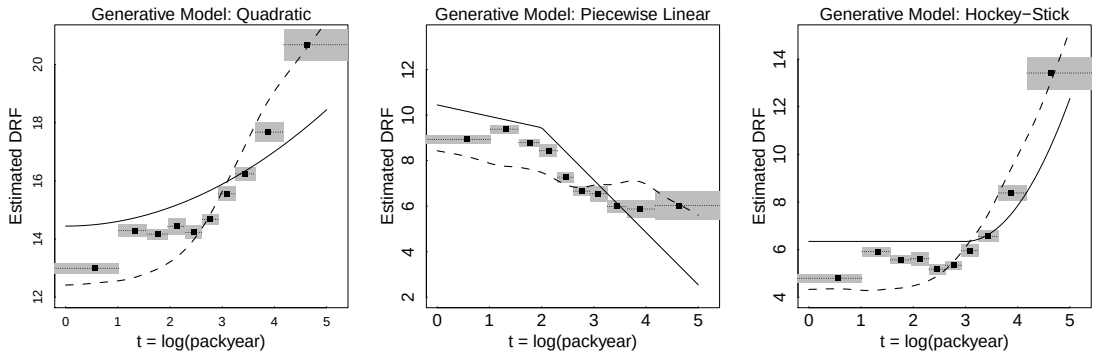

Figure 10: Estimated DRF for the Simulation Based on Smoking Data Using the Covariance Adjustment GPS Method. In all plots the dotted lines with bullets correspond to the fitted DRF. The true DRF is plotted as solid lines while dashed lines represent the fitted SCM of $\log(Y)$ on T, unadjusted for the 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.

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 5%, 15%, ..., 95% quantiles of $\log(\texttt{packyear})$. 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).

References

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24**, 295–313.

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

Kang, J. D. and Schafer, J. L. (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.