

Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects*

Matias Busso
IDB, IZA

John DiNardo
University of Michigan
and NBER

Justin McCrary
University of California, Berkeley
and NBER

June 9, 2009

Abstract

We explore the finite sample properties of several semiparametric estimators of average treatment effects, including propensity score reweighting, matching, double robust, and control function estimators. When there is good overlap in the distribution of propensity scores for treatment and control units, reweighting estimators are preferred on bias grounds and are quite close to the semiparametric efficiency bound even for samples of size 100. Pair matching exhibits similarly good performance in terms of bias, but has notably higher variance. Local linear and ridge matching are competitive with reweighting in terms of bias and variance, but only once $n = 500$. Nearest-neighbor, kernel, and blocking matching are not competitive. When overlap is close to failing, none of the estimators examined perform well and \sqrt{n} -asymptotics may be a poor guide to finite sample performance. Trimming rules are commonly used in the face of problems with overlap. Only some of these rules seem to be effective, and then only in settings with homogeneous treatment effects.

JEL Classification: C14, C21, C52.

Keywords: Treatment effects, propensity score, overlap, irregular identification, semiparametric efficiency.

*For comments that improved the paper, we thank Alberto Abadie, David Card, Marie Davidian, Keisuke Hirano, Guido Imbens, Jack Porter, and Elie Tamer, but in particular Matias Cattaneo, Jeff Smith and Serena Ng. We would also like to thank Markus Frölich for providing us copies of the code used to generate the results in his paper.

I Introduction

This paper explores the finite sample properties of semiparametric estimators of average treatment effects. Such estimators are standard in the program evaluation literature and have become increasingly popular in the applied microeconomic literature. These estimators rely on two assumptions. The first assumption is that assignment to treatment is randomized conditional on a set of observed covariates. The second assumption is more technical and asserts that no value of the observed covariates assures treatment assignment.¹ Intuitively, these assumptions allow for treatment to covary with observed characteristics, but require that there be some unexplained variation in treatment assignment left over after conditioning and that the unexplained aspect of treatment resembles an experiment.²

Estimation of program impacts under these assumptions could proceed using traditional parametric estimation methods such as maximum likelihood. However, an early result of Rosenbaum and Rubin (1983) is that if treatment is randomized conditionally on the observed covariates, then it is randomized conditional on the (scalar) propensity score, the conditional probability of treatment given the observed covariates. Influenced by this result, the subsequent econometric and statistical literatures have focused on semiparametric estimators that eschew parametric assumptions on the relationship between the outcome and observed covariates. Empirical literatures, particularly in economics, but also in medicine, sociology and other disciplines, feature an extraordinary number of program impact estimates based on such semiparametric estimators.

Perhaps surprisingly in light of their ubiquity in empirical work, formal large sample results for these estimators have only recently been derived in the literature. Heckman, Ichimura and Todd (1997) report large sample properties of estimators based on kernel and local linear matching on the true and an estimated propensity score. Hirano, Imbens and Ridder (2003) report large sample properties of a reweighting estimator that uses a nonparametric estimate of the propensity score. This is essentially the same reweighting estimator that was introduced to the economics literature by DiNardo, Fortin and Lemieux (1996) and Dehejia and Wahba (1997), and it is related to an estimator due to Horvitz and Thompson (1952). Importantly, Hirano et al. (2003) establish that their version of a reweighting estimator achieves the semiparametric efficiency bound (SEB) established by Hahn (1998) for this problem. Robins,

¹Selection on observed variables is defined in Section II, as is the second assumption, which is typically referred to as an overlap assumption. In Section III, we emphasize the correct interpretation of selection on observed variables using a specific parametric model.

²In other words, there exists an instrument which is unobserved by the researcher.

Rotnitzky and Zhao (1994) and Robins and Rotnitzky (1995) establish large sample properties and the efficiency of a regression-adjusted reweighting estimator that uses the estimated propensity score. Finally, Abadie and Imbens (2006) establish the large sample properties and near-efficiency of k th nearest-neighbor matching using the true propensity score.³

To date, no formal finite sample properties have been established for any of the estimators discussed, and there is limited simulation evidence on their performance. It is generally desirable to learn about the finite sample properties of estimators used in empirical research, since not all data sets are big enough for asymptotic theory to be a useful guide to estimator properties. It is particularly desirable to learn about the finite sample properties of semiparametric estimators of average treatment effects, given the literature's substantive focus on treatment effect heterogeneity.⁴ In the face of heterogeneity, treatment effects must effectively be estimated for various subsamples.⁵ For many economic data sets, these subsamples are modest in size, perhaps numbering in the hundreds or even dozens, where asymptotic theory may be a particularly poor guide to finite sample performance.

In this paper, we examine the relative performance of several leading semiparametric estimators of average treatment effects in samples of size 100 and 500.⁶ We focus on the performance of propensity score reweighting and matching estimators for estimating the average treatment effect (ATE) and the average effect of treatment on the treated (TOT). We consider a range of matching strategies, including nearest neighbor, kernel, local linear, and ridge matching, and blocking. We also consider several varieties of reweighting estimators, the so-called double robust estimator (Robins et al. 1994), and a specific version of Hahn's (1998) general estimator, which we term a control function estimator. We consider settings with good overlap in the distribution of propensity scores for treatment and control units, as well as settings of poor overlap. In settings of poor overlap, we investigate the performance of various trimming

³It deserves mention that Chen, Hong and Tarozzi (2008) study the large sample properties and efficiency of sieve estimators in this setting. We do not study the finite sample properties of these estimators due to space constraints.

⁴Understanding the sources of treatment effect heterogeneity is critical if the analyst hopes to extrapolate from the findings of a given study to broader forecasts of the likely impacts of policies not yet implemented. These issues are a key focus of the program evaluation literature (see, for example, Heckman and Vytlačil 2005 and Heckman, Urzua and Vytlačil 2006).

⁵Importantly, the intrinsic dimensionality of treatment effect heterogeneity cannot be massaged by appealing to the dimension reduction of the propensity score. The Rosenbaum and Rubin (1983) result that a conditionally randomized treatment is randomized conditional on the scalar propensity score has been interpreted as justification for matching on the propensity score rather than on the full set of covariates. However, the Rosenbaum and Rubin result does not imply that units with the same value of the propensity score have the same treatment effect. Examples of empirical investigation of treatment effect heterogeneity along dimensions different from the propensity score include Card (1996), Katz, Kling and Liebman (2001), Haviland and Nagin (2005) and Kent and Hayward (2008), for example.

⁶This issue has been previously taken up by Lunceford and Davidian (2004), Frölich (2004), Zhao (2004), Zhao (2008), and Freedman and Berk (2008).

methods proposed and used in the literature. Finally, we consider the implications for performance of misspecification of the propensity score, both in terms of an incorrect parametric model for treatment as well as conditioning on the wrong set of covariates.

A summary of our findings is as follows. First, reweighting is approximately unbiased and semiparametrically efficient, even for sample sizes of 100. Our assessment is that reweighting exhibits the best overall finite sample performance of any of the estimators we consider. Second, pair matching shares the good bias performance of reweighting, but has a variance that is roughly 30 percent greater than that of reweighting. Third, k th nearest-neighbor matching, with k chosen by leave-one-out cross-validation, does reduce the excessive variance of pair matching, but at the cost of substantially greater bias. Fourth, kernel, local linear, and ridge matching perform similarly to k -th nearest neighbor matching in exhibiting little variance but much bias when $n = 100$. Once $n = 500$, ridge and local linear matching are both competitive with reweighting on bias and variance grounds.⁷ Fifth, both in terms of bias and variance, the popular blocking matching estimator performs neither as badly as k th nearest-neighbor and kernel matching, nor as well as local linear and ridge matching, and is generally dominated by reweighting. Sixth, the double robust estimator is competitive with reweighting, but appears to be slightly more variable and slightly more biased. Seventh, the control function estimator is approximately unbiased, even for samples of size 100, and is approximately semiparametrically efficient once $n = 500$. Eighth, when there is misspecification of the propensity score either due to parametric assumptions or the lack of availability of important covariates, the relative performance of the estimators is approximately as described above. However, in that context, if problems with bias are suspected and variance is less important, pair matching is the preferred estimator.

The above conclusions hold when the propensity score model is correctly specified and when there is good overlap in the distribution of propensity scores for treatment and control units. Our investigations highlight the problems with semiparametric estimators of average treatment effects when overlap is poor. Khan and Tamer (2007) emphasize this point from a theoretical perspective, noting that when overlap is poor the semiparametric efficient bound derived by Hahn (1998) for this problem can be infinite, leading to a failure of \sqrt{n} -consistency. Consistent with this conclusion, our results indicate that when overlap is poor, none of the estimators studied work well. In cases where overlap is poor, although technically sufficient

⁷All three of these kernel-based estimators use leave-one-out cross-validation to select a bandwidth, and this model selection issue may be an important aspect of performance for smaller sample sizes.

to guarantee \sqrt{n} -consistency, we document poor performance for $n = 100$, but adequate performance for $n = 500$. This suggests that larger sample sizes may be needed for threshold cases.

A standard empirical approach to problems with overlap is to trim observations with extreme values of the propensity score. We investigate four of the trimming strategies used in the literature. Our simulations suggest that some of these procedures can be effective but only in situations in which the treatment effect is similar for all the observations in the sample. Finally, we provide evidence that as problems with overlap arise, the limiting distribution of semiparametric estimators becomes nonstandard.

Our conclusions run contrary to those of the existing literature on the finite sample performance of reweighting and matching. Our simulations indicate that reweighting is a generally robust estimator whose performance in small samples is as effective as in large samples, where it has been shown to be nearly optimal in a certain sense. The matching methods we consider work poorly for samples of size 100, although some of the methods become effective for samples of size 500. In contrast to these findings, the existing finite sample literature is generally negative regarding reweighting and tends to conclude that matching estimators are best. We review this literature. We show that nearly all of the results from the existing finite sample literature are based on data generating processes (DGPs) for which \sqrt{n} -consistent semiparametric estimators do not exist, or DGPs where \sqrt{n} -consistency is close to failing. Our own investigations are unusual, in that we focus on DGPs where semiparametric estimators are expected to perform well. We show that this difference in DGPs accounts for our different conclusions.⁸

The remainder of the paper is organized as follows. Section II sets notation, defines estimators, discusses estimands and efficiency bounds, and emphasizes the connections among the many estimators we consider by casting them in the common framework of weighted regression. In particular, this section provides a 3-step interpretation of matching that clarifies the conceptual similarities and differences between the two approaches. In Section III, we describe our benchmark DGP. This DGP is chosen so that semiparametric estimates of average treatment effects are \sqrt{n} -consistent. Results for the benchmark DGP are presented in Section IV. In Section V, we take up the issue of DGPs for which \sqrt{n} -consistency may be compromised. Results for such DGPs are presented in V. Section VII concludes.

⁸To be clear, we do not advocate the use of reweighting estimators—or any of the estimators studied here—in settings of failure and near failure of \sqrt{n} -consistency of semiparametric estimators of average treatment effects. At present, relatively little is known about appropriate estimation and testing procedures in these settings.

II Notation and Background

Let $Y_i(1)$ denote the outcome for unit i that would obtain under treatment and $Y_i(0)$ the outcome that would obtain under control. Treatment is denoted by the binary variable T_i . We observe $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$, but never the pair $(Y_i(0), Y_i(1))$. The data $(X_i, Y_i, T_i)_{i=1}^n$ are taken to be independent across i , but are potentially heteroscedastic. Let the propensity score, the conditional probability of treatment, be denoted $p(x) \equiv P(T_i = 1 | X_i = x)$. Let the covariate-specific average treatment effect be denoted $\tau(x) = E[Y_i(1) - Y_i(0) | X_i = x]$.

We focus on the relative performance of semiparametric estimators of population averages of $\tau(x)$. Intuitively, semiparametric in this context means an estimator that models the relationship between the probability of receiving treatment and the covariates X_i , but remains agnostic regarding the relationship between the counterfactual outcomes $(Y_i(0), Y_i(1))$ and the covariates.

Semiparametric estimators of treatment effects are justified by an appeal to (1) selection on observed variables and (2) sufficient overlap. Selection on observed variables means that treatment is randomized given X_i , or that $(Y_i(0), Y_i(1), Z_i) \perp\!\!\!\perp T_i | X_i$, where Z_i is any characteristic of the individual that is not affected by treatment assignment (e.g. pre-program earnings).⁹ This assumption has traditionally been referred to as selection on observed variables in the economics literature (e.g., Heckman and Robb 1985). In the statistics and more recent econometrics literature this assumption is instead referred to as ignorability or unconfoundedness (e.g., Rosenbaum and Rubin 1983, Imbens 2004).¹⁰

Selection on observed variables is not by itself sufficient to semiparametrically identify average treatment effects. The DGPs we focus on in Section IV are consistent with both selection on observed variables and a *strict overlap* assumption: $\xi < p(x) < 1 - \xi$ for almost every x in the support of X_i , for some $\xi > 0$. This assumption is stronger than the *standard overlap* assumption that $0 < p(x) < 1$ for almost every x in the support of X_i (e.g., Rosenbaum and Rubin 1983, Heckman et al. 1997, Hahn 1998, Wooldridge 2002, Imbens 2004, Todd 2007), but is also common in the literature (e.g., Robins et al. 1994, Abadie and Imbens 2006, 2008, Crump, Hotz, Imbens and Mitnik 2007a,b). Both the standard overlap and the strict overlap assumptions are strong. Khan and Tamer (2007) emphasize that something akin to the strict overlap assumption is needed to deliver \sqrt{n} -consistency of semiparametric estimators in this context. We

⁹Notice that some pre-program covariates may be affected by anticipation of treatment.

¹⁰Lechner (2005) shows that some control variables may be influenced by the treatment. However, this endogeneity does not matter for consistency of the treatment effect estimator, as long as the usual formulation of the conditional independence assumption holds.

take up the issue of DGPs that violate strict overlap, but satisfy standard overlap, in Sections V and VI, below.

A Estimands and Estimators

As noted, we focus on the performance of estimators for target parameters that are averages of $\tau(x)$. The specific averages we consider are the average treatment effect $\alpha = E[\tau(X_i)]$ and the average treatment effect on the treated $\theta = E[\tau(X_i)|T = 1]$. We refer to these estimands as ATE and TOT, respectively. Although we focus on the performance of estimators for these estimands, we emphasize that ATE and TOT are not the only estimands of interest. However, the performance of these estimators for ATE and TOT is likely to be similar to the performance of these estimators when adapted to estimation of other averages of $\tau(x)$.

We consider fourteen estimators: nine matching estimators, three reweighting estimators (sometimes termed inverse propensity score weighting estimators, or IPW), one control function estimator, and the so-called double robust estimator.¹¹ Each of these estimators are two-step estimators relying on a first-step estimate of the propensity score. The nine matching estimators include pair matching, k th nearest neighbor matching, kernel matching, local linear matching, ridge regression matching, and blocking. Aside from pair matching, each of these matching strategies employs a cross-validation method for choosing a tuning parameter. Kernel, local linear, and ridge matching all further require the choice of a kernel. Following Frölich (2004), we consider both the Epanechnikov kernel and the Gaussian kernel.

The three reweighting estimators include a reweighting estimator in which the sum of the weights is allowed to be stochastic (IPW1), a reweighting estimator in which the sum of the weights is forced to be 1 (IPW2), and an asymptotically optimal combination of the former two estimators (IPW3) that is due to Lunceford and Davidian (2004).

We also consider the so-called double robust estimator due to Robins and Rotnitzky (1995), which has recently received a good deal of attention in the literature (e.g., Imbens 2004). This procedure can be thought of as a regression-adjusted version of reweighting. The regression adjustments are more similar in spirit to an older approach to the problem of estimating treatment effects. We complete our analysis by studying the performance of a control function estimator. This estimator is essentially the same as the

¹¹The breadth of coverage arises from an attempt to encompass many of the estimators used in the literature as well as to be consistent with previous finite sample evidence on the topic. Nonetheless, there are of course other potentially effective estimators whose performance is not covered by the analysis here.

double robust estimator, but is unweighted. The version of the control function estimator we implement models the regression function of the outcome given the covariates and treatment status as a polynomial in the estimated propensity score, with additive and possibly interacted terms for treatment status. This procedure is described in Wooldridge (2002) for the case of ATE, and is in the spirit of Oaxaca (1973) and Blinder (1973) decompositions and Hahn’s (1998) general estimator.

While it is true that at least some versions of reweighting and matching are believed to be semiparametrically efficient in large samples, and while both approaches are based on the same first-step propensity score estimate, it is far from clear that the two approaches would perform similarly in finite samples. First, most matching estimators rely on tuning parameters. It is possible that use of tuning parameters could improve finite sample performance relative to reweighting. Second, the approaches take advantage of very different properties of the propensity score. Matching requires of the estimated propensity score only that it be a balancing score (Rosenbaum and Rubin 1983). In contrast, reweighting requires that the propensity score be a conditional probability. For example, matching on the square root of the propensity score should work just as well as matching on propensity score; in contrast, reweighting with the square root of the propensity score should do badly.

B Weighted Least Squares as a Unifying Framework

Both matching and reweighting estimators of average treatment effects can be understood as the coefficient on the treatment indicator in a weighted regression, with weighting functions that differ by estimator. This common structure clarifies that the essential difference between the estimators is the weighting function implicitly used.

That reweighting estimators have this form is widely understood. A general notation for reweighting estimators for the TOT and ATE is

$$\hat{\theta} = \frac{1}{n_1} \sum_{i=1}^n T_i Y_i - \frac{1}{n_0} \sum_{j=1}^n (1 - T_j) Y_j \bar{w}(j), \quad (1)$$

$$\hat{\alpha} = \frac{1}{n_1} \sum_{i=1}^n T_i Y_i \bar{w}_1(i) - \frac{1}{n_0} \sum_{j=1}^n (1 - T_j) Y_j \bar{w}_0(j). \quad (2)$$

The weights in equations (1) and (2) only add up to one for some versions of reweighting estimators. When the TOT weights add up to one in the sense of $\frac{1}{n_0} \sum_{j=1}^n (1 - T_j) \bar{w}(j) = 1$, the TOT estimate can be obtained

using standard statistical software from the coefficient on treatment in a regression of the outcome on a constant and a treatment indicator using weights $W = T + (1 - T)\bar{w}(\cdot)$. When the weights do not add up to one, the TOT estimate can be calculated directly using equation (1). When the ATE weights add up to one in the sense that $\frac{1}{n_0} \sum_{j=1}^n (1 - T_j)\bar{w}_0(j) = 1$ and $\frac{1}{n_1} \sum_{j=1}^n T_j\bar{w}_1(j) = 1$, the ATE estimate can be obtained from the same regression described, but with weights $W = T\bar{w}_1(\cdot) + (1 - T)\bar{w}_0(\cdot)$. The reweighting estimators we consider are characterized below by enumerating the weighting functions used.

WEIGHTS USED FOR REWEIGHTING ESTIMATORS

Effect	Treatment, t	Estimator	Weighting Function $\bar{w}_t(j)$
TOT	1	IPW1	$\frac{\hat{p}(X_j)}{1 - \hat{p}(X_j)} \Big/ \frac{\hat{p}}{1 - \hat{p}}$
TOT	1	IPW2	$\frac{\hat{p}(X_j)}{1 - \hat{p}(X_j)} \Big/ \frac{1}{n_0} \sum_{k=1}^n \frac{(1 - T_k)\hat{p}(X_k)}{1 - \hat{p}(X_k)}$
TOT	1	IPW3	$\frac{\hat{p}(X_j)}{1 - \hat{p}(X_j)} (1 - C_j) \Big/ \frac{1}{n_0} \sum_{k=1}^n \frac{(1 - T_k)\hat{p}(X_k)}{1 - \hat{p}(X_k)} (1 - C_k)$
ATE	0	IPW1	$\frac{1 - \hat{p}}{1 - \hat{p}(X_j)}$
ATE	1	IPW1	$\frac{\hat{p}}{\hat{p}(X_j)}$
ATE	0	IPW2	$\frac{1}{1 - \hat{p}(X_j)} \Big/ \frac{1}{n_0} \sum_{k=1}^n \frac{1 - T_k}{1 - \hat{p}(X_k)}$
ATE	1	IPW2	$\frac{1}{\hat{p}(X_j)} \Big/ \frac{1}{n_1} \sum_{k=1}^n \frac{T_k}{\hat{p}(X_k)}$
ATE	0	IPW3	$\frac{1}{1 - \hat{p}(X_j)} (1 - C_j^0) \Big/ \frac{1}{n_0} \sum_{k=1}^n \frac{1 - T_k}{1 - \hat{p}(X_k)} (1 - C_k^0)$
ATE	1	IPW3	$\frac{1}{\hat{p}(X_j)} (1 - C_j^1) \Big/ \frac{1}{n_1} \sum_{k=1}^n \frac{T_k}{\hat{p}(X_k)} (1 - C_k^1)$

Note: $\hat{p} \equiv \frac{n_1}{n}$, $A_i = \frac{1 - T_i}{1 - \hat{p}(X_i)}$, $B_i = \frac{T_i}{\hat{p}(X_i)}$, $C_i = \frac{\left(1 - \frac{\hat{p}(X_i)}{\hat{p}}\right) A_i \frac{1}{n} \sum_{j=1}^n \left(1 - \frac{\hat{p}(X_j)}{\hat{p}}\right) A_j}{\frac{1}{n} \sum_{j=1}^n \left(1 - \frac{\hat{p}(X_j)}{\hat{p}}\right) A_j}$,

$C_i^0 = \frac{\frac{1}{1 - \hat{p}(X_i)} \frac{1}{n} \sum_{j=1}^n (A_j \hat{p}(X_j) - T_i)}{\frac{1}{n} \sum_{j=1}^n (A_j \hat{p}(X_j) - T_i)^2}$ and $C_i^1 = \frac{\frac{1}{\hat{p}(X_i)} \frac{1}{n} \sum_{j=1}^n (B_j (1 - \hat{p}(X_j)) - (1 - T_i))}{\frac{1}{n} \sum_{j=1}^n (B_j (1 - \hat{p}(X_j)) - (1 - T_i))^2}$, which are

IPW3 correction factors that are small when the propensity score model is well specified.

The functional form given by IPW1 can be found in many treatments in the literature (e.g., Dehejia and Wahba 1997, Wooldridge 2002, Hirano et al. 2003). IPW2 is advocated by Johnston and DiNardo (1996) and Imbens (2004). Since most applied work is based on regression software, which naturally rescales weights, most estimates in the empirical literature are probably IPW2. With a well-specified propensity score model, the weights used in IPW1 should nearly add up to one and IPW1 and IPW2 should not differ dramatically. This is because, ignoring estimation error in $\hat{p}(X_i)$ and \hat{p} , iterated expectations shows that $E[W_i] = 1$ for both TOT and ATE. However, in finite samples for some DGPs, the sum of the weights can depart substantially from 1. Unlike IPW2, IPW3 is not commonly implemented in the empirical

literature. This estimator, derived by Lunceford and Davidian (2004) for the case of ATE, is the (large sample) variance-minimizing linear combination of IPW1 and IPW2.¹²

While it is widely understood that reweighting estimators can be implemented as a weighted regression, it is less widely understood that matching estimators share this property.¹³ We demonstrate that matching estimators are weighted regressions for the case of TOT.¹⁴ A general notation for a matching estimator of the TOT is (cf., Smith and Todd 2005, eq. 10)

$$\hat{\theta} = \frac{1}{n_1} \sum_{i \in I_1} \left\{ Y_i - \sum_{j \in I_0} w(i, j) Y_j \right\}, \quad (3)$$

where $w(i, j)$ is the weight that the control observation j is assigned in the formation of an estimated counterfactual for the treated observation i , I_1 is the set of n_1 treated units and I_0 is the set of n_0 control units. The weights $w(i, j)$ are in general a function of the distance in the covariates. In the case of propensity score matching, that distance is measured by the difference in the estimated propensity scores. We now describe the matching estimators we consider by enumerating the TOT weighting functions $w(i, j)$.¹⁵

WEIGHTS USED FOR MATCHING ESTIMATORS FOR TOT

Estimator	Weighting Function $w(i, j)$
kth Nearest Neighbor	$\frac{1}{k} \mathbf{1}(\hat{p}(X_j) \in \mathcal{J}_k(i))$
Kernel	$K_{ij} / \sum_{j \in I_0} K_{ij}$
Local Linear	$(K_{ij} L_i^2 - K_{ij} \hat{\Delta}_{ij} L_i^1) / \sum_{j \in I_0} (K_{ij} L_i^2 - K_{ij} \hat{\Delta}_{ij} L_i^1 + r_L)$
Ridge	$K_{ij} / \sum_{j \in I_0} K_{ij} + \tilde{\Delta}_{ij} / \sum_{j \in I_0} (K_{ij} \tilde{\Delta}_{ij}^2 + r_R h \tilde{\Delta}_{ij})$
Blocking	$\sum_{m=1}^M \mathbf{1}(\hat{p}(X_i) \in B_m) \mathbf{1}(\hat{p}(X_j) \in B_m) / \sum_{m=1}^M \mathbf{1}(\hat{p}(X_j) \in B_m)$

All of the matching estimators enumerated can be understood as the coefficient on the treatment indicator

¹²The TOT version of IPW3 is new, but follows straightforwardly, if tediously, from the approach outlined by those authors.

¹³However, there are clear antecedents in the literature. For example equations (3) and (4) of Abadie and Imbens (2006) clarify this common structure.

¹⁴The case of ATE then follows since a matching estimator for the ATE is a convex combination of the average treatment effect for the treated and for the untreated, with convex parameter equal to the fraction treated.

¹⁵The notation is as follows: $\mathcal{J}_k(i)$ is the set of k estimated propensity scores among the control observations that are closest to $\hat{p}(X_i)$, $\hat{\Delta}_{ij} = \hat{p}(X_i) - \hat{p}(X_j)$, $K_{ij} = K(\hat{\Delta}_{ij}/h)$ for $K(\cdot)$ a kernel function and h a bandwidth, $L_i^p = \sum_{j \in I_0} K_{ij} \hat{\Delta}_{ij}^p$, for $p = 1, 2$, $\tilde{\Delta}_{ij} = \hat{p}(X_j) - \bar{p}(X_i)$, $\bar{p}_i = \sum_{j \in I_0} p_j K_{ij} / \sum_{j \in I_0} K_{ij}$, r_L is an adjustment factor suggested by Fan (1993), r_R is an adjustment factor suggested by Seifert and Gasser (2000), B_m is an interval such as $[0, 0.2]$ that gives the m th block for the blocking estimator, and M is the number of blocks used. For a Gaussian kernel, $r_L = 0$ and for an Epanechnikov kernel, $r_L = 1/n^2$. For a Gaussian kernel, $r_R = 0.35$ and for an Epanechnikov kernel, $r_R = 0.31$.

in a weighted regression. To see this, rewrite

$$\begin{aligned}
\hat{\theta} &= \frac{1}{n_1} \sum_{i=1}^n T_i \left\{ Y_i - \sum_{j=1}^n w(i, j)(1 - T_j)Y_j \right\} \\
&= \frac{1}{n_1} \sum_{i=1}^n T_i Y_i - \sum_{j=1}^n (1 - T_j) Y_j \frac{1}{n_1} \sum_{i=1}^n w(i, j) T_i \\
&\equiv \frac{1}{n_1} \sum_{i=1}^n T_i Y_i - \frac{1}{n_0} \sum_{j=1}^n (1 - T_j) Y_j \bar{w}(j),
\end{aligned} \tag{4}$$

where $\bar{w}(j) = \frac{n_0}{n_1} \sum_{i=1}^n w(i, j) T_i$ is proportional to the average weight that a control observation is given, on average across all treatment observations.¹⁶ Viewing matching estimators as weighted least squares is useful as a means of understanding the relationships among the various estimators used in the literature.

For example, the weight used by kernel matching can be written as

$$\bar{w}(j) = \frac{n_0}{n_1} \sum_{i=1}^n T_i w(i, j) = \frac{\sum_{i=1}^n T_i K_{ij} / \sum_{i=1}^n K_{ij}}{\sum_{j=1}^n (1 - T_j) K_{ij} / \sum_{i=1}^n K_{ij}} \bigg/ \frac{\hat{p}}{1 - \hat{p}}.$$

Ignoring estimation error in the propensity score, $\sum_{i=1}^n T_i K_{ij} / \sum_{i=1}^n K_{ij}$ is a kernel regression estimate of $P(T_i = 1 | p(X_i) = p(X_j))$, which is equivalent to $p(X_j)$.¹⁷ If the kernel in question is symmetric, then $\sum_{i=1}^n (1 - T_i) K_{ij} / \sum_{i=1}^n K_{ij}$ is similarly a kernel regression estimate of $P(T_i = 0 | p(X_i) = p(X_j))$, which is equivalent to $1 - p(X_j)$. Thus, for kernel matching with a symmetric kernel, we have

$$\bar{w}(j) \approx \frac{p(X_j)}{1 - p(X_j)} \bigg/ \frac{p}{1 - p},$$

which is the same as the target parameter of the TOT weight used by reweighting.

This result provides a 3-step interpretation to symmetric kernel matching for the TOT:

¹⁶The matching estimators proposed in the literature require no normalization on the weights involved in the second sum in equation (4). This follows because the matching estimators that have been proposed define the weighting functions $w(i, j)$ in such a way that $\sum_{j \in I_0} w(i, j) Y_j$ is a legitimate average of the controls, in that sense that for every treated unit i , $\sum_{j \in I_0} w(i, j) = \sum_{j=1}^n w(i, j)(1 - T_j) = 1$. This has the important implication that the weights in the second sum of equation (4) automatically add up to one:

$$\frac{1}{n_0} \sum_{j=1}^n (1 - T_j) \bar{w}(j) = \frac{1}{n_0} \sum_{j=1}^n \left\{ (1 - T_j) \left[\frac{n_0}{n_1} \sum_{i=1}^n w(i, j) T_i \right] \right\} = \frac{1}{n_1} \sum_{i=1}^n \left\{ T_i \left[\sum_{j \in I_0} w(i, j) \right] \right\} = \frac{1}{n_1} \sum_{i=1}^n T_i = 1.$$

¹⁷Generally, if X and Y are random variables such that $m(X) = E[Y|X]$ exists, then $E[Y|m(X)] = m(X)$ by iterated expectations.

1. Estimate the propensity score, $\hat{p}(X_i)$
2. For each observation j in the control group, compute $\bar{p}(X_j) = \sum_{i=1}^n T_i K_{ij} / \sum_{i=1}^n K_{ij}$. In words, this is the fraction treated among those with propensity scores near $\hat{p}(X_j)$. Under smoothness assumptions on $p(X_i)$, this will be approximately $\hat{p}(X_j)$.
3. Form the weight $\bar{w}(j) = (\bar{p}(X_j)/(1 - \bar{p}(X_j))) / (\hat{p}/(1 - \hat{p}))$ and run a weighted regression of Y_i on a constant and T_i with weight $W_i = T_i + (1 - T_i)\bar{w}(i)$.

Reweighting differs from this procedure in that, in step 2, it directly sets $\bar{p}(X_j) = \hat{p}(X_j)$. The simulation suggests that this shortcut is effective at improving small sample performance.

C Mixed Methods

We also consider the performance of a “double robust” estimator that combines reweighting with more traditional regression techniques. This procedure is discussed by Robins and Rotnitzky (1995) in the related context of imputation for missing data. Imbens (2004) provides a good introductory treatment.¹⁸

To describe the intuition behind this estimator, we first return to a characterization of reweighting. The essential idea behind reweighting is that in large samples, reweighting ensures orthogonality between the treatment indicator and any possible function of the covariates. That is, for any bounded continuous function $g(\cdot)$,

$$\begin{aligned}
 E[g(X_i)|T_i = 1] &= E\left[g(X_i)\frac{p(X_i)}{1-p(X_i)}\bigg/\frac{p}{1-p}\bigg|T_i = 0\right], \\
 E\left[g(X_i)\frac{p}{p(X_i)}\bigg|T_i = 1\right] &= E\left[g(X_i)\frac{1-p}{1-p(X_i)}\bigg|T_i = 0\right] = E[g(X_i)].
 \end{aligned}
 \tag{5}$$

This implies that the joint distribution of X_i is equal in weighted subsamples defined by $T_i = 1$ and $T_i = 0$, using either TOT or ATE weights.¹⁹ This in turn implies that in the reweighted sample, treatment is *unconditionally* randomized, and estimation can proceed by computing the (reweighted) difference in means, as described in subsection B, above. A standard procedure in estimating the effect of an unconditionally randomized treatment is to include covariates in a regression of the outcome on a constant and a treatment indicator. It is often argued that this procedure improves the precision of estimated treat-

¹⁸See also the recent work by Egel, Graham and Pinto (2008), in which a different and possibly more effective double robust estimator is proposed.

¹⁹Given the standard overlap assumption, this result follows from iterated expectations. A proof for the case of TOT is given in McCrary (2007, fn. 35).

ment effects. By analogy with this procedure, a reweighting estimator may enjoy improved precision if the weighted regression of the outcome on a constant and a treatment indicator is augmented by covariates.

The estimator just described is the double robust estimator. Reweighting computes average treatment effects by running a weighted regression of the outcome on a constant and a treatment indicator. Double robust estimation computes average treatment effects by running a weighted regression of the outcome on a constant, a treatment indicator, and some function of the covariates such as the propensity score.

The gain in precision associated with moving from a reweighting estimator to a double robust estimator is likely modest with economic data.²⁰ However, a potentially important advantage is that the estimator is more likely to be consistent, in a particular sense. Suppose that the model for the treatment equation is misspecified, but that the model for the outcome equation is correctly specified. Then the double robust estimator would retain consistency, despite the misspecification of the treatment equation model.²¹ We implement the double robust estimator by including the estimated propensity score linearly into the regression model, for both ATE and TOT.²²

The double robust estimator is related to another popular estimator that we call a control function estimator. For the case of ATE, the control function estimator is the slope coefficient on a treatment indicator in a regression of the outcome on a constant, the treatment indicator, and functions of the covariates X_i . For the case of TOT, we obtain the control function estimator by running a regression of the outcome on a constant and a cubic in the propensity score, separately by treatment status.²³ For each model, we form predicted values, and compute the average difference in predictions, among the treated observations. This procedure is in the spirit of the older Oaxaca (1973) and Blinder (1973) procedure and is related to the general estimator proposed by Hahn (1998).

²⁰Suppose the goal is to obtain a percent reduction of q in the standard error on the estimated treatment effect. Approximate the standard error of the treatment effect by the spherical variance matrix least squares formula. Then reducing the standard error of the estimated treatment effect by q percent requires reducing the regression root mean squared error by q percent, since the “matrix part” of the standard error is affected only negligibly by the inclusion of covariates, due to the orthogonality noted in equation (5). This requires reducing the regression mean squared error (MSE) by roughly $2q$ percent when q is small. A $2q$ percent reduction in the regression MSE requires that the F -statistic on the exclusion of the added covariates be a very large $2qn/K$, where n is the overall sample size and K is the number of added covariates. Consider one of the strongest correlations observed in economic data, that between log-earnings and education. In a typical U.S. Census file with 100,000 observations, the t-ratio on the education coefficient in a log-earnings regression is about 100 (cf., Card 1999). The formula quoted suggests that including education as a covariate with an outcome of log earnings would improve the standard error on a hypothetical treatment indicator by only 5 percent.

²¹In the case described, the double robust estimator would be consistent, but inefficient relative to a regression-based estimator with no weights, by the Gauss-Markov Theorem.

²²We include the $\hat{p}(X_i)$ rather than X_i because the outcome equation in our DGPs is a function of $p(X_i)$.

²³In simulations not shown, we computed the MSE for the control function estimator in which the propensity score entered in a polynomial of order 1,...,5. The cubic polynomial had the lowest MSE on average across contexts.

D Tuning Parameter Selection

The more complicated matching estimators require choosing tuning parameters. Kernel-based matching estimators require selection of a bandwidth, nearest-neighbor matching requires choosing the number of neighbors, and blocking requires choosing the blocks.

In order to select both the bandwidth h to be used in the kernel-based matching estimators and the number of neighbors to be utilized in nearest neighbor matching, we implement a simple leave-one-out cross-validation procedure that chooses h as

$$h^* = \arg \min_{h \in H} \sum_{i \in I_0} [Y_i - \hat{m}_{-i}(p(X_i))]^2,$$

where $\hat{m}_{-i}(p(X_i))$ is the predicted outcome for observation i , computed with observation i removed from the sample, and $m(\cdot)$ is the non-parametric regression function implied by each matching procedure. For kernel, local linear and ridge matching the bandwidth search grid H is $0.01 \times |\kappa| 1.2^{g-1}$ for $g = 1, 2, \dots, 29, \infty$. For nearest-neighbor matching the grid H is $\{1, 2, \dots, 20, 21, 25, 29, \dots, 53, \infty\}$ for a sample size smaller than 500 and $\{1, 2, 5, 8, \dots, 23, 28, 33, \dots, 48, 60, 80, 100, \infty\}$ for 500 or more observations.²⁴

For the blocking estimator, we first stratify the sample into M blocks defined by intervals of the estimated propensity score. We continue to refine the blocks until within each block we cannot reject the null that the expected propensity score among the treated is equal to the expected propensity score among the controls (Rosenbaum and Rubin 1983, Dehejia and Wahba 1999). In order to perform this test we used a simple t -test with a 99 percent confidence level. Once the sample is stratified, we can compute the average difference between the outcome of treated and control units that belong to each block, $\hat{\tau}_m$. Finally, the blocking estimator computes the weighted average of $\hat{\tau}_m$ across M blocks, where the weights are the proportion of observations in each block, either overall (ATE) or among the treated only (TOT).

E Efficiency Bounds

In analyzing the performance of the estimators we study, it useful to have an idea of a lower bound on the variance of the various estimators for a given model. Estimators which attain a variance lower bound are best, in a specific sense.

We consider two variants of efficiency bounds. The first of these is the Cramér-Rao lower bound, which

²⁴For more details on this procedure see Stone (1974) and Black and Smith (2004) for an application.

can be calculated given a fully parametric model. The semiparametric models motivating the estimators under study in this paper do not provide sufficient detail on the putative DGP to allow calculation of the Cramér-Rao bound. Nonetheless, since we assign the DGP in this study, we can calculate the Cramér-Rao bound using this knowledge. This forms a useful benchmark. For example, we will see that in some models, the variance of a semiparametric estimator is only slightly greater than the Cramér-Rao bound. These are then models in which there is little cost to discarding a fully parametric model in favor of a semiparametric model.

The second efficiency bound we calculate is the semiparametric efficiency bound. This efficiency bound can be understood as the supremum of the Crámer-Rao lower bounds associated with regular parametric submodels.²⁵ If $\check{\theta}$ is an estimator that is regular, \sqrt{n} -consistent for TOT, and semiparametrically efficient, then $\sqrt{n}(\check{\theta} - \theta) \xrightarrow{d} N(0, \text{SEB})$. If $\dot{\theta}$ is an estimator that is regular, \sqrt{n} -consistent for TOT, and does not utilize (correct) parametric knowledge of the joint density for (Y_i, T_i, X_i) , then $\sqrt{n}(\dot{\theta} - \theta) \xrightarrow{d} N(0, V)$ with $V \geq \text{SEB}$. An introductory discussion of the SEB concept is given in Newey (1990).

Hahn (1998, Theorems 1, 2) shows that under selection on observed variables and standard overlap, the SEB is given by

$$SEB_{k/u}^{ATE} = E \left[\frac{\sigma_1^2(X_i)}{p(X_i)} + \frac{\sigma_0^2(X_i)}{1 - p(X_i)} + (\tau(X_i) - \alpha)^2 \right], \quad (6)$$

$$SEB_u^{TOT} = E \left[\frac{\sigma_1^2(X_i)p(X_i)}{p^2} + \frac{\sigma_0^2(X_i)p(X_i)^2}{p^2(1 - p(X_i))} + \frac{p(X_i)}{p^2} (\tau(X_i) - \theta)^2 \right], \quad (7)$$

$$SEB_k^{TOT} = E \left[\frac{\sigma_1^2(X_i)p(X_i)}{p^2} + \frac{\sigma_0^2(X_i)p(X_i)^2}{p^2(1 - p(X_i))} + \frac{p(X_i)^2}{p^2} (\tau(X_i) - \theta)^2 \right], \quad (8)$$

where the subindex $l = k, u$ indicates whether the propensity score is known or unknown and $\sigma_t^2(X_i)$ is the conditional variance of $Y_i(t)$ given X_i .

Reweighting using a nonparametric estimate of the propensity score achieves the bounds in equations (6) and (7), as shown by Hirano et al. (2003) for both the ATE and TOT case. Nearest neighbor matching on covariates using a Euclidean norm also achieves these bounds when the number of matches is large.

²⁵A regular parametric submodel consists of a parametric specification of the DGP. As noted in Hahn (1998), in the context of average treatment effects, for a parameter vector η and a set of functions $f_t(y|x, \eta)$, $p(x, \eta)$, and $f(x, \eta)$ corresponding to the conditional density of $Y_i(t)$ given $X_i = x$, the propensity score, and the marginal density of X_i , the data (Y_i, T_i, X_i) are assumed to be a set of n realizations from a distribution with joint density function $q(y, t, x, \eta_0)$, where

$$q(y, t, x, \eta) = [f_1(y|x, \eta)p(x, \eta)]^t [f_0(y|x, \eta)(1 - p(x, \eta))]^{1-t} f(x, \eta)$$

The supremum is taken over $q(\cdot)$ and is finite under strict overlap and conditional independence (Khan and Tamer 2007).

Abadie and Imbens (2006, Theorem 5) demonstrate this for the case of ATE and the case of TOT follows from the machinery they develop.²⁶ However, nearest neighbor matching is inconsistent when there is more than one continuous covariate to be matched. Efficiency results for other matching estimators are not yet available in the literature.

III Data Generating Process

The DGPs we consider are all special cases of the latent index model

$$T_i^* = \eta + \kappa X_i - u_i, \quad (10)$$

$$T_i = 1(T_i^* > 0), \quad (11)$$

$$Y_i = \beta T_i + \gamma m(p(X_i)) + \delta T_i m(p(X_i)) + \varepsilon_i, \quad (12)$$

where u_i and ε_i are independent of X_i and of each other, $m(\cdot)$ is a curve to be discussed, and $p(X_i)$ is the propensity score implied by the model, or the probability of treatment given X_i . The covariate X_i is taken to be distributed standard normal. Our focus is on cross-sectional settings, so ε_i is independent across i , but potentially heteroscedastic. This is achieved by generating e_i as an independent and identically distributed standard normal sequence and then generating $\varepsilon_i = \psi(e_i p(X_i) + e_i T_i) + (1 - \psi)e_i$, where ψ is a parameter that controls heteroskedasticity.

We consider several different distributional assumptions for the treatment assignment equation residual, u_i . As we discuss in more detail in Sections V and VI below, the choice of distribution for u_i can be relevant to both the finite and large sample performance of average treatment effect estimators. Let the distribution function for u_i be denoted by $F(\cdot)$. Then the propensity score is given by

$$p(X_i) \equiv P(T_i^* > 0) = F(\eta + \kappa X_i). \quad (13)$$

The model given in equations (10) through (12) nests a basic latent index regression model, in which

²⁶Using their equation (13) and the results of the unpublished proof of their Theorem 5, it is straightforward to derive the large sample variance of the k th nearest-neighbor matching estimator for the TOT as

$$SEB_u^{TOT} + \frac{1}{2k} E \left[\frac{\sigma_0^2(X_i)}{p^2} \left(\frac{1}{1-p(X_i)} - (1-p(X_i)) \right) \right] \xrightarrow{k \rightarrow \infty} SEB_u^{TOT} \quad (9)$$

treatment effects vary with X_i but are homogeneous, residuals are homoscedastic, and the conditional expectation of the outcome under control is white noise. The model is flexible, however, and can also accommodate heterogeneous treatment effects, heteroscedasticity, and nonlinear response functions.

Heterogeneity of treatment effects is controlled by the parameter δ in equation (12). When $\delta = 0$, the covariate-specific treatment effects are constant: $\tau(x) = \beta$ for all x . Thus under this restriction the average treatment effect (ATE) and the average effect of treatment on the treated (TOT) both equal β in the population and in the sample.²⁷ When $\delta \neq 0$, the covariate-specific treatment effect, given by $\tau(x) = \beta + \delta m(p(x))$, depends on the covariate and ATE and TOT may differ.²⁸ Heteroscedasticity is controlled by the parameter ψ . When $\psi = 0$, we obtain homoscedasticity. When $\psi \neq 0$, the residual variance depends on treatment as well as on the propensity score. The function $m(\cdot)$ and the parameter γ manipulate the non-linearity of the outcome equation that is common to both treated and non-treated observations.²⁹

We assess the relative performance of the estimators described in Section II in a total of twenty-four different contexts. These different contexts are characterized by four different settings, three different designs, and two different regression functions. We now describe these contexts in greater detail.

The four settings we consider correspond to four different combinations of the parameters in equation (12): β , γ , δ , and ψ . In each of these four settings, we set $\beta = 1$ and $\gamma = 1$. However, we vary the values of the parameters δ and ψ , leading to four combinations of homogeneous and heterogeneous treatment effects and homoscedastic and heteroscedastic error terms. The specific configurations of parameters used in these four settings are summarized below:

Setting	β	γ	δ	ψ	Description
I	1	1	0	0	homogeneous treatment, homoscedastic
II	1	1	1	0	heterogeneous treatment, homoscedastic
III	1	1	0	2	homogeneous treatment, heteroscedastic
IV	1	1	1	2	heterogeneous treatment, heteroscedastic

The two regression functions we consider, $m(\cdot)$, correspond to the functional forms used by Frölich (2004). The first curve considered is a simple linear function. The second curve is nonlinear and rises from

²⁷For a discussion of the distinction between sample and population estimands, see Imbens (2004), for example.

²⁸For a discussion of other estimands of interest, see Heckman and Vytlačil (2005), for example.

²⁹Note that we only consider DGPs in which $\gamma \neq 0$. When $\gamma = 0$, all estimators of the TOT for which the weighting function $\bar{w}(j)$ adds up to 1 can analytically be shown to be finite sample unbiased. When $\gamma \neq 0$, no easy analytical finite sample results are available, and simulation evidence is much more relevant.

around 0.7 at $q = 0$ to 0.8 near $q = 0.4$, where the curve attains its peak, before declining to 0.2 at $q = 1$. The precise equations used for these two regression functions are summarized below:

Curve	Formula	Description
1	$m_1(q) = 0.15 + 0.7q$	Linear
2	$m_2(q) = 0.2 + \sqrt{1-q} - 0.6(0.9-q)^2$	Nonlinear

Finally, the three designs we consider correspond to different combinations of the parameters in equation (10): η and κ . These parameters control degrees of overlap between the densities of the propensity score of treated and control observations as well as different ratios of control to treated units. The specific configurations of parameter values for η and κ are different in Sections IV and VI and are enumerated in those sections.

IV Results: Benchmark Case

We begin by focusing on the case of X_i distributed standard normal and u_i distributed standard Cauchy. As we discuss in more detail below, an initial focus on this DGP allows us to sidestep some important technical issues that arise with poor overlap in propensity score distributions between treatment and control units. We defer discussion of these complications until Sections V and VI. The specific configurations of the parameters η and κ used in these three designs are summarized below:

Design	η	κ	Treated-to-Control Ratio
A	0	0.8	1:1
B	0.8	1	2:1
C	-0.8	1	1:2

An important feature of these DGPs is the behavior of the conditional density functions of the propensity score, $p(X_i)$, conditional on treatment. Figure 1A displays the conditional density of the propensity score given treatment status. This figure features prominently in our discussion, and we henceforth refer to such a figure as an *overlap plot*.

The figures point to several important features of our benchmark DGPs. First, for all three designs considered, the strict overlap assumption is satisfied. As noted by Khan and Tamer (2007), this is a sufficient assumption for \sqrt{n} -consistency of semiparametric treatment effects estimators. Second, the ratio

of the treatment density height to that for control gives the treatment-to-control sample size ratio. From this we infer that it is more challenging to estimate the TOT in design C than in designs A or B. Third, design A is symmetric and estimation of the ATE is no more difficult than estimation of the TOT.

We turn next to an analysis of the results of the simulation. In Section IV.A we assume that the propensity score model is correctly specified, and estimation proceeds using a maximum likelihood Cauchy binary choice model that includes X_i as the sole covariate. In Section IV.B we study the impact of misspecification of the propensity score model on performance.

In both sections IV.A and IV.B, and throughout the paper, we report separate estimates of the bias and the variance of the estimators. In addition, for each estimator we test the hypothesis that the bias is equal to zero, and we test the hypothesis that the variance is equal to the SEB. These choices reflect our view that it is difficult to bound the bias a researcher would face, across the possible DGPs the researcher might confront, unless the estimator is unbiased or nearly so. Bounding the bias is desirable under an objective of minimizing the worst case scenario performance of the estimator, across possible DGPs.

A Correct Parametric Specification of Treatment Assignment

Table 1 examines the performance of our 14 estimators in the Normal-Cauchy model for $n = 100$ and $n = 500$. For ease of exposition, we do not show estimates of the bias and variance for all twenty-four contexts.³⁰ Instead, we summarize these estimates by presenting the simulated root mean square bias (RMSB) and average variance, both overall across the twenty-four contexts and separately for the settings described in Section III.³¹ There are 14 columns, one for each estimator under consideration.

Estimates of the RMSB are presented in the first and second panels of Table 1 for $n = 100$ and $n = 500$, respectively. As an aid to summarizing the results, we additionally perform F-tests of the null hypothesis that the bias is zero jointly across the twenty-four contexts and jointly across the designs and curves in any given setting.³² The value of the F-statistic for the joint test across twenty-four contexts is reported below

³⁰As described above, a *context* here means a bundle of setting, design, and curve. We consider four settings, three designs and two curves.

³¹In the main text, we focus on TOT and report summary tables. A series of appendix tables present summary tables for ATE. Detailed tables for both TOT and ATE, as well as Stata data sets containing all of the replication results, are available at <http://www.econ.berkeley.edu/~jmccrary>.

³²Practically, these tests are implemented as Wald tests using a feasible generalized least squares model for the 240,000 replications less their (context-specific) target parameters. To keep the power of these tests constant across sample sizes, we keep nR constant at one million, where R is the number of replications. This implies 10,000 replications for $n = 100$ and 2,000 replications for $n = 500$. This also spares significant computational expense.

the setting-specific RMSB estimates, and p-values for these F-tests are reported in brackets.³³ The values of the F-statistics for the setting-specific tests are suppressed in the interest of space. For these tests, we place an asterisk next to the RMSB when the hypothesis is rejected at the 1% significance level.

Average variances are presented in the third and fourth panels of Table 1 for $n = 100$ and $n = 500$, respectively. We provide a reference point for these variances using the SEB. Below the average variances we report the percentage difference between the estimated variance and the SEB on average across all twenty-four contexts. We also perform a F-test of the equality of the variance estimates and the SEB, jointly across all twenty-four contexts and separately for each setting.³⁴ The F-statistic for the joint test across all twenty-four contexts is presented below the average percent discrepancy between the variances and the SEBs. For the setting-specific test, we suppress the value of the statistic in the interest of space. For these tests, we place an asterisk next to the average variance when the hypothesis is rejected at the 1% significance level.

We turn now to a discussion of the results, beginning with the evidence on bias for $n = 100$. The results suggest several important conclusions. First, the pair matching, reweighting, double robust, and control function estimators are all approximately unbiased. Of these, IPW1 and IPW2 are probably the least biased, performing even better than pair matching. Double robust seems to acquire slightly greater bias in settings with treatment effect heterogeneity, whereas the other unbiased estimators acquire slightly less. The F-statistics reject the null of zero bias at the 5% level of significance for all estimators except IPW1, IPW2, and control function. Second, all matching estimators that rely upon tuning parameters are noticeably biased. We suspect that this is due to the difficulty of accurate estimation of nonparametric tuning parameters.³⁵ Of these estimators, ridge matching performs best, particularly when the Epanechnikov kernel is used.

For $n = 500$, pair matching, reweighting, double robust and control function remain approximately unbiased. In terms of bias, these estimators perform remarkably similarly for this sample size. For the

³³Logical equivalence of null hypotheses implies that these F-tests can be viewed as (i) testing that all twenty-four biases are zero, (ii) testing that all four setting-specific RMSB are zero, or (iii) testing that the overall RMSB is zero.

³⁴Practically, these tests are implemented as Wald tests using a generalized least squares model for the twenty-four estimated variances less their (context-specific) SEB. The variance of the variance can be approximated quite accurately under an auxiliary assumption that the estimates of the TOT are distributed normally. In that case, the variance of the variance is approximately $2\hat{V}^2/(R - 1)$, where \hat{V} is the sample variance itself and R is the number of replications. See Wishart (1928) and Muirhead (2005, Chapter 3).

³⁵Loader (1999) reports that the rates of convergence of cross validation is $O_p(n^{-1/10})$ which could explain the bad performance of these estimators in small samples. See also, Galdo, Smith and Black (2007) for further discussion on alternative cross-validation methods.

more complicated matching estimators, we see reduced bias in all cases as expected, and local linear and ridge matching become competitive with reweighting with the larger sample size. Although we can still reject the null of no bias, blocking becomes much less biased. The bias of nearest-neighbor and kernel matching remains high in all settings.

When analyzing the performance within settings (see appendix tables) we observe similar patterns of relative performance. First, reweighting, double robust, and control function estimators are all unbiased regardless of the shape of the overlap plots and regardless of the ratio of treated to control observations. Second, treatment effect heterogeneity, homoscedasticity, and nonlinearity of the regression response function all affect relative performance negligibly.

We next discuss the variance results, presented in the bottom half of Table 1. These results reveal several important findings. First, pair matching presents the largest variance of all the estimators under consideration in all four settings, for both $n = 100$ and $n = 500$. Second, for $n = 100$, IPW2, IPW3 and double robust have the lowest variance among unbiased estimators. Once $n = 500$, the SEB is essentially attained by all of the unbiased estimators except for pair matching. Compared to the SEB, IPW3 has on average a variance for $n = 100$ that is 3.5% in excess, IPW2 a variance that is 4% in excess, and double robust a variance that is 6.4% in excess. Once $n = 500$, these percentages decline to 1%, 1.2%, and 1.4%, respectively.³⁶ Third, among the biased estimators, those with highest bias (nearest-neighbor and kernel matching) are the ones that present the lowest variance. On average the variance of these estimators is below the SEB. One interpretation of the results in this regard is that these estimators have a variance which approaches the SEB from below as the sample size increases. This conjecture is particularly plausible since local linear and ridge matching, the least biased among the matching estimators, exhibit variance similar to that of the reweighting estimators.

In sum, our analysis indicates that when good overlap is present and misspecification is not a concern, there is little reason to use an estimator other than IPW2 or perhaps IPW3. These estimators are trivial to program, typically requiring 3 lines of computer code, appear to be subject to minimal bias, and are minimal variance among approximately unbiased estimators. A further consideration is that standard

³⁶Although IPW1 does notably worse in terms of variance than IPW2, its performance is not as bad as has been reported in other studies. For instance, Frölich (2004) reports that in a homoscedastic and homogeneous setting IPW1 has an MSE that is between 150% and 1518% higher than that of pair-matching. The good performance of IPW1 documented in Table 1 is due to the fact that, in the Normal-Cauchy model, there is a vanishingly small probability of having an observation with a propensity score close to 1. It is propensity scores near 1 that generate extreme weights, and it is extreme weights that lead to large variance of weighted means.

errors for these estimators are easy to obtain and accurate (Busso 2008). The same cannot be said of the matching estimators.

B Incorrect Specification of Treatment Assignment

We investigate two different types of misspecification of the propensity score. First, we assume that $p(X_i) = F(\eta + \kappa X_i)$ when in fact the true DGP is $p(X_i) = F(\eta + \kappa X_{1i} + X_{2i} + X_{3i})$ where X_{ji} follows a standard normal distribution and $F(\cdot)$ is a Cauchy distribution. We call this a misspecification in terms of covariates, X_i . This kind of misspecification occurs when the researcher fails to include all confounding variables in the propensity score model. Second, we proceed with estimation as if $p(X_i) = \tilde{F}(\eta + \kappa X_i)$ when in fact the true DGP is $p(X_i) = F(\eta + \kappa X_i)$. In particular, we keep $F(\cdot)$ as the distribution function for the standard Cauchy, but estimate the propensity score with a probit—that is, we assume that $p(X_i) = \Phi(\eta + \kappa X_i)$. We call this a misspecification in terms of the treatment equation residual, u_i .

Results of these investigations are displayed in Table 2. The structure of this table is similar to that of Table 1. Table 2 presents the RMSB and average variance for the 14 estimators in a sample size of 100 under the two types of misspecifications. Covariate misspecification is treated in panels 1 and 3, and distributional misspecification is treated in panels 2 and 4.

The first panel shows that covariate misspecification leads every estimator to become biased in every setting. This is expected and emphasizes the central role of the assumption of selection on observed variables. Unless the unexplained variation in treatment status resembles experimental variation, treatment effects estimators cannot be expected to produce meaningful estimates. These estimators may continue to play a role as descriptive tools, however. The third panel shows that the average variances are always below the SEB, typically by 20% to 30%. Thus, the exclusion of relevant covariates from the propensity score model may lead to precise estimates of the wrong parameter.

We turn next to the results on distributional misspecification, where the DGP continues to have a Cauchy residual on the treatment assignment equation, but the researcher uses a probit model for treatment. The second panel presents results for the bias in this case. In this situation, only pair matching and control function remain unbiased. Double robust is approximately unbiased only in settings of homogeneous treatment effects. The reweighting estimators become biased but are always less biased than the matching estimators. The fourth panel shows that none of the estimators achieve the SEB. Unfortunately, the most robust estimators to misspecification of the propensity score, that is pair matching and control

function, are the ones with the largest variance. Ridge matching and IPW3 are closest to the SEB, differing only by 4% to 6%.

V Problems with Propensity Scores Near Boundaries

The model given in equations (10) to (12) assumes selection on observed variables. As has been noted by many authors, selection on observed variables is a strong assumption. It is plausible in settings where treatment is randomized conditional on the function of the X_i given in (12). However, it may not be plausible otherwise.³⁷ We feel that practitioners appreciate the importance of this assumption.

However, perhaps less widely appreciated than the importance of the selection on observed variables assumption is the importance of overlap assumptions. As emphasized by Khan and Tamer (2007), the model outlined in equations (10) to (12)—while quite general and encompassing all of the existing simulation evidence on performance of estimators for ATE and TOT under unconfoundedness of treatment—does not necessarily admit a \sqrt{n} -consistent semiparametric estimator for ATE or TOT. In particular, the standard overlap assumption that $0 < p(X_i) < 1$ is not sufficient to guarantee \sqrt{n} -consistency, whereas the strict overlap assumption that $\xi < p(X_i) < 1 - \xi$ for some $\xi > 0$ is. However, the strict overlap assumption can be violated by the model in equations (10) to (12). For example, Khan and Tamer (2007) note that \sqrt{n} -consistency is violated in the special case of X_i and u_i both distributed standard normal, with $\eta = 0$ and $\kappa = 1$. The following proposition sharpens this important result.

Proposition. *Under the model specified in equations (10) to (12), with X_i and u_i distributed standard normal, boundedness of the conditional variance of e_i given X_i , and boundedness of the function $m(\cdot)$, \sqrt{n} -consistent semiparametric estimators for ATE and TOT are available when $-1 < \kappa < 1$. For $|\kappa| \geq 1$, no \sqrt{n} -consistent semiparametric estimator can exist.*

The proof of this result (available from the authors upon request) proceeds by showing that the SEB is finite if and only if $|\kappa| < 1$. The computations are tedious but elementary.³⁸

The intuition behind this result is that when κ approaches 1, an increasing mass of observations have propensity scores near 0 and 1. This leads to fewer and fewer comparable observations and an “effective” sample size smaller than n . This is important, because it implies potentially poor finite sample properties

³⁷We have emphasized the strength of this assumption by writing the selection on observed variables assumption differently than is typical in the literature (see Section II; cf., Imbens (2004)).

³⁸The idea is to use the bounds $\frac{t}{1+t^2}\phi(t) < 1 - \Phi(t) < \frac{1}{t}\phi(t)$, valid for any $t > 0$ and highly accurate for t above, say, 4. We thank Matias Cattaneo for suggesting the use of this approach to solve a related problem.

of semiparametric estimators, in contexts where κ is *near* 1. This is confirmed by the simulation results presented in Section VI, below.

Assuming both X_i and u_i are distributed continuous, the extent to which the propensity score fluctuates near 0 and 1 is given by the functional form of the density of the propensity score

$$f_{p(X_i)}(q) = \frac{1}{|\kappa|} g\left(\frac{(F^{-1}(q) - \eta)}{\kappa}\right) / f(F^{-1}(q)), \quad (14)$$

where $F(\cdot)$ and $f(\cdot)$ are the distribution and density functions, respectively, for u_i , and $g(\cdot)$ is the density function for X_i .³⁹ For q near one (zero), $F^{-1}(q)$ is of extremely large magnitude and positive (negative) sign. Thus, the functional form given makes it clear that when η and κ take on modest values, the density of $p(X_i)$ is expected to be zero at one (zero) when the positive (negative) tail of $f(\cdot)$, the density for the residual, is fatter than that of $g(\cdot)$, the density for the covariate. When the tails of the density for the residual are too thin relative to those of the covariate, the density of $p(X_i)$ near zero can take on positive values, in which case the SEB is guaranteed to be infinite and \sqrt{n} -consistency is lost.

This is a useful insight, because the behavior of the propensity score density near the boundary can be inferred from data. In fact, many economists already analyze density estimates for the estimated propensity score, separately by treatment status (see, for example, Figure 1 of Black and Smith (2004)). As discussed above, we refer to this graphical display as an *overlap plot*. The unconditional density function is simply a weighted average of the two densities presented in an overlap plot. Thus, the behavior of the unconditional density near the boundaries can be informally assessed using a graphical analysis that is already standard in the empirical literature.⁴⁰ When the overlap plot shows no mass near the corners, semiparametric estimators enjoy \sqrt{n} -consistency. When the overlap plot shows strictly positive height of the density functions at 0 (for ATE) or 1 (for ATE or TOT), no \sqrt{n} -consistent semiparametric estimator exists. In the intermediate case, where the overlap plot shows some mass near the corners, but where the height of the density at 0 or 1 is nonetheless zero, \sqrt{n} -consistent estimators may or may not be available.⁴¹

³⁹The equation in the display also holds when X_i is a vector. In that case, the density of a linear combination of the vector X_i plays the role of the scalar X_i considered here. Suppose the linear combination has distribution function $G(\cdot)$ and density function $g(\cdot)$. Then the density for the propensity score is as is given in the display, with $\kappa = 1$. Note as well that the density of the propensity score among the treated and control is given by $f_{p(X_i)|T_i=1}(q) = \frac{q}{p} f_{p(X_i)}(q)$ and $f_{p(X_i)|T_i=0}(q) = \frac{1-q}{p} f_{p(X_i)}(q)$, respectively.

⁴⁰Because the behavior of the density at the boundaries is the object of primary interest, it is best to avoid standard kernel density routines in favor of histograms or local linear density estimation (see McCrary (2008) for references).

⁴¹As the proposition above clarifies, \sqrt{n} -consistency is available, despite mass near the corners, when the covariate and treatment equation residuals are distributed standard normal. It is not yet known whether \sqrt{n} -consistency is always attainable when there is mass near the corners, but zero height to the density function of $p(X_i)$ in the corners.

To appreciate the problems with applying standard asymptotics to the semiparametric estimators studied here in situations with propensity scores near the boundaries, we turn now to a sequence of DGPs indexed by κ and inspired by the Proposition. Let the DGP be given by equations (10) to (12), with X_i , e_i , and u_i each distributed mutually independent and standard normal, with $\gamma = \delta = \psi = \eta = 0$, with κ ranging from 0.25 to 1.75. This DGP has homogeneous treatment effects, homoscedastic residuals of variance 1, and probability of treatment equal to 0.5.

For this DGP, $\gamma = 0$ and IWP2 for TOT is finite sample unbiased, but inefficient. The efficient estimator is the coefficient on treatment in a regression of the outcome on a constant and the treatment indicator. It is thus easy to show that the Cramér-Rao bound is 4, regardless of the value of κ . When the SEB is close to the Cramér-Rao bound, there is little cost to using a semiparametric estimator. When there is quite good overlap, such as $\kappa = 0.25$, the SEB is in fact scarcely larger than 4 and there is little cost associated with avoiding parametric assumptions on the outcome equation. However, as problems with overlap worsen, the discrepancy between the SEB and the Cramér-Rao bound diverges. The cost of avoiding parametric assumptions on the outcome equation thus becomes prohibitive as κ increases in magnitude.

To convey a sense of the way in which an infinite SEB would manifest itself in an actual data set, Figure 2 shows the evolution of the overlap plot as κ increases. When $\kappa = 1$, the conditional densities are straight lines akin to a supply-demand graph from an undergraduate textbook. For $\kappa < 1$, the values of the conditional densities at the corners are zero. For $\kappa > 1$, the values of the conditional densities at the corners are positive and grow in height as κ increases.

Applying standard asymptotics to this sequence of DGPs suggests that, for $\kappa < 1$, IPW2 and pair matching estimates of the TOT have normalized large sample variances of

$$nV_{IPW2} = \frac{1}{p} + \frac{1}{p} E \left[\frac{p(X_i)^2}{p(1-p(X_i))} \right] > 4, \quad (15)$$

$$nV_{PM} = nV_{IPW2} + \frac{1}{2} \left(1 + \frac{1}{p} E \left[\frac{p(X_i)}{1-p(X_i)} \right] \right) > 4 + \frac{3}{2}. \quad (16)$$

The variance expressions are close to 4 and $4+3/2$ for moderate values of κ but are much larger for large values of κ .⁴² Indeed, the Proposition implies that both nV_{IPW2} and nV_{PM} diverge as κ approaches 1.⁴³

We next examine the accuracy of these large sample predictions by estimating the variance of IPW2

⁴²The inequalities follow from Jensen's inequality and from the fact that $p = 0.5$ for these DGPs.

⁴³The percent increase of nV_{PM} over nV_{IPW2} is between 37.5 percent (when $\kappa = 0$) and 25 percent (when κ approaches 1) and declines monotonically in the magnitude of κ .

and pair matching for each value of κ .⁴⁴ Figure 3 presents the estimated standard deviation of these estimators as a function of κ and show that the quality of the large sample predictions depends powerfully on the value of κ .⁴⁵ For example, for κ below 0.7, the large sample predicted variances are generally accurate, particularly for IPW2. However, for $\kappa = 0.9$, the large sample predicted variances are markedly above the empirical variances for both estimators and the discrepancy grows rapidly as κ approaches 1, with the large sample variances diverging despite modest empirical variances. Roughly speaking, viewed as a function of κ , the standard deviations of IPW2 and pair matching are both linear to the right of $\kappa = 0.7$, with different slopes. The pattern of the variances is consistent with what would be expected if the variance of pair matching and IPW2 were proportional to the inverse of $n^{c_1+c_2\kappa}$, with possibly different coefficients c_1 and c_2 for the two estimators. Under this functional form restriction on the variances, it is possible to estimate c_1 and c_2 using regression. Define $Y_{g\kappa}$ as $\ln(\hat{V}_{100}/\hat{V}_{500})/\ln(5)$ for $g = 1$ and as $\ln(\hat{V}_{500}/\hat{V}_{1000})/\ln(2)$ for $g = 2$, where \hat{V}_n is the estimated variance for sample size n . Then note that under the functional form restriction on the variances, $Y_{g\kappa} \approx c_1 + c_2\kappa$. Thus, a simple method for estimating c_1 and c_2 is a regression of $Y_{g\kappa}$ on a constant and κ .⁴⁶ For both IPW2 and pair matching, we have 26 observations on $Y_{g\kappa}$, 13 for $g = 1$ and 13 for $g = 2$. For IPW2, the regression described has an R-squared of 0.93 and constant and slope coefficients (standard errors) of 1.19 (0.02) and -0.39 (0.02), respectively. For pair matching, the R-squared is 0.94 and the constant and slope coefficients (standard errors) are 1.15 (0.02) and -0.33 (0.02), respectively. We report these results not because we believe that the scaling on the variance is of the form $n^{c_1+c_2\kappa}$, but to emphasize our sense that the correct scaling is smooth in κ .⁴⁷

These results create a strong impression that the asymptotic sequences used in the large sample literature may be accurate in settings of good overlap, but are likely inaccurate in settings of poor overlap. The performance of these two estimators does not seem to degrade discontinuously when κ exceeds one, but rather seems to degrade smoothly as κ approaches one.

Failure to satisfy the strict overlap assumption can also lead to bias in semiparametric estimators. The

⁴⁴We use 2,000 replications.

⁴⁵Interestingly, large sample predictions appear much more accurate for IPW2 than for pair matching.

⁴⁶Weights improve power since the outcome is more variable for $g = 2$ than for $g = 1$. In particular, the delta method and Wishart approximations suggest that the standard deviation of the outcome is approximately $\sqrt{4/2000}/\ln(5)$ for $g = 1$ and $\sqrt{4/2000}/\ln(2)$ for $g = 2$.

⁴⁷However, it is interesting to note that these regressions can be viewed as minimum chi-square estimates. This approach allows for a statistical test of the functional form restriction that the variances are proportional to the inverse of $n^{c_1+c_2\kappa}$. The test takes the form of the minimized quadratic form, or in this case the (weighted) sum of squared residuals. The test statistic is distributed chi-square with 24 degrees of freedom. For IPW2, this test statistic is 28.4 and for pair matching it is 20.5 (95 percent critical value 36.4).

sign and magnitude of the bias will be difficult to infer in empirical work. Consider again the model in equations (10) to (12), with $\eta = 0$, $\beta = 1$, $\gamma = 0$, and $m(q) = q$. In this DGP, when $\delta = 0$, IPW2 for ATE is finite sample unbiased regardless of the value of κ . When $\delta = 1$, the treatment effect is positively correlated with the propensity score and IPW2 for ATE may be biased. Similarly, when $\delta = -1$, the treatment effect is negatively correlated with the propensity score and IPW2 for ATE may be biased.

Figure 4 shows the bias of IPW2 for ATE as a function of κ for $\delta = 0$, $\delta = 1$, and $\delta = -1$. The figure confirms that when $\delta = 0$, large values of κ do not compromise the unbiasedness of IPW2. However, when $\delta \neq 0$, large values of κ lead to bias. Importantly, when overlap is good, IPW2 is unbiased regardless of the value of δ .

VI Results: Boundary Problems

In order to focus attention on how the estimators perform when the strict overlap condition is close to being violated, we turn now to an analysis of a DGP that is a minor modification of that described in Section IV, above. Instead of generating u_i as independent draws from the standard Cauchy distribution, we generate u_i as independent draws from the standard normal distribution. We manipulate the parameters η and κ in the treatment equation (10) to mimic three designs from the influential study of Frölich (2004). These parameters are summarized below:

Design	η	κ	Treated-to-Control Ratio
A	0	0.95	1:1
B	0.3	-0.8	3:2
C	-0.3	0.8	2:3

Figure 1B shows the overlap plot implied by these designs. Each of these designs is consistent with standard overlap, but none are consistent with strict overlap. This figure shows that having many control observations per treated observations does not imply the validity of the strict overlap condition. For example, design A is closer to violating the strict overlap assumption than design C is, even though the ratio of treated to control observation is higher in the former than in the latter.

A Simulation Results with Boundary Problems

In Table 3 we explore estimator performance in DGPs that are close to violating the strict overlap condition. The structure of the table is identical to that of Table 1, but the DGPs correspond to the Normal-Normal model, rather than the Normal-Cauchy model.

The results in the table support several conclusions. First, when $n = 100$, nearly all estimators are biased in all settings. The exceptions are the control function and double robust estimators in homogeneous treatment effect settings. These two estimators impose parametric assumptions on the outcome equation. This allows for extrapolation from the region of common support to the region over which there are treated observations but no controls. Second, although reweighting estimators are biased with $n = 100$, they become unbiased when $n = 500$. This raises the possibility that, for a good finite sample performance, a larger sample size is required for DGPs with poor overlap that is nonetheless technically sufficient to guarantee \sqrt{n} -consistency. Third, aside from pair matching, the magnitude of the bias of matching estimators is between two and five times that of the reweighting estimators, and they remain biased even for $n = 500$. Pair matching is biased for $n = 100$ and nearly unbiased for $n = 500$.⁴⁸ The third and fourth panel show that the variance of all estimators is much higher than in the case in which we satisfy the strict overlap assumption, even though none of the designs imply an infinite SEB. For all estimators we reject the null that the variance equals the SEB in every setting. Contrary to the case of strict overlap analyzed in the preceding section, this holds true for both $n = 100$ and $n = 500$. The variance of all the estimators is on average below the SEB.

In sum, in settings of poor overlap, semiparametric estimators of average treatment effects do not perform well for $n = 100$. Once $n = 500$, the pair matching, reweighting, double robust, and control function estimators show acceptable bias, but only IPW1 has bias small enough that we fail to reject the null of zero bias. The variance of semiparametric estimators is hard to assess in settings of poor overlap, since neither the SEB nor other large sample approximations form acceptable benchmarks. However, considering both bias and variance and performance for $n = 100$ and $n = 500$, the best estimators in settings with poor overlap appear to be IPW2, IPW3, and double robust.

⁴⁸The sign of the bias of the TOT depends on the shape of the outcome equation. An outcome equation that is increasing (decreasing) in the propensity score like curve 1 (curve 2) implies that the bias will be more positive (negative) the closer we are to violating the strict overlap condition because we have too many treated observations and too few controls at the right end of the distribution of the propensity score (see appendix). The bias is not related to the overall ratio of treated per controls units in the sample. The bias of all the estimators tends to be of the same order of magnitude in the three designs.

B Trimming

In many empirical applications, researchers encounter a subset of observations whose propensity scores do not have common support. Such a finding is expected when the strict overlap condition is violated, although it can also occur in finite samples when strict overlap is satisfied in the population. Confronted by lack of common support, many researchers resort to trimming rules. These sample selection rules involve dropping individuals from the treatment group who have no counterparts in the control group with similar propensity scores (for TOT).⁴⁹ Trimming aims at ensuring validity of the common support assumption in the subset of observations that are not trimmed. See Heckman, Ichimura and Todd (1998a), Smith and Todd (2005), and Crump, Hotz, Imbens and Mitnik (2007a) for discussion. There are several trimming methods that have been proposed in the literature. Little is known about their effect on the performance of semiparametric estimators.

As noted by Heckman et al. (1998a), reweighting and matching *at best* correct for bias for the subsample of individuals whose propensity scores have common support. For this reason, trimming is only expected to work in situations of treatment effect homogeneity, simply because the treatment effect can be estimated anywhere on the support of the propensity score. Dropping observations will make the estimator more inefficient but the bias is expected to decrease because we will be estimating the counterfactual mean only in regions in which both treated and control units are available. However, if the treatment effect is heterogeneous, and more importantly, if the heterogeneity occurs precisely in the part of the support for which we do not have both treated and control observations, then trimming will not be a solution.⁵⁰ In those type of situations the researcher might need to redefine the estimand (see Crump et al. 2006) paying a cost in terms of having a result with less external validity or resort to fully parametric models—which will typically only be effective if the full parametric model is correctly specified.

We analyze the effectiveness of the four trimming rules reviewed in Crump et al. (2006):

1. Let $D_i^{ATE} = \mathbf{1}(\hat{a} < \hat{p}(X_i) < \hat{b})$ and $D_i^{TOT} = \mathbf{1}(\hat{p}(X_i) < \hat{b})$ setting \hat{b} to be the k th largest propensity score in the control group and \hat{a} to be the k th smallest propensity score in the treatment group. Then we compute the estimators on the subsample for which $D_i^{TOT} = 1$ (or $D_i^{ATE} = 1$). This rule was proposed by Dehejia and Wahba (1999).
2. Heckman et al. (1996, 1998) and Heckman, Ichimura, Smith and Todd (1998) propose discarding

⁴⁹Trimming in the case of estimation of the ATE is similar, but individuals from both the treatment and the control group are deleted.

⁵⁰An alternative to trimming is to compute bounds for the treatment effects. This possibility was advocated by Lechner (2001) in the context of matching estimators of treatment effects.

observations for which the conditional density of the propensity score is below some threshold. Let $D_{0i}(c) = \mathbf{1}(\widehat{f}_{\widehat{p}(X_i)|T_i=0} < c)$ and $D_{1i}(c) = \mathbf{1}(\widehat{f}_{\widehat{p}(X_i)|T_i=1} < c)$ where c is a tuning parameter, and $\widehat{f}_{\widehat{p}(X_i)|T_i=1}$ and $\widehat{f}_{\widehat{p}(X_i)|T_i=0}$ are kernel density estimates (with Silverman’s rule as a bandwidth selector). Then, following Smith and Todd (2005), fix a quantile $q = 0.02$ and consider the J observations with positive densities $\widehat{f}_{\widehat{p}(X_i)|T_i=1}$ and $\widehat{f}_{\widehat{p}(X_i)|T_i=0}$. Rank all the values of $\widehat{f}_{\widehat{p}(X_i)|T_i=1}$ and $\widehat{f}_{\widehat{p}(X_i)|T_i=0}$ and drop units with a density less than or equal to c_q , where c_q is the largest real number such that $\frac{1}{2J} \sum_{i=1}^J [D_{0i}(c_q) + D_{1i}(c_q)] \leq q$ for the ATE. For the TOT we can proceed in a similar fashion but only using $\widehat{f}_{\widehat{p}(X_i)|T_i=1}$.

3. Ho, Imai, King and Stuart (2007) define the common support region as the convex hull of the propensity scores used by pair matching.
4. Finally, Crump et al. (2007a) propose discarding all units with an estimated propensity score outside the interval $[0.1, 0.9]$ for the ATE and $[0, 0.9]$ for the TOT.

In Table 4 we study whether, in a DGP that is close to violating the strict overlap assumption, trimming succeeds in reducing the bias. As expected, the double robust and control function estimators stay unbiased in homogeneous settings, but trimming increases the bias of those estimators in heterogeneous settings. Trimming rules 1 and 4 seem to lead to unbiasedness of reweighting and pair matching in settings with a homogeneous treatment effect. These rules also reduce the bias of all the matching estimators. Trimming rule 3 only works with pair matching and to a lesser extent with ridge matching. Trimming rule 2 does not seem to work with $n = 100$. This may not be surprising since this rule requires estimating the conditional density of the propensity score with very few observations.

In Table 5 we present the effect of trimming on the variance of the estimators. Rules 1 and 4 reduce the variance of IPW estimators and of local linear and ridge matching. Surprisingly, the variance of the other matching estimators seem to be basically unaffected by any of the trimming rules.

VII Conclusion

In this paper, we assess the finite sample properties of semiparametric estimators of treatment effects using simulated cross-sectional data sets of size 100 and 500. These estimators include many currently popular approaches, including reweighting, double robust, control function, and matching.

The simulation evidence suggests that when there is good overlap in the distribution of propensity scores for treatment and control units, reweighting estimators are preferred on bias grounds and attain the semiparametric efficiency bound, even for samples of size 100. The double robust estimator can be thought of as regression adjusted reweighting and performs slightly worse than reweighting when there is

good overlap, but slightly better when there is poor overlap. Control function estimators perform well only for samples of size 500. Matching estimators perform worse than reweighting if preferences over bias and variance are lexicographic and if good performance for $n = 100$ is required.⁵¹ If there is enough data, then local linear or ridge matching may be competitive with reweighting. The difficulty of the more complicated matching estimators is potentially related to the difficulty of accurate finite sample selection of tuning parameters.

When overlap in the distribution of propensity scores for treatment and control units is close to failing, the semiparametric estimators studied here do not perform well. This difficulty can be inferred from the available large sample results in the literature (Hirano et al. 2003, Abadie and Imbens 2006, Khan and Tamer 2007). We also show that the standard asymptotic arguments used in the large sample literature provide poor approximations to finite sample performance in cases of near failure of overlap. However, our qualitative conclusion is the same as that reached by Khan and Tamer (2007), who note that the semiparametric estimators considered here are on a sound footing only when there is strict overlap in the distribution of propensity scores.

In empirical applications, economists confronting problems with overlap often resort to trimming schemes, in which some of the data are discarded after estimation of the propensity score. We simulate the performance of the estimators studied in conjunction with four trimming rules discussed in the literature. None of these procedures yield good performance unless there is homogeneity in treatment effects along the dimension of the propensity score.

What is then to be done in empirical work in which problems with overlap are suspected? First, to assess the quality of overlap, we recommend a careful examination of the overlap plot, possibly focused on histograms and possibly involving smoothing using local linear density estimation. Second, if overlap indeed appears to be a problem, we recommend analysis of subsamples based on *covariates* to determine if there are subsamples with good overlap. For example, in some settings, it could occur that problems with overlap stem from one particular subpopulation that is not of particular interest. Analyzing subsamples based on *covariates* is likely to work better than analyzing subsamples based on quantiles of an estimated

⁵¹If preferences over bias and variance are not lexicographic, then some of the biased matching estimators may be preferred to reweighting. We caution, however, that the data generating processes we consider may not represent those facing the economist in empirical applications. In empirical applications, the bias could be of lesser, or greater, magnitude than suggested here, in which case the economist's preference ranking over estimators could be different than that suggested by a literal interpretation of the simulation evidence. Our own preferences over bias and variance lean towards lexicographic because we have a taste for estimators that minimize the maximum risk over possible data generating processes.

propensity score. Third, if there is no obvious subpopulation displaying good overlap, we recommend that the economist consider parametric assumptions on the outcome equation. Semiparametric estimators work well in this context when there is good overlap. When overlap is poor, however, these estimators are highly variable, biased, and subject to nonstandard asymptotics. In settings with poor overlap, the motivation for semiparametric estimation is poor and the most effective methods are likely parametric approaches such as those commonly employed in the older Oaxaca (1973) and Blinder (1973) literature.

References

- Abadie, Alberto and Guido W. Imbens, "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, January 2006, 74 (1), 235–267.
- and — , "On the Failure of the Bootstrap for Matching Estimators," *Econometrica*, forthcoming 2008.
- Black, Dan A. and Jeffrey A. Smith, "How Robust is the Evidence on the Effects of College Quality? Evidence From Matching," *Journal of Econometrics*, July-August 2004, 121 (1-2), 99–124.
- Blinder, Alan S., "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, Fall 1973, 8, 436–455.
- Busso, Matias, "A GMM Estimation Approach to Reweighting Estimation," *Unpublished manuscript, University of Michigan*, 2008.
- Card, David, "The Effect of Unions on the Structure of Wages: A Longitudinal Analysis," *Econometrica*, 1996, 64, 957–979.
- Card, David E., "The Causal Effect of Education on Earnings," in Orley Ashenfelter and David E. Card, eds., *The Handbook of Labor Economics*, Vol. 3A, Amsterdam: Elsevier, 1999.
- Chen, Xiaohong, Han Hong, and Alessandro Tarozi, "Semiparametric Efficiency in GMM Models with Auxiliary Data," *Annals of Statistics*, forthcoming 2008.
- Crump, Richard K., V. Joseph Hotz, Guido W. Imbens, and Oscar A. Mitnik, "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Unpublished manuscript, UCLA* 2007.
- , — , — , and — , "Nonparametric Tests for Treatment Effect Heterogeneity," *Review of Economics and Statistics*, forthcoming 2007.
- Dehejia, Rajeev H. and Sadek Wahba, "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs," in "Econometric Methods for Program Evaluation," Cambridge: Rajeev H. Dehejia, Ph.D. Dissertation, Harvard University, 1997, chapter 1.
- and — , "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, December 1999, 94 (448), 1053–1062.
- DiNardo, John E., Nicole M. Fortin, and Thomas Lemieux, "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, September 1996, 64 (5), 1001–1044.
- Egel, Daniel, Bryan S. Graham, and Cristine Campos de Xavier Pinto, "Inverse Probability Tilting and Missing Data Problems," *NBER # 13981*, April 2008.
- Fan, Jianqing, "Local Linear Regression Smoothers and Their Minimax Efficiencies," *Annals of Statistics*, March 1993, 21 (1), 196–216.
- Freedman, David A. and Richard A. Berk, "Weighting Regressions by Propensity Scores," *Evaluation Review*, 2008, 32 (4), 392–409.
- Frölich, Markus, "Finite-Sample Properties of Propensity-Score Matching and Weighting Estimators," *Review of Economics and Statistics*, February 2004, 86 (1), 77–90.

- Galdo, Jose, Jeffrey A. Smith, and Dan Black, “Bandwidth Selection and the Estimation of Treatment Effects with Unbalanced Data,” Unpublished manuscript, University of Michigan 2007.
- Hahn, Jinyong, “On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects,” *Econometrica*, March 1998, *66* (2), 315–331.
- Haviland, Amelia M. and Daniel S. Nagin, “Causal Inferences with Group Based Trajectory Models,” *Psychometrika*, September 2005, *70* (3), 1–22.
- Heckman, James J. and Edward Vytlacil, “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” *Econometrica*, 2005, *73* (3), 669–738.
- and R. Robb, “Alternative Methods for Evaluating the Impact of Interventions,” in James J. Heckman and R. Singer, eds., *Longitudinal Analysis of Labor Market Data*, Cambridge University Press Cambridge 1985.
- , Hidehiko Ichimura, and Petra Todd, “Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme,” *Review of Economic Studies*, October 1997, *64* (4), 605–654.
- , —, and —, “Matching as an Econometric Evaluation Estimator,” *Review of Economic Studies*, April 1998, *65* (2), 261–294.
- , —, Jeffrey Smith, and Petra Todd, “Characterizing Selection Bias Using Experimental Data,” *Econometrica*, September 1998, *66* (5), 1017–1098.
- , Sergio Urzua, and Edward Vytlacil, “Understanding Instrumental Variables in Models with Essential Heterogeneity,” *Review of Economics and Statistics*, August 2006, *88* (3), 389–432.
- Hirano, Keisuke, Guido W. Imbens, and Geert Ridder, “Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score,” *Econometrica*, July 2003, *71* (4), 1161–1189.
- Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart, “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference,” *Political Analysis*, August 2007, *15*, 199–236.
- Horvitz, D. and D. Thompson, “A Generalization of Sampling Without Replacement from a Finite Population,” *Journal of the American Statistical Association*, 1952, *47*, 663–685.
- Imbens, Guido W., “Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,” *Review of Economics and Statistics*, February 2004, *86* (1), 4–29.
- Johnston, Jack and John E. DiNardo, *Econometric Methods*, McGraw-Hill, 1996.
- Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman, “Moving to Opportunity in Boston: Early Results of a Randomized Mobility Experiment,” *Quarterly Journal of Economics*, May 2001, *116* (2), 607–654.
- Kent, David and Rodney Hayward, “Subgroup analyses in clinical trials.,” *New England Journal of Medicine*, Mar 2008, *358* (11), 1199.
- Khan, Shakeeb and Elie Tamer, “Irregular Identification, Support Conditions, and Inverse Weight Estimation,” Unpublished manuscript, Northwestern University 2007.

- Lechner, Michael, "A Note on the Common Support Problem in Applied Evaluation Studies," Discussion Paper N2001-01, Universität St. Gallen 2001.
- , "A Note on Endogenous Control Variables in Evaluation Studies," Discussion Paper N2005-16, Universität St. Gallen 2005.
- Loader, Clive R., "Bandwidth Selection: Classical or Plug-In?," *The Annals of Statistics*, Apr 1999, *27* (2), 415–438.
- Lunceford, Jared K. and Marie Davidian, "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study," *Statistics in Medicine*, 15 October 2004, *23* (19), 2937–2960.
- McCrary, Justin, "The Effect of Court-Ordered Hiring Quotas on the Composition and Quality of Police," *American Economic Review*, March 2007, *97* (4), 318–353.
- , "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, February 2008, *142* (2).
- Muirhead, Robb J., *Aspects of Multivariate Statistical Theory*, Hoboken: John Wiley and Sons, 2005.
- Newey, Whitney, "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, April-June 1990, *5* (2), 99–135.
- Oaxaca, Ronald, "Male–Female Wage Differentials in Urban Labor Markets," *International Economic Review*, 1973, *14*, 693–709.
- Robins, James M. and Andrea Rotnitzky, "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, March 1995, *90* (429), 122–129.
- , —, and Lue Ping Zhao, "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, September 1994, *89* (427), 846–866.
- Rosenbaum, Paul R. and Donald B. Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, April 1983, *70* (1), 41–55.
- Seifert, Burkhardt and Theo Gasser, "Data Adaptive Ridging in Local Polynomial Regression," *Journal of Computational and Graphical Statistics*, June 2000, *9* (2), 338–360.
- Smith, Jeffrey A. and Petra Todd, "Does Matching Overcome Lalonde's Critique of Nonexperimental Estimators?," *Journal of Econometrics*, September 2005, *125* (1–2), 305–353.
- Stone, Mervyn, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society, Series B*, 1974, *36* (2), 111–147.
- Todd, Petra, "Matching Estimators," in P. Newman, M. Milgate, and J. Eatwell, eds., *The New Palgrave—A Dictionary of Economics*, Vol. forthcoming, New York: Macmillan, 2007.
- Wishart, John, "The Generalised Product Moment Distribution in Samples from a Normal Multivariate Population," *Biometrika*, July 1928, *20A* (1/2), 32–52.
- Wooldridge, Jeffrey M., *Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press, 2002.

Zhao, Zong, “Using Matching to Estimate Treatment Effects: Data Requirements, Matching Metrics, and Monte Carlo Evidence,” *Review of Economics and Statistics*, February 2004, 86 (1), 91–107.

— , “Sensitivity of Propensity Score Method to the Specifications,” *Economics Letters*, 2008, 98, 309–319.

Figure 1.A: Overlap Plots, by design (Normal-Cauchy model)

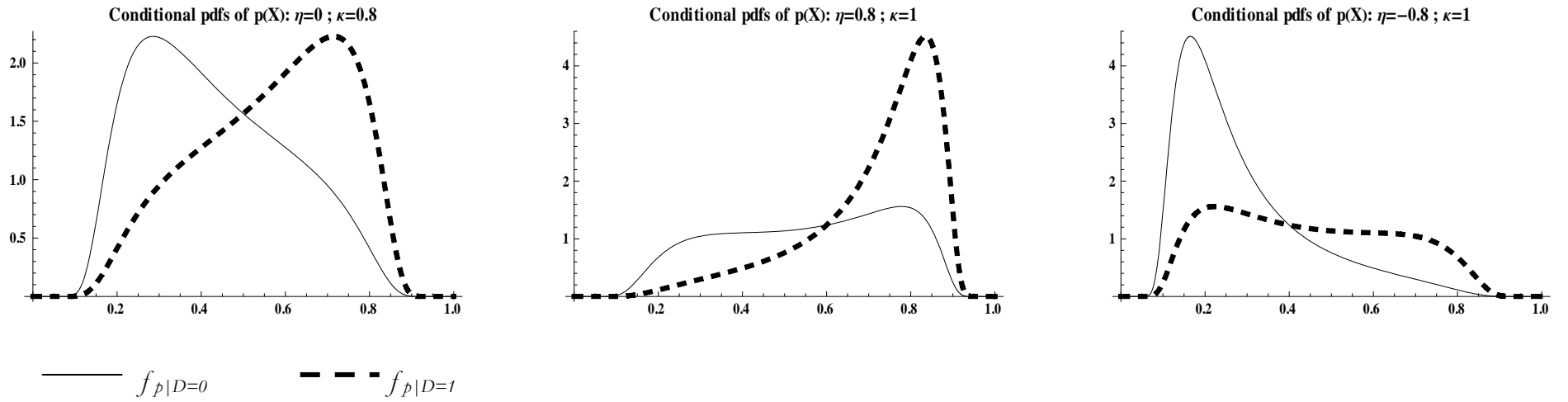


Figure 1.B: Overlap Plots, by design (Normal-Normal model)

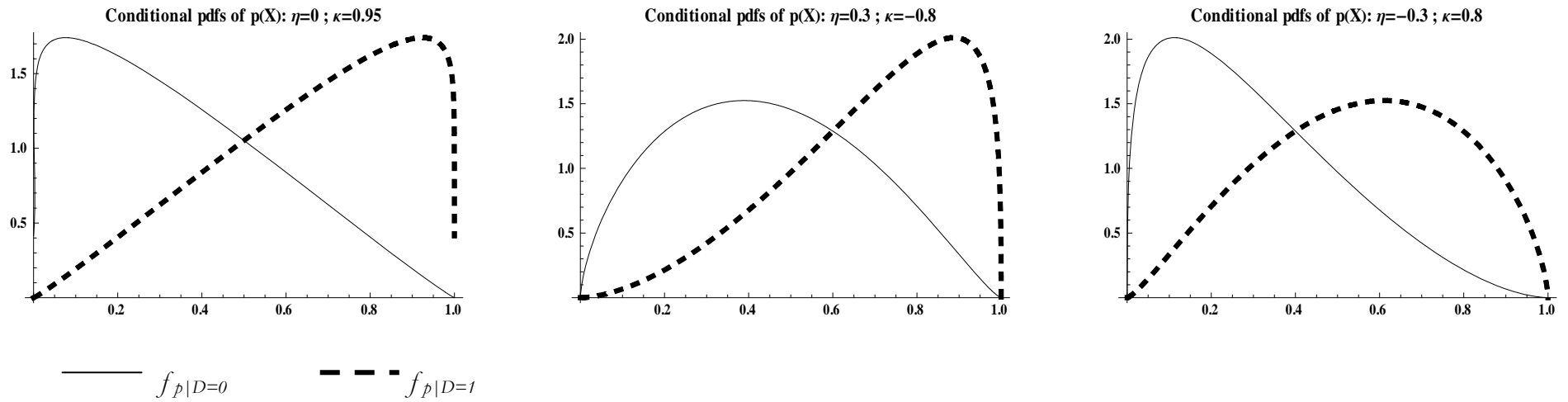


Figure 2: Overlap Plots, by design (Normal-Normal model)

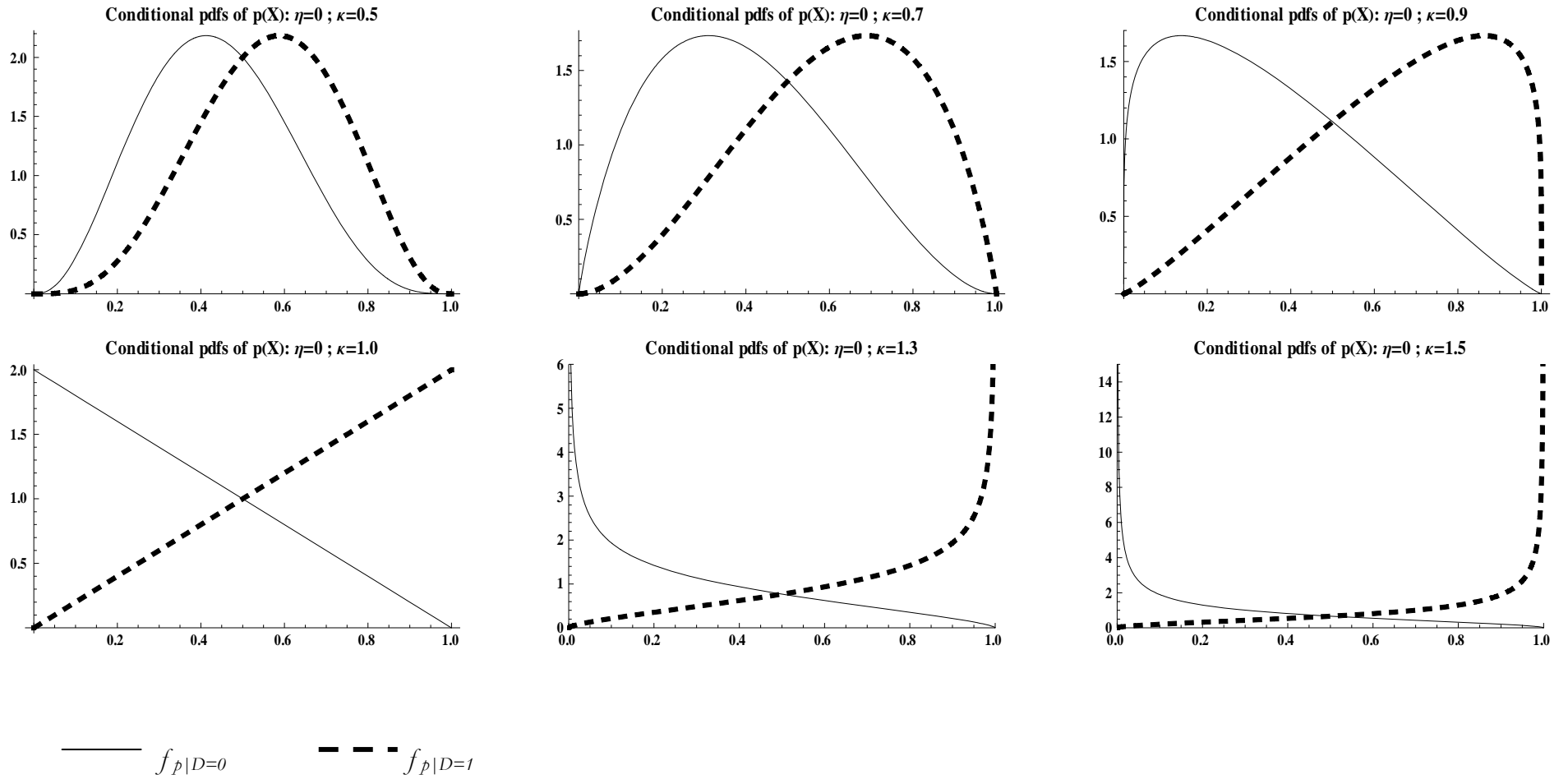
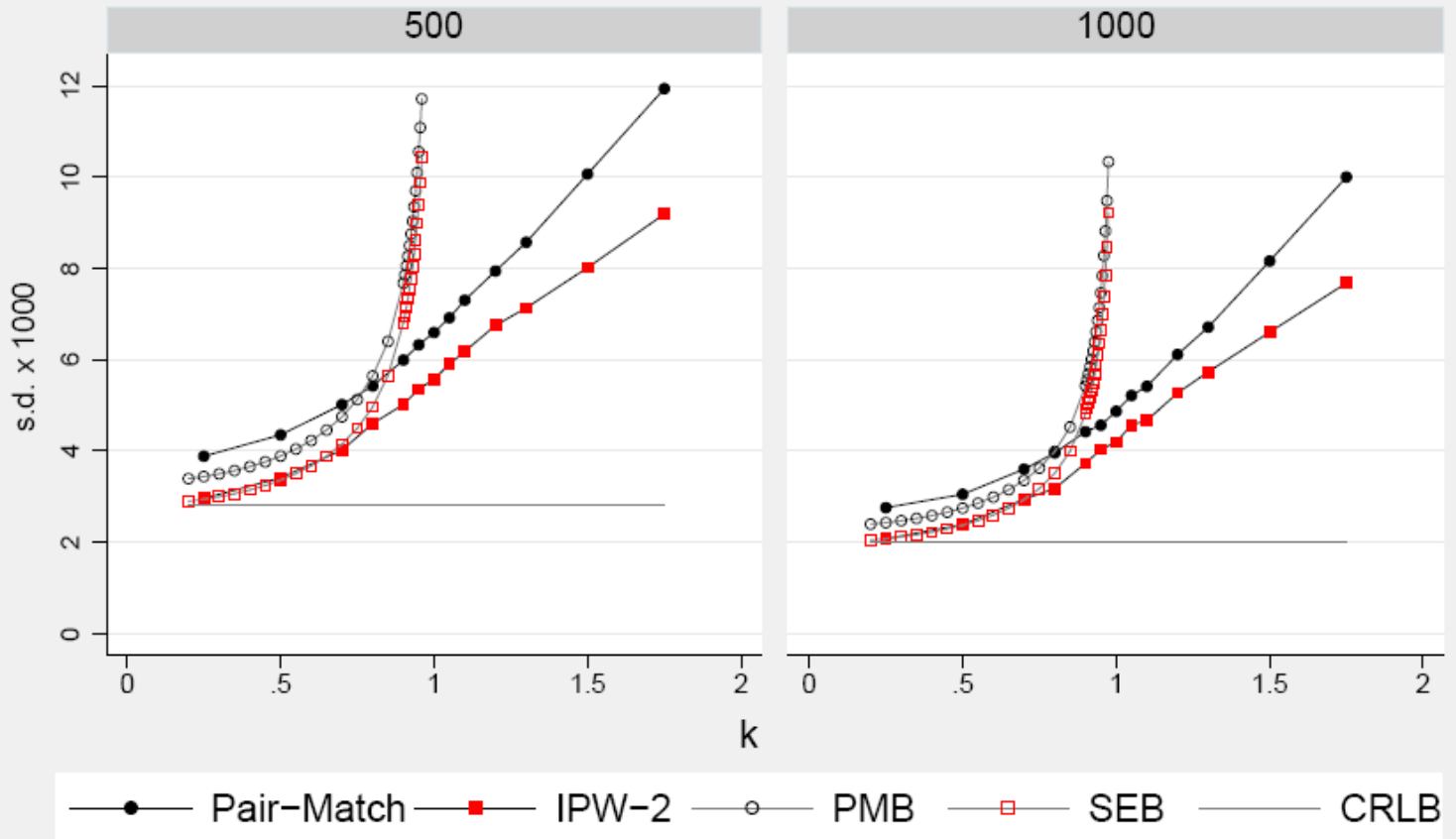


Figure 3: Breakdown of Standard Asymptotics as k grows
 Actual and Expected s.d. of IPW2 & PM in Normal–Normal model



Graphs by Observations

Figure 4: Bias of IPW2

Under different degrees of correlation between TE and $p(x)$

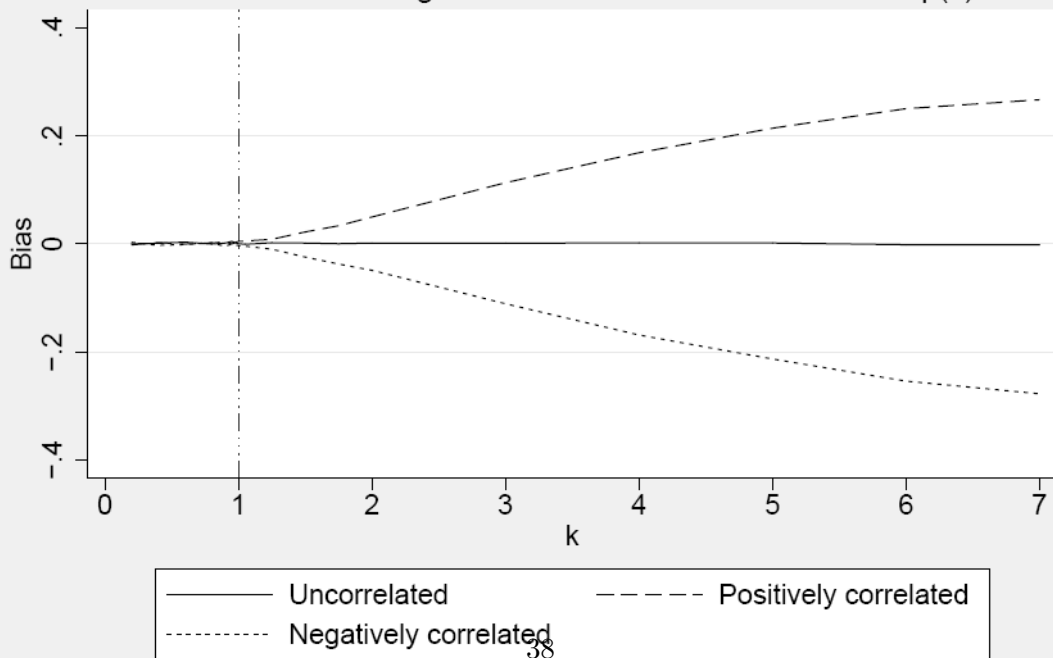


Table 1: Bias and Variance of the Estimated Treatment Effect on the Treated (TOT)

Normal-Cauchy Model

	Sample Size	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
A. Simulated Root Mean Squared Bias (x 1000)	100	I. Homog.-Homosk.	5.2	25.1*	42.5*	35.5*	39.5*	39.0*	39.5*	9.2*	12.8*	3.2	4.0	4.9*	2.5	3.0
		II. Heterog.-Homosk.	2.8	44.4*	41.9*	34.9*	39.2*	38.4*	39.3*	7.8*	12.0*	1.4	2.4	3.1	5.0*	2.0
		III. Homog.-Heterosk.	5.0	26.9*	35.3*	26.8*	28.6*	11.2*	13.0*	8.5*	10.7*	3.5	4.1	4.8	4.1	3.8
		IV. Heterog.-Heterosk.	3.3	42.6*	34.0*	25.0*	26.9*	10.9*	13.2*	6.1*	8.3*	2.2	2.1	2.5	6.5*	2.3
		All	4.2	35.8*	38.6*	30.9*	34.0*	28.5*	29.4*	8.0*	11.1*	2.7	3.3	4.0*	4.8*	2.9
		F-stat (no bias)	37.7	3505.5	5276.1	3309.3	4163.6	2144.7	2384.1	183.4	379.0	19.3	30.9	46.0	58.5	18.8
		[p-value]	[0.037]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.734]	[0.155]	[0.004]	[0.000]
	500	I. Homog.-Homosk.	2.4	7.6*	36.0*	30.2*	33.4*	2.9	2.9	3.1	4.9*	2.3	2.3	2.2	1.9	1.7
		II. Heterog.-Homosk.	2.2	7.7*	32.9*	27.8*	31.1*	2.1	1.8	1.7	2.9	1.9	1.9	1.8	2.6	2.0
		III. Homog.-Heterosk.	2.4	6.8*	23.2*	16.2*	17.9*	2.4	2.4	2.6	3.2	2.3	2.3	2.2	2.3	2.4
		IV. Heterog.-Heterosk.	2.3	9.8*	25.0*	17.5*	19.4*	2.4	2.7	3.2	4.3	2.2	2.2	2.4	1.6	2.2
		All	2.3	8.0*	29.8*	23.7*	26.4*	2.5	2.5	2.7	3.9*	2.2	2.2	2.2	2.2	2.1
		F-stat (no bias)	12.1	182.4	3094.5	2043.2	2548.8	17.6	18.1	20.6	45.8	13.7	13.9	13.5	14.0	11.9
		[p-value]	[0.979]	[0.000]	[0.000]	[0.000]	[0.000]	[0.824]	[0.800]	[0.662]	[0.005]	[0.953]	[0.949]	[0.958]	[0.946]	[0.981]
B. Simulated Average Variance (x 1000)	100	I. Homog.-Homosk.	103.3*	68.9*	53.5*	56.2*	54.5*	82.4*	78.5*	67.4*	64.0*	72.2*	65.8*	65.5*	65.8*	89.3*
		II. Heterog.-Homosk.	106.2*	74.9*	54.9*	57.6*	55.7*	84.1*	80.0*	68.9*	65.4*	73.0*	66.9*	66.8*	67.2*	91.4*
		III. Homog.-Heterosk.	119.0*	110.1*	108.0	109.0	108.8	113.7*	113.0*	111.7	111.1	118.4*	112.6*	112.0*	117.6*	115.9*
		IV. Heterog.-Heterosk.	118.1*	113.7*	107.0*	107.8*	107.6*	112.9*	112.1	110.5	110.0	117.6*	111.7	110.9	117.0*	115.5*
		Average (V-SEB)/SEB	0.376	0.083	-0.078	-0.052	-0.067	0.180	0.145	0.050	0.020	0.116	0.040	0.035	0.064	0.247
		F-stat (V = SEB)	418.6	45.4	73.0	30.1	54.5	158.5	116.4	17.5	4.0	60.1	9.7	8.3	19.5	243.7
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	500	I. Homog.-Homosk.	21.1*	12.8	11.2*	11.2*	11.1*	13.5*	12.9	13.0	12.6	13.3*	12.6	12.6	12.6	12.9
		II. Heterog.-Homosk.	21.5*	13.3*	11.6*	11.4*	11.3*	13.6*	13.2*	13.2*	12.8	13.7*	12.9	12.9	13.0	13.3*
		III. Homog.-Heterosk.	24.1*	21.8	21.2	21.5	21.5	22.2	22.0	22.0	21.8	22.6	21.9	21.8	21.9	21.9
		IV. Heterog.-Heterosk.	24.1*	22.3	21.5	21.8	21.8	22.5	22.3	22.3	22.2	23.0	22.3	22.2	22.4	22.3
		Average (V-SEB)/SEB	0.398	0.021	-0.060	-0.055	-0.062	0.048	0.025	0.027	0.009	0.059	0.012	0.010	0.014	0.025
		F-stat (V = SEB)	88.8	1.6	11.3	9.5	12.3	3.3	1.4	1.5	0.7	3.9	0.8	0.7	0.9	1.6
		[p-value]	[0.000]	[0.032]	[0.000]	[0.000]	[0.000]	[0.000]	[0.097]	[0.064]	[0.817]	[0.000]	[0.793]	[0.817]	[0.584]	[0.032]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). **Models:** Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, $RMSB = \sqrt{(1/6)(b_1 + \dots + b_6)}$ and the $AV = (1/6)(v_1 + \dots + v_6)$ where b_i ($i=1, \dots, 6$) is the square of the bias and v_i ($i=1, \dots, 6$) is the variance of one of the six combinations of the two curves and the three designs. “All” is the RMSB across all 24 contexts. Average (V-SEB)/SEB is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present two F-tests and their p-values: (i) $H_0: bias=0$, (ii) $H_0: V=SEB$ (see section IV for details). One star means that we reject the null at the 1%.

Table 2: Bias and Variance of the Estimated Treatment Effect on the Treated (TOT) under Misspecification

Misspecification of the Propensity Score in the Normal-Cauchy Model (sample size 100)

	Misspec. Type	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
A. Simulated Root Mean Squared Bias (x 1000)	Xs	I. Homog.-Homosk.	127.2*	123.2*	142.4*	141.1*	142.6*	155.4*	153.3*	128.7*	130.4*	125.4*	125.6*	125.9*	125.2*	124.9*
		II. Heterog.-Homosk.	126.4*	120.9*	143.1*	141.6*	143.1*	156.3*	154.0*	128.8*	130.5*	125.3*	125.7*	126.2*	123.7*	125.1*
		III. Homog.-Heterosk.	125.5*	122.7*	140.8*	139.3*	140.6*	131.4*	132.3*	127.7*	129.4*	124.6*	124.9*	125.2*	124.5*	124.2*
		IV. Heterog.-Heterosk.	126.1*	120.1*	142.2*	140.7*	142.1*	131.8*	132.8*	128.4*	130.4*	125.5*	125.9*	126.2*	123.2*	124.9*
		All	126.3*	121.7*	142.1*	140.7*	142.1*	144.2*	143.5*	128.4*	130.2*	125.2*	125.5*	125.9*	124.1*	124.8*
		F-stat (no bias)	36397.5	46067.5	69490.6	67500.6	70217.2	56723.0	58158.0	51241.9	54298.3	49436.3	50884.0	51261.2	48656.0	43699.7
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	Dist of u.	I. Homog.-Homosk.	5.4	27.3*	42.6*	37.0*	40.7*	39.5*	41.1*	9.5*	12.9*	22.6*	7.6*	6.5*	2.4	2.1
		II. Heterog.-Homosk.	3.0	47.3*	42.3*	36.3*	40.1*	39.1*	40.3*	8.4*	12.1*	19.5*	4.7*	3.9	7.3*	4.0
		III. Homog.-Heterosk.	5.4	28.7*	36.1*	29.4*	31.8*	12.8*	15.5*	9.3*	11.5*	21.8*	7.9*	7.1*	4.0	3.8
		IV. Heterog.-Heterosk.	3.6	46.1*	34.7*	27.9*	30.3*	12.7*	15.8*	7.1*	9.2*	21.0*	7.9*	6.1*	8.6*	3.1
		All	4.5	38.5*	39.1*	32.9*	36.0*	29.2*	30.8*	8.6*	11.5*	21.3*	7.2*	6.0*	6.1*	3.3
		F-stat (no bias)	42.4	4067.8	5378.4	3728.1	4597.8	2209.3	2587.9	212.0	402.1	744.2	130.1	98.5	99.0	23.8
		[p-value]	[0.012]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.474]
B. Simulated Average Variance (x 1000)	Xs	I. Homog.-Homosk.	87.4*	55.0*	48.3*	48.8*	47.7*	67.2*	64.3*	55.8*	53.6*	54.9*	53.0*	52.9*	53.8*	64.9*
		II. Heterog.-Homosk.	88.0*	59.0*	49.9*	50.2*	49.1*	68.8*	65.6*	57.0*	54.8*	56.0*	54.3*	54.2*	55.0*	66.3*
		III. Homog.-Heterosk.	132.7*	121.7*	121.4*	121.5*	121.3*	125.5*	124.8*	123.2*	122.5*	124.4*	122.6*	122.5*	128.3*	126.3*
		IV. Heterog.-Heterosk.	132.5*	124.6*	121.8*	121.9*	121.7*	125.6*	125.2*	123.5*	123.0*	124.4*	122.9*	122.8*	128.7*	126.6*
		Average (V-SEB)/SEB	-0.066	-0.266	-0.314	-0.311	-0.318	-0.197	-0.216	-0.268	-0.282	-0.270	-0.285	-0.286	-0.261	-0.208
		F-stat (V = SEB)	36.0	1079.7	2031.3	1955.7	2150.8	460.2	580.3	1095.5	1323.2	1186.4	1387.9	1401.2	1258.8	512.1
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	Dist of u.	I. Homog.-Homosk.	103.1*	68.6*	53.1*	55.4*	53.9*	82.9*	78.9*	67.2*	63.8*	93.3*	72.2*	67.3*	68.5*	93.0*
		II. Heterog.-Homosk.	105.9*	74.9*	54.5*	56.7*	55.1*	84.1*	80.0*	68.5*	65.3*	93.8*	73.3*	68.9*	70.1*	96.5*
		III. Homog.-Heterosk.	118.7*	110.2*	107.5*	108.8	108.7	114.1*	113.9*	111.7*	111.2	138.9*	118.7*	114.8*	121.4*	118.7*
		IV. Heterog.-Heterosk.	117.8*	114.4*	106.4*	107.7*	107.4*	113.4*	113.0*	110.6	110.1	145.4*	118.5*	114.0*	121.4*	117.6*
		Average (V-SEB)/SEB	0.373	0.084	-0.084	-0.060	-0.073	0.184	0.150	0.047	0.019	0.383	0.118	0.063	0.105	0.293
		F-stat (V = SEB)	415.7	46.4	77.1	40.0	62.1	162.5	119.1	16.1	3.5	418.4	80.3	27.9	61.0	296.0
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). **Models:** Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, $RMSB = \sqrt{\{(1/6)(b_1 + \dots + b_6)\}}$ and the $AV = (1/6)(v_1 + \dots + v_6)$ where b_i ($i=1, \dots, 6$) is the square of the bias and v_i ($i=1, \dots, 6$) is the variance of one of the six combinations of the two curves and the three designs. "All" is the RMSB across all 24 contexts. Average (V-SEB)/SEB is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present two F-tests and their p-values: (i) $H_0: bias=0$, (ii) $H_0: V=SEB$ (see section IV for details). One star means that we reject the null at the 1%.

Table 3: Bias and Variance of the Estimated Treatment Effect on the Treated (TOT)

Normal-Normal Model

	N	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
A. Simulated Root Mean Squared Bias (x 1000)	100	I. Homog.-Homosk.	19.3*	45.1*	88.8*	70.5*	76.7*	54.2*	55.4*	26.9*	32.1*	15.8*	16.7*	16.8*	3.3	7.8*
		II. Heterog.-Homosk.	16.5*	66.3*	87.4*	68.6*	75.1*	53.3*	55.0*	25.4*	30.4*	12.8*	14.8*	15.4*	16.9*	7.6*
		III. Homog.-Heterosk.	14.4*	43.4*	76.1*	55.4*	58.7*	21.6*	25.4*	24.4*	27.5*	15.2*	13.8*	13.9*	3.7	6.6
		IV. Heterog.-Heterosk.	14.6*	67.4*	74.8*	54.2*	57.9*	21.1*	24.3*	23.0*	26.7*	15.0*	12.4*	12.9*	19.5*	5.7
		All	16.3*	56.7*	82.0*	62.6*	67.7*	40.9*	42.8*	24.9*	29.3*	14.7*	14.5*	14.8*	13.2*	7.0*
		F-stat (no bias)	428.2	6163.6	21354.4	11216.0	13439.8	3071.5	3564.4	1385.7	1989.8	267.3	422.9	466.7	335.6	69.9
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	500	I. Homog.-Homosk.	7.8*	17.1*	72.1*	52.8*	58.4*	12.0*	14.2*	14.1*	17.4*	3.1	6.3	7.7*	4.0	6.4*
		II. Heterog.-Homosk.	5.9	16.4*	68.7*	50.1*	55.4*	9.8*	12.1*	11.5*	15.4*	4.7	4.1	4.5	8.5*	3.5
		III. Homog.-Heterosk.	4.1	15.0*	60.3*	36.9*	40.5*	7.3*	8.4*	11.0*	13.1*	4.4	3.8	5.4	2.7	3.5
		IV. Heterog.-Heterosk.	7.3	17.4*	60.7*	37.2*	40.7*	9.6*	9.9*	12.2*	14.1*	4.6	6.7	7.5*	6.4	5.0
		All	6.4*	16.5*	65.7*	44.9*	49.4*	9.8*	11.4*	12.3*	15.1*	4.3	5.4*	6.4*	5.8*	4.7
		F-stat (no bias)	52.8	487.7	12564.8	5392.8	6719.7	167.8	251.3	285.5	461.6	23.6	43.6	71.8	68.1	42.0
		[p-value]	[0.001]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.482]	[0.009]	[0.000]	[0.000]	[0.013]
B. Simulated Average Variance (x 1000)	100	I. Homog.-Homosk.	139.6*	104.8*	58.0*	66.5*	64.9*	117.1*	108.8*	88.7*	84.6*	166.8*	98.1*	93.1*	92.7*	166.2*
		II. Heterog.-Homosk.	143.6*	122.1*	59.2*	68.5*	66.7*	119.0*	111.4*	91.4*	87.1*	162.3*	101.5*	96.0*	96.5*	169.4*
		III. Homog.-Heterosk.	152.0*	141.2*	120.1*	127.9*	128.1*	140.8*	139.3*	133.4*	133.0*	207.5*	149.1*	142.2*	152.8*	162.2*
		IV. Heterog.-Heterosk.	151.7*	156.0*	120.0*	127.1*	126.8*	139.6*	138.6*	132.7*	132.1*	195.2*	149.0*	142.2*	153.6*	161.6*
		Average (V-SEB)/SEB	-0.280	-0.375	-0.570	-0.533	-0.537	-0.369	-0.392	-0.459	-0.472	-0.165	-0.404	-0.432	-0.410	-0.190
		F-stat (V = SEB)	6454.1	8872.2	46732.8	31680.5	33472.8	9986.0	11525.1	17199.1	18762.0	2701.3	11956.3	13919.1	12969.7	4169.5
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	500	I. Homog.-Homosk.	33.7*	21.4*	13.2*	15.0*	14.6*	22.1*	20.2*	19.8*	18.7*	39.8*	24.5*	21.0*	21.4*	21.9*
		II. Heterog.-Homosk.	33.9*	22.8*	13.2*	14.9*	14.5*	22.0*	20.1*	19.7*	18.7*	36.3*	24.2*	20.8*	21.4*	21.6*
		III. Homog.-Heterosk.	35.7*	29.8*	24.2*	26.9*	26.8*	30.6*	29.8*	28.7*	28.4*	44.3*	34.4*	30.6*	31.7*	29.7*
		IV. Heterog.-Heterosk.	36.4*	30.8*	24.5*	27.1*	26.9*	30.9*	29.7*	28.9*	28.6*	86.9*	34.2*	30.7*	31.7*	29.7*
		Average (V-SEB)/SEB	-0.162	-0.379	-0.547	-0.503	-0.510	-0.375	-0.409	-0.423	-0.439	-0.011	-0.322	-0.393	-0.380	-0.387
		F-stat (V = SEB)	627.6	1754.0	7311.1	4590.5	4981.7	1688.7	2165.0	2390.6	2688.4	244.4	1087.3	1904.1	1704.4	1877.6
		[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). **Models:** Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, $RMSB = \sqrt{(1/6)(b_1 + \dots + b_6)}$ and the $AV = (1/6)(v_1 + \dots + v_6)$ where b_i ($i=1, \dots, 6$) is the square of the bias and v_i ($i=1, \dots, 6$) is the variance of one of the six combinations of the two curves and the three designs. "All" is the RMSB across all 24 contexts. Average (V-SEB)/SEB is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present two F-tests and their p-values: (i) $H_0: bias=0$, (ii) $H_0: V=SEB$ (see section IV for details). One star means that we reject the null at the 1%.

Table 4: Simulated Root Mean Squared Bias (x 1000) of the Estimated Treatment Effect on the Treated (TOT)

Trimming Results in the Normal-Normal Model (Sample size 100)

Trimming	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
Rule 1	I. Homog.-Homosk.	6.2*	34.3*	56.5*	46.5*	51.8*	49.5*	48.6*	11.0*	15.4*	8.9*	5.4*	4.2	2.2	4.1
	II. Heterog.-Homosk.	44.8*	69.2*	23.0*	24.6*	24.3*	53.5*	54.4*	39.7*	35.2*	47.8*	45.0*	46.9*	55.4*	48.3*
	III. Homog.-Heterosk.	5.4	30.7*	49.5*	37.9*	40.7*	17.9*	20.1*	10.9*	14.1*	11.7*	5.4	4.3	4.0	3.8
	IV. Heterog.-Heterosk.	43.5*	66.0*	19.1*	21.8*	19.9*	45.2*	46.6*	37.9*	34.6*	47.5*	44.1*	45.7*	53.3*	47.5*
	All	31.5*	53.1*	40.4*	34.2*	36.5*	43.8*	44.4*	28.5*	26.8*	34.5*	31.7*	32.9*	38.5*	34.0*
	F-stat (no bias)	1750.2	5867.8	4868.6	3309.1	3921.1	4082.7	4373.9	1891.3	1729.6	2662.4	2343.6	2461.9	3404.0	2156.7
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 2	I. Homog.-Homosk.	16.7*	44.6*	89.3*	70.2*	77.2*	53.5*	54.6*	25.7*	31.3*	13.0*	20.0*	23.2*	1.8	5.7
	II. Heterog.-Homosk.	22.2*	64.1*	96.0*	77.0*	83.9*	60.6*	62.2*	30.7*	36.2*	18.3*	26.3*	29.3*	15.0*	11.4*
	III. Homog.-Heterosk.	16.1*	41.8*	80.8*	59.1*	62.7*	23.4*	27.7*	25.9*	29.5*	14.2*	20.9*	23.6*	3.9	5.7
	IV. Heterog.-Heterosk.	21.9*	64.8*	87.3*	66.3*	70.3*	30.4*	33.8*	32.4*	36.2*	21.9*	28.7*	30.9*	15.4*	11.9*
	All	19.4*	54.9*	88.5*	68.4*	73.9*	44.7*	46.8*	28.8*	33.5*	17.2*	24.2*	27.0*	11.0*	9.2*
	F-stat (no bias)	569.1	5531.9	23283.6	12403.4	14937.3	3423.2	3971.0	1708.3	2405.7	344.7	1085.5	1481.9	230.6	114.0
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 3	I. Homog.-Homosk.	2.3	3.3	49.8*	41.6*	47.2*	42.3*	40.3*	5.3*	9.2*	139.4*	34.7*	12.7*	1.8	1.7
	II. Heterog.-Homosk.	50.3*	48.3*	22.3*	27.3*	26.2*	54.5*	54.4*	47.2*	42.6*	155.5*	83.9*	62.0*	35.7*	49.7*
	III. Homog.-Heterosk.	4.2	4.6	42.9*	33.0*	36.2*	14.6*	16.4*	7.1*	9.5*	137.2*	35.4*	14.6*	5.3	4.6
	IV. Heterog.-Heterosk.	48.9*	45.8*	20.5*	27.0*	24.0*	50.7*	51.3*	45.6*	42.4*	151.5*	82.2*	60.8*	34.6*	48.7*
	All	35.1*	33.4*	36.2*	32.8*	34.7*	43.4*	43.3*	33.1*	30.8*	146.1*	63.7*	44.5*	25.0*	34.9*
	F-stat (no bias)	2364.1	2587.8	3980.7	3168.6	3728.1	4519.7	4571.4	2742.5	2449.6	23016.7	7638.3	4646.7	1332.2	3071.5
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 4	I. Homog.-Homosk.	5.6	37.8*	59.3*	48.2*	53.8*	47.6*	46.8*	10.1*	14.9*	7.4*	2.4	3.5	2.3	2.9
	II. Heterog.-Homosk.	38.0*	70.8*	31.9*	27.2*	29.5*	51.8*	52.3*	32.8*	28.6*	41.4*	40.6*	39.4*	45.5*	40.7*
	III. Homog.-Heterosk.	5.4	35.6*	53.2*	40.1*	42.8*	17.6*	19.9*	11.4*	14.8*	5.7	4.2	4.8	3.6	3.1
	IV. Heterog.-Heterosk.	38.5*	69.4*	27.4*	22.5*	21.7*	40.8*	42.4*	32.9*	29.8*	42.5*	41.6*	40.1*	45.6*	42.0*
	All	27.3*	56.0*	45.1*	36.0*	39.0*	41.6*	42.2*	24.4*	23.2*	30.0*	29.2*	28.3*	32.3*	29.3*
	F-stat (no bias)	1345.5	6791.7	6219.9	3780.3	4623.4	3790.0	4049.8	1400.9	1309.3	1931.8	1973.8	1868.4	2387.4	1708.9
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). **Models:** Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, $RMSB = \sqrt{\{(1/6)(b_1 + \dots + b_6)\}}$ and the $AV = (1/6)(v_1 + \dots + v_6)$ where b_i ($i=1, \dots, 6$) is the square of the bias and v_i ($i=1, \dots, 6$) is the variance of one of the six combinations of the two curves and the three designs. "All" is the RMSB across all 24 contexts. Average $(V-SEB)/SEB$ is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present two F-tests and their p-values: (i) $H_0: bias=0$, (ii) $H_0: V=SEB$ (see section IV for details). One star means that we reject the null at the 1%.

Table 5: Simulated Average Variance (x 1000) of the Estimated Treatment Effect on the Treated (TOT)

Trimming Results in the Normal-Normal Model (Sample size 100)

Trimming	Setting	Pair match	Blocking	k-NN	Kernel match (Epa)	Kernel match (Gauss)	LLR match (Epa)	LLR match (Gauss)	Ridge match (Epa)	Ridge match (Gauss)	IPW1	IPW2	IPW3	Double robust	Control function
Rule 1	I. Homog.-Homosk.	125.7*	86.7*	61.1*	66.2*	63.5*	102.7*	97.6*	81.6*	77.4*	85.2*	79.0*	81.8*	79.5*	111.3*
	II. Heterog.-Homosk.	128.0*	99.6*	62.8*	67.7*	65.4*	105.2*	100.0*	83.9*	79.4*	87.3*	81.6*	84.2*	81.2*	114.0*
	III. Homog.-Heterosk.	141.9*	134.1*	127.2*	130.0*	130.1*	136.6*	136.1*	132.7*	132.3*	141.5*	134.9*	136.1*	143.1*	141.7*
	IV. Heterog.-Heterosk.	141.9*	144.5*	128.5*	130.9*	130.9*	137.0*	136.8*	133.6*	133.0*	142.9*	136.5*	137.2*	145.1*	141.6*
	Average (V-SEB)/SEB	-0.331	-0.441	-0.545	-0.526	-0.533	-0.411	-0.426	-0.476	-0.490	-0.451	-0.480	-0.471	-0.462	-0.380
	F-stat (V = SEB)	9263.8	15514.0	40067.8	32825.1	35463.8	13319.6	14776.0	20859.7	23178.4	17567.5	21230.5	19582.3	20380.1	10186.5
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 2	I. Homog.-Homosk.	149.0*	111.0*	60.6*	70.3*	68.1*	124.0*	114.7*	94.1*	89.3*	170.7*	99.4*	91.0*	96.1*	176.1*
	II. Heterog.-Homosk.	150.5*	128.4*	61.6*	71.6*	69.5*	124.7*	116.7*	96.0*	91.4*	172.2*	102.5*	93.3*	99.8*	176.5*
	III. Homog.-Heterosk.	157.6*	147.0*	124.5*	132.1*	132.4*	145.7*	144.0*	138.1*	137.6*	214.2*	153.8*	143.8*	159.7*	167.3*
	IV. Heterog.-Heterosk.	157.0*	162.7*	124.7*	132.0*	132.1*	145.3*	144.5*	137.9*	137.6*	213.7*	152.9*	143.8*	159.8*	167.6*
	Average (V-SEB)/SEB	-0.249	-0.348	-0.554	-0.515	-0.520	-0.344	-0.369	-0.437	-0.451	-0.123	-0.394	-0.437	-0.391	-0.162
	F-stat (V = SEB)	5497.7	7484.9	42164.2	27865.0	29746.0	8474.7	9822.4	14735.8	16190.7	2447.2	11114.9	14281.3	11368.1	3468.4
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 3	I. Homog.-Homosk.	111.0*	79.3*	59.3*	62.1*	59.2*	88.6*	85.7*	72.8*	68.7*	202.0*	102.6*	77.9*	82.2*	71.9*
	II. Heterog.-Homosk.	112.8*	82.6*	60.9*	63.8*	61.2*	90.7*	88.1*	75.0*	70.9*	202.1*	105.2*	79.6*	84.7*	73.6*
	III. Homog.-Heterosk.	139.7*	135.6*	127.5*	130.0*	129.7*	135.6*	134.9*	133.0*	132.2*	226.8*	158.4*	140.1*	164.8*	132.4*
	IV. Heterog.-Heterosk.	139.3*	137.6*	128.3*	130.3*	130.1*	135.7*	135.3*	133.2*	132.6*	226.7*	157.7*	140.4*	165.0*	132.7*
	Average (V-SEB)/SEB	-0.374	-0.476	-0.550	-0.538	-0.546	-0.452	-0.461	-0.501	-0.514	-0.021	-0.377	-0.475	-0.415	-0.506
	F-stat (V = SEB)	11858.0	20730.0	42775.8	37569.1	41420.8	18083.0	19281.0	26259.2	29626.0	1414.1	10112.0	21478.8	16769.2	26786.4
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Rule 4	I. Homog.-Homosk.	121.3*	84.2*	59.2*	64.0*	61.4*	100.0*	95.1*	79.5*	75.1*	86.9*	78.8*	78.0*	78.4*	103.0*
	II. Heterog.-Homosk.	122.5*	96.6*	59.9*	64.9*	62.5*	101.8*	97.0*	81.1*	76.7*	88.4*	80.6*	79.8*	80.4*	105.3*
	III. Homog.-Heterosk.	140.4*	131.1*	125.5*	128.4*	128.3*	135.4*	134.7*	131.5*	130.9*	141.8*	135.1*	133.8*	142.7*	139.0*
	IV. Heterog.-Heterosk.	139.6*	142.1*	125.7*	128.3*	128.2*	135.4*	134.9*	131.4*	130.9*	142.1*	135.3*	134.1*	143.4*	138.8*
	Average (V-SEB)/SEB	-0.347	-0.454	-0.557	-0.538	-0.545	-0.420	-0.435	-0.486	-0.500	-0.447	-0.481	-0.486	-0.466	-0.398
	F-stat (V = SEB)	10121.3	16728.4	43560.5	35892.8	38878.5	14475.8	15970.7	22559.7	25166.4	17014.9	21910.2	22396.7	21177.8	13108.1
	[p-value]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]

NOTES: Replications: 10,000 for N=100 and 2000 for N=500. **Estimators:** See sections II.B and II.C. The numbers of neighbors in k-NN and the bandwidth of kernel-based matching were selected using leave-one-out cross validation (see section II.D). **Models:** Normal-Cauchy uses a treatment equation with a Cauchy distributed error term. Normal-Normal has an error term which is standard Normal. (see section III and section IV). **Settings:** Simulations were done for 24 different contexts which combine two outcome curves, three treatment designs, and four settings (homogeneous treatment - homoskedastic outcome error, homogenous-heteroskedastic, etc.) See section III and section IV.

Statistics (RMSB, AV and SEB): We summarize results by showing simulated root mean square bias (RMSB) and the average variance (AV) for each setting. For a given setting, $RMSB = \sqrt{\{(1/6)(b_1 + \dots + b_6)\}}$ and the $AV = (1/6)(v_1 + \dots + v_6)$ where b_i ($i=1, \dots, 6$) is the square of the bias and v_i ($i=1, \dots, 6$) is the variance of one of the six combinations of the two curves and the three designs. "All" is the RMSB across all 24 contexts. Average (V-SEB)/SEB is the average percentage difference between the variance and the semiparametric efficiency bound (SEB). See section II.E and Appendix II. **Stars, tests and p-values:** We present two F-tests and their p-values: (i) $H_0: bias=0$, (ii) $H_0: V=SEB$ (see section IV for details). One star means that we reject the null at the 1%.