Center for Evaluation and Development
WORKING PAPER SERIES

The Finite Sample Performance of Semi- and Nonparametric Estimators for Treatment Effects and Policy Evaluation

Working Paper 2015/4

Markus Frölich
Martin Huber
Manuel Wiesenfarth

ABSTRACT

This paper investigates the finite sample performance of a comprehensive set of semi- and nonparametric estimators for treatment and policy evaluation. In contrast to previous simulation studies which mostly considered semiparametric approaches relying on parametric propensity score estimation, we also consider more flexible approaches based on semi- or nonparametric propensity scores, nonparametric regression, and direct covariate matching. In addition to (pair, radius, and kernel) matching, inverse probability weighting, regression, and doubly robust estimation, our studies also cover recently proposed estimators such as genetic matching, entropy balancing, and empirical likelihood estimation. We vary a range of features (sample size, selection into treatment, effect heterogeneity, and correct/misspecification) in our simulations and find that several nonparametric estimators by and large outperform commonly used treatment estimators using a parametric propensity score. Nonparametric regression, nonparametric doubly robust estimation, nonparametric IPW, and one-to-many covariate matching perform best.

JEL Classification: C1

Keywords: treatment effects, policy evaluation, simulation, empirical Monte Carlo study, propensity score, semi- and nonparametric estimation

Corresponding author:

Markus Frölich
University of Mannheim
L7, 3-5
68131 Mannheim, Germany
E-mail: froelich@uni-mannheim.de

1 Introduction

Estimators for the evaluation of binary treatments (or policy interventions) in observational studies under a 'selection-on-observables' assumption (see for instance Imbens (2004)) are widely applied in empirical economics, social sciences, epidemiology and other fields. (Similar estimators are also used in instrumental variable settings, as we discuss later.) In most cases, researchers using these methods aim at estimating the average causal effect of the treatment (e.g. a new medical treatment or assignment to a training program) on an outcome of interest (e.g. health or employment) by controlling for differences in observed covariates across treated and non-treated sample units. More generally, such estimators may be used for any problem in which the means of an outcome variable in two subsamples should be purged from differences due to other observed variables, including wage gap decompositions (see for instance Frölich (2007b) and Ñopo (2008)) and further applications not necessarily concerned with the estimation of causal effects.

Most treatment effect estimators control for the treatment propensity score, i.e. the conditional probability to receive the treatment given the covariates, rather than directly for the covariates. Popular approaches include propensity score matching (see for instance Rosenbaum and Rubin (1985), Heckman, Ichimura, and Todd (1998a), and Dehejia and Wahba (1999)) and inverse probability weighting (henceforth IPW, Horvitz and Thompson (1952) and Hirano, Imbens, and Ridder (2003)). A further class constitute the so-called doubly robust estimators (henceforth DR, Robins, Mark, and Newey (1992), Robins, Rotnitzky, and Zhao (1995), and Robins and Rotnitzky (1995)), which rely on models for both the propensity score and the conditional mean outcome, and are consistent if one or the other (or both) are correctly specified. In almost all applications of such estimators, the propensity score is modelled parametrically based on probit or logit specifications.

A first reason for the wide use of propensity score methods appears to be the avoidance of the 'curse of dimensionality' that could arise when directly controlling for multidimensional covariates. That is, it may not be possible to find observations across treatment states that are com-

parable in terms of covariates for all combinations of covariate values in the data, while finding comparable observations in terms of propensity scores is easier because distinct combinations of covariates may still yield similar propensity scores. A second reason for the particular popularity of semiparametric estimators – i.e. parametric propensity score estimation combined with nonparametric treatment effect estimation – may be the ease of implementation. Specifically, probit or logit estimation of the propensity score does not require the choice of any bandwidth or other tuning parameters. The latter would, however, be necessary under semiparametric (see for instance Klein and Spady (1993) and Ichimura (1993)) or nonparametric binary choice models for the treatment.

The price to pay for the convenience of a parametric propensity score is that misspecifying the latter may entail an inconsistent treatment effect estimator. While some methods are more sensitive to the use of incorrect propensity scores than others (see for instance the comparison of IPW and matching in Waernbaum (2012) or the results of Zhao (2008)), none is robust to arbitrary specification errors in general. In this light, investigating and comparing the finite sample behavior of *nonparametric* treatment estimators appears interesting, but is lacking in previous simulation studies, which predominantly focus on (subsets of) semiparametric estimators (see Frölich (2004), Zhao (2004), Lunceford and Davidian (2004), Busso, DiNardo, and McCrary (2009), and Huber, Lechner, and Wunsch (2013)).

This paper aims at closing this gap by analysing the so far most comprehensive set of semi- and nonparametric estimators of the average treatment effect on the treated (ATET) in a large scale simulation study based on empirical labor market data from Switzerland first investigated by Behncke, Frölich, and Lechner (2010a,b). By using empirical (rather than arbitrarily modelled) associations between the treatment, the covariates, and the outcomes, we hope that our simulation design more closely mimics real world evaluation problems. The only other such 'empirical Monte Carlo study' focussing on treatment effect estimators we are aware of is Huber, Lechner, and Wunsch (2013), who, however, consider substantially fewer (and in particular no nonparametric) estimators. We vary several empirically relevant design features in our simulations, namely the

sample size, selection into treatment, effect heterogeneity, and correct versus misspecification. Furthermore, we consider estimation with and without trimming observations with (too) large propensity scores, considering seven different trimming rules.

Our analysis includes a range of propensity score methods, namely pair, kernel, and radius matching, as well as IPW and DR. In contrast to previous work, we consider four different approaches to propensity score estimation, which for the first time sheds light on the sensitivity of the various ATET estimators to the choice of the propensity score method: probit estimation, semiparametric maximum likelihood estimation of Klein and Spady (1993) (based on a parametric index model and a nonparametric distribution of the errors), nonparametric local constant kernel regression, and estimation based on the 'covariate balancing propensity score' method (CBPS) of Imai and Ratkovic (2014). The latter is an empirical likelihood approach which obtains exact balancing of particular moments of the covariates in the sample and is thus somewhat in the spirit of inverse probability tilting (IPT) suggested in Graham, Pinto, and Egel (2012) and Graham, Pinto, and Egel (2011), which is also implemented in our study as a weighting estimator. Our analysis also includes several nonparametric ATET estimators not requiring propensity score estimation: pair, radius, or one-to-many matching (directly) on the covariates via the Mahalanobis distance metric, nonparametric regression (which can be regarded as kernel matching on the covariates), the genetic matching algorithm of Diamond and Sekhon (2013), and entropy balancing as suggested by Hainmueller (2012). Finally, parametric regression among nontreated outcomes is also considered.

In our simulations, we find that several *nonparametric* estimators – in particular nonparametric regression, nonparametric doubly robust (DR) estimation as suggested by Rothe and Firpo (2013), nonparametric IPW, and one-to-many covariate matching – by and large outperform all ATET estimators based on a parametric propensity score. These results are quite robust across various simulation features and estimation methods with or without trimming.[1] Our results sug-

[1] Among the semiparametric methods investigated, IPW based on the (overidentified version of) CBPS of Imai and Ratkovic (2014) is best and slightly dominates the overall top performing nonparametric methods in a subset of the scenarios.

gest that nonparametric ATET estimators can be quite competitive even in moderate samples. However, not all nonparametric approaches perform equally well. A puzzling finding is that nonparametric propensity score estimation on the one hand entails very favorable IPW and DR estimators, but on the other hand leads to inferior matching estimators when compared to matching on a parametric or semiparametric propensity score. Another interesting outcome, which is in line with Zhao (2004), is that the best covariate matching estimators clearly dominate the top propensity score matching methods.

The remainder of this paper is organized as follows. Section 2 introduces the ATET and propensity score estimators considered in this paper. Section 3 discusses our Swiss labor market data and the simulation design. Section 4 presents the results for the various ATET and propensity score estimates across all simulation settings with and without trimming, as well as separately for particular simulation features such as sample size and effect heterogeneity. Section 5 concludes.

2 Estimators

2.1 Overview

In our treatment evaluation framework, let D denote the binary treatment indicator, Y the outcome, and X a vector of observed covariates. The aim of the methods discussed below is to compare the mean outcome of the treated group ($D = 1$) to that of the non-treated group ($D = 0$), after making the latter group comparable to the former in terms of the X covariates. More formally, the parameter of interest is

$$\Delta = E[Y|D = 1] - E[E[Y|D = 0, X]|D = 1]. \tag{1}$$

Δ corresponds to the average treatment effect on the treated (ATET) if the so-called 'selection on observables' or 'conditional independence' assumption (CIA) is invoked (see for instance Imbens

(2004)), which rules out the existence of (further) confounders that jointly influence D and Y conditional on X. This parameter has received much attention in the program evaluation literature, for instance when assessing active labor market policies or health interventions. However, the econometric methods may also be applied (as a descriptive tool) in non-causal contexts such as wage gap decompositions, which frequently make use of endogenous X variables (see the discussion in Huber (2014)).

ATET estimators may either make use of X directly, or of the conditional treatment probability $\Pr(D = 1|X)$ instead, henceforth referred to as propensity score. In Section 2.4 we present the estimators that directly control for X. In Section 2.3, on the other hand, we introduce the estimators of ATET that (semi- or nonparametrically) make use of the propensity score (IPW, IPT, DR, propensity score matching). All these estimators themselves depend on the plug-in estimation of the propensity score. The various propensity score estimators are discussed in Section 2.2, namely parametric, empirical likelihood-based, semiparametric, and nonparametric estimation. Hence, the propensity score estimators we examine in the simulation study are a combination of one estimator of Section 2.2 and one estimator of Section 2.3, whereas the estimators in Section 2.4 do not require plug-in estimates.

In our simulation study we focus on estimation of ATET. We expect the main lessons to also carry over to the estimation of the average treatment effect ATE, which has a structure similar to (1) and is defined as

$$ATE = E[E[Y|D = 1, X]] - E[E[Y|D = 0, X]]. \tag{2}$$

Also estimators of distributional or quantile treatment effects have a similar structure. Under a selection on observables assumption the distribution function $F_{Y^1}(a)$ of the potential outcome Y^1 is identified as

$$E[E[I(Y \leq a)|D = 1, X]].$$

This distribution function can be inverted to obtain the quantile function. Analogously the

distribution and quantile function of the potential outcome Y^0 can be obtained. Given this similar structure, we would therefore expect that estimators that perform better with respect to the estimation of the ATET should in tendency also do well for distributional treatment effects.

Furthermore, also several IV estimators have a structure similar to (1) and (2). E.g. the nonparametric IV estimator in Frölich (2007a), which exploits an instrumental variable Z that may only be conditionally valid, can be represented as a ratio of two matching estimators. Let Z be a binary instrumental variable that satisfies the usual instrument independence and exclusion restrictions, then the local average treatment effect is identified as

$$LATE = \frac{E[E[Y|Z=1,X]] - E[E[Y|Z=0,X]]}{E[E[D|Z=1,X]] - E[E[D|Z=0,X]]}. \tag{3}$$

The numerator and denominator are each of a structure like the right hand side of (2). Hence, estimators that perform better with respect to (2) should in tendency also do better in estimating numerator and denominator of (3).

Hence, although our simulation study will target the estimation of Δ, which is usually the focus under a 'selection on observables' assumption, we expect our main results to also roughly carry over to instrumental variable estimation (of LATE). This is relevant particularly in economic applications, where the 'selection on observables' assumption is often deemed to be too restrictive and instrumental variable estimation is being resorted to.

2.2 Estimation of propensity score

Define the propensity score as a real valued function of x as $p(x) = \Pr(D=1|X=x)$ and define $P \equiv p(X) = \Pr(D=1|X)$ as the corresponding random variable. Rosenbaum and Rubin (1983) showed that the propensity score possesses the so-called 'balancing property'. That is, conditioning on the one-dimensional P equalizes the distribution of the (possibly high dimensional)

covariates X across D, so that

$$\Delta = E[Y|D=1] - E[E[Y|D=0,P]|D=1] \tag{4}$$

is an alternative way to obtain the parameter of interest. As a practical matter, controlling for the propensity score rather than the full vector of covariates avoids the curse of dimensionality in finite samples, which is a major reason for the popularity of propensity score methods. The downside is that in practice, the (unkown) propensity score needs to be estimated by an adequate model. We investigate four estimation approaches in our simulations: probit regression, the empirical likelihood-based method of Imai and Ratkovic (2014), semiparametric estimation as suggested in Klein and Spady (1993), and nonparametric kernel regression. These estimators of the propensity score are described in the following subsections. In all settings we assume an i.i.d. sample of size N containing $\{Y_i, D_i, X_i\}$ for each observation i.

2.2.1 Parametric probit estimation of the propensity score

As it is standard in the vast majority of empirical studies, we consider parametric modelling as one option to estimate the propensity score, in our case based on a probit specification. The probit estimator of $p(X_i)$ is given by

$$\hat{p_i} \equiv p(X_i;\hat{\beta}) \equiv \Phi((1,X_i')\hat{\beta}), \tag{5}$$

where the coefficient estimates $\hat{\beta}$ are obtained by maximizing the log-likelihood function

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{N} \{D_i \log[\hat{p}(X_i;\beta)] + (1-D_i)\log[1-\hat{p}(X_i;\beta)]\}, \tag{6}$$

where Φ denotes the cumulative distribution function (cdf) of the standard normal distribution.

2.2.2 Empirical likelihood estimation of the propensity score by CBPS

The previous probit estimator assumed the parametric model to be correctly specified. However, if the propensity score model is misspecified, $\hat{p}(X)$ may not balance X, which generally biases estimation of Δ. As a second approach, we therefore consider the 'covariate balancing propensity score' (CBPS) procedure of Imai and Ratkovic (2014), an empirical likelihood (EL) method which models treatment assignment and at the same time optimizes covariate balance. More specifically, a set of moment conditions that are implied by the covariate balancing property (e.g., mean independence between the treatment and covariates after IPW) is exploited to estimate the propensity score, while also considering the maximum likelihood (ML) score condition of the propensity score model. The main idea behind CBPS is that a single model determines the treatment assignment mechanism and the covariate balancing weights across treatment groups, so that both the moment conditions related to the balancing property as well as the score condition can be used as moment conditions.

Formally, the covariate balancing property is operationalized by the following covariate balancing moment conditions for Δ implied by IPW:

$$E\left[D\tilde{X} - \frac{p(X)(1-D)\tilde{X}}{1-p(X)}\right] = 0, \tag{7}$$

where $\tilde{X}$ is a possibly multidimensional function of X, because the true propensity score must balance any function of X as long as the expectation exists (for instance, mean, variance, or higher moments). The sample analogue of (7) used in the Imai and Ratkovic (2014) procedure is

$$\frac{1}{N}\sum_{i=1}^{N} w(D_i, X_i; \tilde{\beta}) \cdot \tilde{X}_i \qquad \text{with} \qquad w(D_i, X_i; \tilde{\beta}) = \frac{N}{N_1}\frac{D_i - \tilde{p}(X_i)}{1-\tilde{p}(X_i)}, \tag{8}$$

where N_1 is the number of treated observations. Denoting the propensity score estimate by $\tilde{p}(X_i) = p(X_i; \tilde{\beta})$ highlights that it may differ from $\hat{p}(X_i)$ by adjusting the coefficients $\hat{\beta}$ of the initial propensity score model to $\tilde{\beta}$ such that they exactly balance $\tilde{X}_i$ in the sample, i.e. such that (8) equals zero. In the simulations, we set $\tilde{X}_i = X_i$, so that the first moment of each covariate is

exactly balanced when using CBPS for propensity score estimation. We follow Imai and Ratkovic (2014) and use a logit model for the propensity score, i.e. $p(X_i;\beta) = \frac{\exp((1,X_i')\beta)}{1+\exp((1,X_i')\beta)}$. When using only X_i as moments conditions, the CBPS is exactly identified, and we will label this the *just identified CBPS* estimator.

However, the moment condition (7) can also be combined with the first order condition of the maximum likelihood estimator, which leads to the *overidentified CBPS*. Let $s(D_i, X_i;\beta) = \frac{D_i \frac{\partial p(X_i;\beta)}{\partial \beta}}{p(X_i;\beta)} - \frac{(1-D_i)\frac{\partial p(X_i;\beta)}{\partial \beta}}{1-p(X_i;\beta)}$ denote the score function of the ML estimator of β. Following Imai and Ratkovic (2014) we use the GMM estimator

$$\tilde{\beta} = \arg\min_{\bar{\beta}} \overline{g}(D, X;\bar{\beta})' \cdot \Xi(D, X;\bar{\beta})^{-1} \cdot \overline{g}(D, X;\bar{\beta}) \tag{9}$$

where $\overline{g}(D, X;\bar{\beta}) = N^{-1}\sum_{i=1}^{N} g(D_i, X_i;\bar{\beta})$ is the sample mean of the moment conditions $g(D_i, X_i;\bar{\beta}) = \begin{pmatrix} s(D_i, X_i;\bar{\beta}) \\ w(D_i, X_i;\bar{\beta})\tilde{X}_i \end{pmatrix}$, which combine the balancing conditions and the score condition. Finally, $\Xi(D, X;\bar{\beta})$ denotes the covariance matrix of $g(D_i, X_i;\bar{\beta})$

$$\Xi(D, X;\bar{\beta}) = N^{-1}\sum_{i=1}^{N} \begin{pmatrix} p(X_i;\bar{\beta})\{1-p(X_i;\bar{\beta})\}X_iX_i' & Np(X_i;\bar{\beta})X_i\tilde{X}_i'/N_1 \\ Np(X_i;\bar{\beta})\tilde{X}_iX_i'/N_1 & N^2p(X_i)/[N_1^2\{1-p(X_i;\bar{\beta})\}]\tilde{X}_i\tilde{X}_i' \end{pmatrix}$$

In the simulations, we consider the just identified CBPS method for almost all propensity score based estimators. Only for IPW, also overidentified CBPS is investigated to compare both methods.[2]

From an applied perspective, one attractive feature of such EL approaches as well as the entropy balancing method outlined in Section 2.4.3 over conventional propensity score estimation is that iterative balance checking and searching for propensity score specifications that entail balancing is not required.

[2]Examining overidentified CBPS for all treatment effect estimators would have been computationally too demanding. Neither just identified, nor overidentified CBPS is used in the case of inverse probability tilting (see Section 2.3.1), which constitutes yet another EL method for exact balancing.

2.2.3 Semiparametric estimation of the propensity score

As a third approach to propensity score estimation, we apply the semiparametric binary choice estimator suggested by Klein and Spady (1993). The latter assumes the propensity score to be a nonparametric function of a linear index of the covariates:

$$p(X_i) = p(X_i; \beta) = \eta(X_i'\beta), \tag{10}$$

where the link function η is unknown. This is more general than fully parametric models, that specify the link function (e.g., $\eta = \Phi$) and therefore assume a particular distribution of the error terms, which is not the case here. What remains parametrically specified is the linear index. Estimation is based on the following ML kernel regression approach:

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{N} \{D_i \log[\hat{p}(X_i; \beta] + (1 - D_i) \log[1 - \hat{p}(X_i; \beta)]\}, \tag{11}$$

where

$$\hat{p}(X_i; \beta) = \frac{\sum_{j=1}^{N} D_j K\left(\frac{X_i'\beta - X_j'\beta}{h}\right)}{\sum_{j=1}^{N} K\left(\frac{X_i'\beta - X_j'\beta}{h}\right)} \tag{12}$$

is the propensity score estimate under a particular (candidate) coefficient value β. The estimated propensity score for observation i is therefore $\hat{p}(X_i; \hat{\beta})$. $K(\cdot)$ denotes the kernel function, in the case of our simulations the Epanechnikov kernel. h is the kernel bandwidth, which is chosen through cross-validation by maximizing the leave-one-out log likelihood function jointly with respect to the bandwidth and the coefficients, as implemented in the 'np' package for the statistical software R by Hayfield and Racine (2008).

2.2.4 Nonparametric estimation of the propensity score

Our final propensity score estimator relies on local constant (Nadaraya-Watson) kernel regression, and is therefore fully nonparametric, because a linear index is no longer assumed:

$$\hat{p}_i = \hat{p}(X_i) = \frac{\sum_{j=1}^{N} D_j K\left(\frac{X_i - X_j}{h}\right)}{\sum_{j=1}^{N} K\left(\frac{X_i - X_j}{h}\right)}. \tag{13}$$

To be concise, we use the kernel regression method of Racine and Li (2004), which allows for both continuous and discrete regressors and is implemented in the 'np' package of Hayfield and Racine (2008). $K(\cdot)$ now denotes a product kernel (i.e., the product of several kernel functions), because X is multidimensional. For continuous elements in X, the Epanechnikov kernel is used, while for ordered and unordered discrete regressors, the kernel functions are based on Wang and van Ryzin (1981) and Aitchison and Aitken (1976), respectively. The bandwidth h is selected via Kullback-Leibler cross-validation, see Hurvich, Simonoff, and Tsai (1998). While the nonparametric propensity score estimator is most flexible in terms of functional form assumptions, it may have a larger variance than (semi)parametric methods in finite samples.

2.3 Propensity score-based estimators of ATET

In the previous subsection we discussed various estimators of the propensity score. These estimates of the propensity score, which we henceforth denote as $\hat{p}_i$ or as $\hat{p}(X_i)$, are now being used as plug-in estimates in the following estimators of the ATET. All the estimators of Δ discussed in this section make use of the estimated propensity scores. (In Section (2.4) we will examine estimators of Δ that do not use the propensity score.)

2.3.1 Inverse probability weighting

Inverse probability weighting (IPW) bases estimation on weighting observations by the inverse of their propensity scores and goes back to Horvitz and Thompson (1952). For our parameter of

interest Δ, it is the non-treated outcomes that are reweighted in order to control for differences in the propensity scores between treated and non-treated observations. Hirano, Imbens, and Ridder (2003) discuss the properties of IPW estimators of average treatment effects, which can attain the semiparametric efficiency bound derived by Hahn (1998) if the propensity score is nonparametrically estimated (which is generally not the case for parametric propensity scores).[3] In our simulations, we consider the following normalized IPW estimator:

$$\hat{\Delta}_{\text{IPW}} = N_1^{-1} \sum_{i=1}^{N} D_i Y_i - \sum_{i=1}^{N} (1 - D_i) Y_i \left\{ \frac{\frac{\hat{p}_i}{1-\hat{p}_i}}{\sum_{j=1}^{N} \frac{(1-D_j)\hat{p}_j}{1-\hat{p}_j}} \right\}, \tag{14}$$

where the normalization $\sum_{j=1}^{N} \frac{(1-D_j)\hat{p}_j}{1-\hat{p}_j}$ ensures that the weights sum up to one, see Imbens (2004) for further discussion. This estimator was very competitive in the simulation study of Busso, DiNardo, and McCrary (2009).

IPW has the advantages that it is easy to implement, computationally inexpensive, and does not require choosing any tuning parameters (other than for propensity score estimation). However, it also has potential shortcomings. Firstly, estimation is likely sensitive to propensity scores that are 'too' close to one, as suggested by simulations in Frölich (2004) and Busso, DiNardo, and McCrary (2009) and discussed in Khan and Tamer (2010) on theoretical grounds. Secondly, IPW may be less robust to propensity score misspecification than matching (which merely uses the score to match treated and non-treated observations, rather than plugging it directly into the estimator), see Waernbaum (2012).

2.3.2 Inverse probability tilting

A variation of IPW is inverse probability tilting (IPT) as suggested in Graham, Pinto, and Egel (2012), an empirical likelihood (EL) approach entailing exact balancing of the covariates, which is therefore somewhat related to the CBPS procedure of Imai and Ratkovic (2014). The IPT method

[3] See also Ichimura and Linton (2005) and Li, Racine, and Wooldridge (2009) for a discussion of the asymptotic properties of IPW when using nonparametric kernel regression for propensity score estimation, rather than series estimation as in Hirano, Imbens, and Ridder (2003).

appropriate for estimating Δ and also considered in our simulations is the so-called 'auxiliary to study' tilting, see Graham, Pinto, and Egel (2011), which (in contrast to Imai and Ratkovic (2014)) estimates separate propensity scores for the treated and non-treated observations. The method of moments estimator of the propensity scores of the non-treated is based on adjusting the coefficients of the initial (parametric) propensity score $\hat{p}(X_i)$ such that the following, efficiency-maximizing moment conditions are satisfied:

$$\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}(1-D_i)\,\frac{\hat{p}(X_i)}{1-\tilde{p}_0(X_i)}\cdot\frac{1}{\frac{1}{N}\sum_{j=1}^{N}\hat{p}(X_j)}\\(1-D_i)\,\tilde{X}_i\frac{\hat{p}(X_i)}{1-\tilde{p}_0(X_i)}\cdot\frac{1}{\frac{1}{N}\sum_{j=1}^{N}\hat{p}(X_j)}\end{pmatrix}=\frac{1}{N}\sum_{i=1}^{N}\begin{pmatrix}1\\\hat{p}_i\tilde{X}_i\cdot\frac{1}{\frac{1}{N}\sum_{j=1}^{N}\hat{p}(X_j)}\end{pmatrix}, \tag{15}$$

where $\tilde{X}_i$ is a (possibly multidimensional) function of the covariates X_i, and where $\hat{p}(X_i) = p(X_i;\hat{\beta})$ is the (initial) ML-based propensity score and $\tilde{p}_0(X_i) = p(X_i;\tilde{\beta})$ the modified propensity score. That is, the coefficients $\tilde{\beta}$ are chosen such that the reweighted moments of the covariates among non-treated observations on the left hand side of (15) are numerically identical to the efficiently estimated moments among the treated on the right hand side. Analogously, the IPT propensity score among the treated $\tilde{p}_1(X_i)$ is estimated by replacing $1-D_i$ with D_i and $1-\tilde{p}_0(X_i)$ with $\tilde{p}_1(X_i)$ in (15) such that the moments on the left hand side, which now refer to the treated, again coincide with the right hand side. Δ is then estimated by

$$\hat{\Delta}_{\text{IPT}} = \sum_{i=1}^{N}\frac{D_i}{\tilde{p}_1(X_i)}\frac{\hat{p}(X_i)}{\sum_{j=1}^{N}\hat{p}(X_j)}Y_i - \sum_{i=1}^{N}\frac{1-D_i}{1-\tilde{p}_0(X_i)}\frac{\hat{p}(X_i)}{\sum_{j=1}^{N}\hat{p}(X_j)}Y_i. \tag{16}$$

In the simulations, we consider IPT only for probit-based estimation of the (initial) propensity score $\hat{p}(X_i)$ and set $\tilde{X}_i = X_i$ so that the covariate means are balanced. In contrast, IPW is analyzed for all propensity score methods outlined in Section 2.2 and represents under the (just and overidentified) CBPS method of Imai and Ratkovic (2014) yet another EL approach.

2.3.3 Doubly robust estimation

Doubly robust (DR) estimation combines IPW with a model for the conditional mean outcome (as a function of the treatment and the covariates). It owes its name to the fact that it remains consistent if either the propensity score or the conditional mean outcome are correctly specified, see for instance Robins, Mark, and Newey (1992) and Robins, Rotnitzky, and Zhao (1995). If both models are correct, DR is semiparametrically efficient, as shown in Robins, Rotnitzky, and Zhao (1994) and Robins and Rotnitzky (1995). Kang and Schafer (2007) discuss various approaches how to implement DR estimation in practice. Despite the theoretical attractiveness of the double robustness property, their simulation results suggest that DR may (similar to IPW) be sensitive to misspecifications of the propensity score if some propensity scores estimates are close to the boundary.

In our simulations, we consider the following DR estimator, which is based on the sample analog of the semiparametrically efficient influence function, see Rothe and Firpo (2013):

$$\hat{\Delta}_{\text{DR}} = \frac{1}{N_1} \sum_{i=1}^{N} \left(D_i(Y_i - \hat{\mu}(0, X_i)) - \hat{p}(X_i) \frac{(1 - D_i)(Y_i - \hat{\mu}(0, X_i))}{1 - \hat{p}(X_i)} \right), \tag{17}$$

where $\hat{\mu}(D, X)$ is an estimate of the conditional mean outcome $\mu(D, X) = E(Y|D, X)$.

We consider *five* different versions of DR, depending on how the conditional mean outcomes and propensity scores are estimated. The standard approach in the applied literature is estimating both $\hat{p}(X_i)$ and $\hat{\mu}(0, X_i)$ based on parametric models, in our case by probit and linear regression of Y_i on a constant and X_i among the non-treated, respectively. We also combine linear outcome regression with propensity score estimation by (just identified) CBPS (Imai and Ratkovic (2014)) and semiparametric regression (Klein and Spady (1993)). Finally, we follow Rothe and Firpo (2013) who suggest nonparametrically estimating both the propensity score *and* the conditional mean outcome. For the latter, we use local linear kernel regression[4] of Y on X (using the

[4]Local linear regression is superior to local constant estimation in terms of boundary bias (which is for local linear regression the same as in the interior, see Fan (1993)). For nonparametric propensity score estimation we nevertheless use local constant regression to prevent the possibility of predictions outside the theoretical bounds of zero and one.

'np' package of Hayfield and Racine (2008)) among the non-treated. We consider estimating (17) based on two different bandwidth choices for the outcome regression: First, we use the bandwidth suggested by least squares cross-validation, which we will refer to as *crossval bandwidth*. Second, we divide this bandwidth by two, which we will refer to as *undersmoothed bandwidth*. The latter is motivated by the general finding in the literature that $\sqrt{n}$-consistent estimation of the ATET requires undersmoothing. The kernel-based estimation of the propensity score proceeds as outlined in Section 2.2.4.

Strictly speaking the label 'DR' is misleading for the Rothe and Firpo (2013) estimator, because (17) would already be consistent under the nonparametric estimation of *either* $\hat{p}(X_i)$ *or* $\hat{\mu}(0, X_i)$. However, Rothe and Firpo (2013) show that by estimating both models nonparametrically, (17) has a lower first order bias and second order variance than either IPW (using a nonparametric propensity score) or treatment evaluation based on nonparametric outcome regression (see Section 2.4.2 below). Furthermore, its finite sample distribution is less dependent on the accuracy of $\hat{p}(X_i)$ and $\hat{\mu}(0, X_i)$ – and thus on the bandwidth choice in kernel regression – and it can be approximated more precisely by first order asymptotics. For these reasons, DR may appear relatively more attractive from an applied perspective, even though by first-order asymptotics, DR, IPW, and regression are all normally distributed and equally efficient, attaining the semiparametric efficiency bound of Hahn (1998) for appropriate bandwidth choices.

2.3.4 Propensity score matching

Propensity score matching is based on finding for each treated observation one or more non-treated units that are comparable in terms of the propensity score. The average difference in the outcomes of the treated and the (appropriately weighted) non-treated matches yields an estimate of Δ. As discussed in Smith and Todd (2005), matching estimators have the following general form:

$$\hat{\Delta}_{\text{match}} = N_1^{-1} \sum_{i:D_i=1} \left(Y_i - \sum_{j:D_j=0} W_{i,j} Y_j \right), \tag{18}$$

where $W_{i,j}$ is the weight the outcome of a non-treated unit j is given when matched to some treated observation i. In the simulations we consider three classes of propensity score matching estimators: pair matching, radius matching, and kernel matching.

The prototypical pair-matching (also called one-to-one matching) estimator with respect to the propensity score and with replacement,[5] see for instance Rosenbaum and Rubin (1983), matches to each treated observation exactly the non-treated observation that is most similar in terms of the propensity score. The weights in (18) therefore are

$$W_{i,j} = I\left(|\hat{p}(X_j) - \hat{p}(X_i)| = \min_{l:D_l=0} |\hat{p}(X_l) - \hat{p}(X_i)| \right), \tag{19}$$

where $I\{\cdot\}$ is the indicator function which is one if its argument is true and zero otherwise. I.e. all weights are zero except for that observation j that has smallest distance to i in terms of the estimated propensity score.

Pair matching is not efficient, because only one non-treated observation is used for each treated one, irrespective of the sample size and of how many potential matches with similar propensity scores are available. On the other hand, it is likely more robust to propensity score misspecification than for instance IPW, in particular if the misspecified propensity score model is only a monotone transformation of the true model, see Zhao (2008) and Millimet and Tchernis (2009) for some affirmative results. In the simulations, we include pair matching based on all four estimation approaches of the propensity scores discussed in Section 2.2.

In contrast to pair matching, radius matching (see for instance Rosenbaum and Rubin (1985) and Dehejia and Wahba (1999)) uses *all* non-treated observations whith propensity scores within a predefined radius around that of the treated reference observation. This may increase efficiency if several good potential matches are available (at the cost of a somewhat higher bias). In the simulations, we consider the radius matching algorithm of Lechner, Miquel, and Wunsch (2011),

[5] 'With replacement' means that a non-treated observation may serve several times as a match, whereas estimation 'without replacement' requires that it is not used more than once. The latter approach is only feasible when there are substantially more non-treated than treated observations and is not frequently applied in econometrics. It is not considered in our simulations either, which consider shares of 50% treated and 50% non-treated.

which performed overall best in Huber, Lechner, and Wunsch (2013). The Lechner, Miquel, and Wunsch (2011) estimator combines distance-weighted radius matching with an OLS regression adjustment for bias correction (see Rubin (1979) and Abadie and Imbens (2011)). Furthermore and as suggested in Rosenbaum and Rubin (1985), it includes the option to directly match on additional covariates in addition to the propensity score (which are, however, also included in the propensity score) based on the Mahalanobis distance metric (defined in equation (21)). The first estimation step consists of radius matching either on the propensity score or the Mahalanobis metric based on the score and the additional covariates, respectively. Distance-weighting implies that non-treated within the radius are weighted proportionally to the inverse of their distance to the treated reference observation, so that this approach can also be regarded as kernel matching (see below) using a truncated kernel. Secondly, the matching weights are used in a weighted linear regression to remove small sample bias due to mismatches. (This bears some similarities to the doubly robust approach, albeit using a linear model.) For a detailed description of the (algorithm of the) estimator, we refer to Huber, Lechner, and Steinmayr (2014).

An important question is how to determine the radius size, for which no well-established algorithm exists. We follow Lechner, Miquel, and Wunsch (2011) and define it as a function of the distribution of distances between treated and matched non-treated observations in pair matching. In our simulations, we define the radius size to be either $\frac{1}{3}$, 1, or 3 times the 0.95 quantile of the matching distances.[6] If not even a single non-treated observation is within the radius, the closest observation is taken as the match just as in pair-matching. As in Huber, Lechner, and Wunsch (2013), we investigate the performance when matching (i) on the propensity score only as well as (ii) on both the propensity score and on two important confounders (that also enter the propensity score) directly, based on the Mahalanobis metric outlined in equation (21). This hybrid of propensity score and direct matching (see Section 2.4.1) allows assuring that such important confounders are given priority in terms of balancing their distributions across treatment states.[7]

[6]Basing the choice on a particular quantile may be more robust to outliers than simply taking the maximum distance in pair matching.

[7]In our case, 'number of unemployment spells in the last two years prior to treatment' and the interaction

All in all, we consider 24 different radius matching estimators based on the four different propensity score estimators and three radius sizes for each propensity score and Mahalanobis distance matching. The latter, however, generally performs worse than radius matching on the propensity score alone in our simulations. Therefore, the results of (Mahalanobis) matching on the propensity score and further confounders are not presented, so that the number of Lechner, Miquel, and Wunsch (2011)-type radius matching estimators discussed in the paper reduces to 12. (The omitted results are available from the authors upon request.)

2.3.5 Propensity score kernel regression

Propensity score kernel regression is based on first estimating the conditional mean outcome given the propensity score without treatment, $m(0,\rho) = E(Y|D=0, p(X)=\rho)$, by kernel regression of the outcome on the estimated propensity score among the non-treated and then averaging the estimates according to the propensity score distribution of the treated. Formally,

$$\hat{\Delta}_{\text{kernmatch}} = N_1^{-1} \sum_{i:D_i=1} \left(Y_i - \hat{m}(0, \hat{p}(X_i))\right), \tag{20}$$

where $\hat{m}(0,\hat{p}(X_i))$ is an estimate of $m(0,p(X_i))$. This estimator, which has been discussed in Heckman, Ichimura, and Todd (1998a) and Heckman, Ichimura, Smith, and Todd (1998), again satisfies the general structure of (18) with $\hat{m}(0,\hat{p}(X_i)) = \sum_{j:D_j=0} W_{i,j} Y_j$. $W_{i,j}$ now reflects the kernel-based weights related to the difference $\hat{p}(X_j) - \hat{p}(X_i)$, which are provided in Busso, DiNardo, and McCrary (2009) for various kernel methods. Frölich (2004) investigated the finite sample performance of several kernel matching approaches and found estimation based on ridge regression, which extends local linear regression by adding a (small) ridge term to the estimator's denominator to prevent division by values close to zero (see Seifert and Gasser (1996)), to perform best. In the simulations, we use local linear regression and the Epanechnikov kernel

between 'French speaking region' and 'jobseeker gender' - see Section 3.2 for a discussion of the covariates in the data - are used as matching variables besides the propensity score, which are important predictors of both treatment and outcome. Yet, the propensity score is given five times as much weight as either covariate in the Mahalanobis metric, to account for the fact that it represents all covariates.

for the estimation of $m(0,\rho)$ based on the 'np' package, which also includes a form of ridging. Three different kernel bandwidths are considered: The bandwidth suggested by cross-validation is labelled cross-validation bandwidth. In addition, we examine the case where we double this bandwidth (referred to as oversmoothing) and where we take half that bandwidth (referred to as undersmoothing).[8] As the procedures are implemented based on all four propensity score approaches, all in all twelve kernel matching estimators are included in the simulations.

2.4 Entropy balancing and matching/regression (directly) on the covariates

This section discusses methods that do *not* work through a propensity score model, but rather condition on the covariates directly. I.e. we estimate the conditional mean $E[Y|D=0,X]$ by alternative methods and then average within the $D=1$ subpopulation, according to equation (1).

2.4.1 Pair, one-to-many and radius matching on the covariates

Matching on the covariates directly rather than the propensity score is rarely considered in applied work. Technically, estimation is nevertheless straightforward to implement, once a distance metric has been chosen that weights the differences in the various covariates between treated and non-treated matches in a specific way. Among the most commonly used metrics is the Mahalanobis distance metric, which is defined for some covariate vectors X_i and X_j as follows:

$$||X_i - X_j|| = \sqrt{(X_i - X_j)'C^{-1}(X_i - X_j)}, \tag{21}$$

where C denotes the covariance matrix of the covariates.

Building on this distance metric, we examine various estimators. *Pair matching on the co-*

[8] It is worth noting that cross validation aims at determining the crossval bandwidth for the estimation of the conditional mean function $m(0,\rho)$, not the actual parameter of interest, Δ. Even though Frölich (2005) provides a plug-in method for choosing the bandwidth that is optimal for kernel matching based on an approximation of the mean squared error, his simulations suggest that the approximation is not sufficiently accurate for the sample sizes considered. On the other hand, Frölich (2005) finds conventional cross-validation to perform rather well and we therefore follow this latter approach.

variates is defined as (18), with weights

$$W_{i,j} = I\left(||X_j - X_i|| = \min_{l:D_l=0} ||X_l - X_i|| \right). \tag{22}$$

(This estimator is thus similar to propensity score matching with the only difference that the distance metric is $||X_j - X_i||$ instead of $|\hat{p}(X_j) - \hat{p}(X_i)|$.) In the simulations, we consider pair matching both with and without regression adjustment for bias correction (see Abadie and Imbens (2011)) using the 'Matching' package for R of Sekhon (2011).

Secondly, we also investigate the performance of *one-to-many* matching, implying that the M closest non-treated observations (in terms of the Mahalanobis distance) are matched to each treated. In other words, for each treated observation i we find the M nearest neighbours among the non-treated observations, where nearness is measured by Mahalanobis distance to i. Then each of the M nearest neighbours receives a weight of $\frac{1}{M}$, whereas all other non-treated observations receive a weight of zero. The pair matching estimator is included as a special case when setting $M = 1$.

Increasing M reduces the variance compared to pair matching, at the cost of increasing the bias due to relying on more and potentially worse matches. In the simulations, we set $M = 5$ and perform one-to-many matching estimation with and without bias correction. Finally, we also consider a particular form of radius matching on the covariates, again based on the 'Matching' package for R. The radius is defined such that only non-treated observations satisfying that all their covariate values are not more than 0.25 standard deviations of the respective covariate away from the treated reference observation are used. Again, radius matching with and without bias correction is included in the simulations.[9]

Finally, we consider the Genetic Matching algorithm of Diamond and Sekhon (2013), a generalization of Mahalanobis distance matching on the covariates. It aims at optimizing covariate balance according to a range predefined balance metrics (which is in spirit somewhat related to

[9]In contrast to the Lechner, Miquel, and Wunsch (2011) propensity score-based radius matching estimator discussed in Section 2.3.4, no distance weighting is applied within the radius.

the EL methods and entropy balancing, see Section 2.4.3 below). This is obtained by using a weighted version of the Mahalanobis distance metric, where the weights are chosen such that a particular loss function reflecting overall imbalance is minimized. The procedure's default loss function, which is the one considered in our simulations, requires the algorithm to minimize the largest discrepancies for all elements in X according to the p-values from Kolmogorov-Smirnov (KS) tests (for equalities in covariate distributions) and paired t-tests (for equalities in covariate means). Note that as the algorithm is based on p-values (rather than test statistics directly), the outcomes of the different tests (e.g. KS and t-tests) can be compared on the same scale. Formally, the generalized version of the Mahalanobis distance metric is defined as

$$||X_i - X_j||_W = \sqrt{(X_i - X_j)'(C^{-1/2})'WC^{-1/2}(X_i - X_j)}, \tag{23}$$

where $C^{-1/2}$ is the Cholesky decomposition of C, the covariance matrix of the covariates. W denotes a (positive definite) weighting matrix of the same dimension as C, and is chosen iteratively until overall imbalance is minimized. For a more detailed discussion of the Genetic Matching, we refer to the flow chart of the algorithm in Figure 2 of Diamond and Sekhon (2013). In the simulations Genetic Matching both with and without bias adjustment is considered.

2.4.2 Regression

We also consider estimation of Δ based on parametric and nonparametric regression of the outcome on the covariates under non-treatment:

$$\hat{\Delta}_{\text{reg}} = N_1^{-1} \sum_{i:D_i=1} \left(Y_i - \hat{\mu}(0, X_i)\right), \tag{24}$$

where $\hat{\mu}(D, X)$ is an estimate of the conditional mean outcome $\mu(D, X) = E(Y|D, X)$. In the parametric case, $\hat{\mu}(0, X_i)$ is predicted based on the coefficients of an OLS regression among the non-treated. $\hat{\Delta}_{\text{reg}}$ then corresponds to what is called the unexplained component in the linear decomposition of Blinder (1973) and Oaxaca (1973).

In the nonparametric case, we apply local linear regression using the 'np' package of Hayfield and Racine (2008). Analogosuly to DR estimation, see equation (17), the performance of the estimator under two different bandwidths is investigated: We use the bandwidth obtained by least squares cross validation and alternatively this bandwidth divided by half (which we refer to as undersmoothing). Note that the kernel regression-based method, which may attain the semiparametric efficiency bound of Hahn (1998), can also be interpreted as kernel matching estimator on X_i (rather than on the propensity score as in the kernel matching procedure of Section 2.3). To see this, note that (24) satisfies the general notation for matching estimators in (18), with $W_{i,j}$ representing the kernel-based weights related to the difference $X_j - X_i$.

2.4.3 Entropy balancing

Entropy balancing (EB) as suggested in Hainmueller (2012) aims at balancing the covariates across treatment groups based on a maximum entropy reweighting scheme which does not rely on a propensity score model. That is, it calibrates the weights of the non-treated observations so that exact balance in prespecified covariate moments (e.g. the mean) is obtained for the reweighted non-treated group and the treated. Even though this is similar in spirit to the EL methods discussed in Sections 2.2.2 and 2.3.2, one difference is that the latter start out from a specific propensity score model, while EB relies on user-provided (initial) base weights (e.g., uniform weights). The weights finally estimated are computed such that the Kullback-Leibler divergence from the baseline weights is minimized, subject to the balancing constraints. Hainmueller (2012) points out that similar to (conventional) IPW, the estimator may have a large variance when only few non-treated observations obtain large weights due to weak overlap in the covariate distributions across treatment states.

Technically, the weights for the nontreated are chosen by minimizing the following loss function, while at the same time balancing $\tilde{X}_i$, which is a (possibly multidimensional) function of the

covariate vector X_i

$$\min_{\omega_i} \sum_{\{i:D_i=0\}} h(\omega_i) \tag{25}$$

subject to the balance constraint

$$\sum_{i:D_i=0} \omega_i \tilde{X}_i = \frac{1}{N_1} \sum_{i:D_i=1} \tilde{X}_i \tag{26}$$

and the normalizing constraints

$$\sum_{\{i:D_i=0\}} \omega_i = 1 \quad \text{and} \quad \omega_i \geq 0 \quad \text{for all} \quad i \quad \text{with} \quad D_i = 0. \tag{27}$$

ω_i denotes the estimated weight for observation i and $h(\cdot)$ is a distance metric. Hainmueller (2012) proposes using the directed Kullback (1959) entropy divergence defined by $h(\omega_i) = \omega_i \log(\omega_i/q_i)$, with q_i denoting the (initial) base weight. The loss function $\sum_{\{i:D_i=0\}} h(\omega_i)$ measures the distance between the distributions of the estimated weights $\omega_1, \ldots, \omega_{N_0}$ and the base weights $q_1, \ldots, q_{n_0}$. The balance constraint (26) equalizes $\tilde{X}$ between the treatment and the reweighted non-treated group. The normalizing constraints (27) ensure that the weights sum up to unity and do not take on negative values. Hainmueller (2012) shows that a tractable and unique solution (if one exists) can be obtained based on the Lagrange multiplier, see his Section 3.2. In our simulations, EB uses the default option of uniform base weights, $q_i = N_0^{-1}$ where N_0 is the number of non-treated observations. Furthermore, $\tilde{X}$ is set to X, so that the covariate means are balanced across treatment groups.

3 Data and simulation design

3.1 Overview

Inspired by Huber, Lechner, and Wunsch (2013), we base our simulation design as much as possible on empirical data rather than on data generating processes (DGP) that are fully artificial

and prone to the arbitrariness of the researcher. By using empirical (rather than simulated) associations between the treatment, the covariates, and (in the case of effect heterogeneity) the outcomes, we hope to more closely mimic real world evaluation problems. The 'empirical Monte Carlo study' (EMCS) design nevertheless requires calibrating several important simulation parameters, namely the strength of selection into the treatment, treatment effect heterogeneity, and sample size, in order to consider a range of different scenarios. As in the simulation study by Huber, Lechner, and Mellace (2014a) (however, on the methodologically different framework of mediation analysis), our EMCS uses a large-scale Swiss labor market data set with linked jobseeker-caseworker information first analyzed by Behncke, Frölich, and Lechner (2010a,b).

The simulation design is as follows. First, we match to each treated observation that non-treated observation which is most similar in terms of the covariates. Matching proceeds without replacement. The latter matches serve as (pseudo-)treated 'population' for our simulations, the remaining unmatched subjects as non-treated 'population'. Second, we repeatedly draw simulation samples with replacement out of the 'populations', which consist of 50% pseudo-treated and 50% non-treated. By definition, the true treatment effects are zero (as not even the pseudo-treated actually received any treatment) and therefore homogeneous.

Additionally, in order to investigate estimator performance under heterogeneous effects, we also model the outcome as a function of the treatment and the covariates in our initial data (before generating the pseudo-treatments). We consider two different treatment variables that differ in terms of selection into treatment: participation in an active labor market program and assignment to a cooperative or noncooperative caseworker, see Behncke, Frölich, and Lechner (2010b) and Huber, Lechner, and Mellace (2014b) for empirical investigations of these variables. Furthermore, we vary the sample size (750 and 3000 observations) and whether estimation controls for all confounders (correct specification) or omits some of them (misspecification), a case likely to occur in empirical applications. All in all, the various simulation parameters entail 16 different scenarios. Our EMCS includes the so far most comprehensive set of treatment effect estimators (in particular, also a range of fully nonparametric methods) and assesses their (relative) performance

in terms of the mean squared error, both with and without trimming (i.e. discarding) observations with 'too' extreme treatment propensity scores.

We subsequently present the details of how the EMCS is implemented. The next section describes the data sources and the definitions of the treatments, outcomes, and covariates and also provides descriptive statistics. Section 3.3 outlines the simulation design with the various simulation parameters (selection, heterogeneity, misspecification, sample size) and also discusses the various trimming rules considered.

3.2 Data and definition of treatments, outcomes, and covariates

As for Huber, Lechner, and Mellace (2014a), the data used in our EMCS include individuals who registered at Swiss regional employment offices anytime during the year 2003. Detailed jobseeker characteristics are available from the unemployment insurance system and social security records, including gender, mother tongue, qualification, information on registration and deregistration of unemployment, employment history, participation in active labor market programs, and an employability rating by the caseworker in the employment office. Regional (labour market relevant) variables such as the cantonal unemployment rate were also matched to the jobseeker information. These administrative data were linked to a caseworker survey based on a written questionnaire that was sent to all caseworkers in Switzerland who were employed at an employment office in 2003 and still active in December 2004 (see Behncke, Frölich, and Lechner (2010b) for further details). The questionnaire included questions about aims, strategies, processes, and organisation of the employment office and the caseworkers. The definition of the jobseeker sample ultimately used for our simulations closely follows the sample selection criteria in Behncke, Frölich, and Lechner (2010b) so that we refer to their paper for further details.[10]

The outcome variable Y in the simulations is defined as the cumulative months an individual

[10]Our final sample size differs slightly from theirs because we exclude, following Huber, Lechner, and Mellace (2014a), individuals who were registered in the Italian-speaking part of Switzerland in order to reduce the number of language interaction terms to be included in the model. We also deleted 102 individuals who registered with the employment office before 2003. The final sample therefore consists of 93,076 unemployed persons (rather than 100,222 as in Behncke, Frölich, and Lechner (2010b)).

was a jobseeker between (and including) month 10 and month 36 after start of the unemployment spell. Figure A.1 in Appendix A.1 displays the distribution of the semi-continuous outcome, which has a large probability mass at zero months. We consider two distinct treatment variables D. The first is defined in terms of participation in an active labour market program within the 9 months after the start of the unemployment spell. Possible program participation states in the data include job search training, personality course, language skill training, computer training, vocational training, employment program or internship. The alternative is non-participation in any program. For the simulations, the treatment state is one if an individual participates in any program in the 9 months window and zero otherwise. In the data, 26,062 observations (or 28%) participate in at least one program, while 67,014 or (72%) do not. The second treatment comes from the caseworker questionnaire and is defined in terms of how important the caseworker considers cooperation with the jobseeker, i.e., whether the aim is to satisfy wishes of the jobseeker or whether the caseworker's strategy is rather independent of the jobseeker's preferences. As in the main specification of Behncke, Frölich, and Lechner (2010b), the treatment D is defined to be one if the caseworker reports to pursue a noncooperative strategy (43,669 observations or 47%) and zero otherwise (49,407 observations or 53%).

Table 1: Descriptive statistics under various treatment states

	program participation				noncooperative caseworker			
	D=1		D=0		D=1		D=0	
	mean	std	mean	std	mean	std	mean	std
female jobseeker	0.465	0.499	0.430	0.495	0.430	0.495	0.449	0.497
foreign mother tongue	0.317	0.465	0.319	0.466	0.324	0.468	0.314	0.464
unskilled*	0.220	0.414	0.215	0.411	0.224	0.417	0.209	0.407
semiskilled*	0.155	0.362	0.165	0.372	0.163	0.370	0.162	0.368
skilled, no degree*	0.043	0.203	0.042	0.201	0.044	0.205	0.041	0.199
unemployment spells last 2 years	0.320	0.837	0.657	1.294	0.570	1.205	0.556	1.183
fraction employed last yr	0.807	0.261	0.797	0.248	0.799	0.251	0.800	0.252
low employability**	0.132	0.338	0.139	0.346	0.142	0.349	0.133	0.339
medium employab.**	0.772	0.419	0.748	0.434	0.754	0.431	0.755	0.430
female caseworker	0.444	0.497	0.411	0.492	0.420	0.494	0.421	0.494
caseworker above vocational+	0.451	0.498	0.444	0.497	0.422	0.494	0.467	0.499
caseworker higher education+	0.238	0.426	0.244	0.429	0.239	0.427	0.245	0.430
cantonal unemployment rate	3.680	0.828	3.694	0.870	3.708	0.883	3.675	0.836
French speaking region	0.218	0.413	0.269	0.443	0.234	0.424	0.273	0.445
obs.	26,062		67,014		43,669		49,407	

Note: 'mean' and 'sd' give the means and standard deviations of the covariates in the respective treatment groups. *: Reference category is 'skilled with degree'. **: Reference category is 'high employability'. +: Reference category is 'caseworker has vocational training or lower'.

The covariates X which serve as confounders in our EMCS have been used as control variables in Behncke, Frölich, and Lechner (2010b) and Huber, Lechner, and Mellace (2014b) to control for selection into program participation and/or assignment to a noncooperative caseworker, respectively. As jobseeker characteristics, we include gender, a dummy for whether the mother tongue is not one of the official languages in Switzerland, the level of qualification (unskilled, semiskilled, skilled without degree, skilled with degree), the previous employment history (namely the number of unemployment spells in last two years and the proportion of time employed in the last year), and an employability rating by the caseworker (low, medium, high). We also use several caseworker characteristics coming from the questionnaire, namely gender, education (vocational training or lower, above vocational, higher education). Concerning regional characteristics, the cantonal unemployment rate and a dummy for a French speaking region enter the simulation design. Table 1 provides the means and standard deviations of the covariates across the various treatment states of the two treatment variables 'program participation' and 'noncooperative caseworker'.

Table 2: Probit regressions of treatments on covariates in the data

	program participation				noncooperative caseworker			
	coef	se	z-val	p-val	coef	se	z-val	p-val
constant	-0.779	0.031	-24.740	0.000	-0.139	0.029	-4.749	0.000
female jobseeker	0.038	0.010	3.663	0.000	-0.033	0.010	-3.328	0.001
foreign mother tongue	0.045	0.012	3.767	0.000	-0.027	0.011	-2.425	0.015
unskilled	0.048	0.012	3.836	0.000	0.056	0.012	4.802	0.000
semiskilled	-0.001	0.013	-0.102	0.919	0.025	0.012	2.076	0.038
skilled, no degree	0.057	0.023	2.495	0.013	0.066	0.021	3.095	0.002
unemployment spells last 2 years	-0.169	0.004	-37.707	0.000	0.004	0.004	1.126	0.260
fraction employed last yr	0.074	0.018	4.138	0.000	-0.024	0.017	-1.429	0.153
low employability	0.089	0.019	4.730	0.000	0.067	0.017	3.912	0.000
medium employability	0.102	0.015	6.926	0.000	0.039	0.014	2.848	0.004
female caseworker	0.066	0.010	6.349	0.000	-0.025	0.010	-2.563	0.010
caseworker above vocational	0.033	0.010	3.191	0.001	-0.149	0.010	-15.344	0.000
caseworker higher education	0.040	0.012	3.244	0.001	-0.086	0.011	-7.575	0.000
cantonal unemployment rate	0.018	0.006	3.333	0.001	0.049	0.005	9.469	0.000
French speaking region	-0.120	0.018	-6.764	0.000	-0.166	0.016	-10.288	0.000
French × female jobseeker	0.080	0.021	3.825	0.000	-0.064	0.019	-3.321	0.001
French × mother tongue	-0.134	0.023	-5.891	0.000	0.097	0.021	4.711	0.000
French × female caseworker	-0.066	0.022	-3.033	0.002	0.034	0.020	1.722	0.085

Note: 'coef', 'se', 'z-val', and 'p-val' give the coefficient estimates, standard errors, z-values, and p-values, respectively.

Table 2 presents probit regressions of the treatments on the original covariates as well as interaction terms between French region and jobseeker gender, mother tongue, and caseworker

gender to verify the magnitude of selection with respect to the various variables in our data. Note that the very same set of regressors is used in the parametric (namely probit and EL) and semi-parametric propensity score specifications in the simulations, while the interaction terms need not be included in the nonparametric propensity score specifications. We see that most regressors are highly significant in predicting the treatments, in particular for 'program participation'. Yet, many coefficients are actually rather small, so that treatment selectivity in the data is actually weaker than a mere look on the p-values would suggest. However, our particular EMCS design entails a considerably stronger selection in the actual simulation samples, see the pseudo-R^2 reported for either treatment in Section 3.3.1.

3.3 Simulation design

3.3.1 Generation of the population and selection into treatment

To generate the 'population' out of which the simulation samples are drawn, we make use of a pair matching algorithm to the data introduced in Section 3.2. To each treated observation in our data (either defined upon 'program participation' or 'noncooperative caseworker'), the nearest non-treated observation in terms of the Mahalanobis distance with respect to the covariates in Table 1 is matched without replacement. The treated observations are then discarded and do not play any further role in the simulation design, while the matched non-treated subjects become the pseudo-treated 'population' used for the simulations, while the remaining (i.e. unmatched) non-treated observations constitute the non-treated 'population'. As neither the (unmatched) non-treated, nor the (pseudo-)treated have actually received any treatment, any treatment effect is zero by definition and thus homogeneous in the 'population'. However, as the pseudo-treated mimic the covariate distributions of the discarded treated, they differ from the remaining non-treated in terms of X. Using pair matching without replacement therefore nonparametrically creates selection into the (pseudo-)treatment, and is therefore agnostic about the functional form of the 'true' propensity score model. This stands in contrast to the parametric specification of the 'true' propensity score model in the EMCS of Huber, Lechner, and Wunsch (2013), which we

avoid to prevent favoring a particular (parametric) propensity score estimator in the simulations, which would jeopardize a fair assessment of the nonparametric propensity score methods.

We apply this methodology to the treatments 'program participation' and 'noncooperative caseworker' in order to produce two different scenarios with respect to selection into the pseudo-treatment. For 'program participation', the initially 26,062 treated and 67,014 non-treated observations entail a 'population' of 26,062 (pseudo-)treated and 40,952 (=67,014-26,062) unmatched non-treated observations. Concerning the treatment 'noncooperative caseworker', the initially 43,669 treated and 49,407 non-treated observations entail a 'population' of 43,669 (pseudo-)treated and 5,738 unmatched non-treated observations. Table A.1 in Appendix A.2 reports probit regressions of either treatment variable on the same covariates and interaction terms as in Table 2, however, after creating the respective 'population', which considerably increases treatment selectivity. Still, for both 'program participation' and 'noncooperative caseworker', the overlap in the distributions of the (probit-based) propensity score estimates across the treated and non-treated 'populations' is quite satisfactory, as shown in Figure A.2 in Appendix A.1.

In each simulation replication, half of the observations are drawn with replacement from the pseudo-treated 'population' and half from the non-treated 'population', so that the treatment share is 50% in any simulation. Running a probit regression of the pseudo-treatment 'program participation' on the covariates in all simulations yields an average maximum likelihood pseudo R^2 of roughly 0.08, which points to moderate selection. For 'noncooperative caseworker' the pseudo R^2 is roughly 0.24, indicating a more pronounced selection into the pseudo-treatment.

3.3.2 Effect heterogeneity

As already mentioned, the treatment effects under the design of Section 3.3.1 are zero and thus homogeneous. To also generate a scenario with heterogeneous effects, we necessarily require a model for the outcome as a function of the covariates and the treatment. The advantage to allow for heterogeneous effects therefore comes with the disadvantage to put more (non-empirical)

structure on the simulations. To mitigate the latter concern, we base the outcome model on the statistical associations found in our initial data, i.e. prior to the generation of the 'population'. More specifically, in our initial data we run an OLS regression of the outcomes under either treatment state on all variables and interaction terms entering the probit model in Table 2 (as also used for the parametric and semiparametric propensity score models), as well as on second and third order terms of number of unemployment spells in last two years, the second order terms of share of time in employment in last year and cantonal unemployment rate, and interaction terms between French speaking regions and levels of qualification and the unemplyoment rate and its second order term (which come in addition to the interactions with jobseeker gender, mother tongue, and caseworker gender already used in the propensity score model).

Table A.2 in Appendix A.2 reports the 'true' OLS coefficients for the outcome models conditional on $D = 1$ or $D = 0$ under the treatments 'program participation' and 'noncooperative caseworker'. Note that modelling the outcomes separately by treatment state allows for arbitrary interactions between the respective treatment and each of the covariates in the outcome model. In each simulation, the coefficients under $D = 1$ and $D = 0$ as well as the simulation draw-specific values of the covariates and the corresponding higher order/interaction terms are then used to predict the potential outcomes under treatment and non-treatment. To these predictions, normally distributed error terms with variances that correspond to the estimated error variances in the outcome models under treatment and non-treatment are added to create the actual potential outcomes with and without treatment. Finally, any negative potential outcome values in each simulation draw are set to zero, in order to respect the non-negativity of the original outcome variable cumulative months in job search).

3.3.3 Misspecification, sample sizes, and number of simulations

As a further simulation parameter, we consider including all confounders entering the treatment selection process vs. misspecification due to omitting some covariates in the various estimators. In the latter case, the regional variables 'cantonal unemployment rate' and 'French speaking region'

are omitted, as well as any interactions with 'French speaking region' whenever applicable (e.g. in the parametric propensity score specifications). This allows investigating how robust the various estimators are to not conditioning on all confounders, a case that is likely to be very relevant in empirical applications.

As a final simulation feature, we consider two different sample sizes (N): 750 and 3000 observations, so that either 375 or 1500 each treated and non-treated subjects are drawn with replacement from the 'population' in the simulations. This allows investigating how the relative performances of semi- and the fully nonparametric methods change as the number of observations increases from a fairly moderate sample of less than 1000 subjects to several 1000 observations.

All in all, the combination of two treatments (with distinct selection into treatment), effect homogeneity vs. heterogeneity, controlling for all confounders vs. misspecification due to omitted variable bias, and two different sample sizes yields 16 different simulation designs. Concerning the number of simulation draws in our EMCS, the latter should ideally be as large as possible to minimize simulation noise, which negatively depends on the number of draws and positively on the variance of the estimators. Note that the estimators' asymptotic variance approximately doubles when the sample size is reduced by half (as all methods considered are possibly $\sqrt{n}$-consistent) and simulation noise is doubled when the number of replications is reduced by half (at least for averages over the i.i.d. simulations). We therefore follow Huber, Lechner, and Wunsch (2013) and make the number of simulation draws inversely proportional to the sample size, using 4000 simulations for $N = 750$ and 1000 for $N = 3000$, as the larger sample size is computationally more expensive and has less variability of the results across different simulation samples than the smaller one.

3.3.4 Trimming

Several simulation studies point to the importance of trimming observations with too large propensity scores or weights in treatment effect estimation, see for instance Huber, Lechner, and Wunsch (2013), Lechner and Strittmatter (2014), and Pohlmeier, Seiberlich, and Uysal (2013),

while the conclusions in Busso, DiNardo, and McCrary (2009) are more ambiguous. Besides estimation without trimming, we consider seven different trimming rules in our simulations. Firstly, we discard all treated observations with propensity scores greater than the maximum propensity score among the non-treated to enforce common support, as suggested by Dehejia and Wahba (1999). Secondly, we drop subjects with propensity scores higher than either 0.99 or 0.95, as frequently applied in estimation based on IPW. Thirdly, as suggested in Imbens (2004) and discussed in more detail in Huber, Lechner, and Wunsch (2013), we discard any non-treated observation whose IPW-based relative weight (as a proportion of the total of non-treated weights) surpasses a particular threshold, in our case 1% (of all non-treated weights). Finally, we combine the common support procedure of Dehejia and Wahba (1999) with the second and third trimming approaches, respectively: the Dehejia and Wahba (1999) common support restriction (CS), removal of propensity scores larger than 0.99 (ps0.99) or 0.95 (ps0.95), removal of relative weights larger than 0.01 (w0.01), (CS) combined with (ps0.99), (CS) combined with (ps0.95), and (CS) combined with (w0.01). In the discussion of the results under trimming, see Section 4.4, we only focus on CS, which turned out to be the overall best performing trimming rule in terms of average mean squared error reduction. For the other trimming procedures, the outcomes are available from the authors upon request.

4 Results

4.1 Overview

In this section, we first discuss the overall performance of the different methods across all DGPs without trimming. Secondly, we analyze the results separately for the two treatment definitions, sample sizes, effect homogeneity vs. heterogeneity, and correct specification vs. omitted variables. Thirdly, we reassess overall performance of the estimators when applying the propensity score trimming rule that performs best on average across all simulations, namely the common support restriction of Dehejia and Wahba (1999). Fourthly, we investigate how the various approaches

to propensity score estimation affect the performance of certain propensity score-based ATET estimators (IPW, DR, and matching). Finally, we compare the performance of propensity score matching (using a parametric propensity score model) to various versions of (direct) covariate matching. We derive our conclusions from analyzing the mean squared error (MSE) of the estimators. Tables A.4 and A.5 in Appendix A.2 contain further results concerning the squared bias and the variance of the estimators.

4.2 Overall results without trimming

This section presents the overall results of the various estimators *without trimming* averaged over all 16 features of the DGPs. Table 3 reports the average MSE of the estimators sorted from the lowest to the highest. It also provides the relative MSE difference in percent of each method to the lowest average MSE (i.e. of the best performing estimator), as a measure of relative performance in terms of average MSE. Finally, the estimators' average rank in terms of MSE across all 16 settings is also reported, a performance measure that is more robust to outliers in MSE in particular DGPs. Note that the results are only shown for 51 out of the initially 63 estimators investigated. As mentioned in Section 2.3.4, twelve versions of radius matching on both the propensity score and additional confounders are omitted from the discussion because they are always outperformed by the respective radius matching estimators on the propensity score alone.

It is remarkable that in terms of average MSE, the top five performing estimators are all nonparametric methods. The overall winner is the nonparametric outcome regression estimator (or kernel matching directly on the covariates, see Section 2.4.2) using the crossval bandwidth. In second place comes nonparametric DR estimation as suggested by Rothe and Firpo (2013), using crossval bandwidths for nonparametric propensity score and conditional mean outcome estimation. The average MSE of the latter is 8.5% larger than that of the winner, but has an even better average rank (5.6 vs. 8.5) across all simulations. The third best method is again nonparametric regression, this time using undersmoothing, followed by IPW based on the nonparametric

Table 3: Average MSE over all settings without trimming

	MSE	relative difference	rank
nonpara regression (crossval bandwidth)	0.25	0.0	8.5
doubly robust (nonpara; crossval bandwidth)	0.27	8.5	5.6
nonpara regression (undersmoothing)	0.29	19.5	10.4
IPW (nonpara pscore)	0.30	22.2	10.2
doubly robust (nonpara; undersmoothed outcomes)	0.30	23.9	10.8
IPW (overidentified CBPS)	0.31	26.3	5.2
direct 1:5 match	0.33	34.0	12.8
direct 1:5 match with bias correction	0.33	34.0	12.6
IPW (para pscore)	0.34	38.5	8.2
IPW (semipara pscore)	0.34	38.9	14.0
para regression	0.34	39.0	9.4
inverse probability tilting (IPT)	0.34	39.1	11.2
doubly robust (para pscore and outcome)	0.36	45.2	8.7
doubly robust (semipara pscore, para outcome)	0.37	50.0	12.0
kernel match (para pscore; oversmooth)	0.38	56.1	13.6
kernel match (semipara pscore; oversmooth)	0.40	61.6	17.8
IPW (just identified CBPS)	0.41	67.9	12.3
entropy balancing (EB)	0.41	68.0	13.2
doubly robust (just identified CBPS, para outcome)	0.41	68.0	12.7
kernel match (just identified CBPS; oversmooth)	0.42	71.8	16.1
direct pair match with bias correction	0.49	99.9	31.8
direct pair match	0.49	99.9	32.0
genetic match	0.53	117.2	35.8
genetic match with bias correction	0.53	117.2	35.7
radius match (semipara pscore; large radius)	0.55	125.9	28.9
radius match (semipara pscore; medium radius)	0.58	136.0	31.8
pair match (semipara pscore)	0.59	138.9	31.8
radius match (semipara pscore; small radius)	0.61	148.0	34.6
radius match (para pscore; large radius)	0.63	154.8	30.4
kernel match (just identified CBPS; crossval bandw)	0.63	155.2	26.1
kernel match (semipara pscore; crossval bandw)	0.67	173.5	31.4
radius match (para pscore; medium radius)	0.68	175.3	32.9
radius match (para pscore; small radius)	0.72	193.3	36.5
pair match (para pscore)	0.72	193.3	36.8
kernel match (para pscore; crossval bandw)	0.80	224.4	22.9
radius match (just identified CBPS; large radius)	0.94	284.5	36.2
kernel match (para pscore; undersmooth)	1.08	341.5	26.9
radius match (just identified CBPS; medium radius)	1.22	398.3	39.1
pair match (just identified CBPS)	1.31	432.1	42.1
radius match (just identified CBPS; small radius)	1.38	463.6	42.4
direct radius match	1.55	529.8	52.9
direct radius match with bias correction	1.55	529.8	52.8
kernel match (just identified CBPS; undersmooth)	1.89	669.6	39.3
radius match (nonpara pscore; large radius)	3.30	1244.5	52.9
pair match (nonpara pscore)	3.32	1254.8	46.9
radius match (nonpara pscore; medium radius)	4.61	1781.0	54.9
radius match (nonpara pscore; small radius)	6.15	2408.9	56.8
kernel match (nonpara pscore; oversmooth)	9.56	3798.7	43.7
kernel match (semipara pscore; undersmooth)	13.18	5271.2	51.0
kernel match (nonpara pscore; crossval bandw)	>100	>10000	51.1
kernel match (nonpara pscore; undersmooth)	>100	>10000	60.4

Note: 'MSE' gives the average MSE, 'relative difference' is in percent and provides the relative MSE difference to the lowest average MSE (of the best performing estimator), 'rank' gives the average rank of the estimators in terms of MSE across all simulations. All radius matching estimators on the propensity score ('radius m.') include bias correction.

propensity score and nonparametric DR with undersmoothed estimation of the conditional mean outcome. IPW using the overidentified CBPS method of Imai and Ratkovic (2014) comes in sixth place in terms of average MSE (and is therefore the strongest not fully nonparametric method), but is actually the best performing estimator with respect to the average rank (5.2). After that we have yet another nonparametric method, namely one-to-many matching on the covariates (in our case one-to-five matching using the Mahalanobis metric) without and with bias correction. It is followed by several methods whose performance is almost identical, namely IPW using a parametric and semiparametric propensity score, parametric regression among nontreated observations as outlined in (2.4.2), and IPT of Graham, Pinto, and Egel (2011).

To the best of our knowledge, none of the top five estimators have been investigated in previous simulation studies, which predominantly focussed on (subsets of) parametric or semiparametric estimators (with parametric propensity scores). Busso, DiNardo, and McCrary (2009), for instance, find IPW to be competitive in DGPs where no common support issues arise, but do not consider fully nonparametric IPW or nonparametric regression. Lunceford and Davidian (2004) conclude that DR performs well in a very broad class of DGPs, but at the time of their simulation the Rothe and Firpo (2013) DR estimator had not even been suggested yet. It is also noteworthy that in contrast to the simulation studies of Huber, Lechner, and Wunsch (2013) and Frölich (2004), no propensity score matching method is among the best performing methods, no matter whether parametric or semi-/nonparametric propensity scores are used. Furthermore, using the nonparametric propensity score entails a substantially larger MSE than the (semi-)parametric scores in the case of matching, in particular due to an explosion in the variance (see Table A.5 in Appendix A.2) and quite contrary to IPW and DR. For instance, kernel matching on the nonparametric propensity score is generally the worst estimator in the simulations, while oversmoothed kernel matching on parametric and semiparametric propensity scores performs best among all propensity score matching algorithms (yet, it is nowhere near the top).

A general pattern among kernel and radius matching on the propensity score that was also

found in Huber, Lechner, and Wunsch (2013) and Huber, Lechner, and Steinmayr (2014) is that a larger bandwidth (oversmoothing) or a larger radius reduces the MSE of the respective estimators. This is somewhat in contrast to the theoretical finding that one should rather undersmooth, compared to conventional cross-validation bandwidth choice, see e.g. Heckman, Ichimura, and Todd (1998b). One needs to keep in mind, though, that the 'undersmoothing' recommended by econometric theory refers to the convergence *rate* of the bandwidth and not to the bandwidth value for a given sample size. Hence, for a particular sample size 'oversmoothing' may be appropriate, but the degree of 'oversmoothing' should decrease for cross-validation bandwidth choice when the sample size increases.

Within the class of all matching estimators, a further interesting result is that the best covariate matching algorithms outperform the best propensity score matching methods, see also Section 4.6 for more detailed results. Specifically, nonparametric outcome regression (which may be regarded as kernel matching on the covariates) and direct 1:5 matching outperform (oversmoothed) propensity score kernel matching. While covariate matching was not considered at all in Huber, Lechner, and Wunsch (2013) and Frölich (2004), our findings are in line with the simulation results of Zhao (2004) (who, however, considers fewer covariates than our setup). There, covariate matching based on the Mahalanobis distance dominates propensity score matching in a range of different (artificial) DGPs.

4.3 Simulation results by treatments and other DGP features

Tables 4 and 5 present the results separately for the treatments 'program participation' and 'noncooperative caseworker', respectively. For the ease of exposition, only the top twelve estimators are included in the tables.[11] When considering the treatment 'program participation', the nonparametric methods are less dominating than in the overall results. Here, IPW with overidentified CBPS and with probit-based propensity scores come in first and second place, respectively. They are very closely followed by nonparametric DR with the crossval bandwidth,

[11] The complete results are available from the authors upon request.

DR using parametric models for the propensity score and the outcome, IPW and DR based on the just identified CBPS, and entropy balancing of Hainmueller (2012), which performs considerably better than in the overall results. However, it is worth noting that differences in the MSEs of the top 23 methods are moderate (less than 15% in terms of relative MSE), so that we conclude that a wide range of semi- and nonparametric treatment effect estimators is similarly competitive under the treatment 'program participation'.

Table 4: Average MSE for treatment 'program participation' without trimming

	MSE	relative difference	rank
IPW (overidentified CBPS)	0.19	0.0	4.5
IPW (para pscore)	0.19	0.2	4.8
doubly robust (nonpara; crossval bandwidth)	0.19	0.3	7.6
doubly robust (para pscore and outcome)	0.19	0.6	6.4
IPW (just identified CBPS)	0.19	0.8	7.1
doubly robust (just identified CBPS, para outcome)	0.19	0.8	7.5
entropy balancing (EB)	0.19	0.8	8.2
para regression	0.20	1.3	9.5
inverse probability tilting (IPT)	0.20	2.2	11.4
kernel match (para pscore; oversmooth)	0.20	2.5	11.8
doubly robust (semipara pscore, para outcome)	0.20	2.8	10.8
kernel match (just identified CBPS; oversmooth)	0.20	4.4	13.6

Note: 'MSE' gives the average MSE, 'relative difference' is in percent and provides the relative MSE difference to the lowest average MSE (of the best performing estimator), 'rank' gives the average rank of the estimators in terms of MSE across all simulations. All radius matching estimators on the propensity score ('radius m.') include bias correction.

For the treatment 'noncooperative caseworker', however, differences in MSEs are more pronounced. The nonparametric methods dominate similarly as in Section 4.2, and also the ranking among the top five is the same: nonparametric regression based on the crossval bandwidth is the overall winner, followed by nonparametric DR estimation using crossval bandwidths, undersmoothed nonparametric regression, nonparametric IPW, and DR with undersmoothed estimation of conditional mean outcome. While the relative performance of estimators is not exactly identical across the two treatments, some estimators perform similarly well in either case, e.g., nonparametric DR with crossval bandwidths.

In a next step, we look at further features of the DGPs, namely sample size, effect homogeneity vs. heterogeneity (across both treatments), and absence or prevalence of omitted variable bias. Table 6 reports the top five estimators for each of the scenarios considered. It is interesting

Table 5: Average MSE for treatment 'non-cooperative caseworker' without trimming

	MSE	relative difference	rank
nonpara regression (crossval bandwidth)	0.29	0.0	2.2
doubly robust (nonpara; crossval bandwidth)	0.34	17.5	3.5
nonpara regression (undersmoothing)	0.37	29.1	5.4
IPW (nonpara pscore)	0.39	34.1	5.5
doubly robust (nonpara; undersmoothed outcomes)	0.40	37.1	6.4
IPW (overidentified CBPS)	0.43	48.0	6.0
direct 1:5 match	0.45	56.4	10.4
direct 1:5 match with bias correction	0.45	56.4	10.2
IPW (semipara pscore)	0.47	64.3	10.4
inverse probability tilting (IPT)	0.49	68.4	11.0
IPW (para pscore)	0.49	68.7	11.6
para regression	0.49	68.9	9.4

Note: 'MSE' gives the average MSE, 'relative difference' is in percent and provides the relative MSE difference to the lowest average MSE (of the best performing estimator), 'rank' gives the average rank of the estimators in terms of MSE across all simulations. All radius matching estimators on the propensity score ('radius m.') include bias correction.

to see that nonparametric methods dominate both under smaller and larger sample size. This result suggests that nonparametric methods might already work well in moderate samples, given that the number of (continuous) confounders is not too large. Secondly, those five nonparametric estimators that dominate the overall results (see Section 4.2) appear in the very same order under both effect homogeneity and heterogeneity. A notable change occurs, however, when looking at settings with *correct specifications*, in the sense that no confounders are omitted from estimation. In this case, a semiparametric method performs best, namely IPW based on the overidentified CBPS of Imai and Ratkovic (2014). Also IPW using the parametric propensity score makes it to the top five, while the remaining three methods are again fully nonparametric. Even though nonparametric regression performs worse than semiparametric CBPS, its MSE is only 8.9% larger than that of the best estimator.

Under *misspecification* due to omitted regional variables, it is again the same nonparametric estimators with the overall smallest average MSEs that are in lead. The nonparametric methods therefore appear to be less vulnerable to omitted variable bias than the semiparametric estimators, while at the same time being very competitive under correct specification.

Table 6: Average MSE for various DGP features without trimming

$N = 750$	MSE	relative difference	rank
nonpara regression (crossval bandwidth)	0.35	0.0	6.6
doubly robust (nonpara; crossval bandwidth)	0.38	7.0	5.6
nonpara regression (undersmoothing)	0.42	17.9	11.4
direct 1:5 match with bias correction	0.43	21.1	9.6
direct 1:5 match	0.43	21.1	10.0
$N = 3000$	MSE	relative difference	rank
nonpara regression (crossval bandwidth)	0.14	0.0	10.4
IPW (nonpara pscore)	0.15	8.3	7.8
doubly robust (nonpara; crossval bandwidth)	0.15	12.2	5.5
nonpara regression (undersmoothing)	0.17	23.5	9.5
doubly robust (nonpara; undersmoothed outcomes)	0.18	29.5	9.0
effect homogeneity	MSE	relative difference	rank
nonpara regression (crossval bandwidth)	0.28	0.0	7.0
doubly robust (nonpara; crossval bandwidth)	0.31	9.9	4.8
nonpara regression (undersmoothing)	0.35	22.2	9.1
IPW (nonpara pscore)	0.36	25.7	11.2
doubly robust (nonpara; undersmoothed outcomes)	0.36	26.5	10.2
effect heterogeneity	MSE	relative difference	rank
nonpara regression (crossval bandwidth)	0.21	0.0	10.0
doubly robust (nonpara; crossval bandwidth)	0.22	6.6	6.4
nonpara regression (undersmoothing)	0.24	15.8	11.8
IPW (nonpara pscore)	0.24	17.3	9.2
doubly robust (nonpara; undersmoothed outcomes)	0.25	20.4	11.4
correct specification	MSE	relative difference	rank
IPW (overidentified CBPS)	0.21	0.0	2.6
doubly robust (nonpara; crossval bandwidth)	0.23	6.1	7.9
nonpara regression (crossval bandwidth)	0.23	8.9	14.2
IPW (para pscore)	0.24	11.3	4.6
direct 1:5 match with bias correction	0.24	12.6	12.8
misspecification	MSE	relative difference	rank
nonpara regression (crossval bandwidth)	0.26	0.0	2.8
doubly robust (nonpara; crossval bandwidth)	0.31	18.5	3.2
nonpara regression (undersmoothing)	0.33	27.8	5.5
IPW (nonpara pscore)	0.34	32.9	6.6
doubly robust (nonpara; undersmoothed outcomes)	0.35	36.6	7.9

Note: 'MSE' gives the average MSE, 'relative difference' is in percent and provides the relative MSE difference to the lowest average MSE (of the best performing estimator), 'rank' gives the average rank of the estimators in terms of MSE across all simulations.

4.4 Results with trimming

This section presents the overall results for the overall best performing trimming rule among the seven investigated, which is the common support restriction of Dehejia and Wahba (1999). The latter reduces the average MSE across all DGPs by 29.88% (or by 7.07% for the treatment 'program particiaption' and by 31.43% for the treatment 'noncooperative caseworker'). However, it has to be emphasized that the improvement is largely driven by those estimators which had performed very poorly without trimming, in particular matching on the nonparametric propensity score. The best performing estimators, on the other hand, are hardly affected and frequently even

do slightly worse when imposing common support than without trimming. Note that the common support restriction was applied to each of the propensity score methods considered. Therefore, the number of observations discarded after trimming may differ depending on which propensity score method is used. In particular, one might suspect that a lack of support (and thus, trimming) more likely occurs for nonparametric propensity scores than under probit estimation (which has a lower variance). Table A.3 in Appendix A.2 presents the average number of propensity scores dropped due to the common support restriction and confirms this expectation. On average, 161.44 nonparametric, but only 14.89 parametric propensity scores are trimmed across all settings, with all remaining methods (semiparametric propensity score estimation, just/overidentified CBPS) lying in between.

Table 7: Average MSE over all settings with trimming (common support rule)

	MSE	relative difference	rank
nonpara regression (crossval bandwidth)	0.25	0.0	9.0
doubly robust (nonpara; crossval bandwidth)	0.28	14.1	6.9
nonpara regression (undersmoothing)	0.29	17.2	10.6
doubly robust (nonpara; undersmoothed outcomes)	0.31	24.0	11.2
direct 1:5 match	0.32	28.6	11.8
direct 1:5 match with bias correction	0.32	28.6	11.6
IPW (overidentified CBPS)	0.32	29.6	7.2
IPW (nonpara pscore)	0.32	29.9	13.3
para regression	0.32	30.3	8.4
inverse probability tilting (IPT)	0.33	34.1	11.1
doubly robust (para pscore and outcome)	0.34	37.2	8.9
kernel match (para pscore; oversmooth)	0.35	40.1	13.3

Note: 'MSE' gives the average MSE, 'relative difference' is in percent and provides the relative MSE difference to the lowest average MSE (of the best performing estimator), 'rank' gives the average rank of the estimators in terms of MSE across all simulations.

Table 7 reports the results for the twelve best estimators across all DGPs under trimming. As in Section 4.2, the three best performing estimators are nonparametric regression (or kernel matching on the covariates) using the crossval bandwidth, nonparametric DR using crossval bandwidths for both nonparametric propensity score and outcome estimation, and nonparametric regression using undersmoothing. They are followed by three further nonparametric methods (DR with undersmoothing in conditional outcome estimation and one-to-five covariate matching without and with bias adjustment) and IPW using the overidentified CBPS of Imai and Ratkovic (2014). Only slightly behind are nonparametric IPW

and parametric regression. As in Section 4.2, nonparametric propensity score matching methods are at the bottom of the ranking (not reported), even though trimming considerably reduces their relative differences to the better performing methods (and therefore importantly drive the average MSE improvement of trimming).

4.5 Impact of propensity score methods on estimator performance

In this section, we investigate how different types of propensity score estimators affect the performance of different types of (propensity score-based) treatment effects estimators. In Table 8, the methods are ordered according to their average MSE over all settings without trimming when using a parametric propensity score. The first column in the table gives the respective MSEs. The remaining columns give the MSEs under alternative propensity score methods along with the relative increase or decrease of the MSE (in percent) when compared to parametric propensity score estimation. The just identified CBPS method of Imai and Ratkovic (2014) increases MSE over probit estimation for almost all estimators. The same conclusion holds when separately looking at the treatments 'program participation' or 'noncooperative caseworker', see Tables 9 and 10, with the relative MSE increase generally being much larger in the latter case. As a word of caution, however, the same (negative) result need not necessarily hold for overidentified CBPS, which in our simulations was only considered for IPW. In fact, across all settings, IPW based on the overidentified CBPS (MSE: 0.31) outperforms both just identified CBPS (MSE: 0.41) and probit estimation (MSE: 0.34). For program participation, IPW with overidentified CBPS is even best performing estimator among all propensity score based methods (MSE: 0.19). In any scenario (overall and either treatment), overidentified CBPS performs better than the competing EL method IPT, see Tables 3, 4, and 5.

Semiparametric propensity score estimation based on Klein and Spady (1993) reduces the MSE of almost all matching estimators compared to probit estimation, with the exception of kernel matching on propensity score with over- or undersmoothing. In particular under smaller bandwidth, the use of the more flexible propensity score method is detrimental to kernel matching.

Table 8: Average MSE over all settings without trimming for various propensity score methods

	MSE para pscore	MSE CBPS	relative change	MSE semipara pscore	relative change	MSE nonpara pscore	relative change
IPW	0.34	0.41	21.2	0.34	0.3	0.30	-11.8
doubly robust	0.36	0.41	15.7	0.37	3.3	0.27*	-25.3*
kernel match (oversmooth)	0.38	0.42	10.1	0.40	3.5	9.56	2398.1
radius match (large radius)	0.63	0.94	50.9	0.55	-11.4	3.30	427.6
radius match (medium radius)	0.68	1.22	81.0	0.58	-14.3	4.61	583.1
radius match (small radius)	0.72	1.38	92.2	0.61	-15.5	6.15	755.4
pair match	0.72	1.31	81.4	0.59	-18.6	3.32	361.8
kernel match (crossval bandw)	0.80	0.63	-21.3	0.67	-15.7	182.14	22784.7
kernel match (undersmooth)	1.08	1.89	74.3	13.18	1116.7	> 1000	> 100000

Note: The MSE of IPW with overidentified CBPS is 0.31. *: doubly robust (DR) estimation with nonparametric estimation of both the propensity score and the outcome model as suggested in Rothe and Firpo (2013) using crossval bandwidths. (DR results without '*' are based on a parametric outcome model.) The MSE of nonparametric DR estimation of Rothe and Firpo (2013) based on undersmoothed estimation of the conditional mean outcome is 0.30.

The average MSE of DR estimation with a parametric outcome model is slightly increased when using the Klein and Spady (1993) estimator, too, while IPW is barely affected. Also when looking at the treatment 'program participation', semiparametric propensity score estimation entails a larger MSE in most cases, but the changes are generally rather minor with the exception of kernel matching. Under the treatment 'non-cooperative caseworker', the results are qualitatively more in line with the overall effects, as the MSEs of most matching estimators are reduced, that of IPW decreases slightly, and that of DR (with a parametric outcome model) increases moderately.

Finally, nonparametric propensity score estimation considerably increases the average MSEs of matching methods (as already discussed in Section 4.2). Yet, it reduces the MSE of IPW. Also DR estimation using nonparametric propensity score and outcome models (see Rothe and Firpo (2013)) and crossval bandwidths (MSE: 0.27) considerably improves upon DR based on parametric models (MSE: 0.36). So does the Rothe and Firpo (2013) estimator using undersmoothing for outcome estimation (MSE: 0.30). For the treatment 'program participation' alone, propensity score estimation very slightly reduces the MSE of DR with crossval bandwidths, but increases that of IPW moderately and that of matching estimators profoundly in most cases. Under

Table 9: Average MSE for treatment 'program participation' without trimming for various propensity score methods

	MSE para pscore	MSE CBPS	relative change	MSE semipara pscore	relative change	MSE nonpara pscore	relative change
IPW	0.19	0.19	0.5	0.21	7.4	0.21	10.0
doubly robust	0.19	0.19	0.1	0.20	2.1	0.19*	-0.3*
kernel match (oversmooth)	0.20	0.20	1.8	0.21	8.4	0.32	63.3
kernel match (crossval bandw)	0.21	0.22	3.4	0.40	88.4	0.45	111.8
kernel match (undersmooth)	0.22	0.32	47.6	1.49	593.6	> 1000	> 100000
radius match (large radius)	0.27	0.27	1.7	0.28	2.5	0.35	31.2
radius match (medium radius)	0.28	0.29	1.6	0.28	1.2	0.36	28.7
radius match (small radius)	0.29	0.30	1.3	0.30	0.6	0.37	25.8
pair match	0.30	0.30	1.0	0.29	-2.7	0.32	5.0

Note: The MSE of IPW with overidentified CBPS is 0.19. *: doubly robust (DR) estimation with nonparametric estimation of both the propensity score and the outcome model as suggested in Rothe and Firpo (2013) using crossval bandwidths. (DR results without '*' are based on a parametric outcome model.) The MSE of nonparametric DR estimation of Rothe and Firpo (2013) based on undersmoothed estimation of the conditional mean outcome is 0.21.

the 'noncooperative caseworker' treatment, the deterioration of matching is even much more pronounced, while both IPW and DR improve considerably under nonparametric propensity score estimation. Overall, the latter is the optimal propensity score method for IPW and DR, while semiparametric estimation is best for radius and pair matching. For kernel matching with the crossval bandwidth the CBPS method is preferred, while the parametric propensity score entails the smallest MSE for kernel matching using under- or oversmoothing. The starkly contrary impact of nonparametric propensity score estimation on matching estimators and IPW/DR is as interesting as puzzling and may deserve further attention in future research.

4.6 Comparison propensity score vs. covariate matching

We subsequently more thoroughly compare matching on the parametric propensity score, as it is standard in the treatment evaluation literature, to (less frequently used) covariate matching. Table 11 reports the average MSEs of parametric propensity score matching across all settings without trimming in the first column of the table, while the first row contains

Table 10: Average MSE for treatment 'noncooperative caseworker' without trimming for various propensity score methods

	MSE para pscore	MSE CBPS	relative change	MSE semipara pscore	relative change	MSE nonpara pscore	relative change
IPW	0.49	0.63	29.5	0.47	-2.6	0.39	-20.5
doubly robust	0.52	0.63	21.5	0.54	3.8	0.34*	-34.6*
kernel match (oversmooth)	0.57	0.64	12.9	0.58	1.9	18.81	3209.9
radius match (large radius)	0.98	1.61	64.3	0.83	-15.1	6.24	536.0
radius match (medium radius)	1.07	2.16	101.8	0.87	-18.3	8.87	728.7
pair match	1.14	2.31	102.6	0.88	-22.8	6.33	456.0
radius match (small radius)	1.14	2.47	115.5	0.92	-19.6	11.94	942.8
kernel match (crossval bandw)	1.38	1.03	-25.1	0.95	-31.6	363.83	26238.2
kernel match (undersmooth)	1.95	3.46	77.3	24.86	1174.4	> 1000	> 100000

Note: The MSE of IPW with overidentified CBPS is 0.43. *: doubly robust (DR) estimation with nonparametric estimation of both the propensity score and the outcome model as suggested in Rothe and Firpo (2013) using crossval bandwidths. (DR results without '*' are based on a parametric outcome model.) The MSE of nonparametric DR estimation of Rothe and Firpo (2013) based on undersmoothed estimation of the conditional mean outcome is 0.40.

the figures for various covariate matching approaches. The values in the matrix in between the first column and first row give the relative increase or decrease in MSE in percent when switching from the respective propensity score matching to the respective covariate matching method. As already mentioned in Section 4.2, the general picture is that covariate matching algorithms outperform most propensity score approaches, which is in line with the findings in Zhao (2004). An exception is the very poorly performing direct radius matching estimator, for which, however, only one (rather small) radius was considered in the simulations (while results for kernel and radius matching on the propensity score showed that larger bandwidths/radii are to be preferred). All other direct matching approaches outperform any of the propensity score methods, apart from oversmoothed kernel matching on the propensity score. The latter is better than direct pair matching and genetic matching, but still outperformed by one-to-many matching and nonparametric regression (or direct kernel matching), which is the overall winner.

Table 11: Propensity score vs. direct covariate matching in all settings without trimming

	MSE	direct pair match	dir. p. match with bias correction	dir. 1:5 match	dir. 1:5 match with bias cor.	dir. radius match	dir. radius match with bias cor.	nonpara reg. (crossval bandw)	nonp. reg. (undersmooth)	genetic match	genetic match with bias cor.
MSE		0.49	0.49	0.33	0.33	1.55	1.55	0.25	0.29	0.53	0.53
pscore kernelmatch (oversmooth)	0.38	28.10	28.10	-14.11	-14.11	303.54	303.54	-35.93	-23.45	39.19	39.19
pscore radiusmatch (large radius)	0.63	-21.55	-21.55	-47.40	-47.40	147.14	147.14	-60.76	-53.12	-14.76	-14.76
pscore radiusmatch (medium radius)	0.68	-27.39	-27.39	-51.32	-51.32	128.74	128.74	-63.68	-56.61	-21.11	-21.11
pscore radiusmatch (small radius)	0.72	-31.83	-31.83	-54.30	-54.30	114.73	114.73	-65.91	-59.27	-25.94	-25.94
pscore pairmatch	0.72	-31.84	-31.84	-54.31	-54.31	114.70	114.70	-65.91	-59.27	-25.95	-25.95
pscore kernelmatch (crossval bandw)	0.80	-38.37	-38.37	-58.68	-58.68	94.13	94.13	-69.18	-63.18	-33.04	-33.04
pscore kernelmatch (undersmooth)	1.08	-54.71	-54.71	-69.64	-69.64	42.67	42.67	-77.35	-72.94	-50.79	-50.79

Note: All values are relative changes in % when switching from propensity score matching to (direct) covariate matching, except in the first row and first column of the data matrix, where the average MSEs of the various propensity score and direct matching estimators are given.

5 Conclusion

This paper investigates the finite sample performance of a comprehensive set of semi- and nonparametric treatment effect estimators using a simulation design that is based on a large empirical data set from Switzerland. In contrast to previous simulation studies which mostly considered semiparametric approaches relying on parametric propensity score estimation, we also investigate more flexible approaches based on semi- or nonparametric propensity scores, nonparametric regression, and direct matching. In addition to (pair, radius, and kernel) matching, inverse probability weighting, regression, and doubly robust estimation, our study also covers recently proposed estimators such as genetic matching, entropy balancing, and weighting based on an empirical likelihood approach. We vary a range of relevant features, such as sample size, selection into treatment, effect heterogeneity, and correct versus misspecification in our simulations. We find that several nonparametric estimators – in particular nonparametric regression, nonparametric DR estimation as suggested by Rothe and Firpo (2013), nonparametric IPW, and one-to-many

covariate matching – by and large outperform commonly used treatment estimators that rely on a parametric propensity score. Among the semiparametric methods investigated, IPW based on the overidentified CBPS of Imai and Ratkovic (2014) is best.

Not all nonparametric approaches work equally well. A maybe surprising result is that nonparametric propensity score estimation on the one hand leads to very competitive IPW and DR estimators, but on the other hand entails a deterioration of matching estimators when compared to matching on a parametric or semiparametric propensity score. Another interesting outcome (in line with Zhao (2004)) with respect to matching is that the best covariate matching estimators, nonparametric regression (which is kernel matching on the covariates) and one-to-many covariate matching, clearly dominate the top propensity score matching methods, namely kernel matching on the parametric propensity score. We also looked at various methods for trimming too large propensity scores but found it to have little impact on the best performing estimators of our simulations, while it crucially improved some of the worst performing matching methods (using the nonparametric propensity score).

Our findings contribute to the literature on the finite sample performance of treatment effect estimators, because a range of methods analysed have (to the best of our knowledge) not been considered in previous studies, which concerns in particular the top performing nonparametric estimators in our study.

References

Abadie, A., and G. W. Imbens (2011): "Bias-Corrected Matching Estimators for Average Treatment Effects," *Journal of Business & Economic Statistics*, 29, 1–11.

Aitchison, J., and C. Aitken (1976): "Multivariate binary discrimination by the kernel method," *Biometrika*, 63, 413–420.

Behncke, S., M. Frölich, and M. Lechner (2010a): "A Caseworker Like Me - Does The Similarity Between The Unemployed and Their Caseworkers Increase Job Placements?," *The Economic Journal*, 120, 1430–1459.

——— (2010b): "Unemployed and their caseworkers: should they be friends or foes?," *Journal of the Royal Statistical Society: Series A*, 173, 67–92.

Blinder, A. (1973): "Wage Discrimination: Reduced Form and Structural Estimates," *Journal of Human Resources*, 8, 436–455.

Busso, M., J. DiNardo, and J. McCrary (2009): "Finite Sample Properties of Semiparametric Estimators of Average Treatment Effects," *forthcoming in the Journal of Business and Economic Statistics*.

Dehejia, R. H., and S. Wahba (1999): "Causal Effects in Non-experimental Studies: Reevaluating the Evaluation of Training Programmes," *Journal of American Statistical Association*, 94, 1053–1062.

Diamond, A., and J. S. Sekhon (2013): "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies," *Review of Economics and Statistics*, 95(3), 932–945.

Fan, J. (1993): "Local Linear Regression Smoothers and their Minimax Efficiency," *Annals of Statistics*, 21, 196–216.

Frölich, M. (2004): "Finite Sample Properties of Propensity-Score Matching and Weighting Estimators," *The Review of Economics and Statistics*, 86, 77–90.

——— (2005): "Matching Estimators and Optimal Bandwidth Choice," *Statistics and Computing*, 15/3, 197–215.

——— (2007a): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139, 35–75.

——— (2007b): "Propensity score matching without conditional independence assumption - with an application to the gender wage gap in the United Kingdom," *Econometrics Journal*, 10, 359–407.

Graham, B., C. Pinto, and D. Egel (2011): "Efficient Estimation of Data Combination Models by the Method of Auxiliary-to-Study Tilting (AST)," *NBER Working Paper No. 16928*.

——— (2012): "Inverse probability tilting for moment condition models with missing data," *Review of Economic Studies*, 79, 1053–1079.

Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66(2), 315–331.

Hainmueller, J. (2012): "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies," *Political Analysis*, 20(1), 25–46.

Hayfield, T., and J. Racine (2008): "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27, 1–32.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998): "Characterizing selection bias using experimental data," *Econometrica*, 66, 1017–1098.

Heckman, J. J., H. Ichimura, and P. Todd (1998a): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

Heckman, J. J., H. Ichimura, and P. Todd (1998b): "Matching as an econometric evaluation estimator," *Review of Economic Studies*, 65, 261–294.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189.

Horvitz, D., and D. Thompson (1952): "A Generalization of Sampling without Replacement from a Finite Population," *Journal of American Statistical Association*, 47, 663–685.

Huber, M. (2014): "Causal pitfalls in the decomposition of wage gaps," *forthcoming in the Journal of Business and Economic Statistics*.

Huber, M., M. Lechner, and G. Mellace (2014a): "The finite sample performance of estimators for mediation analysis under sequential conditional independence," *University of St. Gallen, Department of Economics Discussion Paper No. 2014-15*.

Huber, M., M. Lechner, and G. Mellace (2014b): "Why do tougher caseworkers increase employment? The role of programme assignment as a causal mechanism," *University of St. Gallen, Department of Economics Discussion Paper No. 2014-14*.

Huber, M., M. Lechner, and A. Steinmayr (2014): "Radius matching on the propensity score with bias adjustment: tuning parameters and finite sample behaviour," *forthcoming in Empirical Economics*.

Huber, M., M. Lechner, and C. Wunsch (2013): "The performance of estimators based on the propensity score," *Journal of Econometrics*, 175, 1–21.

Hurvich, C., J. Simonoff, and C. Tsai (1998): "Smoothing Parameter Selection in Nonparametric Regression using an Improved Akaike Information Criterion," *Journal of Royal Statistical Society, Series B*, 60, 271–293.

Ichimura, H. (1993): "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58, 71–120.

Ichimura, H., and O. Linton (2005): "Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators," in *Identification and Inference for Econometric Models*, ed. by D. Andrews, and J. Stock, pp. 149–170. Cambridge University Press, Cambridge.

Imai, K., and M. Ratkovic (2014): "Covariate balancing propensity score," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 243–263.

Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *The Review of Economics and Statistics*, 86, 4–29.

Kang, J. D. Y., and J. L. Schafer (2007): "Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data," *Statistical Science*, 22, 523–539.

Khan, S., and E. Tamer (2010): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78, 2021–2042.

Klein, R. W., and R. H. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61, 387–421.

Kullback, S. (1959): *Information theory and statistics*. Wiley, New York.

Lechner, M., R. Miquel, and C. Wunsch (2011): "Long-run Effects of Public Sector Sponsored Training in West Germany," *Journal of the European Economic Association*, 9, 742–784.

Lechner, M., and A. Strittmatter (2014): "Practical Procedures to Deal with Common Support Problems in Matching Estimation," *University of St. Gallen, Department of Economics Discussion Paper Discussion Paper no. 2014-10*.

Li, Q., J. Racine, and J. Wooldridge (2009): "Efficient Estimation of Average Treatment Effects With Mixed Categorical and Continuous Data," *Journal of Business and Economics Statistics*, 27, 206–223.

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

Millimet, D., and R. Tchernis (2009): "On the specification of propensity scores, with applications to the analysis of trade policies.," *Journal of Business & Economic Statistics*, 27, 297–315.

Ñopo, H. (2008): "Matching as a Tool to Decompose Wage Gaps," *Review of Economics and Statistics*, 90, 290–299.

Oaxaca, R. (1973): "Male-Female Wage Differences in Urban Labour Markets," *International Economic Review*, 14, 693–709.

Pohlmeier, W., R. R. Seiberlich, and S. D. Uysal (2013): "A Simple and Successful Method to Shrink the Weight," *University of Konstanz, Department of Economics Discussion Working Paper no. 2013-05*.

Racine, J., and Q. Li (2004): "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data," *Journal of Econometrics*, 119, 99–130.

Robins, J., A. Rotnitzky, and L. Zhao (1994): "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89, 846866.

Robins, J. M., S. D. Mark, and W. K. Newey (1992): "Estimating exposure effects by modelling the expectation of exposure conditional on confounders," *Biometrics*, 48, 479–495.

Robins, J. M., and A. Rotnitzky (1995): "Semiparametric Efficiency in Multivariate Regression Models with Missing Data," *Journal of the American Statistical Association*, 90, 122–129.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995): "Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data," *Journal of the American Statistical Association*, 90, 106–121.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70(1), 41–55.

——— (1985): "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score.," *The American Statistician*, 39, 33–38.

Rothe, C., and S. Firpo (2013): "Semiparametric Estimation and Inference Using Doubly Robust Moment Conditions," *IZA Discussion Paper No. 7564*.

Rubin, D. B. (1979): "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74, 318–328.

Seifert, B., and T. Gasser (1996): "Finite-Sample Variance of Local Polynomials: Analysis and Solutions," *Journal of American Statistical Association*, 91, 267–275.

Sekhon, J. S. (2011): "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R," *Journal of Statistical Software*, 42, 1–52.

Silverman, B. (1986): *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Smith, J., and P. Todd (2005): "Does matching overcome LaLonde's critique of nonexperimental estimators?," *Journal of Econometrics*, 125, 305–353.

Waernbaum, I. (2012): "Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation," *Statistics in Medicine*, 31, 1572–1581.

Wang, M., and J. van Ryzin (1981): "A class of smooth estimators for discrete distributions," *Biometrika*, 68, 301–309.

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

——— (2008): "Sensitivity of Propensity Score Methods to the Specifications," *Economics Letters*, 98, 309–319.

A Appendix

A.1 Figures

A.2 Further tables

Figure A.1: Outcome distribution (cumulative months jobseeking betw. months 10 and 36)

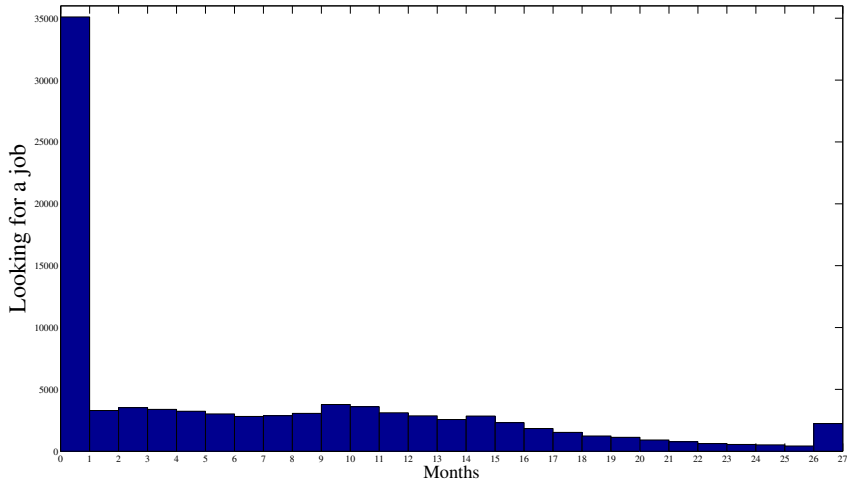
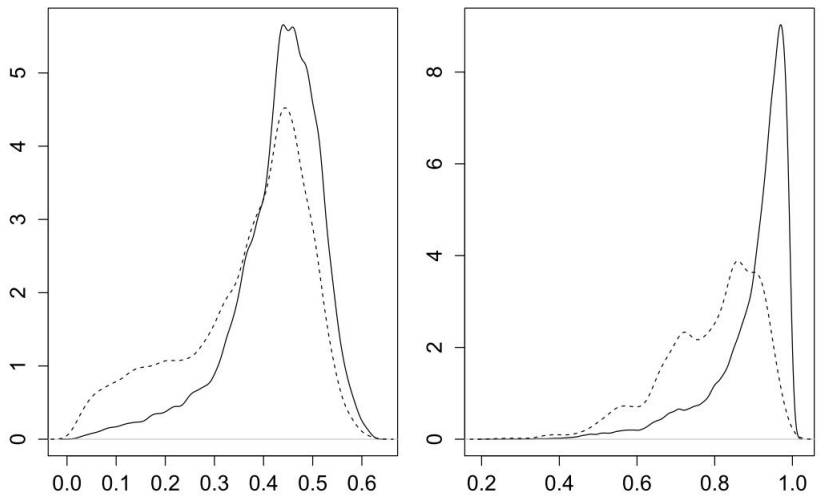


Figure A.2: Propensity score overlap across treatment states in the 'populations'



Note: Propensity score distributions under treatment (solid lines) and non-treatment (dashed lines) in the 'population' used for the simulations, for the treatments 'program participation' (left graph) and 'noncooperative caseworker' (right graph). Results are based on kernel density estimation using a Gaussian kernel, where the Silverman (1986) rule of thumb-based bandwidths for the treated and non-treated groups are 0.009 and 0.014, respectively, in the left graph and 0.008 and 0.020, respectively, in the right graph.

Table A.1: Probit regressions of treatments on covariates in the 'populations' used for the simulations

	treatment: program participation				treatment: noncooperative caseworker			
	coef	se	z-val	p-val	coef	se	z-val	p-val
constant	-0.712	0.036	-19.915	0.000	0.733	0.058	12.632	0.000
female jobseeker	0.0697	0.012	5.837	0.000	-0.291	0.022	-13.538	0.000
foreign mother tongue	0.088	0.014	6.459	0.000	-0.252	0.024	-10.665	0.000
unskilled	0.092	0.014	6.443	0.000	0.453	0.025	18.379	0.000
semiskilled	0.002	0.015	0.156	0.876	0.185	0.023	8.011	0.000
skilled, no degree	0.109	0.026	4.192	0.000	0.582	0.048	12.080	0.000
unemployment spells last 2 years	-0.258	0.005	-52.513	0.000	-0.128	0.006	-20.576	0.000
fraction employed last yr	0.195	0.021	9.436	0.000	0.023	0.032	0.715	0.475
low employability	0.154	0.021	7.317	0.000	0.595	0.035	17.152	0.000
medium employability	0.174	0.016	10.608	0.000	0.302	0.023	13.199	0.000
female caseworker	0.121	0.012	10.172	0.000	-0.297	0.021	-13.842	0.000
caseworker above vocational	0.061	0.012	5.169	0.000	-0.182	0.020	-9.037	0.000
caseworker higher education	0.073	0.014	5.239	0.000	-0.109	0.023	-4.743	0.000
cantonal unemployment rate	0.041	0.006	6.589	0.000	0.263	0.011	23.886	0.000
French speaking region	-0.187	0.020	-9.485	0.000	-1.228	0.030	-40.275	0.000
French × female jobseeker	0.114	0.024	4.851	0.000	-0.136	0.033	-4.149	0.000
French × mother tongue	-0.223	0.025	-8.831	0.000	0.465	0.036	12.819	0.000
French × female casew.	-0.127	0.024	-5.227	0.000	0.375	0.034	11.037	0.000

Note: 'coef', 'se', 'z-val', and 'p-val' give the coefficient estimates, standard errors, z-values, and p-values, respectively.

Table A.2: Outcome regressions on the covariates conditional on treatment state

	program participation=1				program participation=0			
	coef	se	z-val	p-val	coef	se	z-val	p-val
constant	-0.990	1.161	-0.852	0.394	5.166	0.701	7.367	0.000
female jobseeker	3.821	1.691	2.260	0.024	-2.857	1.039	-2.750	0.006
foreign mother tongue	0.903	0.118	7.678	0.000	1.478	0.078	18.956	0.000
unskilled	2.240	0.167	13.398	0.000	1.568	0.105	14.908	0.000
semiskilled	2.038	0.176	11.613	0.000	1.319	0.107	12.282	0.000
skilled, no degree	1.860	0.301	6.173	0.000	1.515	0.191	7.945	0.000
unemployment spells last 2 years	0.973	0.252	3.856	0.000	1.196	0.124	9.641	0.000
unemployment spells last 2 years^2	-0.504	0.150	-3.360	0.001	-0.336	0.062	-5.382	0.000
unemployment spells last 2 years^3	0.074	0.021	3.574	0.000	0.029	0.007	3.917	0.000
frac. employed last yr	6.744	0.683	9.875	0.000	1.446	0.460	3.144	0.002
frac. employed last yr^2	-4.629	0.572	-8.098	0.000	-1.289	0.379	-3.397	0.001
low employability	1.191	0.191	6.244	0.000	1.456	0.118	12.342	0.000
medium employability	0.590	0.152	3.893	0.000	0.727	0.092	7.899	0.000
female caseworker	-0.150	0.101	-1.484	0.138	-0.289	0.068	-4.212	0.000
caseworker above vocational	-0.079	0.104	-0.762	0.446	-0.117	0.067	-1.737	0.082
caseworker higher education	-0.012	0.122	-0.095	0.924	-0.068	0.079	-0.861	0.389
cantonal unemployment rate	2.218	0.647	3.426	0.001	-0.813	0.395	-2.056	0.040
cantonal unemployment rate^2	-0.254	0.089	-2.844	0.004	0.130	0.055	2.372	0.018
French speaking region	1.466	0.187	7.855	0.000	1.914	0.112	17.105	0.000
French × female jobseeker	-0.036	0.227	-0.158	0.875	0.057	0.142	0.401	0.689
French × mother tongue	-0.335	0.237	-1.413	0.158	-0.285	0.139	-2.055	0.040
French × female caseworker	-0.168	0.222	-0.754	0.451	0.611	0.136	4.499	0.000
French × unskilled	0.176	0.222	0.794	0.427	0.394	0.145	2.718	0.007
French × semiskilled	-0.257	0.253	-1.017	0.309	0.369	0.161	2.288	0.022
French × skilled, no degree	0.138	0.443	0.311	0.756	0.464	0.293	1.582	0.114
French × cantonal unemployment rate	-2.279	0.964	-2.364	0.018	1.372	0.602	2.281	0.023
French × cantonal unemployment rate^2	0.308	0.133	2.322	0.020	-0.167	0.084	-1.991	0.046
	noncooperative caseworker=1				noncooperative caseworker=0			
	coef	se	z-val	p-val	coef	se	z-val	p-val
constant	4.083	0.872	4.684	0.000	2.280	0.835	2.730	0.006
female jobseeker	-1.608	1.275	-1.261	0.207	-0.914	1.241	-0.737	0.461
foreign mother tongue	1.421	0.095	14.936	0.000	1.255	0.090	13.988	0.000
unskilled	1.617	0.129	12.529	0.000	1.897	0.124	15.306	0.000
semiskilled	1.437	0.133	10.816	0.000	1.574	0.127	12.372	0.000
skilled, no degree	1.700	0.230	7.388	0.000	1.570	0.228	6.899	0.000
unemployment spells last 2 years	0.901	0.159	5.682	0.000	1.059	0.152	6.965	0.000
unemployment spells last 2 years^2	-0.223	0.081	-2.739	0.006	-0.361	0.078	-4.602	0.000
unemployment spells last 2 years^3	0.018	0.010	1.832	0.067	0.036	0.010	3.796	0.000
frac. employed last yr	2.597	0.560	4.634	0.000	3.117	0.523	5.954	0.000
frac. employed last yr^2	-1.987	0.464	-4.281	0.000	-2.210	0.434	-5.093	0.000
low employability	1.305	0.147	8.860	0.000	1.487	0.138	10.789	0.000
medium employability	0.637	0.117	5.431	0.000	0.790	0.106	7.422	0.000
female caseworker	-0.227	0.083	-2.727	0.006	-0.211	0.079	-2.684	0.007
caseworker above vocational	-0.097	0.083	-1.166	0.244	-0.085	0.078	-1.087	0.277
caseworker higher education	-0.012	0.097	-0.124	0.901	-0.049	0.092	-0.531	0.595
cantonal unemployment rate	-0.136	0.493	-0.277	0.782	0.475	0.467	1.017	0.309
cantonal unemployment rate^2	0.039	0.068	0.566	0.571	-0.030	0.065	-0.461	0.645
French speaking region	1.573	0.142	11.093	0.000	1.883	0.132	14.291	0.000
French × female jobseeker	0.095	0.178	0.535	0.593	-0.033	0.166	-0.202	0.840
French × mother tongue	-0.279	0.178	-1.572	0.116	-0.317	0.162	-1.958	0.050
French × female caseworker	0.455	0.175	2.601	0.009	0.326	0.155	2.096	0.036
French × unskilled	0.270	0.176	1.534	0.125	0.413	0.169	2.448	0.014
French × semiskilled	0.154	0.200	0.767	0.443	0.273	0.186	1.465	0.143
French × skilled, no degree	0.103	0.357	0.290	0.772	0.554	0.339	1.636	0.102
French × cantonal unemployment rate	0.626	0.742	0.844	0.399	0.385	0.710	0.542	0.588
French × cantonal unemployment rate^2	-0.070	0.103	-0.680	0.497	-0.039	0.098	-0.402	0.687

Note: 'coef', 'se', 'z-val', and 'p-val' give the coefficient estimates, standard errors, z-values, and p-values, respectively.

Table A.3: Average number of trimmed obs. under the common support restriction

	overall	$N = 750$	$N = 3000$	correct spec.	misspec.	D: program	D: noncoop. caseworker
probit-based pscore	14.89	13.20	16.59	24.16	5.63	2.91	26.87
just identified CBPS	44.09	34.28	53.89	78.76	9.41	3.77	84.40
overidentified CBPS	20.39	18.88	21.90	33.81	6.97	3.50	37.28
semipara pscore	31.50	19.87	43.13	43.44	19.55	15.19	47.80
nonpara pscore	161.44	58.72	264.15	215.20	107.68	10.19	312.69

Table A.4: Average squared biases over all settings without trimming

	squared bias	relative difference	rank
nonpara regression (crossval bandwidth)	0.04	0.0	20.1
doubly robust (nonpara; crossval bandwidth)	0.06	46.1	14.4
nonpara regression (undersmoothing)	0.07	80.2	21.4
doubly robust (nonpara; undersmoothed outcomes)	0.08	100.3	18.8
IPW (nonpara pscore)	0.09	123.9	33.1
IPW (overidentified CBPS)	0.11	176.7	23.5
direct 1:5 match	0.11	184.3	23.5
direct 1:5 match with bias correction	0.11	184.3	23.2
inverse probability tilting (IPT)	0.12	203.0	25.9
IPW (semipara pscore)	0.12	204.7	38.8
kernel match (semipara pscore; oversmooth)	0.12	207.3	34.4
IPW (just identified CBPS)	0.12	212.3	26.3
entropy balancing (EB)	0.12	212.3	27.5
doubly robust (just identified CBPS, para outcome)	0.12	212.3	26.8
doubly robust (para pscore and outcome)	0.12	215.0	27.7
kernel match (para pscore; oversmooth)	0.12	216.4	27.6
IPW (para pscore)	0.13	217.7	26.9
kernel match (just identified CBPS; oversmooth)	0.13	218.5	28.4
para regression	0.13	218.7	31.4
doubly robust (semipara pscore, para outcome)	0.13	222.9	25.7
direct pair match	0.13	236.0	32.9
direct pair match with bias correction	0.13	236.0	32.8
genetic match	0.14	244.4	31.3
genetic match with bias correction	0.14	244.4	31.5
kernel match (para pscore; crossval bandw)	0.14	263.5	33.9
kernel match (semipara pscore; crossval bandw)	0.14	265.5	42.5
kernel match (just identified CBPS; undersmooth)	0.15	278.8	40.2
kernel match (just identified CBPS; crossval bandw)	0.15	283.6	39.5
kernel match (para pscore; undersmooth)	0.15	284.0	39.3
radius match (semipara pscore; large radius)	0.16	297.8	33.9
radius match (semipara pscore; medium radius)	0.16	304.0	35.6
radius match (semipara pscore; small radius)	0.16	310.3	37.8
radius match (para pscore; large radius)	0.16	314.2	27.8
pair match (nonpara pscore)	0.16	315.1	39.7
radius match (para pscore; medium radius)	0.17	322.6	29.2
pair match (semipara pscore)	0.17	323.8	44.5
radius match (para pscore; small radius)	0.17	330.9	31.6
direct radius match	0.17	335.0	31.2
direct radius match with bias correction	0.17	335.0	31.2
pair match (para pscore)	0.18	353.6	35.9
radius match (just identified CBPS; large radius)	0.18	360.7	32.2
radius match (just identified CBPS; medium radius)	0.19	383.7	34.6
radius match (just identified CBPS; small radius)	0.19	393.8	35.9
radius match (nonpara pscore; large radius)	0.20	396.4	36.8
pair match (just identified CBPS)	0.20	409.7	41.5
radius match (nonpara pscore; medium radius)	0.20	419.0	37.9
radius match (nonpara pscore; small radius)	0.21	436.8	38.4
kernel match (semipara pscore; undersmooth)	0.22	454.2	44.2
kernel match (nonpara pscore; oversmooth)	0.22	460.2	38.1
kernel match (nonpara pscore; crossval bandw)	0.38	868.0	39.4
kernel match (nonpara pscore; undersmooth)	>1000	>100000	49.2

Note: 'squared bias' gives the average squared bias, 'relative difference' is in percent and provides the relative difference to the lowest average squared bias (of the best performing estimator), 'rank' gives the average rank of the estimators across all simulations. All radius matching estimators on the propensity score ('radius m.') include bias correction.

Table A.5: Average variances over all settings without trimming

	variance	relative difference	rank
IPW (overidentified CBPS)	0.20	0.0	3.3
nonpara regression (crossval bandwidth)	0.21	2.5	10.5
doubly robust (nonpara; crossval bandwidth)	0.21	3.8	8.8
IPW (nonpara pscore)	0.21	5.3	6.9
IPW (para pscore)	0.21	6.9	5.0
para regression	0.22	7.3	3.9
direct 1:5 match with bias correction	0.22	8.0	16.4
direct 1:5 match	0.22	8.0	16.6
IPW (semipara pscore)	0.22	9.8	13.4
inverse probability tilting (IPT)	0.22	10.5	10.9
nonpara regression (undersmoothing)	0.22	10.6	14.2
doubly robust (nonpara; undersmoothed outcomes)	0.23	12.0	15.2
doubly robust (para pscore and outcome)	0.23	15.5	6.8
doubly robust (semipara pscore, para outcome)	0.24	19.9	12.1
kernel match (para pscore; oversmooth)	0.26	28.6	13.0
kernel match (semipara pscore; oversmooth)	0.28	37.1	20.4
IPW (just identified CBPS)	0.29	43.9	10.4
entropy balancing (EB)	0.29	43.9	10.5
doubly robust (just identified CBPS, para outcome)	0.29	43.9	10.2
kernel match (just identified CBPS; oversmooth)	0.30	47.3	15.7
direct pair match	0.36	78.3	33.8
direct pair match with bias correction	0.36	78.3	33.8
genetic match with bias correction	0.40	97.8	35.9
genetic match	0.40	97.8	35.9
radius match (semipara pscore; large radius)	0.40	97.9	28.6
pair match (semipara pscore)	0.42	108.6	29.9
radius match (semipara pscore; medium radius)	0.42	109.0	31.2
radius match (semipara pscore; small radius)	0.45	122.4	34.9
radius match (para pscore; large radius)	0.46	130.0	28.6
kernel match (just identified CBPS; crossval bandw)	0.47	136.4	22.7
radius match (para pscore; medium radius)	0.51	153.4	31.4
kernel match (semipara pscore; crossval bandw)	0.53	162.3	32.9
pair match (para pscore)	0.54	169.3	34.4
radius match (para pscore; small radius)	0.55	173.7	35.6
kernel match (para pscore; crossval bandw)	0.65	224.9	20.4
radius match (just identified CBPS; large radius)	0.76	279.2	33.4
kernel match (para pscore; undersmooth)	0.93	363.8	25.2
radius match (just identified CBPS; medium radius)	1.03	413.6	36.5
pair match (just identified CBPS)	1.10	449.8	39.2
radius match (just identified CBPS; small radius)	1.19	491.4	40.8
direct radius match	1.37	583.8	53.9
direct radius match with bias correction	1.37	583.8	53.8
kernel match (just identified CBPS; undersmooth)	1.74	765.4	37.2
radius match (nonpara pscore; large radius)	3.10	1444.3	52.3
pair match (nonpara pscore)	3.16	1472.8	46.6
radius match (nonpara pscore; medium radius)	4.41	2094.8	54.4
radius match (nonpara pscore; small radius)	5.94	2858.1	56.6
kernel match (nonpara pscore; oversmooth)	9.34	4550.4	45.7
kernel match (semipara pscore; undersmooth)	12.96	6349.3	49.4
kernel match (nonpara pscore; crossval bandw)	181.76	90359.1	52.1
kernel match (nonpara pscore; undersmooth)	>1000	>100000	60.1

Note: 'variance' gives the average variance, 'relative difference' is in percent and provides the relative difference to the lowest average variance (of the best performing estimator), 'rank' gives the average rank of the estimators across all simulations. All radius matching estimators on the propensity score ('radius m.') include bias correction.