FINITE-SAMPLE PROPERTIES OF PROPENSITY-SCORE MATCHING AND WEIGHTING ESTIMATORS

Markus Frölich*

Abstract—The finite-sample properties of matching and weighting estimators, often used for estimating average treatment effects, are analyzed. Potential and feasible precision gains relative to pair matching are examined. Local linear matching (with and without trimming), *k*-nearest-neighbor matching, and particularly the weighting estimators performed worst. Ridge matching, on the other hand, leads to an approximately 25% smaller MSE than does pair matching. In addition, ridge matching is least sensitive to the design choice.

I. Introduction

MATCHING estimators are frequently applied in treatment evaluation to estimate average treatment effects. In this paper, the finite-sample properties of different matching and weighting estimators are analyzed. When selection is on the observables, for estimating treatment effects it is necessary to adjust for the different distributions of the observed characteristics in the treated and the nontreated population. By doing so, the causal effect is separated from the effect due to nonrandom selection into treatment. Several nonparametric estimators have been suggested for this.

Pair matching, which is most frequently used in applications, proceeds by finding for each treated observation a nontreated observation with identical (or very similar) characteristics. This aligns the distributions of the characteristics in the treated and the matched comparison sample. Finding matches that are identical with respect to all relevant covariates, however, can be difficult if the number of covariates is large. Nevertheless, as proposed by Rosenbaum and Rubin (1983), matching on the (one-dimensional) *propensity score* suffices to adjust for the differences in the observed covariates.

Matching only one nontreated observation to each treated observation is obviously inefficient, and alternative matching estimators have been suggested. Heckman, Ichimura, and Todd (1997, 1998) proposed local polynomial matching, which includes kernel matching and local linear matching. An alternative approach to matching is based on weighting the nontreated observations by the propensity score, as suggested in Imbens (2000).

Asymptotic properties of these matching and weighting estimators have been examined by Hahn (1998), Heckman et al. (1998), Abadie and Imbens (2001), Ichimura and Linton (2002), and Hirano, Imbens, and Ridder (2003). The finite-sample properties of these estimators, however, have not been subject to extensive investigation. For their practical use, it would be important to know which estimator performs best under which conditions. In addition, many of these estimators require the choice of bandwidth parameters, and guidance is needed on how to select these.

In this paper the finite-sample properties are investigated for pair matching, kernel matching, local linear matching, *k*-nearest-neighbor (*k*-NN) matching, and the weighting estimator. Moreover, the effect of trimming, as suggested by Heckman et al. (1997, 1998), is examined. In addition, an alternative matching estimator based on nonparametric ridge regression is proposed and analyzed. This *ridge matching* estimator is motivated by the observation that local linear regression can be very unreliable in regions of sparse data. Trimming the observations in regions with few comparisons, as in Heckman, Ichimura, and Todd (1997, 1998), is one way to sidestep this problem. Ridge matching, on the other hand, seeks to stabilize the estimator through ridge regression (Seifert and Gasser, 1996, 2000).

The Monte Carlo simulations consist of two parts: First, the mean squared error of the various estimators is simulated at their optimal bandwidth values. These results show the potential of each estimator, independent of the particular bandwidth selector. They also indicate how large the improvement in mean squared error vis-à-vis the benchmark pair-matching estimator could be at most. Thereafter, the mean squared error is examined when the bandwidth is chosen by cross-validation. Cross-validation is a convenient approach to selecting the bandwidth value and turns out to work reasonably well for some estimators. Ridge matching very often performs best among all estimators (with a mean squared error approximately 25% lower than for pair matching) and is robust to the simulation design. Local linear matching (with and without trimming) and *k*-NN matching are more susceptible to the simulation design and can often be worse than pair matching. The weighting estimator is very unreliable and has a larger MSE than pair matching in all simulations.

II. Matching Estimators

When evaluating the effect of a treatment, interest often lies in the estimation of average treatment effects. Let Y_i^0 and Y_i^1 denote the *potential* outcomes, where Y_i^1 is the outcome if the individual receives treatment, and Y_i^0 the outcome if he or she does not. Let $D_i \in \{0, 1\}$ indicate treatment receipt. The average treatment effect on the treated is

$$E[Y^1|D = 1] - E[Y^0|D = 1].$$

Received for publication August 2, 2001. Revision accepted for publication July 29, 2003.

* University of St. Gallen.

The author is also affiliated with the Institute for the Study of Labor (IZA), Bonn. He would like to thank Yuanhua Feng, Bernd Fitzenberger, Hidehiko Ichimura, Michael Lechner, Jeff Smith, two anonymous referees, and participants at the Econometric Society Meeting in Venice (2002) and at seminars at the University of Konstanz and Mannheim for helpful comments and suggestions. Financial support from the Swiss National Science Foundation (Project NSF 4043-058311) is gratefully acknowledged.

Whereas the first term can be estimated by the mean outcome among the treated individuals, the term $E[Y^0|D = 1]$ is counterfactual. Identifying this *counterfactual mean* is fundamental to treatment evaluation. Generally, $E[Y^0|D = 1]$ is not equal to $E[Y^0|D = 0]$ if treatment selection is nonrandom. Nevertheless, conditional on all confounding factors X, that is, all factors that influenced the potential outcome *and* the decision to participate in treatment, D is independent of Y^0:

$$Y^0 \perp\!\!\!\perp D|X, \tag{1}$$

and the counterfactual mean is identified as $E[Y^0|D = 1] = E[E[Y^0|X, D = 0]|D = 1]$, provided the support of X among the treated is contained in the support of X among the nontreated. If X contains only a few variables, the counterfactual mean can be estimated using nonparametric regression of Y on X in the nontreated sample. However, if X is high-dimensional, nonparametric regression can be difficult. Rosenbaum and Rubin (1983) have shown that the dimension of the estimation problem can be reduced substantially in that the counterfactual mean is also identified as

$$E[Y^0|D = 1] = \int E[Y|p(X) = \rho, D = 0] \cdot f_{p|D=1}(\rho)\, d\rho, \tag{2}$$

where $p(x) = P(D = 1|X = x)$ is the *one-dimensional propensity score*, and $f_{p|D=1}$ is the density of $p(X)$ in the treated population.

Consider a sample of i.i.d. observations $\{Y_i, D_i, p_i\}_{i=1}^n$, where $p_i = p(X_i)$. Denote the number of treated observations by n_1, and the number of control observations by n_0. A variety of propensity-score matching estimators have been proposed for estimating the counterfactual mean, which can be characterized as

$$\widehat{E[Y^0|D = 1]} = \frac{1}{n_1} \sum_{i:D_i=1} \hat{m}(p_i), \tag{3}$$

where $\hat{m}(\rho)$ is an estimate of the mean outcome in the nontreated population conditional on the propensity score: $m(\rho) = E[Y|p(X) = \rho, D = 0]$. The matching estimators differ in how they estimate $m(\rho)$. The most prevalent estimator is *pair matching*, which proceeds by finding for each treated observation a control observation with identical (or very similar) value of p, that is, it uses first-nearest-neighbor regression to estimate $m(\rho)$.

Heckman et al. (1997, 1998) suggested estimating $m(\rho)$ by local polynomial regression: *Kernel matching* estimates $m(\rho)$ by Nadaraya-Watson regression, whereas *local linear matching* is based on a local linear regression estimator. Heckman, Ichimura, and Todd (1997) advocated local linear regression for its well-known optimality properties.[1]

In small samples, however, local linear regression often leads to a very rugged curve in regions of sparse or clustered data (Seifert & Gasser, 1996). To deal with this erratic behavior, Heckman et al. (1997, 1998) implement a trimming procedure to discard the nonparametric regression results in regions where the density of the propensity score in the nontreated population is small. The *trimmed matching estimator* is

$$\widehat{E[Y^0|D = 1]} = \frac{\sum_{i:D_i=1} \hat{m}(p_i) \cdot 1[\hat{f}_{p|D=0}(p_i) > \tau]}{\sum_{i:D_i=1} 1[\hat{f}_{p|D=0}(p_i) > \tau]}, \tag{4}$$

where τ is set so that, for example, 2% or 5% of the treated observations are trimmed.

Trimming, however, is a very rough solution for the small-sample problems of local linear regression. Various approaches to stabilize the local linear estimator in finite samples have been developed. A simple but promising approach is local linear ridge regression, which modifies the local linear estimator by adding a ridge term to its denominator to avoid near-zero denominators. The regression estimator of Seifert and Gasser (1996, 2000) is

$$\hat{m}(\rho) = \frac{T_0}{S_0} + \frac{T_1 \cdot (\rho - \bar{p})}{S_2 + rh|\rho - \bar{p}|}, \tag{5}$$

where

$$S_a(\rho) = \sum_{j:D_j=0} (p_j - \bar{p})^a K\left(\frac{p_j - \rho}{h}\right),$$

$$T_a(\rho) = \sum_{j:D_j=0} Y_j (p_j - \bar{p})^a K\left(\frac{p_j - \rho}{h}\right),$$

and

$$\bar{p} = \sum_{j:D_j=0} p_j K\left(\frac{p_j - \rho}{h}\right) \Big/ \sum_{j:D_j=0} K\left(\frac{p_j - \rho}{h}\right).$$

The bandwidth value h converges to 0 with growing sample size, and $K(\cdot)$ is a kernel weighting function. The constant r is the ridge parameter that ensures nonzero denominators. According to the rule of thumb of Seifert and Gasser (2000), r is set to $\frac{5}{16}$ for the Epanechnikov kernel and to $[4\sqrt{2\pi} \int \phi^2(u)\, du]^{-1} \approx 0.35$ for the Gaussian kernel. Inserting equation (5) in (3) gives the *ridge matching* estimator.

In the Monte Carlo simulations, 11 different matching estimators are compared: pair matching, kernel matching (with Epanechnikov and Gaussian kernel), local linear matching and trimmed local linear matching (with Epanechnikov

[1] See, for example, Fan (1992, 1993), Hastie and Loader (1992) and Fan et al. (1997).

and Gaussian kernel), ridge matching (with Epanechnikov and Gaussian kernel), and two variants of k-NN matching. In the first variant of k-NN matching, the standard k-NN regression estimator is used [which estimates $m(\rho)$ by the average outcome of the k selected neighbors]. In the second variant, $m(\rho)$ is estimated by a *weighted* average outcome of the k neighbors, where the neighbors are weighted according to their distance to ρ (using Epanechnikov weights).

An alternative approach to estimating the counterfactual mean is based on weighting by the propensity score ratio, as suggested in Imbens (2000). The *weighting estimator* is

$$\widehat{E[Y^0|D=1]} = \frac{1}{n_1} \sum_{i:D_i=0} Y_i \frac{p_i}{1-p_i}. \tag{6}$$

Because the ratio $p_i/(1-p_i)$ can become very large for values of p_i close to 1, some form of trimming or capping is necessary in finite samples. In the simulations, a ceiling $\bar{c}$ for the ratio $p_i/(1-p_i)$ is examined (that is, ratios larger than $\bar{c}$ are set to $\bar{c}$). In addition, a trimming rule, viz. deleting the observations with the largest ratios, was considered, but performed worse.

III. Monte Carlo Study

A. Simulation Design

The design of the Monte Carlo study consists of two parts: the distributions of the propensity score in the treated population ($f_{p|D=1}$) and the nontreated population ($f_{p|D=0}$) (see figure 1), and the specification of $m(p)$, that is, the conditional expectation function of Y given the propensity score (figure 2). The distribution of the propensity scores in figure 1 is driven by the distribution of X and the specification of the propensity score $p(x)$. To have a simple design that, at the same time, allows one to generate very different shapes of $f_{p|D=1}$ and $f_{p|D=0}$, the covariate X is chosen to be one-dimensional and drawn from the Johnson S_B distribution (see figure A1), and the propensity score is specified as a linear function $p(x) = \alpha + \beta x$. (Details are given in the appendix.) Through the choice of different values for α and β, very different shapes of $f_{p|D=1}$ and $f_{p|D=0}$ can be generated, as demonstrated in Figure 1. An increase in α shifts the average value of the propensity score upwards, so that the number of treated relative to the number of nontreated increases. Through the parameter β, the spread of the propensity score values is controlled.

Figure 1 shows the five different distributions of the propensity score that are used in the Monte Carlo simulations. [The support of the propensity score is $(\alpha, \alpha + \beta)$ and thus varies with α and β. Figure 1 displays the densities of the *rescaled* propensity score, which is scaled by $(p - \alpha)/\beta$ so that its support is always (0,1). This ensures that the support is compatible with the regression curves in figure 2 for all designs. In the simulations, this rescaled propensity score is used.] The designs are chosen to illustrate different degrees of overlapping density mass and to represent different ratios of control to treated observations.[2] In the first three designs, the population mean of the propensity score is 0.5, that is, the expected ratio of control to treated observations is 1 : 1. In the fourth design the ratio is 4 : 1; in the last design it is 1 : 4. The fourth design is most pertinent to the estimation of the average treatment effect on the treated when the pool of control observations is large. However, if interest lies also in the average treatment effect or in the average treatment effect on the nontreated, or if multiple treatments are to be evaluated, the estimation of the counterfactual means involves settings where the numbers of treated and control observations are of similar magnitude and/or where the treated greatly exceed the controls.

The first three designs, where the control-treated ratio is 1 : 1, vary in the discrepancy between the two densities $f_{p|D=1}$ and $f_{p|D=0}$. The differences in the two densities determine how challenging the estimation setting is for the matching estimator. If the two densities are rather different, the matching estimator (3) needs to estimate $m(p)$ often in regions where there are only very few control observations. In design 1, for example, a substantial amount of the probability mass of the treated is located to the right of 0.8, where there is only little probability mass of the nontreated. Hence nonparametric regression in these regions is rather imprecise. In design 3, on the other hand, the two densities are very similar, and nonparametric regression should be more precise at all relevant locations. Design 2 is midway between design 1 and design 3 and will allow us to determine whether the measures of closeness reported below are monotonic in the shape.

A useful measure of the distance between these two densities is the Kullback-Leibler information criterion (KLIC), which is defined as (Kullback and Leibler, 1951; Kullback, 1959)

$$\text{KLIC} = \int \left(\ln \frac{f_{p|D=1}(\rho)}{f_{p|D=0}(\rho)} \right) f_{p|D=1}(\rho)\, d\rho,$$

where the integral is taken over the common support of $f_{p|D=0}$ and $f_{p|D=1}$. The KLIC is equal to 0 if the two densities are identical, and it increases with the discrepancy between the two distributions. The KLIC is attractive here because it weights the discrepancy in the densities by the probability mass among the treated, analogously to the weighting for the counterfactual mean (2). Hence regions where $f_{p|D=0}$ is small (and $f_{p|D=1}/f_{p|D=0}$ is large) contribute significantly to an increase in the discrepancy measure only when $f_{p|D=1}$ itself is large, that is, when there are many treated observations for whom $m(p)$ needs to be estimated. In regions where $f_{p|D=1}$ is small, the density ratio hardly matters at all.

[2] The support of $f_{p|D=1}$ and $f_{p|D=0}$ is identical in all designs.

FIGURE 1.—DENSITIES OF THE PROPENSITY SCORE: $f_{p|D=1}$ AND $f_{p|D=0}$

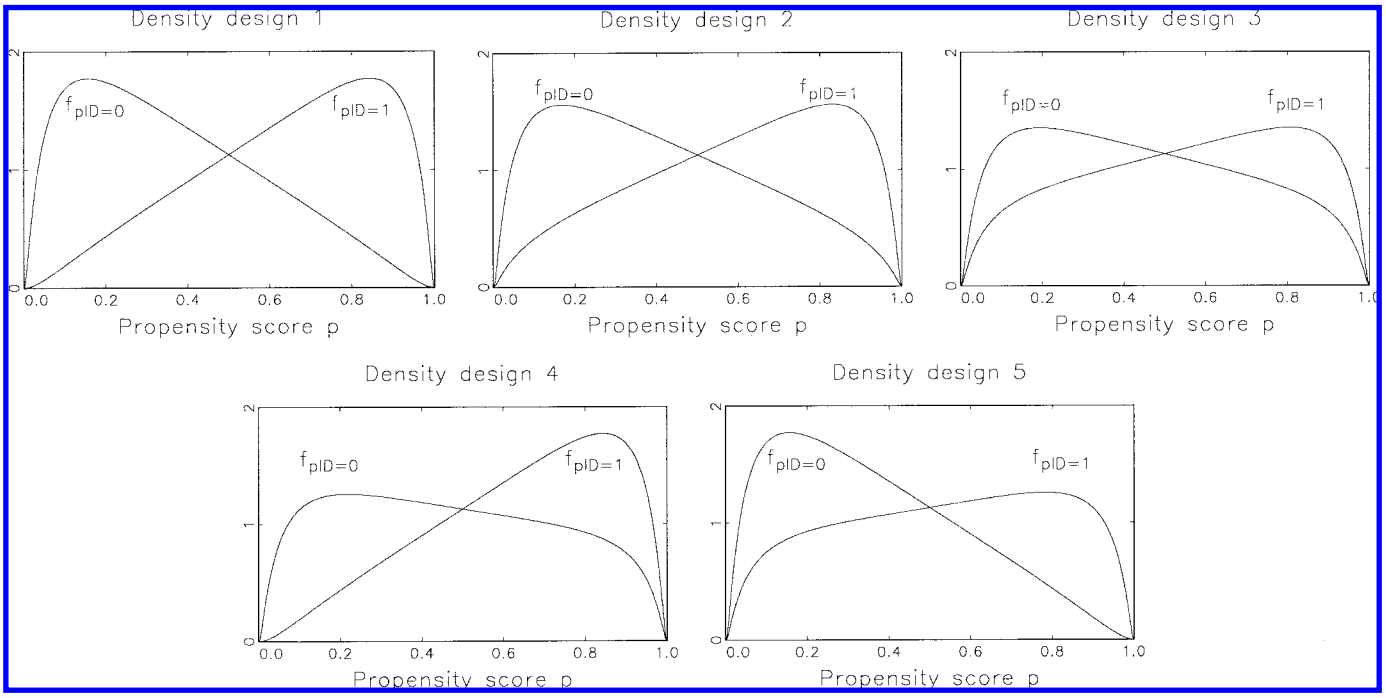


TABLE 1.—KLIC AND THE CONTROL-TREATED RATIO

Design	α	β	Control-treated Ratio	KLIC
1	0	1	1:1	0.73
2	0.15	0.7	1:1	0.42
3	0.30	0.4	1:1	0.23
4	0	0.4	4:1	0.58
5	0.6	0.4	1:4	0.77

Note: The control-treated ratio is the ratio of the expected numbers of controls and treated: $E[n_0]/E[n_1]$.

Table 1 displays the values of the KLIC and the control-treated ratio for the different designs. These two parameters are indicative of the level of difficulty of the estimation setting. Whereas a larger KLIC indicates that the shapes of the two densities are more different, so that the matching estimator more often needs to estimate $m(p)$ in regions with few control observations, the control-treated ratio indicates the absolute number of control observations available. For a given sample size $(n_0 + n_1)$, a larger control-treated ratio implies that more observations on Y are available to estimate $m(p)$. In contrast, if the control-treated ratio is 1 : 4, as in design 5, there are very few nontreated observations to estimate $m(p)$.

The second part of the design of the Monte Carlo simulations concerns the specification of $m(p)$. Six different regression curves with different degrees of nonlinearity are considered (figure 2 and table A1). The first regression curve is a straight line. The second has a conspicuous local nonlinearity,[3] and the third is mildly nonlinear throughout. The fourth and the fifth curve are concave in shape, and the latter has some additional nonlinear structure. The sixth curve is highly nonlinear.[4] A mean-zero, uniform error term with standard deviation 0.1 is added as noise to these curves.

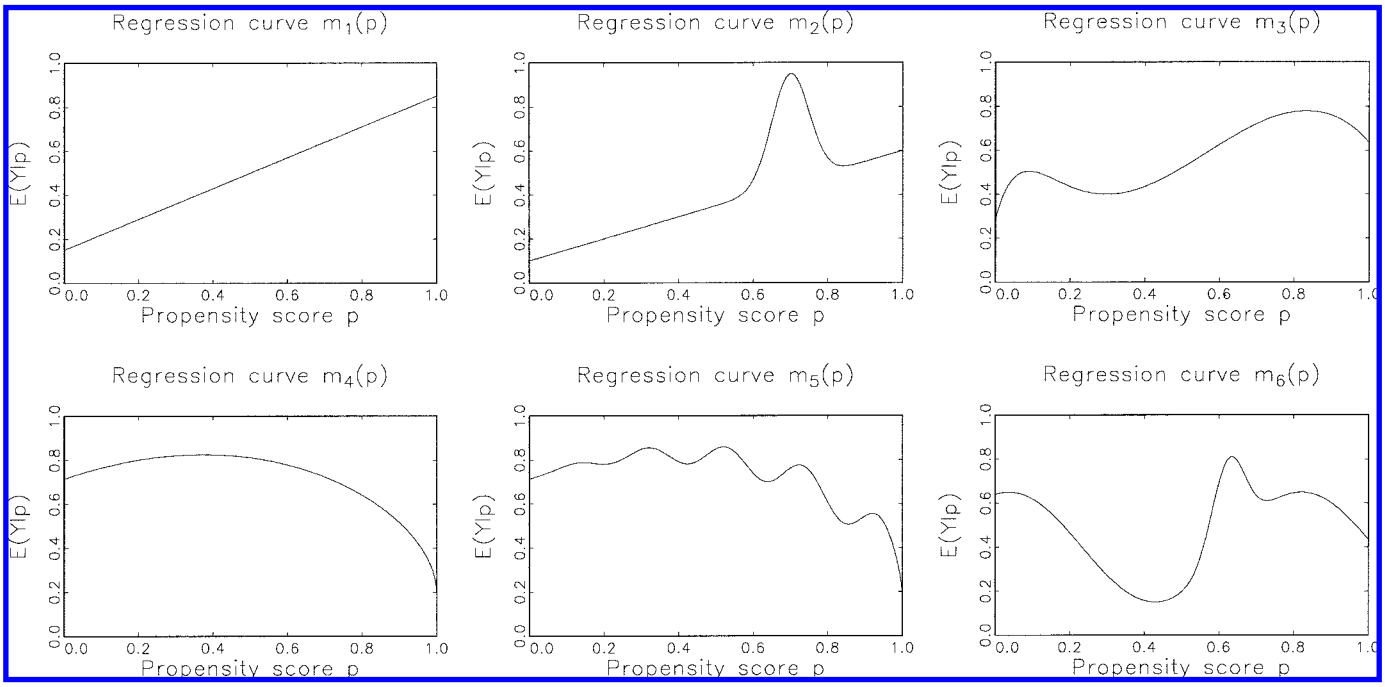
Though it was straightforward to interpret the KLIC and the control-treated ratio, in that a higher KLIC or a lower control-treated ratio makes the estimation setting more demanding (that is, increases the MSE), such an interpretation is less obvious for the regression curve $m(p)$. Although curves with greater curvature imply a larger local bias for the nonparametric regression estimator, these biases may partly cancel when the matching estimator takes the average of the estimated $\hat{m}(p_i)$. Hence, more pronounced nonlinearities in itself will not necessarily increase the MSE. They may favor one estimator over the other, though. The straight line m_1, for instance, is more suited to local linear regression.

Rather than the type of the nonlinearities, it might be more important where they are located. If they are situated in regions where the density in the treated population is low, they are less relevant. To analyze the effect of nonlinearities on the relative performance of the various estimators, the regression curves are chosen such that the nonlinearities are located mostly to the right of the center, because that is the

[3] This regression curve might represent a situation where the potential outcome depends discontinuously on some confounding variables X that are itself strongly related to the propensity score. Consider, for example, a treatment whose expected potential outcome is discontinuous in age. (For example, the treatment might consist of three programs: one for 10- to 15-year-olds, one for 16- to 17-year-olds, and one for 18- to 20-year-olds, and only the program for the 16- to 17-year-olds, is beneficial.) If age is also the main determinant of the participation decision, a shape like regression curve 2 could result.

[4] Regression curves 2 and 6 are taken from Fan and Gijbels (1992).

FIGURE 2.—REGRESSION CURVES $m(p)$



region where much of the density of the treated lies but rather little of the density of the controls.

In the following sections the Monte Carlo results for these different designs are presented. The mean squared error is simulated for the various estimators for three different sample sizes: 100, 400, and 1600 observations. Pair matching is considered as the benchmark estimator, and the MSE of the other estimators is analyzed relative to that of pair matching. Because all other matching estimators depend on the choice of a bandwidth value, the Monte Carlo results consist of two parts. In section IIIB the potential efficiency gains are considered: The MSE of the matching estimators at their *optimal* bandwidth values is compared with the MSE of pair matching. This gives an indication by how much the MSE could be reduced if the optimal bandwidth were known. Because the optimal bandwidth is unknown in practice, section IIIC examines the finite-sample performance when the bandwidth is chosen by cross-validation. This gives a feasible estimator that is easy to implement.

B. *Potential Efficiency Gains*

Table 2 provides the simulated MSE for sample size 100. In the first column the MSE of pair-matching is given. The first seven rows provide the results for the first density design (first graph in figure 1), the second seven rows represent the second density design, and so on. Within each block, the first six rows correspond to the results for the six regression curves m_1 to m_6, and the seventh row shows the average result for these six curves. Examining the results for the first three density designs (where the control-treated ratio is 1 : 1), the MSE of pair matching decreases from design 1 to design 3 for all regression curves. Hence, it is smaller when the shapes of $f_{p|D=1}$ and $f_{p|D=0}$ are more similar (and the KLIC is smaller). On the other hand, the MSE becomes larger when the control-treated ratio differs from 1 : 1, particularly when the number of treated observations predominates (design 5).

The dependence of the MSE of pair matching on the density design and the regression curve, however, is not of particular concern in this paper. It is rather of interest how the choice of the best estimator depends on the density design and the regression curve. Therefore, the MSE of the other estimators is always presented relative to the MSE of pair matching (in percent), that is, a value above 100 indicates that pair matching is more precise, and a value below 100 indicates the reverse. The subsequent columns of table 2 provide the results for the different matching estimators: kernel matching, local linear matching, and ridged local linear matching, each with Epanechnikov kernel and with Gaussian kernel. These are followed by k-NN matching and Epanechnikov-weighted k-NN matching. Finally, the results for the weighting estimator are given. (In each row of Table 2, the smallest entries are underlined.) Table 2 presents the MSE for the matching estimators at their simulated *optimal* bandwidth values: For each estimator and for each Monte Carlo design, the MSE is simulated at 60 different bandwidth values, and the minimum of these simulated MSE is given.[5]

[5] For kernel and (ridged) local linear matching the bandwidth grid is $0.01\sqrt{1.2^{g-2}}$ for $g = 1, \ldots, 59$ and ∞. For k-NN matching the

TABLE 2.—MEAN SQUARED ERROR OF ESTIMATED COUNTERFACTUAL MEAN OUTCOME AT OPTIMAL BANDWIDTH VALUE, SAMPLE SIZE 100

			MSE Relative to That of Pair Matching (%)									
		1000× MSE for Pair	Kernel Matching		Local Linear Matching		Ridge Matching		k-NN Matching		Weighting Estimator	
Design	Curve	Matching	Epa	Gauss	Epa	Gauss	Epa	Gauss	—	Epa	—	Opt.
1	m_1	1.77	90.4	83.9	58.6	58.2	58.5	57.6	84.6	80.4	2555	1175
	m_2	2.86	67.6	83.8	109.6	105.9	64.1	81.4	98.7	100.8	922	520
	m_3	1.54	69.3	66.5	65.2	64.5	64.9	61.3	68.8	66.5	2528	1337
	m_4	1.81	92.9	83.9	67.6	63.0	69.4	71.3	93.4	89.7	958	633
	m_5	1.96	83.3	82.6	73.7	70.5	73.1	76.6	96.0	91.2	1033	623
	m_6	2.15	67.2	71.1	100.6	80.5	74.0	77.3	81.1	81.1	1111	671
		2.01	78.5	78.6	79.2	73.7	67.3	70.9	87.1	85.0	1518	827
2	m_1	1.30	77.7	77.0	70.0	71.8	70.5	71.8	82.4	79.8	812	654
	m_2	1.83	81.8	80.9	116.4	94.8	74.5	80.1	97.4	89.9	435	387
	m_3	1.11	66.9	67.3	63.4	63.4	63.8	62.6	73.2	70.7	885	755
	m_4	1.06	73.4	70.0	69.0	66.6	68.7	64.6	85.2	83.9	449	413
	m_5	1.07	73.5	72.4	73.3	69.6	71.3	68.9	85.2	81.6	475	422
	m_6	1.57	73.5	71.8	68.0	62.8	62.4	69.7	85.3	85.6	452	396
		1.33	74.4	73.2	76.7	71.5	68.5	69.6	84.8	81.9	585	504
3	m_1	1.19	73.7	74.0	74.0	76.0	68.0	70.8	74.6	75.4	326	312
	m_2	1.58	81.3	81.7	92.0	86.5	74.7	77.8	89.8	87.5	224	227
	m_3	0.95	66.4	65.3	64.7	64.8	62.9	62.6	72.1	69.8	371	370
	m_4	0.83	64.9	64.1	61.0	61.2	62.9	60.1	73.1	70.4	157	153
	m_5	0.84	64.5	63.5	63.6	61.3	63.0	60.2	76.8	74.6	163	161
	m_6	1.40	71.5	70.0	59.2	57.6	57.6	62.9	84.3	81.6	210	204
		1.13	70.4	69.8	69.1	67.9	64.8	65.7	78.4	76.6	242	238
4	m_1	2.08	67.3	68.3	71.1	70.7	68.3	66.3	68.2	70.5	198	190
	m_2	3.03	58.8	57.7	59.9	57.3	57.1	56.7	78.6	68.5	106	108
	m_3	1.75	56.3	57.8	46.1	45.7	43.8	43.9	64.6	63.9	226	224
	m_4	1.56	53.3	52.6	36.1	34.7	35.8	33.6	61.0	59.9	126	125
	m_5	1.57	51.1	51.7	35.1	35.8	31.5	34.5	63.5	59.1	131	132
	m_6	2.85	42.6	41.8	28.9	28.8	29.3	27.9	69.0	63.7	107	105
		2.14	54.9	55.0	46.2	45.5	44.3	43.8	67.5	64.3	149	147
5	m_1	2.23	105.8	88.7	61.2	61.6	65.3	64.3	97.5	91.9	2243	804
	m_2	4.54	78.8	77.0	81.5	82.3	75.3	77.6	96.4	84.7	667	336
	m_3	1.88	91.8	81.7	93.7	120.2	80.5	82.9	97.8	88.6	2156	952
	m_4	2.29	88.7	87.3	76.4	83.0	76.9	78.1	95.4	95.4	771	424
	m_5	2.49	86.5	87.8	78.8	84.6	76.1	78.9	96.6	92.9	755	392
	m_6	3.07	83.0	89.1	150.1	183.9	75.6	92.4	95.3	92.2	920	441
		2.75	89.1	85.3	90.3	102.6	75.0	79.0	96.5	91.0	1252	558

Note: The first two columns indicate the design of the propensity score densities and the regression curve, respectively. MSE of pair matching is multiplied by 1000. MSE of all other estimators is given relative to MSE of pair matching. In each row, the estimator with the minimum MSE and all estimators with MSE not larger than 0.5 above the minimum are underlined. Epa indicates Epanechnikov kernel; Gauss indicates Gaussian kernel. The MSE of the matching estimators is given at the simulated optimal bandwidth value, which is simulated over a grid of 60 different values (30 values for k-NN matching). Simulations are based on 10,000 replications.

Inspecting these relative MSE across the different designs, it is seen that the relative performance of all matching estimators depends strongly on the density design and, in particular, on the control-treated ratio: Whereas the relative MSE of the matching estimators decreases somewhat when the propensity score densities become more similar (from design 1 to 3), the MSE is much lower when the control-treated ratio is high (design 4) and higher when the ratio of controls to treated is low (design 5). In design 4, the number of control observations is approximately 80, and the number of treated is around 20. Pair matching performs very poorly in this situation (the MSE of the other matching estimators is often 50% lower), because it uses the observations on Y for the 20 matched control observations only, whereas the other matching estimators use all control observations. In design 5, on the other hand, the control-treated ratio is reversed, and pair matching uses the 20 controls multiple times for matching to the 80 treated observations. Consequently, the loss of information due to pair matching is much smaller, and the MSE of the other matching estimators is only 10% to 25% lower than for pair matching.

Whereas the design of the propensity score densities showed a clear effect on the relative precision of alternative matching estimators, the shape of the regression curve seems to be less important. On average, the relative MSE tends to be somewhat larger for regression curve 2 (with the

bandwidth grid contains only 30 values: 1, . . . , 21, 25, 29, . . . , 53, ∞. For bandwidth ∞ no local smoothing takes place and local linear regression corresponds to OLS. For larger sample sizes the bandwidth grid for k-NN matching is adopted to include larger numbers of neighbors.

conspicuous local peak) and slightly smaller for regression curve 3 (with mild nonlinearities). Nevertheless, the alternative estimators perform better than pair matching in all designs. In general, ridge matching performs best, followed by local linear matching. Ridge matching is in most designs the best estimator, and in those designs where it is not, it is usually not much worse than the best estimator. In addition, ridge matching is in all designs better than pair matching. Local linear matching also performs well in many situations, but can be worse than pair matching in other situations. Although it sometimes performs a little better than ridge matching, it can be much worse in other designs. In general, local linear matching seems to be more sensitive to the density design and the regression curve, performing well in some situations (for example, when the regression curve is linear) but poorly in others (for example, for regression curve 2 with the local peak). Kernel matching, on the other hand, is less sensitive to the simulation design, but performs worse than ridge matching in almost all designs. The choice of the kernel function, for these three matching estimators, is of lesser importance. For ridge matching, the Epanechnikov kernel is slightly preferable to the Gaussian kernel. For kernel matching it is the Gaussian kernel.

The *k*-NN matching estimator is in many designs the worst of all matching estimators.[6] In all designs it has a larger MSE than ridge matching. *k*-NN matching with Epanechnikov weights performs somewhat better, but is still always worse than ridge matching. Finally, in the last two columns the relative MSE of the weighting estimator is given. The MSE of the simple weighting estimator is always larger than the MSE of pair matching and can be up to 25 times larger. The MSE of the capped weighting estimator with optimal capping rule is given in the last column. The optimal cap is implemented by simulating the MSE for 60 different caps ($\bar{c} = 1, \ldots, 60$) and choosing the minimum of these. However, even with this optimal cap, the capped weighting estimator is clearly worse than pair-matching in all designs.

Table A2 shows the simulation results for sample size 400. The results are largely similar to those for sample size 100, with all matching estimators becoming somewhat more precise relative to pair-matching. Particularly the *k*-NN matching estimators improve their relative position. Nevertheless, the unweighted *k*-NN matching estimator is still the worst among all matching estimators in many of the designs; and ridge matching is still the best estimator in general. The relative MSE of the weighting estimator worsens with increasing sample size, and this trend is continued for sample size 1600 (not shown in the tables).

In total, these simulation results reveal that the potential precision gains can be substantial. On average, ridge matching is approximately 35% more precise than pair matching in designs 1 to 3, 55% in design 4, and 25% in design 5 (for sample sizes 100 and 400).

C. Cross-validation Bandwidth Choice

The previous section examined the MSE at the optimal bandwidth value. Those results indicated the *potential* for precision gains, that is, the maximum reductions in MSE that could be achieved by the different estimators. In this section, *feasible* precision gains are analyzed, when the bandwidth is chosen by cross-validation. Cross-validation is a widely used bandwidth selector for nonparametric regression, and although it will not lead to asymptotically optimal bandwidth choices for the matching estimator, it is worthwhile to examine its usefulness in this setting. Cross-validation chooses the bandwidth as

$$h^{CV} = \arg\min_{h} \sum_{i:D_i=0} [Y_i - \hat{m}_{-i}(p_i)]^2,$$

where $\hat{m}_{-i}(\rho)$ is the estimate with observation *i* removed from the sample.[7]

Table 3 shows the relative MSE for sample size 100, when the bandwidth is chosen by cross-validation. (The structure of the table is identical to that of table 2.) In addition, the results for *trimmed local linear matching* are shown.[8] In Table 3 it is striking how sensitive the performance of many matching estimators is to the distribution of the propensity score. The relative MSE of kernel, local linear, and *k*-NN matching decreases substantially (and by much more than would be expected from table 2) from design 1 to design 3. The MSE of kernel matching is reduced from approximately 95% in design 1 (where the controls and treated differ sharply in their propensity scores) to approximately 75% in design 3 (where the propensity score densities are very similar). The relative improvement is even more drastic for local linear matching and *k*-NN matching, which both perform very poorly in designs 1 and 2. The particularly poor performance of local linear matching is ameliorated a little when trimming is introduced. Trimming discards treated observations from the sample if they are located in regions where there are few control observations. Table 3 gives the MSE of the trimmed local linear matching estimator (for Epanechnikov and for Gaussian kernel) for three different trimming rules: The 2%, 5%, or 10%, respectively, of the treated observations with the smallest values of $f_{p|D=0}(p_i)$ are deleted.[9] The trimmed local

[6] It cannot be worse than pair matching, for pair-matching is equal to *k*-NN matching with $k = 1$.

[7] For kernel and (ridged) local linear matching the bandwidth search grid is $0.01 \times 1.2^{g-1}$ for $g = 1, \ldots, 29$ and ∞. For *k*-NN matching the grid is $1, \ldots, 21, 25, 29, \ldots, 53, \infty$.

[8] In the Monte Carlo simulations, trimming was implemented for all the other matching estimators as well. Here the trimming results are shown only for local linear matching, because for all other estimators trimming always led to an increase in MSE. The other results are available from the author.

[9] In the simulations, trimming has been carried out on the basis of the true density function $f_{p|D=0}$ and on the basis of the estimated density function $\hat{f}_{p|D=0}$. The results were similar and often slightly better for

linear matching estimator has a significantly lower MSE in designs 1 and 2 than the nontrimmed estimator. However, its MSE is still very often higher than for pair matching. In addition, it is not clear how the trimming level should be chosen in practice. Whereas a trimming level of 10% would often be the best choice in design 1 and design 5, a trimming level of 2% or 5% would be preferable in designs 2 and 3. In design 4, a trimming level of 10% would lead to substantially worse results than no trimming at all. Ridge matching, in contrast, is notably less sensitive to design choice. Its MSE decreases only slightly, from approximately 80% in design 1 to approximately 75% in design 3. In design 4 it is approximately 70%, and in design 5 approximately 90%.

In design 4 (where the number of controls greatly exceeds the number of treated), all estimators perform well and their relative MSE are very similar: approximately 70%. In such a situation, trimming seems not to be useful for local linear matching (particularly in larger samples; see tables A3 and A4). In design 5, on the other hand, where the control observations are scarce, only ridge matching consistently dominates pair matching. All other estimators are usually worse than pair matching. Ridge matching with Epanechnikov kernel is, in fact, the only estimator that is never worse than pair matching in any of the designs. In addition, ridge matching is often the best estimator in designs 1, 2, and 5, and where it is not, it is usually not much worse. Indeed, this holds not only for the more demanding designs 1, 2, and 5, but also for the more favorable designs 3 and 4.

Whereas the density design strongly influences the relative performance of all matching estimators, the pattern is less clear-cut with respect to the shape of the regression curve $m(p)$. For regression curve m_2, ridge matching is the best estimator in all density designs. Nevertheless, the relative MSE of ridge matching is, in general, only very little affected by the shape of the regression curve. The (nontrimmed) local linear matching estimators perform, as expected, relatively better for the straight regression line (m_1), whereas the k-NN matching estimators perform relatively better for regression curves m_3 and m_6. These patterns, however, are weak and much less pronounced than the dependence on the density design. The differences with respect to the choice of the kernel function are also not very conclusive. In general, the Epanechnikov kernel is somewhat better suited for ridge matching, and the Gaussian works slightly better for kernel matching.

Tables A3 and A4 provide the simulation results for sample size 400 and 1600, respectively. With increasing sample size all estimators, except k-NN matching, become more precise in comparison with pair matching. The relative MSE of k-NN matching decreases for designs 2 and 3, but increases substantially for some regression curves in designs 1 and 5. Local linear matching (without trimming) improves significantly with larger sample size, and with 1600 observations it comes close to ridge matching and kernel matching in designs 2, 3, and 4. In the more demanding designs 1 and 5, however, it is still often substantially worse than pair matching. Trimming improves the local linear matching estimator in designs 1 and 5, but not enough to dominate the pair-matching estimator. In addition, the optimal trimming level needs to be found in practice, as the appropriate trimming depends on the sample size. Whereas 5% trimming is best in designs 1 and 5 for sample size 400, it would lead to worse results than no trimming for sample size 1600. With 1600 observations, 2% trimming would be appropriate for designs 1 and 5, and no trimming for designs 2, 3, and 4. But even with appropriate trimming, local linear matching is still usually worse than ridge matching.

The relative MSE of kernel matching decreases when the sample size is increased from 100 to 400. A further increase to sample size 1600 affects its MSE only a little, except for design 5. Kernel matching is the best estimator for various regression curves in designs 2, 3, and 4, and it has the lowest MSE in designs 3 and 4 with respect to the average over all six regression curves. On the other hand, kernel matching can be worse than pair matching in designs 1 and 5, even with 1600 observations.

Finally, ridge matching is still the most appealing of all estimators. Although its merits are less compelling in larger samples than they were with 100 observations, it is still most often the estimator with the lowest MSE. Even when it is worse than kernel matching, the difference is usually not large. In addition, ridge matching is least sensitive to the design choice and the shape of the regression curve $m(p)$. As a rough measure of its robustness to the simulation design, the standard deviation of the relative MSE for the 30 different simulation designs (5 density designs times 6 regression curves) has been computed from Table A3. This standard deviation is only 6.1 for ridge matching with Epanechnikov kernel, whereas it is 15.1 for kernel matching with Gaussian kernel and far higher for all other estimators. Also, the range between the smallest and the largest MSE is much smaller for ridge matching. Hence, the performance of ridge matching relative to pair matching is much more stable than it is for the other estimators. In addition, it is the only estimator whose MSE never exceeded the MSE of pair matching in any simulation. In total, ridge matching is approximately 25% more precise than pair matching. When the number of control observations is much larger than the number of treated (design 4), the MSE of ridge matching is approximately 30% lower than the MSE of pair matching. In the reverse situation, when the number of treated predominates (design 5), the precision gains are smaller and depend on the sample size. If the sample is small (approximately 80 treated and 20 control observations), the precision gains are around 10%, and they increase to approximately 30% for sample size 1600 (with approximately 320

$\hat{f}_{p|D=0}$. Table 3 reports only the results based on the estimated density function.

TABLE 3.—MEAN SQUARED ERROR OF ESTIMATED COUNTERFACTUAL MEAN OUTCOME (RELATIVE TO MSE OF PAIR MATCHING): BANDWIDTH VALUE CHOSEN BY CROSS VALIDATION, SAMPLE SIZE 100

		MSE Relative to That of Pair Matching (%)													
		Kernel Matching		Local Linear Matching with Epanechnikov Kernel				Local Linear Matching with Gauss Kernel				Ridge Matching		*k*-NN Matching	
Design	Curve	Epa	Gauss	—	Trim 2%	Trim 5%	Trim 10%	—	Trim 2%	Trim 5%	Trim 10%	Epa	Gauss	—	Epa
1	m_1	104.8	96.5	105.8	104.1	107.9	126.0	109.2	103.8	105.6	121.5	76.2	71.4	162.8	146.1
	m_2	73.5	92.6	168.2	158.0	146.6	129.8	258.8	222.1	188.8	150.0	70.1	91.9	137.3	119.1
	m_3	78.3	74.9	212.0	191.3	171.6	150.1	276.1	234.1	199.5	164.5	78.4	77.3	106.5	99.3
	m_4	119.4	118.1	107.5	95.9	90.4	94.7	129.6	109.5	97.9	96.3	84.6	83.6	211.5	196.7
	m_5	112.2	111.3	133.6	116.2	103.1	95.6	180.4	145.8	119.9	102.5	87.9	89.4	173.9	161.6
	m_6	82.4	83.7	197.0	185.5	173.2	155.6	355.3	306.7	263.2	211.6	81.8	86.2	93.0	84.3
		95.1	96.2	154.0	141.8	132.1	125.3	218.2	187.0	162.5	141.1	79.8	83.3	147.5	134.5
2	m_1	84.9	83.4	87.7	89.5	101.2	139.1	83.8	88.1	102.3	144.4	77.2	76.1	106.0	99.9
	m_2	86.3	87.7	141.8	132.1	125.8	123.0	176.6	156.3	142.9	135.2	83.5	86.6	113.6	102.8
	m_3	72.5	72.1	120.4	107.8	102.1	108.3	109.1	97.0	93.5	103.1	73.2	73.3	84.2	82.7
	m_4	87.5	86.3	96.7	78.9	75.4	93.3	100.9	79.6	72.6	85.1	72.9	71.9	138.3	131.3
	m_5	88.5	87.4	122.5	98.8	87.0	92.3	124.5	96.4	81.7	84.8	77.2	78.0	125.7	116.2
	m_6	79.7	81.6	123.5	112.3	105.0	103.2	149.8	128.6	115.4	109.1	79.8	81.7	87.7	84.9
		83.2	83.1	115.4	103.2	99.4	109.9	124.1	107.7	101.4	110.3	77.3	77.9	109.3	103.0
3	m_1	75.5	75.1	80.7	85.2	99.0	137.6	77.8	84.2	100.0	140.9	77.7	75.6	80.7	78.2
	m_2	84.6	84.4	112.7	106.1	105.2	112.0	134.1	125.2	122.6	130.3	82.5	83.6	95.5	89.7
	m_3	70.0	69.7	82.5	78.1	81.4	99.8	73.9	73.7	79.3	100.6	70.2	70.7	74.5	73.7
	m_4	70.9	70.0	93.5	72.3	68.1	86.1	102.2	76.0	65.5	73.5	70.3	69.2	88.6	86.5
	m_5	70.9	69.2	102.5	79.9	71.0	80.7	103.4	78.5	67.4	71.4	74.1	72.8	87.1	81.2
	m_6	79.9	79.2	93.8	86.6	86.6	97.5	94.5	87.1	87.2	98.5	76.5	78.4	85.9	84.2
		75.3	74.6	94.3	84.7	85.2	102.3	97.7	87.4	87.0	102.5	75.2	75.1	85.4	82.2
4	m_1	70.0	70.1	73.1	83.8	90.3	114.9	72.4	83.2	89.5	114.3	72.9	71.9	70.7	71.7
	m_2	77.8	77.5	79.7	81.4	83.0	90.8	82.7	81.8	82.7	89.3	76.6	76.4	80.6	81.0
	m_3	63.8	63.6	65.5	70.6	73.9	88.6	62.7	69.2	72.9	88.9	62.4	62.8	67.0	64.7
	m_4	58.7	57.0	75.1	68.0	71.1	88.3	72.3	63.0	65.0	79.4	65.7	63.0	64.4	61.4
	m_5	60.4	60.0	72.4	67.6	70.5	87.9	70.5	62.1	63.4	75.3	62.3	63.0	64.9	61.8
	m_6	75.6	76.6	74.5	77.1	79.3	90.0	71.5	72.9	74.9	85.7	75.8	75.5	80.0	79.7
		67.7	67.5	73.4	74.7	78.0	93.4	72.0	72.0	74.7	88.8	69.3	68.8	71.3	70.0
5	m_1	126.3	99.1	137.7	133.2	133.9	151.0	90.9	92.1	100.8	129.6	91.5	82.4	148.3	134.4
	m_2	87.0	98.2	118.5	111.9	105.6	101.7	86.8	84.3	83.6	88.9	75.1	93.1	127.4	108.9
	m_3	116.8	102.1	223.7	203.9	184.1	164.2	149.4	144.5	144.9	155.0	97.9	90.7	161.1	142.0
	m_4	106.3	108.3	149.8	128.4	109.3	92.3	114.9	102.6	92.4	83.2	93.6	95.4	142.8	135.5
	m_5	99.4	102.2	156.8	134.6	114.1	94.2	121.1	107.6	95.9	84.3	89.8	93.7	126.3	120.9
	m_6	91.4	94.9	229.1	213.8	195.1	168.5	258.9	240.1	221.4	200.3	94.1	101.7	143.4	121.4
		104.5	100.8	169.3	154.3	140.3	128.7	137.0	128.5	123.2	123.5	90.3	92.8	141.6	127.2

Note: In each row, the estimator with the minimum MSE and all estimators with MSE not larger than 0.5 above the minimum are underlined. Epa indicates Epanechnikov kernel; Gauss indicates Gaussian kernel. Trim 2% means that estimator is trimmed by deleting the 2% of the observations with lowest density. The bandwidth is chosen by leave-one-out cross-validation of the nonparametric regression estimator, over a grid of 30 different values. Simulations are based on 10,000 replications.

non-treated observations). With respect to the shape of the regression curve, no consistent pattern can be detected.

IV. Conclusions

In this paper the finite-sample properties of various propensity-score matching estimators of the counterfactual mean (which is the central ingredient in the computation of average treatment effects) have been analyzed. First, the potential efficiency gains of alternative matching estimators vis-à-vis pair matching were assessed, that is, it was examined by how much the MSE could be reduced if the optimal bandwidth was known. In general, ridge matching demonstrated the largest potential. In the designs where the numbers of treated and control observations were approximately equal, its MSE was approximately 35% lower. It was approximately 55% lower when the control observations exceeded the number of treated by 4 : 1, and approximately 25% lower when the ratio was 1 : 4. Almost regardless of the Monte Carlo design, ridge matching was usually the best estimator, particularly in small samples.

Though these simulations indicated the potential for precision gains, in practice a data-driven bandwidth choice is necessary to select the bandwidth value. A handy and easy-to-implement bandwidth selector is cross validation of the nonparametric regression estimator. Although cross-validation will not lead to asymptotically optimal bandwidth choices, it seems to work well in practice for some estimators, at least for the sample sizes considered. In the rather

facile simulation designs (similar characteristics of treated and control populations, or a large control-to-treated ratio), kernel matching was often the best estimator, immediately followed by ridge matching. In more demanding situations (small sample size, large differences in the characteristics of treated and control populations, or a small number of control observations), however, kernel matching performed usually worse than ridge matching and sometimes even worse than pair-matching. Ridge matching with Epanechnikov kernel was often the estimator with smallest MSE, particularly in the more demanding designs. In addition, its relative MSE was least sensitive to the simulation design, and it was always lower than for pair matching. Hence, for practical guidance on estimator choice, ridge matching seems to be a good choice in most situations. (If the control-treated ratio is large, kernel matching is also an option.) The precision gains of ridge matching did not depend much on the shape of the regression curve and also not much on the shapes of the propensity score densities, but were affected somewhat by the control-to-treated ratio. For approximately equal numbers of controls and treated, ridge matching was approximately 25% more precise than pair matching. For a control-treated ratio of 4 : 1, the reductions in MSE were approximately 30%, but they were approximately 10–30% (depending on sample size) for a ratio of 1 : 4. (To put these figures in perspective, a reduction in MSE of 25% means that pair matching would require approximately 33% more observations to achieve the same precision.) Although these reductions in MSE are substantial, they do not reach the potential precision gains. Particularly in the designs where the control observations exceeded the treated by 4 : 1, the potential for further improvement through a better bandwidth choice seems greatest. In all other designs, the differences between the potential and the feasible precision gains are not very large and the development of a superior bandwidth selector might be difficult.

The performance of local linear matching with cross-validation bandwidth selection, on the other hand, was deceptive. Although trimming improved its MSE, it was still clearly worse than ridge matching in most designs. Hence, trimming seems not to be the best response to the variance problems of the local linear regression estimator. In addition, it might be difficult in practice to determine the optimal trimming level. *k*-NN matching was also not very successful. Finally, the weighting estimator turned out be worst of all. Even with an optimal capping rule, it is far worse than pair matching in all of the designs.

REFERENCES

Abadie, A., and G. Imbens, "Simple and Bias-Corrected Matching Estimators for Average Treatment Effects," Harvard University mimeograph (2001).

Fan, J., "Design-adaptive Nonparametric Regression," *Journal of the American Statistical Association* 87 (1992), 998–1004.

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

Fan, J., T. Gasser, I. Gijbels, M. Brockmann, and J. Engel, "Local Polynomial Regression: Optimal Kernels and Asymptotic Minimax Efficiency," *Annals of the Institute of Mathematical Statistics* 49 (1997), 79–99.

Fan, J., and I. Gijbels, "Variable Bandwidth and Local Linear Regression Smoothers," *Annals of Statistics* 20 (1992), 2008–2036.

Hahn, J., "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (1998), 315–331.

Hastie, T., and C. Loader, "Local Regression. Automatic Kernel Carpentry," *Statistical Science* 8 (1992), 120–143.

Heckman, J., H. Ichimura, and P. Todd, "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies* 64 (1997), 605–654.

——— "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies* 65 (1998), 261–294.

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

Ichimura, H., and O. Linton, "Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators," University College London mimeograph (2002).

Imbens, G., "The Role of the Propensity Score in Estimating Dose-Response Functions," *Biometrika* 87 (2000), 706–710.

Johnson, N., "Systems of Frequency Curves Generated by Methods of Translation," *Biometrika* 36 (1949), 149–176.

Johnson, N., S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, vol. 1, 2nd ed. (New York: Wiley, 1994).

Kullback, J., *Information Theory and Statistics* (New York: Wiley, 1959).

Kullback, J., and R. Leibler, "On Information and Sufficiency," *Annals of Mathematical Statistics* 22 (1951), 79–86.

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

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

——— "Data Adaptive Ridging in Local Polynomial Regression," *Journal of Computational and Graphical Statistics* 9 (2000), 338–360.

APPENDIX

1. Generation of the Propensity-Score Distributions

To have a simple but versatile design for the generation of different distributions of the propensity score ($f_{p|D=1}$ and $f_{p|D=0}$), the propensity score is specified as $p(x) = \alpha + \beta x$, where X is one-dimensional and distributed symmetrically in (0, 1). The parameters α and β are chosen such that $0 < p(x) < 1$. The propensity-score values thus range from α to $\alpha + \beta$. Denote the expected value of the propensity score by P_1; it is given by $P_1 = E[p(X)] = \alpha + \beta/2$. Let P_0 denote the size of the nontreated population: $P_0 = 1 - P_1$. For this specification, the ensuing densities $f_{p|D=1}$ and $f_{p|D=0}$ are thus

$$f_{p|D=1}(\rho) = \frac{\rho}{\beta P_1} \cdot f_X\left(\frac{\rho - \alpha}{\beta}\right),$$

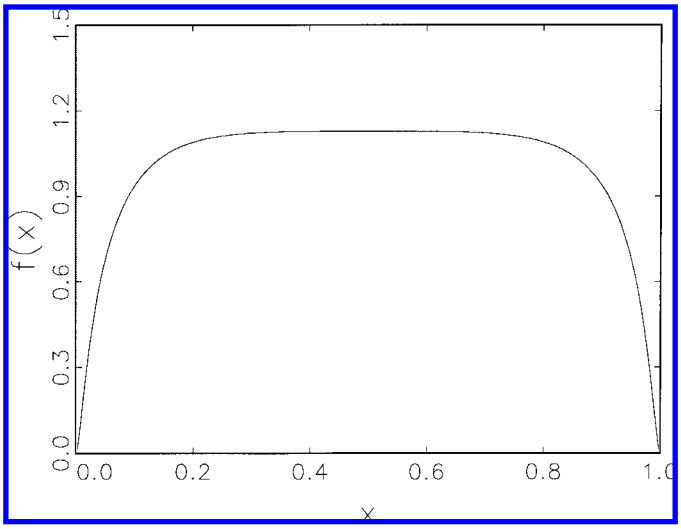
$$f_{p|D=0}(\rho) = \frac{1 - \rho}{\beta P_0} \cdot f_X\left(\frac{\rho - \alpha}{\beta}\right) \quad \text{for} \quad \alpha < \rho < \alpha + \beta,$$

and accordingly the ratio of the two densities is given by

$$\frac{f_{p|D=1}(\rho)}{f_{p|D=0}(\rho)} = \frac{\rho}{1 - \rho}\frac{P_0}{P_1}.$$

Hence in a design where the treated and nontreated populations are of equal size ($P_1 = P_0$), the ratio $f_{p|D=1}/f_{p|D=0}$ is 1 at $\rho = 0.5$ and close to 1 in its neighborhood. To be able to generate large differences between $f_{p|D=1}$ and $f_{p|D=0}$ (and thus a large value of the KLIC, as in the first graph

FIGURE A1.—DENSITY OF f_X



2. The Regression Curves

TABLE A1.—REGRESSION CURVES $m(p)$

$m_1(\rho) = 0.15 + 0.7\rho$
$m_2(\rho) = 0.1 + \frac{\rho}{2} + \frac{1}{2}\exp[-200(\rho - 0.7)^2]$
$m_3(\rho) = 0.8 - 2(\rho - 0.9)^2 - 5(\rho - 0.7)^3 - 10(\rho - 0.6)^{10}$
$m_4(\rho) = 0.2 + \sqrt{1 - \rho} - 0.6(0.9 - \rho)^2$
$m_5(\rho) = 0.2 + \sqrt{1 - \rho} - 0.6(0.9 - \rho)^2 - 0.1\rho\cos(30\rho)$
$m_6(\rho) = 0.4 + 0.25\sin(8\rho - 5) + 0.4\exp[-16(4\rho - 2.5)^2]$

of figure 1) requires that the distribution of X have substantial probability mass in its tails. In addition, to ensure that the support of X is identical in the treated and the nontreated population, $\lim_{x \downarrow 0} f_X(x)$ and $\lim_{x \uparrow 1} f_X(x)$ should be both 0. The Johnson S_B family provides a convenient distribution satisfying these properties, which is depicted in figure A1. Its density is

$$f_X(x) = \frac{1}{2\sqrt{\pi}x(1-x)} \exp\left[-\frac{1}{4}\ln^2\left(\frac{x}{1-x}\right)\right], \quad 0 < x < 1;$$

see Johnson, Kotz, and Balakrishnan (1994, p. 37, $\gamma = 0$, $\delta = \sqrt{0.5}$) or Johnson (1949).

TABLE A2.—MEAN SQUARED ERROR OF ESTIMATED COUNTERFACTUAL MEAN OUTCOME AT OPTIMAL BANDWIDTH VALUE, SAMPLE SIZE 400

			MSE Relative to That of Pair Matching (%)									
		1000× MSE for Pair Matching	Kernel Matching		Local Linear Matching		Ridge Matching		k-NN Matching		Weighting Estimator	
Design	Curve		Epa	Gauss	Epa	Gauss	Epa	Gauss	—	Epa	—	Opt.
1	m_1	0.46	77.1	73.4	55.1	54.7	56.2	59.2	75.6	74.6	2422	1455
	m_2	0.56	76.9	76.9	125.2	104.6	75.2	77.2	78.6	74.8	1192	830
	m_3	0.44	58.2	58.2	53.8	52.6	56.0	53.4	58.7	52.2	2225	1509
	m_4	0.44	89.2	77.7	60.1	59.4	58.1	60.8	89.3	83.0	972	798
	m_5	0.43	75.6	74.0	73.0	72.6	72.2	64.2	73.7	71.6	1160	863
	m_6	0.56	59.8	59.0	79.5	79.2	73.3	73.9	76.1	72.3	1089	806
		0.48	72.8	69.9	74.5	70.5	65.2	64.8	75.3	71.4	1510	1044
2	m_1	0.34	71.3	74.1	73.5	70.2	66.4	71.5	71.5	72.2	786	735
	m_2	0.42	80.6	79.6	75.3	76.8	69.7	74.0	83.2	76.5	467	459
	m_3	0.28	64.7	64.0	63.5	60.7	61.2	58.0	67.1	64.3	878	848
	m_4	0.26	68.3	69.4	68.1	63.8	68.6	63.8	75.5	71.6	459	451
	m_5	0.25	68.7	69.9	71.4	66.3	70.5	64.8	72.0	68.4	483	465
	m_6	0.38	72.4	68.3	63.2	66.0	59.5	70.6	82.7	80.7	461	440
		0.32	71.0	70.9	69.2	67.3	66.0	67.1	75.3	72.3	589	566
3	m_1	0.30	72.7	72.5	70.7	70.2	74.1	71.5	77.2	74.0	324	319
	m_2	0.39	78.6	79.3	71.1	73.7	75.3	76.5	80.7	81.1	225	223
	m_3	0.24	67.0	65.2	60.6	60.1	60.0	59.5	67.0	65.5	365	355
	m_4	0.20	63.0	62.8	58.8	60.6	57.0	59.0	68.8	63.7	154	154
	m_5	0.21	63.2	64.3	60.2	60.8	60.7	57.5	66.1	65.0	161	157
	m_6	0.35	70.0	66.2	59.8	60.7	59.3	64.9	79.1	80.2	210	205
		0.28	69.1	68.4	63.5	64.4	64.4	64.8	73.1	71.6	240	236
4	m_1	0.50	70.4	69.6	71.8	70.0	72.7	68.8	71.1	69.9	198	199
	m_2	0.74	61.0	64.2	58.1	59.1	60.4	60.8	76.7	69.5	110	109
	m_3	0.43	59.2	60.4	46.7	45.9	47.0	50.6	63.8	61.8	228	233
	m_4	0.40	57.4	57.7	33.8	34.1	32.0	34.6	60.4	61.2	128	126
	m_5	0.39	56.8	53.5	33.3	33.7	32.5	33.5	62.2	59.1	131	134
	m_6	0.67	46.4	50.5	27.9	28.4	29.1	27.2	69.7	64.9	109	109
		0.52	58.5	59.3	45.3	45.2	45.6	45.9	67.3	64.4	151	152
5	m_1	0.52	87.5	79.1	59.6	58.0	58.5	59.5	84.4	83.7	2269	1209
	m_2	0.64	77.7	89.2	159.9	130.1	83.9	89.9	98.1	99.7	1192	758
	m_3	0.47	67.4	68.2	68.3	68.8	68.8	65.4	68.2	67.9	2108	1287
	m_4	0.53	92.3	83.5	67.5	68.0	79.1	71.6	91.3	97.5	786	565
	m_5	0.48	96.9	82.2	84.6	86.0	85.7	72.8	94.5	95.8	927	617
	m_6	0.57	75.2	77.6	111.6	93.5	76.5	89.9	79.7	80.5	1157	782
		0.54	82.9	80.0	91.9	84.1	75.4	74.9	86.0	87.5	1407	870

Note: The first two columns indicate the design of the propensity-score densities and the regression curve, respectively. MSE of pair matching is multiplied by 1000. MSE of all other estimators given relative to MSE of pair-matching. In each row, the estimator with the minimum MSE and all estimators with MSE not larger than 0.5 above the minimum are underlined. Epa indicates Epanechnikov kernel; Gauss indicates Gaussian kernel. The MSE of the matching estimators is given at the simulated optimal bandwidth value, which is simulated over a grid of 60 different values (30 values for k-NN matching). Simulations are based on 5000 replications.

TABLE A3.—MEAN SQUARED ERROR OF ESTIMATED COUNTERFACTUAL MEAN OUTCOME (RELATIVE TO MSE OF PAIR MATCHING): BANDWIDTH VALUE CHOSEN BY CROSS-VALIDATION, SAMPLE SIZE 400

		MSE Relative to That of Pair Matching (%)													
		Kernel Matching		Local Linear Matching with Epanechnikov Kernel				Local Linear Matching with Gauss Kernel				Ridge Matching		*k*-NN Matching	
Design	Curve	Epa	Gauss	—	Trim 2%	Trim 5%	Trim 10%	—	Trim 2%	Trim 5%	Trim 10%	Epa	Gauss	—	Epa
1	m_1	92.7	90.3	74.8	76.2	98.1	186.2	76.3	76.7	99.0	189.1	66.0	64.5	196.9	192.1
	m_2	78.3	78.3	136.5	122.9	106.3	93.1	279.1	210.1	151.7	107.4	78.4	80.3	114.7	105.9
	m_3	63.5	64.1	130.5	109.6	94.2	100.4	165.9	125.6	98.0	97.4	74.0	72.4	81.8	73.3
	m_4	119.6	114.0	91.7	76.5	88.3	169.0	90.4	75.5	90.5	179.1	71.3	67.5	327.0	301.5
	m_5	83.9	76.3	109.1	97.7	101.2	161.0	209.1	143.0	113.9	162.7	80.3	75.7	220.7	194.1
	m_6	72.3	78.1	122.0	110.1	94.9	83.3	255.7	189.6	134.8	97.6	73.9	81.7	81.9	77.6
		85.1	83.5	110.8	98.8	97.2	132.2	179.4	136.8	114.6	138.4	74.0	73.7	170.5	157.4
2	m_1	80.7	83.8	75.4	85.8	135.1	304.6	74.0	84.6	132.2	295.7	72.7	72.4	106.6	106.2
	m_2	79.0	79.0	90.0	86.1	89.7	108.5	105.5	93.6	95.8	115.9	78.9	79.2	84.5	83.7
	m_3	66.3	65.4	77.5	75.8	87.3	148.4	79.4	74.2	83.2	139.8	69.3	67.9	76.5	73.7
	m_4	81.4	83.8	86.7	67.1	99.9	259.8	82.0	66.2	99.8	253.1	75.6	72.6	140.0	131.5
	m_5	67.7	67.0	78.6	77.0	114.8	260.5	90.3	75.0	105.9	243.5	70.0	68.6	94.4	88.7
	m_6	77.0	76.7	87.8	82.6	83.3	100.3	103.4	87.0	84.5	102.5	77.1	78.6	82.0	80.5
		75.3	76.0	82.7	79.0	101.7	197.0	89.1	80.1	100.2	191.8	73.9	73.2	97.4	94.0
3	m_1	74.6	75.0	73.7	87.3	141.4	300.0	75.0	90.1	146.1	308.1	75.5	73.4	79.7	79.6
	m_2	79.9	79.7	82.6	85.8	97.9	123.7	86.6	87.7	100.0	126.5	78.9	79.8	81.8	80.4
	m_3	67.5	66.1	67.5	71.3	90.5	166.2	67.5	71.8	92.0	169.9	68.7	66.6	68.7	68.3
	m_4	66.5	65.2	85.5	66.6	109.5	289.5	83.1	63.7	105.3	281.8	79.5	77.4	80.6	75.4
	m_5	63.6	63.5	70.6	73.0	124.5	303.4	76.8	66.6	107.0	271.4	66.7	64.5	67.9	66.6
	m_6	76.0	75.6	79.1	77.8	83.9	114.6	83.2	79.3	86.4	121.9	76.4	74.3	80.4	77.7
		71.3	70.8	76.5	76.9	108.0	216.2	78.7	76.6	106.1	213.2	74.3	72.7	76.5	74.7
4	m_1	70.0	71.0	71.6	80.3	107.9	190.5	69.9	78.8	107.0	191.7	70.9	72.9	73.5	71.3
	m_2	78.0	78.6	76.9	79.8	85.4	96.2	75.6	78.6	83.9	94.7	77.6	78.7	78.7	77.1
	m_3	63.7	63.9	66.2	70.1	79.5	111.1	63.3	67.7	78.0	111.8	63.4	65.9	66.3	65.4
	m_4	58.9	58.8	75.5	68.0	96.8	206.7	68.9	64.8	94.9	205.4	68.5	67.4	64.9	62.6
	m_5	58.9	59.2	62.1	69.0	100.4	194.1	64.0	66.7	95.4	188.1	60.9	59.4	61.0	59.9
	m_6	76.1	75.4	77.9	80.1	85.2	99.4	75.7	77.5	82.5	96.3	77.4	75.5	79.8	79.4
		67.6	67.8	71.7	74.5	92.5	149.7	69.6	72.3	90.3	148.0	69.8	70.0	70.7	69.33
5	m_1	103.7	93.0	96.4	97.9	127.1	247.7	90.3	92.8	124.2	247.5	75.3	71.5	178.5	169.5
	m_2	85.0	91.1	281.2	255.7	219.9	182.7	385.0	315.8	249.5	197.6	82.3	89.0	166.9	129.8
	m_3	74.9	71.0	208.6	176.4	145.3	146.3	217.7	172.1	138.5	144.2	81.1	79.8	101.6	101.0
	m_4	120.2	121.9	108.4	81.4	74.9	119.3	120.4	85.4	75.0	118.4	78.0	73.7	249.3	255.1
	m_5	119.5	104.9	171.5	128.8	95.4	113.3	219.5	146.3	98.8	114.7	89.5	86.0	221.7	199.8
	m_6	80.3	83.3	226.5	200.6	166.1	134.5	404.2	317.9	238.2	176.1	78.4	84.9	89.0	82.8
		97.3	94.2	182.1	156.8	138.1	157.3	239.5	188.4	154.0	166.4	80.8	80.8	167.8	156.3

Note: In each row, the estimator with the minimum MSE and all estimators with MSE not larger than 0.5 above the minimum are underlined. Epa indicates Epanechnikov kernel; Gauss indicates Gaussian kernel. Trim 2% means that estimator is trimmed by deleting the 2% of the observations with lowest density. The bandwidth is chosen by leave-one-out cross-validation of the nonparametric regression estimator, over a grid of 30 different values. Simulations are based on 5000 replications.

TABLE A4.—MEAN SQUARED ERROR OF ESTIMATED COUNTERFACTUAL MEAN OUTCOME (RELATIVE TO MSE OF PAIR MATCHING): BANDWIDTH VALUE CHOSEN BY CROSS VALIDATION, SAMPLE SIZE 1600

		MSE Relative to That of Pair Matching (%)													
		Kernel Matching		Local Linear Matching with Epanechnikov Kernel				Local Linear Matching with Gauss Kernel				Ridge Matching		*k*-NN Matching	
Design	Curve	Epa	Gauss	—	Trim 2%	Trim 5%	Trim 10%	—	Trim 2%	Trim 5%	Trim 10%	Epa	Gauss	—	Epa
1	m_1	93.3	96.0	63.4	81.7	182.9	541.9	54.9	71.9	172.3	535.6	62.7	62.2	327.3	272.9
	m_2	71.5	71.2	101.9	85.1	75.8	89.5	141.1	92.9	84.4	104.3	76.9	75.3	74.7	79.6
	m_3	61.4	62.0	104.6	76.8	74.4	147.9	91.1	72.2	77.4	152.7	73.0	75.0	52.8	65.0
	m_4	118.3	126.1	85.4	77.2	185.5	594.3	74.6	74.0	185.3	588.6	69.3	63.3	399.8	551.4
	m_5	73.9	73.5	104.1	97.6	174.6	457.4	89.5	92.5	178.5	461.3	79.2	72.6	120.9	145.8
	m_6	72.2	71.1	110.2	88.9	72.8	77.0	127.2	82.7	66.2	74.7	73.7	84.0	101.7	112.7
		81.8	83.3	94.9	84.6	127.7	318.0	96.4	81.0	127.3	319.5	72.5	72.1	179.5	204.6
2	m_1	81.9	80.2	75.1	118.4	324.9	1041.5	68.6	98.2	272.4	905.0	73.8	67.5	106.0	118.9
	m_2	76.2	74.1	76.9	84.5	109.1	176.5	80.2	82.4	102.1	160.7	79.7	76.4	82.4	86.4
	m_3	62.9	64.8	69.2	74.9	117.4	335.5	72.1	81.3	126.1	339.0	67.9	64.3	58.5	83.6
	m_4	82.5	81.4	89.2	77.1	279.1	1023.3	94.6	71.7	272.6	1042.9	87.9	87.0	134.2	131.3
	m_5	67.1	66.0	66.1	105.5	289.8	859.9	68.6	107.5	276.3	812.3	64.9	63.4	85.0	65.4
	m_6	80.1	76.4	77.4	77.2	81.6	123.3	77.9	75.8	79.1	123.4	78.0	75.9	103.8	80.0
		75.1	73.8	75.7	89.6	200.3	593.3	77.0	86.2	188.1	563.9	75.4	72.4	95.0	94.3
3	m_1	76.9	77.0	74.2	124.3	358.7	1038.4	75.6	122.4	327.2	919.0	73.0	73.3	63.9	90.0
	m_2	78.5	78.3	79.9	86.3	117.2	182.7	75.5	84.7	115.8	179.8	79.6	81.4	92.3	82.9
	m_3	63.1	66.2	71.9	87.7	157.0	440.1	68.8	83.2	148.5	410.0	66.4	70.1	64.4	62.5
	m_4	69.9	65.6	91.9	89.6	368.0	1271.0	86.5	83.0	331.3	1139.2	86.5	81.6	89.7	78.3
	m_5	62.1	63.9	62.2	121.7	356.0	1076.7	60.3	108.9	311.0	942.9	62.8	60.8	67.4	67.4
	m_6	77.3	76.0	73.9	75.5	84.1	159.4	77.1	78.4	90.3	183.0	72.9	78.5	77.3	82.8
		71.3	71.2	75.7	97.5	240.2	694.7	73.9	93.4	220.7	629.1	73.6	74.3	75.8	77.3
4	m_1	75.9	79.7	76.4	100.1	199.3	523.0	73.7	100.6	205.9	546.8	69.8	70.0	81.1	76.4
	m_2	80.9	76.8	77.6	81.2	89.7	110.2	74.5	77.0	84.5	101.9	77.0	79.0	69.9	73.5
	m_3	65.3	66.4	66.0	70.5	89.3	177.9	69.3	74.9	96.5	191.7	64.7	60.7	53.6	65.8
	m_4	62.5	60.3	77.3	80.1	227.0	738.0	67.5	73.4	226.0	744.6	75.3	79.1	77.1	74.8
	m_5	57.6	58.9	64.7	93.3	210.1	546.4	65.6	96.0	212.2	551.2	61.3	63.0	62.8	62.2
	m_6	80.3	76.5	74.6	77.6	82.7	101.3	80.4	83.1	87.9	106.6	77.0	78.0	82.1	79.5
		70.4	69.8	72.8	83.8	149.7	366.1	71.8	84.2	152.2	373.8	70.9	71.6	71.1	72.0
5	m_1	89.1	90.0	76.4	102.3	243.2	743.1	78.6	98.4	239.8	761.5	65.9	69.2	198.6	215.7
	m_2	76.2	80.5	150.3	125.7	127.4	209.0	281.7	189.7	149.1	209.2	80.0	76.7	100.3	84.9
	m_3	66.1	63.9	130.0	103.1	113.3	294.8	152.6	106.1	110.0	278.7	71.2	74.1	79.2	98.2
	m_4	128.6	118.6	104.4	73.2	152.3	493.0	84.8	68.1	159.9	495.8	72.5	68.6	442.9	323.0
	m_5	86.2	74.1	102.0	92.7	166.4	510.0	177.3	93.3	146.6	481.3	77.3	73.0	166.0	145.8
	m_6	68.6	76.1	117.0	94.1	74.9	113.1	236.8	137.8	86.6	123.3	72.4	81.3	56.3	77.2
		85.8	83.9	113.3	98.5	146.3	393.8	168.7	115.5	148.7	391.6	73.2	73.8	173.9	157.5

Note: In each row, the estimator with the minimum MSE and all estimators with MSE not larger than 0.5 above the minimum are underlined. Epa indicates Epanechnikov kernel; Gauss indicates Gaussian kernel. Trim 2% means that estimator is trimmed by deleting the 2% of the observations with lowest density. The bandwidth is chosen by leave-one-out cross validation of the nonparametric regression estimator, over a grid of 30 different values. Simulations are based on 1000 replications.

This article has been cited by:

1. Stefanie Behncke, Markus Frölich, Michael Lechner. 2010. A Caseworker Like Me - Does The Similarity Between The Unemployed and Their Caseworkers Increase Job Placements?*. *The Economic Journal* **120**:549, 1430-1459. [CrossRef]
2. Christian Volpe Martincus, Jerónimo Carballo. 2010. Entering new country and product markets: does export promotion help?. *Review of World Economics* **146**:3, 437-467. [CrossRef]
3. Weihua An. 2010. BAYESIAN PROPENSITY SCORE ESTIMATORS: INCORPORATING UNCERTAINTIES IN PROPENSITY SCORES INTO CAUSAL INFERENCE. *Sociological Methodology* **40**:1, 151-189. [CrossRef]
4. Anton Flossmann. 2010. Accounting for missing data in M-estimation: a general matching approach. *Empirical Economics* **38**:1, 85-117. [CrossRef]
5. Stefanie Behncke, Markus FrÃ¶lich, Michael Lechner. 2010. Unemployed and their caseworkers: should they be friends or foes?. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **173**:1, 67-92. [CrossRef]
6. Inha Oh, Jeong-Dong Lee, Almas Heshmati, Gyoung-Gyu Choi. 2009. Evaluation of credit guarantee policy using propensity score matching. *Small Business Economics* **33**:3, 335-351. [CrossRef]
7. Annette Bergemann, Bernd Fitzenberger, Stefan Speckesser. 2009. Evaluating the dynamic employment effects of training programs in East Germany using conditional difference-in-differences. *Journal of Applied Econometrics* **24**:5, 797-823. [CrossRef]
8. Michael Rosholm, Lars Skipper. 2009. Is labour market training a curse for the unemployed? Evidence from a social experiment. *Journal of Applied Econometrics* **24**:2, 338-365. [CrossRef]
9. Guido W Imbens, Jeffrey M Wooldridge. 2009. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature* **47**:1, 5-86. [CrossRef]
10. Brian C. Briggeman, Charles A. Towe, Mitchell J. Morehart. 2009. Credit Constraints: Their Existence, Determinants, and Implications for U.S. Farm and Nonfarm Sole Proprietorships. *American Journal of Agricultural Economics* **91**:1, 275-289. [CrossRef]
11. Markus Frölich. 2008. Parametric and Nonparametric Regression in the Presence of Endogenous Control Variables. *International Statistical Review* **76**:2, 214-227. [CrossRef]
12. Pamela Farley Short, Joseph J. Vasey, John R. Moran. 2008. Long-Term Effects of Cancer Survivorship on the Employment of Older Workers. *Health Services Research* **43**:1p1, 193-210. [CrossRef]
13. Jennifer Davis, Heather Lukacs, Marc Jeuland, Alfonso Alvestegui, Betty Soto, Gloria Lizárraga, Alex Bakalian, Wendy Wakeman. 2008. Sustaining the benefits of rural water supply investments: Experience from Cochabamba and Chuquisaca, Bolivia. *Water Resources Research* **44**:12. . [CrossRef]
14. Daniel J. Henderson, Daniel L. Millimet, Christopher F. Parmeter, Le WangFertility and the health of children: A nonparametric investigation **21**, 167-195. [CrossRef]
15. Marco Caliendo, Reinhard Hujer, Stephan L. ThomsenThe employment effects of job-creation schemes in Germany: A microeconometric evaluation **21**, 381-428. [CrossRef]
16. Antonio Bento, Charles Towe, Jacqueline Geoghegan. 2007. The Effects of Moratoria on Residential Development: Evidence from a Matching Approach. *American Journal of Agricultural Economics* **89**:5, 1211-1218. [CrossRef]
17. Markus Frölich. 2007. On the inefficiency of propensity score matching. *AStA Advances in Statistical Analysis* **91**:3, 279-290. [CrossRef]
18. Markus Frölich. 2007. Propensity score matching without conditional independence assumption?with an application to the gender wage gap in the United Kingdom. *The Econometrics Journal* **10**:2, 359-407. [CrossRef]
19. Miana Plesca, Jeffrey Smith. 2007. Evaluating multi-treatment programs: theory and evidence from the U.S. Job Training Partnership Act experiment. *Empirical Economics* **32**:2-3, 491-528. [CrossRef]
20. Marvin A. Titus. 2007. Detecting selection bias, using propensity score matching, and estimating treatment effects: an application to the private returns to a master's degree. *Research in Higher Education* **48**:4, 487-521. [CrossRef]
21. Markus Frölich. 2006. Non-parametric regression for binary dependent variables. *The Econometrics Journal* **9**:3, 511-540. [CrossRef]
22. Markus Frölich. 2006. Semiparametric estimation of conditional mean functions with missing data. *Empirical Economics* **31**:2, 333-367. [CrossRef]

23. Richard Blundell, Lorraine Dearden, Barbara Sianesi. 2005. Evaluating the effect of education on earnings: models, methods and results from the National Child Development Survey. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **168**:3, 473-512. [CrossRef]