Statistics and Computing **15**: 197–215, 2005

Matching estimators and optimal bandwidth choice

MARKUS FRÖLICH

University College London and University of St. Gallen, Bodanstrasse 8, CH-9000 St. Gallen, Switzerland; Institute for the Study of Labor (IZA), Bonn
markus.froelich@unisg.ch, www.siaw.unisg.ch/froelich

Received September 2003 and accepted January 2005

Optimal bandwidth choice for matching estimators and their finite sample properties are examined. An approximation to their MSE is derived, as a basis for a plug-in bandwidth selector. In small samples, this approximation is not very accurate, though. Alternatively, conventional cross-validation bandwidth selection is considered and performs rather well in simulation studies: Compared to standard pair-matching, kernel and ridge matching achieve reductions in MSE of about 25 to 40%. Local linear matching and weighting perform poorly. Furthermore, the scope for developing better bandwidth selectors seems to be limited for ridge matching, but non-negligible for kernel and local linear matching.

Keywords: covariate adjustment, nonparametric regression, propensity score, missing data, counterfactual, treatment effect

1. Introduction

Matching estimators for estimating treatment effects have received a lot of attention in recent years.[1] Optimal bandwidth choice for these estimators, however, has not been analyzed so far. In this paper, a mean squared error approximation for matching estimators is derived and its usefulness for bandwidth selection analyzed. In addition, conventional bandwidth selection methods are examined and their finite-sample efficiency is investigated.

Matching estimators attempt to estimate an expected value for a particular population, by using data from a different population and adjusting nonparametrically for the different distributions of covariates in these two populations. The prototypical example for nonparametric covariate-distribution adjustment is the estimation of average treatment effects in treatment evaluation. Suppose some individuals or units receive a particular treatment, e.g. a medical drug, participation in a training programme or access to subsidized loans, whereas others do not. To assess the effectiveness of treatment, we would like to compare the outcome with treatment (Y_i^1) and the outcome without treatment (Y_i^0) for the same individual. Since we cannot observe any individual at the same time in both states, treatment evaluation has to rely on a comparison of the observed outcomes of the treated individuals with those of the non-treated individuals. Because individuals are often not randomly assigned to treatment but selected on the basis of certain characteristics X (e.g. age, motivation, income), treated and non-treated will usually differ in their covariates, which has to be taken into account. If one knew that the true relationship between the outcome variable and these covariates X was linear, linear projections (least squares regression) could be used to adjust for these differences in the covariates. *Nonparametric matching* estimators, on the other hand, do not rely on such functional form assumptions and proceed by first estimating conditional expectation functions to eliminate selection bias and then averaging these to get the population mean. If D_i denotes whether an individual got treated ($D_i = 1$) or not ($D_i = 0$) and X contains all confounding variables,[2] then $E[Y^0 \mid X, D = 0] = E[Y^0 \mid X, D = 1]$ and the counterfactual mean for the treated in the hypothetical case of not getting treated is identified as $E[E[Y \mid X, D = 0] \mid D = 1]$.

Matching estimators can also be used in many other situations to adjust for differences in the covariates. *Missing data*, for example, poses substantial problems in many survey data sets. Missing data on the outcome variable Y might also be a deliberate part of the sampling design (e.g. in clinical trials) when collecting covariate information is much cheaper than collecting information on the outcome variable. Very often the missingness

is not completely random and for obtaining the population mean of the outcome variable Y, we need to estimate the mean outcome in the non-responding population. If sufficient covariate information X is available such that, conditional on X, data is missing at random (Little and Rubin 1987), adjusting for the differences in the covariates among respondents and non-respondents by a matching estimator gives the mean outcome for the non-responding population.

As another example for the applicability of matching estimators, consider the analysis of *discrimination* according to race or gender. Differences in wages or earnings are partly due to differences in skills (type of education, job experience) between men and women or between whites and blacks. To separate these from the effects of discrimination, we need to adjust for observed skill differences. A matching estimator would deliver, e.g., the mean wage women would receive if they had the skills and human capital endowments of men.

As a final example, matching estimators can be used to simulate the consequences of, e.g., demographic changes in the population on the fertility rate, health care demand or the voting behaviour, in a scenario were individuals do not adjust their behaviour to the changing environment.

The most popular matching estimator is pair-matching, which proceeds by finding for each observation of the first population the most similar observation of the second population. In treatment evaluation, it matches to each treated individual the most similar non-treated individual. For discrimination analysis, it matches to each man the most similar woman, or vice versa. When accounting for missing data, it matches to each observation with missing outcome data the most similar observation with complete data. Similarity can be measured by an index function or a distance metric, such as the Mahalanobis distance. A particularly convenient index function is the *propensity score*, which is the ratio of the density of the covariates in the two populations.[3] Rosenbaum and Rubin (1983) showed that matching with respect to the (one-dimensional) propensity score is consistent. For applications of pair-matching see e.g. Angrist (1998), Dehejia and Wahba (1999), Gerfin and Lechner (2002) or Lechner (1999).

Since pair-matching matches only a single observation to each observation of the other population, it may have a rather high variance. Abadie and Imbens (2001) have shown that pair-matching is inefficient[4] and may not even be $\sqrt{n}$-consistent. As an alternative to pair-matching, Heckman, Ichimura and Todd (1997, 1998) proposed local polynomial matching and showed its $\sqrt{n}$-consistency and asymptotic normality. Local polynomial matching, however, requires the choice of a bandwidth parameter and, as in other fields of nonparametric regression, the estimation results are likely to be rather sensitive to the bandwidth value. Yet, optimal bandwidth choice for matching estimators has not been analyzed so far.

In this paper, an approximation to the mean squared error of local polynomial matching estimators is developed, which could be used for a plug-in bandwidth selector. The accuracy of this MSE approximation in finite samples is then examined, and it turns out to be not very reliable for bandwidth choice in small samples. As an alternative, the usefulness of conventional cross-validation bandwidth selection is examined. Although not being consistent, cross-validation performs rather well in small samples, at least for a particular ridge matching estimator. For other matching estimators, however, there remains scope for improvement.

As a second contribution of this paper, the finite-sample properties of various matching and weighting estimators are compared. Matching based on local linear regression performs rather poorly and is often even worse than pair-matching. Matching based on kernel regression or ridge regression is usually more precise than pair-matching by about 15 to 40%. Ridge matching often performs best and, in addition, is rather robust to the simulation design and the bandwidth value. A weighting estimator, as considered in Horvitz and Thompson (1952), Imbens (2000) and Hirano, Imbens and Ridder (2003), on the other hand, performed much worse than the matching estimators.[5]

Section 2 introduces the matching estimators and develops an approximation to their MSE. Section 3 examines the accuracy of this approximation in finite samples. Section 4 analyzes the finite sample properties of matching estimators with cross-validation bandwidth selection. In Section 5, the finite sample properties of matching on an estimated propensity score are examined. Section 6 concludes.

2. Covariate adjustment and optimal bandwidth choice

Let $Y \in \Re$ be an outcome variable of interest and $X \in \Re$ be a covariate.[6] Suppose that observations on (Y, X) are sampled independently from a *source* population and observations on X are sampled from a target population. In treatment evaluation, the non-treated are the source population and the treated are the target population. We are interested in estimating the mean of Y in the *target* population. Let $f_0(x)$ be the density of X in the source population and $f_1(x)$ be the density in the target population. Let $m(x) = E[Y \mid X = x]$ be the conditional mean function in the source population. Denote the source sample by $\{Y_i^0, X_i^0\}_{i=1}^{n_0}$ and the target sample by $\{X_j^1\}_{j=1}^{n_1}$. The counterfactual mean of Y in the target population is

$$E_1[Y] = \int m(x) \cdot f_1(x)\, dx, \tag{1}$$

which can be estimated by a *matching estimator* (Heckman, Ichimura and Todd 1998) as

$$\widehat{E_1[Y]} = \frac{1}{n_1} \sum_{j=1}^{n_1} \hat{m}\left(X_j^1\right), \tag{2}$$

where $\hat{m}(x)$ is a nonparametric regression estimator of $m(x)$.[7]

The different matching estimators differ in how they estimate the conditional mean function $m(x)$ from the source sample $\{Y_i^0, X_i^0\}_{i=1}^{n_0}$. *Pair-matching* estimates $m(x)$ by first-nearest neighbour regression and is widely used in the statistics

literature.[8] Restricting the estimation to only one neighbour, however, is likely to be inefficient and alternative variants have been considered recently. Heckman, Ichimura and Todd (1998) suggested to estimate $m(x)$ by local polynomial regression and showed $\sqrt{n}$-consistency and asymptotic normality of the *local polynomial matching* estimator.

Local polynomial regression is a class of nonparametric regression estimators, including Nadaraya-Watson regression (=local constant regression)

$$\hat{m}_{NW}(x) = \frac{T_0}{S_0} \tag{3}$$

and local linear regression

$$\hat{m}_{ll}(x) = \frac{T_0}{S_0} + (x - \bar{x})\frac{T_1}{S_2}, \tag{4}$$

where $\bar{x} = \sum X_i^0 K(\frac{X_i^0 - x}{h}) / \sum K(\frac{X_i^0 - x}{h})$ and $S_r = \sum K(\frac{X_i^0 - x}{h})$ $(X_0^i - \bar{x})^r$ and $T_r = \sum Y_i^0 K(\frac{X_i^0 - x}{h})(X_i^0 - \bar{x})^r$ and $K(\cdot)$ is a symmetric kernel function and h a bandwidth parameter.[9] Inserting $\hat{m}_{NW}$ or $\hat{m}_{ll}$ in (2) provides the *kernel matching* and *local linear matching* estimator, respectively.[10]

Local linear regression is known for its optimality properties (Fan 1993, Fan *et al.* 1997). However, in small samples, local linear regression with a fixed bandwidth often leads to a very rugged curve in regions of sparse data, see Seifert and Gasser (1996). The denominator S_2 in (4) can become very small, or even zero with a compact kernel. Therefore, the local linear estimator has infinite unconditional variance and unbounded conditional variance. Seifert and Gasser (1996) also showed that the probability for the occurrence of sparse regions is substantial if the X_i^0 observations are randomly spaced. Their simulation results reveal the gravity of this behaviour, demonstrating that the mean integrated squared error of local linear regression explodes at bandwidth values that are only slightly below the asymptotically optimal bandwidth. For reliable small sample behaviour Seifert and Gasser (1996, 2000) proposed to modify the local linear estimator by adding a ridging term to its denominator. Their modified local linear estimator for the Epanechnikov kernel is

$$\hat{m}_{\text{Ridge}}(x) = \frac{T_0}{S_0} + \frac{(x - \bar{x})T_1}{S_2 + \tau}, \tag{5}$$

with ridge term $\tau = \frac{5}{16} h \cdot |x - \bar{x}|$, see Seifert and Gasser (2000). The ridge regression estimator is a weighted average of the Nadaraya-Watson and the local linear regression estimator:

$$\hat{m}_{\text{Ridge}}(x) = \hat{m}_{ll}(x)\alpha + (1 - \alpha)\hat{m}_{NW}(x),$$

where $\alpha = \frac{S_2}{S_2 + \tau} \in (0, 1]$. For h converging to zero with growing sample size, the ridge term τ goes to zero and $\alpha \to 1$. Hence, the ridge regression estimator converges to local linear regression with growing sample size, but has better variance properties in finite samples.

In this paper, only matching with respect to a *single* covariate is examined. In many applications, however, one needs to adjust for many covariates to estimate the counterfactual mean. For example, in treatment evaluation many confounding factors that influenced treatment assignment have to be taken into account. Nevertheless, the set-up of this paper still applies to those situations, because matching on the *one-dimensional* propensity score is consistent for estimating the counterfactual mean, as shown by Rosenbaum and Rubin (1983).[11] Hence, instead of adjusting for the different distributions of all covariates, it suffices to adjust for the different distribution of a one-dimensional function of the covariates. Let $Z \in \Re^k$ be a k vector of covariates and let $f_{Z|0}$, $f_{Z|1}$ be the density of Z in the source and the target population, respectively. Let P_0/P_1 denote the relative size of the source to the target population. Define the (one-dimensional) random variable X

$$X = \frac{f_{Z|1}(Z)}{f_{Z|1}(Z) + f_{Z|0}(Z)\frac{P_0}{P_1}} \tag{6}$$

as the propensity score.[12] Then the expression (1) is identical to covariate adjustment with respect to all characteristics Z:

$$E_1[Y] = \int E[Y \mid Z = z] \cdot f_{Z|1}(z)\, dz.$$

Hence, matching with respect to the one-dimensional propensity score (=propensity score matching) is often used when differences in the distribution of many covariates need to be accounted for.[13]

2.1. *Optimal bandwidth choice*

Matching requires the choice of a smoothing parameter. In conventional nonparametric regression it is usually attempted to choose the bandwidth value that minimizes the mean integrated squared error

$$MISE(h) = E \int (\hat{m}(x; h) - m(x))^2\, dx, \tag{7}$$

and cross-validation is often used to estimate the optimal bandwidth value, see Loader (1999).

However, minimizing MISE may not lead to optimal bandwidth choices for the matching estimator. The MISE criterion neglects f_0 and f_1, i.e. the location of the source and the target population. Yet, precise estimation of m is particularly important in regions where the target population is concentrated. Moreover, by averaging over the imputed values $\hat{m}$, the matching estimator adds another smoothing step, which might change its sensitivity to the bandwidth value. Instead of minimizing MISE, the bandwidth value should be chosen to minimize the mean squared error of the matching estimate (2).

To derive the asymptotic MSE approximation for the matching estimator, I make use of the asymptotic linear representation of local polynomial regression as developed in Heckman, Ichimura and Todd (1998). A nonparametric estimator $\hat{m}(x)$ of $m(x)$ at a point x where $f_0(x)$ is bounded away from zero, on basis of the sample $\{Y_i^0, X_i^0\}_{i=1}^{n_0}$, is asymptotically linear if it can be written as

$$\hat{m}(x) - m(x) = \frac{1}{n_0} \sum_{i=1}^{n_0} \psi\left(Y_i^0, X_i^0, x\right) + b(x) + R(x), \tag{8}$$

with the properties (i) $E[\psi(Y_i^0, X_i^0, X) \mid X = x] = 0$, (ii) *plim* $n_0^{-\frac{1}{2}} \sum b(X_i^0) < \infty$ and (iii) $n_0^{-\frac{1}{2}} \sum R(X_i^0) = o_p(1)$. The functions ψ, b and R depend on smoothing parameters and therefore depend on the sample size n_0. (This dependence on n_0 is kept implicit in the notation.) The local influence function ψ has mean zero and determines the local variance of the estimate. $b(x)$ is the local bias and $R(x)$ is a remainder term. Under iid sampling, $m(x)$ and $f_0(x)$ being twice continuously differentiable with second derivative Hölder continuous, the bandwidth sequence satisfying $n_0 h / \ln n_0 \to \infty$ and $n_0 h^4 \to c < \infty$ and symmetric, compact and Lipschitz continuous kernel function with $\int K(u)du = 0$, the local constant and the local linear regression estimator are asymptotically linear, see Heckman, Ichimura and Todd (1998).

Inserting (8) in the matching estimator (2) gives the expression

$$\frac{1}{n_1}\sum_{j=1}^{n_1}\hat{m}\big(X_j^1\big) = \frac{1}{n_1}\sum_{j=1}^{n_1}\left(m\big(X_j^1\big) + b\big(X_j^1\big) + R\big(X_j^1\big) + \frac{1}{n_0}\sum_{i=1}^{n_0}\psi\big(Y_i^0, X_i^0, X_j^1\big)\right) \tag{9}$$

The bias of (9) is

$$\begin{aligned} &E\left[\frac{1}{n_1}\sum_{j=1}^{n_1}\left(m\big(X_j^1\big) + b\big(X_j^1\big) + R\big(X_j^1\big) + \frac{1}{n_0}\sum_{i=1}^{n_0}\psi\big(Y_i^0, X_i^0, X_j^1\big)\right)\right] - E_1[Y] \\ &= E\big[m\big(X_j^1\big) + b\big(X_j^1\big) + R\big(X_j^1\big)\big] - \int m(x) f_1(x)\,dx \\ &= \int b(x) f_1(x)\,dx, \end{aligned} \tag{10}$$

where the lower order remainder term is dropped. The variance of (9) is

$$\begin{aligned} &\mathrm{Var}\left(\frac{1}{n_1}\sum_{j=1}^{n_1}\left(m\big(X_j^1\big) + b\big(X_j^1\big) + R\big(X_j^1\big) + \frac{1}{n_0}\sum_{i=1}^{n_0}\psi\big(Y_i^0, X_i^0, X_j^1\big)\right)\right) \\ &= \mathrm{Var}\left(\frac{1}{n_1}\sum_{j=1}^{n_1}\big(m\big(X_j^1\big) + b\big(X_j^1\big) + R\big(X_j^1\big)\big)\right) \\ &\quad + \mathrm{Var}\left(\frac{1}{n_1 n_0}\sum_{j=1}^{n_1}\sum_{i=1}^{n_0}\psi\big(Y_i^0, X_i^0, X_j^1\big)\right) \end{aligned}$$

because the covariance terms $E[(m(X_j^1) + b(X_j^1) + R(X_j^1)) \cdot \psi(Y_k^0, X_k^0, X_l^1)] = E[E[(m(X_j^1) + b(X_j^1) + R(X_j^1)) \cdot \psi(Y_k^0, X_k^0, X_l^1) \mid X_j^1, X_l^1]]$ are zero since $E[\psi(Y_i^0, X_i^0, X) \mid X = x] = 0$.

The first variance term captures the variance due to the variation of m and the local bias b along x. With iid data this term simplifies to after dropping the lower order remainder term.

$$\begin{aligned} &\mathrm{Var}\left(\frac{1}{n_1}\sum_{j=1}^{n_1}\big(m\big(X_j^1\big) + b\big(X_j^1\big) + R\big(X_j^1\big)\big)\right) \\ &= \frac{1}{n_1} Var\big(m\big(X_j^1\big) + b\big(X_j^1\big) + R\big(X_j^1\big)\big) \\ &= \frac{1}{n_1}\left[\int (m(x) + b(x))^2 f_1(x)\,dx - \left(\int (m(x) + b(x)) f_1(x)\,dx\right)^2\right], \end{aligned} \tag{11}$$

The second variance term captures the local variance of the nonparametric regression estimator. Summing up all covariance elements yields

$$\begin{aligned} &\mathrm{Var}\left(\frac{1}{n_1 n_0}\sum_{j=1}^{n_1}\sum_{i=1}^{n_0}\psi\big(Y_i^0, X_i^0, X_j^1\big)\right) \\ &= \frac{1}{n_1^2 n_0^2}\sum_{j=1}^{n_1}\sum_{i=1}^{n_0}\sum_{l=1}^{n_1}\sum_{k=1}^{n_0} E\big[\psi\big(Y_i^0, X_i^0, X_j^1\big)\psi\big(Y_k^0, X_k^0, X_l^1\big)\big]. \end{aligned} \tag{12}$$

The particular expressions for the local influence function ψ are for kernel regression

$$\psi\big(Y_i^0, X_i^0, x\big) = \varepsilon_i \cdot \frac{K\big(\frac{X_i^0 - x}{h}\big)}{E\big[K\big(\frac{X_i^0 - x}{h}\big)\big]},$$

and for local linear regression

$$\begin{aligned} \psi\big(Y_i^0, X_i^0, x\big) &= \varepsilon_i \cdot (1, 0) \\ &\times \begin{bmatrix} E\big[K\big(\frac{X_i^0 - x}{h}\big)\big] & E\big[\frac{X_i^0 - x}{h} K\big(\frac{X_i^0 - x}{h}\big)\big] \\ E\big[\frac{X_i^0 - x}{h} K\big(\frac{X_i^0 - x}{h}\big)\big] & E\big[\big(\frac{X_i^0 - x}{h}\big)^2 K\big(\frac{X_i^0 - x}{h}\big)\big] \end{bmatrix}^{-1} \\ &\times \begin{bmatrix} K\big(\frac{X_i^0 - x}{h}\big) \\ \frac{X_i^0 - x}{h} K\big(\frac{X_i^0 - x}{h}\big) \end{bmatrix}, \end{aligned}$$

where $\varepsilon_i = (Y_i^0 - m(X_i^0))$, see Heckman, Ichimura and Todd (1998, Theorem 3).

Denote the boundaries of the support of X by a and b, where a may be $-\infty$ and b may be ∞. Defining

$$\mu_r(x, h) = \int_{(a-x)/h}^{(b-x)/h} u^r K(u)\,du,$$

and using the approximation

$$
\begin{aligned}
&E\left[\left(\frac{X_i^0-x}{h}\right)^r K\left(\frac{X_i^0-x}{h}\right)\right] \\
&\quad=\int_a^b \left(\frac{X_i^0-x}{h}\right)^r K\left(\frac{X_i^0-x}{h}\right) f_0\big(X_i^0\big)dX_i^0 \\
&\quad= h\int_{(a-x)/h}^{(b-x)/h} u^r K(u) f_0(x+uh)\,du \\
&\quad= h\int_{(a-x)/h}^{(b-x)/h} u^r K(u)(f_0(x)+O(h))\,du \\
&\quad= h\mu_r(x,h) f_0(x)+O(h),
\end{aligned}
$$

where a change in variables and a Taylor series expansion of $f_0(x+uh)=f_0(x)+uhf_0'(x)+o(h)=f_0(x)+O(h)$ has been used, the local influence functions ψ can be written as

$$
\psi\big(Y_i^0, X_i^0, x\big) \approx \frac{\varepsilon_i}{hf_0(x)} K^*\left(\frac{X_i^0-x}{h}, x, h\right),
$$

where

$$
\begin{aligned}
&K_{NW}^*\left(\frac{X_i^0-x}{h}, x, h\right) = \mu_0^{-1}(x,h)\cdot K\left(\frac{X_i^0-x}{h}\right) \\
&K_{ll}^*\left(\frac{X_i^0-x}{h}, x, h\right) \\
&\quad= \frac{\mu_2(x,h)-\mu_1(x,h)\frac{X_i^0-x}{h}}{\mu_0(x,h)\mu_2(x,h)-\mu_1^2(x,h)} K\left(\frac{X_i^0-x}{h}\right)
\end{aligned}
$$

for Nadaraya-Watson and local linear regression, respectively. These equivalence kernels $K^*()$ capture as well the interior as the boundary region. In the interior these expressions simplify considerably, because with a kernel function compact in $[-1, 1]$, the kernel moments $\mu_r(x,h)=\mu_r$ do no longer depend on x if $b-x>h$ and $x-a>h$. Furthermore, with a symmetric kernel, μ_r is zero for r odd, and $\mu_0=1$ for a kernel integrating to one.

With these expressions, the variance term (12) can be further examined. The summation terms in (12) can be distinguished into three groups: $n_0(n_0-1)n_1^2$ terms with $i\neq k$, and $n_0n_1(n_1-1)$ terms with $i=k$ and $j\neq l$, and n_0n_1 terms with $i=k$ and $j=l$. With iid sampling, all terms with $i\neq k$ are zero because the terms ε_i and ε_k factor and integrate to zero conditional on X_i^0 and X_k^0. Second, the $n_0n_1(n_1-1)$ terms with $i=k$ and $j\neq l$ can be expressed as

$$
\begin{aligned}
&\int_a^b\int_a^b\int_a^b \frac{\sigma^2\big(X_i^0\big)}{h^2}\frac{K^*\big(\frac{X_i^0-X_j^1}{h}, X_j^1, h\big)}{f_0\big(X_j^1\big)}\frac{K^*\big(\frac{X_i^0-X_l^1}{h}, X_l^1, h\big)}{f_0\big(X_l^1\big)} \\
&\quad\cdot f_1\big(X_j^1\big) f_1\big(X_l^1\big) f_0\big(X_i^0\big) dX_j^1 dX_l^1 dX_i^0 \\
&= \int_a^b \int_{\frac{a-x}{h}}^{\frac{b-x}{h}}\int_{\frac{a-x}{h}}^{\frac{b-x}{h}} \sigma^2(x)\frac{K^*(u,x-uh,h)}{f_0(x-uh)}\frac{K^*(v,x-vh,h)}{f_0(x-vh)} \\
&\quad\times f_1(x-uh) f_1(x-vh) f_0(x)\,du\,dv\,dx
\end{aligned}
$$

with the change in variables $u=(X_i^0-X_j^1)/h$ and $v=(X_i^0-X_l^1)/h$ and $x=X_i^0$ and $\sigma^2\big(X_i^0\big)=\mathrm{Var}[Y_i\mid X_i^0]$. Assuming that f_0 and f_1 are differentiable, $f_1(x-uh)$ can be approximated by a series as $f_1(x)-uhf_1'(x)+o(h)=f_1(x)+O(h)$. Also the equivalent kernel $K^*(u,x-uh,h)\approx K^*(u,x,h)$ for small values of h. Notice that this latter approximation is only relevant for boundary points, since in the interior $K^*(u,x,h)$ does not depend on x. Hence for small bandwidth values

$$
\begin{aligned}
&\approx \int_a^b \int_{\frac{a-x}{h}}^{\frac{b-x}{h}}\int_{\frac{a-x}{h}}^{\frac{b-x}{h}} \sigma^2(x)K^*(u,x,h)K^*(v,x,h) \\
&\quad\cdot\frac{f_1^2(x)}{f_0(x)}\,du\,dv\,dx = \int_a^b\left(\sigma^2(x)\frac{f_1^2(x)}{f_0(x)}\int_{\frac{a-x}{h}}^{\frac{b-x}{h}} K^*(u,x,h)\,du\right. \\
&\quad\left.\times\int_{\frac{a-x}{h}}^{\frac{b-x}{h}} K^*(v,x,h)\,dv\right)dx = \int\sigma^2(x)\frac{f_1^2(x)}{f_0(x)}dx, \qquad (13)
\end{aligned}
$$

since $\int_{(a-x)/h}^{(b-x)/h} K^*(u,x,h)du=1$ for Nadaraya-Watson as well as for local linear regression.

Finally, the n_0n_1 covariance terms with $i=k$ and $j=l$ can be expressed as

$$
\begin{aligned}
&\int_a^b\int_a^b \frac{\sigma^2\big(X_i^0\big)}{h^2}\frac{K^{*^2}\big(\frac{X_i^0-X_j^1}{h}, X_j^1, h\big)^2}{f_0^2\big(X_j^1\big)}\cdot f_1\big(X_j^1\big) f_0\big(X_i^0\big)\,dX_j^1\,dX_i^0 \\
&= \int_a^b\int_{\frac{a-x}{h}}^{\frac{b-x}{h}} \frac{\sigma^2(x)}{h}\frac{K^{*^2}(u,x-uh,h)}{f_0^2(x-uh)}\cdot f_1(x-uh) f_0(x)\,du\,dx
\end{aligned}
$$

with the change in variables $u=(X_i^0-X_j^1)/h$ and $x=X_i^0$. With $f_1(x-uh)=f_1(x)+O(h)$

$$
\begin{aligned}
&\approx \int_a^b \frac{\sigma^2(x)}{h}\frac{f_1(x)}{f_0(x)}\int_{\frac{a-x}{h}}^{\frac{b-x}{h}} K^{*^2}(u,x,h)\,du\,dx \\
&= n_0\int \mathrm{Var}[\hat m(x)] f_1(x)dx, \qquad (14)
\end{aligned}
$$

with $\mathrm{Var}[\hat m(x)]=(\int_{(a-x)/h}^{(b-x)/h} K^{*^2}(u,x,h)du)\frac{\sigma^2(x)}{n_0hf_0(x)}(1+o_p(1))$, see Ruppert and Wand (1994).

Collecting the $n_0n_1(n_1-1)$ covariance terms with $i=k$ and $j\neq l$ and the n_0n_1 covariance terms with $i=k$ and $j=l$ gives for expression (12)

$$
\begin{aligned}
&\mathrm{Var}\left(\frac{1}{n_1n_0}\sum_{j=1}^{n_1}\sum_{i=1}^{n_0}\psi\big(Y_i^0, X_i^0, X_j^1\big)\right) \\
&\quad\approx \frac{1}{n_0}\int\sigma^2(x)\frac{f_1^2(x)}{f_0(x)}dx+\frac{1}{n_1}\int\mathrm{Var}[\hat m(x)] f_1(x)\,dx.
\end{aligned}
$$

Combining this expression with (11) and the squared bias gives the approximate MSE as a function of h as

$$MSE(h) \approx \left(\int b(x) f_1(x) dx \right)^2 + \frac{1}{n_1} \left[\int (m(x) + b(x))^2 f_1(x) dx - \theta^2 \right] + \frac{1}{n_0} \int \sigma^2(x) \frac{f_1^2(x)}{f_0(x)} dx + \frac{1}{n_1} \int Var[\hat{m}(x)] f_1(x) dx, \quad (15)$$

where $\theta = \int (m(x) + b(x)) f_1(x) dx$. The first term in (15) is the squared bias contribution to the MSE, while the second, third and fourth terms represent the variance of local polynomial matching. The second term stems from the variation of $m(x)$ and $b(x)$ along x, the third term represents the covariances of ψ and the fourth term the variances of ψ. The bias $b(x)$ for local polynomial regression of order p (i.e. $p = 0$ for Nadaraya-Watson regression and $p = 1$ for local linear regression) for a symmetric kernel with compact support in $[-1, 1]$ is given by

$$\begin{aligned} E[\hat{m}(x, h) - m(x)] &= \left(\int_{(a-x)/h}^{(b-x)/h} u^{p+1} K^*(u; x, h)\, du \right) \frac{m^{(p+1)}(x)}{(p+1)!} h^{p+1} \\ &+ \left(\int_{(a-x)/h}^{(b-x)/h} u^{p+2} K^*(u; x, h)\, du \right) \\ &\times \left(\frac{m^{(p+1)}(x)}{(p+1)!} \frac{f'(x)}{f(x)} + \frac{m^{(p+2)}(x)}{(p+2)!} \right) h^{p+2} \\ &- \Lambda(x, h) \frac{m^{(p+1)}(x)}{(p+1)!} \frac{f'(x)}{f(x)} h^{p+2} + o_p(h^{p+2}), \end{aligned} \quad (16)$$

with $\Lambda(x, h) = \frac{\mu_1^2(x,h)}{\mu_0^2(x,h)}$ for $p = 0$ and $\Lambda(x, h) = \frac{\mu_1^2 \mu_2 \mu_3 + \mu_0 \mu_2^2 \mu_3 - \mu_0 \mu_1 \mu_3^2 - \mu_1 \mu_2^3}{(\mu_0 \mu_2 - \mu_1^2)^2}$ for $p = 1$, see Ruppert and Wand (1994). (In the latter expression the dependence of the kernel moments μ on (x, h) is suppressed to ease notation.) This MSE approximation (15) with bias expression (16) is henceforth referred to as the two-terms *approximation*.

Simpler expressions are obtained when approximating local variance and local bias by the respective expressions for interior points and retaining only the first leading term in the bias approximation. Then the approximate MSE is

$$\begin{aligned} MSE(h) \approx & \left(\int \bar{b}(x) f_1(x)\, dx \right)^2 \\ & + \frac{1}{n_1} \int (m(x) + \bar{b}(x))^2 f_1(x)\, dx \\ & - \frac{1}{n_1} \left(\int (m(x) + \bar{b}(x)) f_1(x) dx \right)^2 \\ & + \frac{1}{n_0} \int \sigma^2(x) \frac{f_1^2(x)}{f_0(x)} dx + \frac{\bar{\mu}_0}{n_0 n_1 h} \int \sigma^2(x) \frac{f_1(x)}{f_0(x)}\, dx, \end{aligned} \quad (17)$$

with $\bar{b}(x) = h^2 \mu_2 (\frac{m'(x) f_0'(x)}{f_0(x)} + \frac{m''(x)}{2})$ for Nadaraya-Watson and $\bar{b}(x) = h^2 \mu_2 \frac{m''(x)}{2}$ for local linear regression, where $\mu_r = \int u^r K(u) du$ and $\bar{\mu}_r = \int u^r K^2(u) du$. This simpler expression is henceforth referred to as the *first-term approximation*.

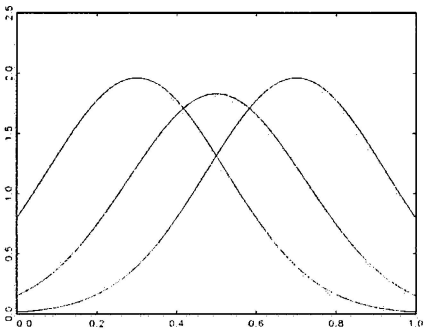
These approximations to the MSE differ substantially from the mean integrated squared error criterion (7). Hence, conventional bandwidth selectors are not consistent for local polynomial matching and bandwidth selection for matching should therefore attempt to minimize MSE (15) through the choice of the bandwidth value. By estimating the components of the MSE approximation (15) or (17), using a pilot bandwidth, a plug-in bandwidth estimator for matching estimators can be developed. For a plug-in bandwidth selector, the first-term approximation (17) is much more convenient than the two-terms approximation (15). However, since the local bias and variance expressions used in (17) are for interior points and since matching estimators are often used in situations where much of the density mass of the target population is located in the boundary region of the source population, a plug-in selector based on (15) might fare better.

Yet, before embarking on developing a data-driven bandwidth selector, it is worthwhile to examine how well the asymptotic MSE approximations accord with the true mean squared error of matching estimators in small samples. If the MSE approximations (15) or (17) fare well in small samples, a plug-in bandwidth selector may outperform conventional bandwidth selectors. On the other hand, if the MSE approximations do not conform well with the true MSE, a plug-in selector based on (15) or (17) would be useful only for rather large datasets.

3. MSE approximation accuracy in finite samples

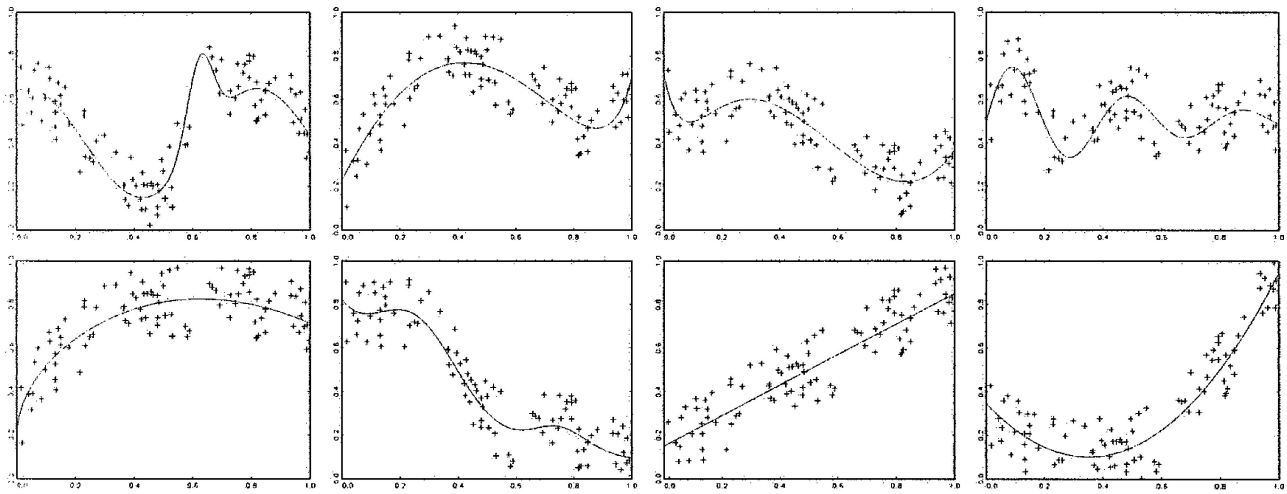
To assess the accuracy of the MSE approximation in small samples, the approximations (15) and (17) are computed for a variety of simulation designs and compared to the true MSE, which is simulated via Monte Carlo replications. The MSE approximations are analyzed for 6 different density combinations (f_0, f_1) and 8 different regression curves m. The six different density combinations are $(f_0, f_1) = (N_1, N_2), (N_1, N_3), (N_2, N_1), (N_2, N_3), (N_3, N_1), (N_3, N_2)$, where N_1, N_2 and N_3 refer to the three truncated normal distributions displayed in Fig. 1. For example, with the density combination (N_1, N_2), X is drawn from N_1 in the source sample and from N_2 in the target sample. The support of X is always $[0, 1]$. With the density combinations (N_1, N_3) and (N_3, N_1), the source and target population are more distinct than with the other density combinations.

The Y observations are sampled from one of the eight regression curves depicted in Fig. 2,[14] with an additive mean-zero, uniform error term with standard deviation 0.1. (The dots in Fig. 2 illustrate the signal-to-noise ratio.) Fig. 3 further shows an exemplary draw from the density combination (N_3, N_1) and regression curve m_4.



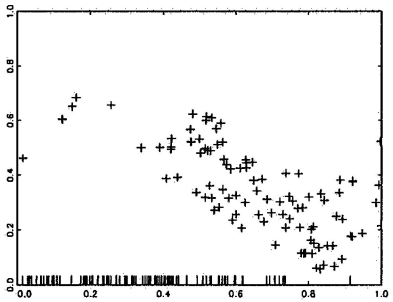
Note: Density distributions from which source and target samples are drawn. From left to right: $N_1(u) = 1.100 \exp\left(-(u-0.3)^2/\left(2\sigma_x^2\right)\right) \cdot 1_{[0,1]}(u)/\sqrt{2\pi\sigma_x^2}$ *and* $N_2(u) = 1.026 \exp\left(-(u-0.5)^2/\left(2\sigma_x^2\right)\right) \cdot 1_{[0,1]}(u)/\sqrt{2\pi\sigma_x^2}$ *and* $N_3(u) = 1.100 \exp\left(-(u-0.7)^2/\left(2\sigma_x^2\right)\right) \cdot 1_{[0,1]}(u)/\sqrt{2\pi\sigma_x^2}$ *with* $\sigma_x^2 = 0.05$.

Fig. 1. *Design densities N_1, N_2, N_3 of* X



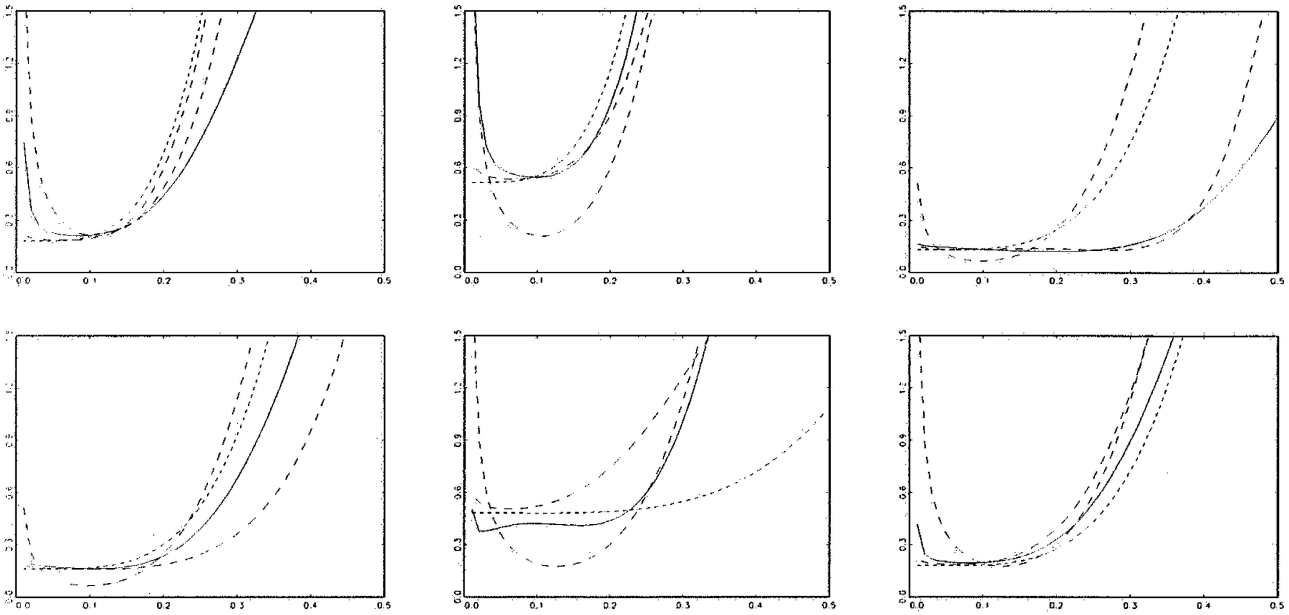
Regression curves m_1 to m_4 from left to right in the upper panel, regression curves m_5 to m_8 in the lower panel.

Fig. 2. *Regression curves m(x).*



Note: The source population is represented by + symbols and combines draws from density N_3 and outcomes $Y = m_3(X) + u$ where u is a mean-zero uniform error term with standard deviation 0.1. The target population is represented by draws from density N_1 and marked along the X-axis. Their outcomes are unobserved.

Fig. 3. *Exemplary draw from density combination (N_3, N_1) and m_3*



*Abscissa: bandwidth h, Ordinate: MSE*1000, MISE*100. Two-term approximation MSE (dotted-dashed), first-term approximation MSE (short-dashed), MISE approximation (long-dashed), and true MSE (solid). Density combinations* (N_1, N_2), (N_1, N_3), (N_2, N_1) *from top left to top right picture, density combinations* (N_2, N_3), (N_3, N_1), (N_3, N_2) *from bottom left to bottom right picture.*

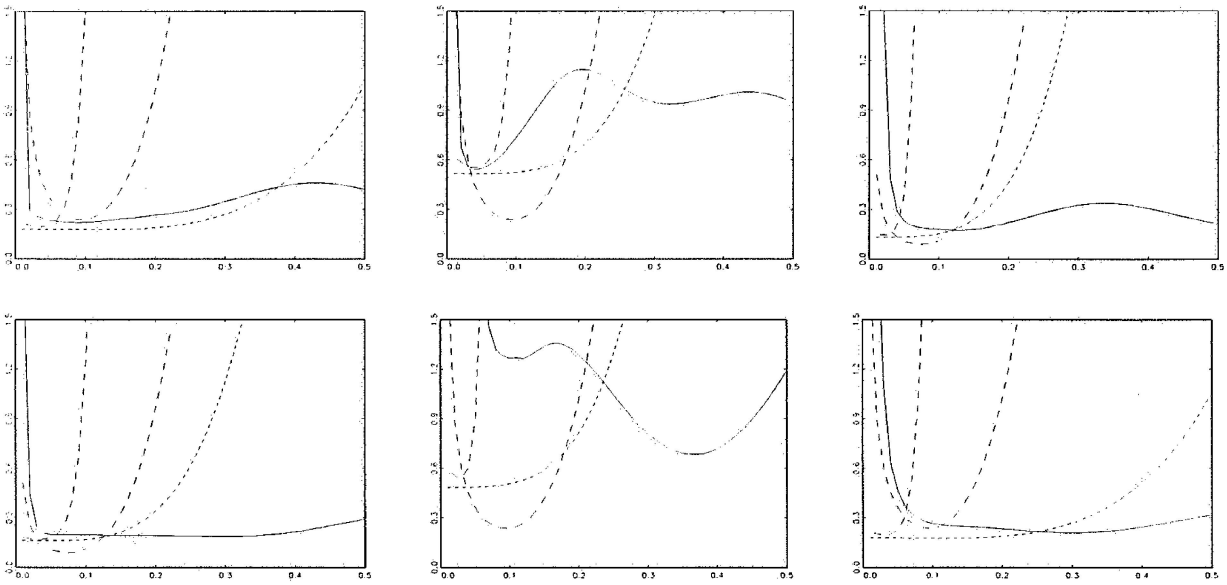
Fig. 4. *Approximations to the MSE for Kernel matching (curve* m_3*, sample size 200)*

3.1. *MSE approximation accuracy in finite samples*

For each of these 48 designs, the MSE approximations are computed for different bandwidth values. The true MSE is simulated by Monte Carlos simulations. The results for regression curve m_3 and sample size $n_0 = n_1 = 200$ are exemplary shown in Fig. 4 (for kernel matching) and Fig. 5 (for local linear matching).[15] The results for the other regression curves and for the sample size $n_0 = n_1 = 1000$ are given in the supplementary appendix.

Fig. 4 shows the true MSE (solid line), the two-term approximation (15, dotted-dashed) and the first-term approximation (17, short-dashed). For comparison, also the MISE approximation is shown (7, long-dashed). The graph in the upper left refers to the density combination (N_1, N_2), the upper middle to (N_1, N_3) and so forth. Hence, the two pictures in the middle correspond to the more difficult estimation settings where the two densities overlap less than in the four other settings (cf. Fig. 1). The MSE is scaled by 10^3. The MISE approximation is scaled by 10^2 to fit into the same graph and thus only its shape and location of its minimum can be interpreted. Since the MISE approximation does not take into account the location of the target population, its graph is identical for the density combinations (N_1, N_2) and (N_1, N_3) as well as for (N_2, N_1) and (N_2, N_3) and for (N_3, N_1) and (N_3, N_2). Hence its graph changes only in every second picture.

At a first impression, the MSE approximations are quite close to the true MSE in relatively large bandwidth regions, particularly for sample size 1000. They respond to the location of the target population and resemble the simulated MSE in level and shape, with the two-term approximation usually being somewhat more precise. Yet, for being useful for bandwidth selection, it is important that the minima of the MSE approximation and of the true MSE largely coincide. From Fig. 4 and the figures in the supplementary appendix, however, we find that both approximations are often rather flat for bandwidth values below 0.10. Particularly the first-term approximation does not rise steeply for small bandwidth values (despite increasing local variance) since the last term in (17) is divided by $n_1 n_0$. Also the two-term approximation is often quite flat for bandwidths smaller than 0.10 and does not increase very much for small bandwidth values. By contrast, the true MSE sometimes explodes for small bandwidth values, particularly with the more difficult density combinations (N_1, N_3) and (N_3, N_1). Thus, a bandwidth selector based on the MSE approximations would likely tend to *undersmooth* in the sense of choosing very low bandwidths at which the true MSE might be very large. Undersmoothing might be even of greater risk if, as usual, the elements of the MSE approximations are not known and need to be estimated. The steep increase of the MSE approximations at large bandwidth values, which always sets in at lower bandwidths than for the true MSE, would prevent choosing a bandwidth too large. But the nearly flat MSE approximation at lower bandwidth values would complicate the bandwidth choice and might often result in choosing too small bandwidth values. Because of this limited sensitivity to low bandwidth values, the MSE approximations (15) and (17) seem not to be very suited as a basis for data-driven bandwidth selection in small samples.



See note below Figure 4

Fig. 5. *Approximations to the MSE for local linear matching (curve m_3, sample size 200)*

The MISE approximation, on the other hand, displays more curvature around its minimum and increases steeply at small and at large bandwidth values. Although neither its shape nor its level resemble the true MSE at all, the true MSE at the MISE-minimizing bandwidth value is often not much higher than at its optimal bandwidth value. Moreover, no obvious pattern of over- or undersmoothing can be detected, suggesting that a conventional bandwidth selector might still be a useful starting point for kernel matching.

The results for local linear matching are given in Fig. 5 and in the supplementary appendix. The MSE approximations, particularly the two-term approximation, are often rather different from the true MSE. The true MSE frequently has local minima while the MSE approximations are always globally convex. Moreover, the MSE approximations would many times suggest a rather small bandwidth value, whereas the minimum of the true MSE is often at very large bandwidths. The MISE approximation, on the other hand, would often fail as well in picking a suited bandwidth value. Hence bandwidth selection seems to be more difficult for local linear matching.

4. Bandwidth choice by cross-validation

The previous simulation results indicated that the derived asymptotic MSE approximations do not approximate the true MSE very accurately in small samples. Hence, a plug-in bandwidth selector based on the asymptotic MSE approximation is unlikely to perform well in small samples. On the other hand, the bandwidth that minimizes MISE appeared often to be closer to the optimal bandwidth value than those suggested by the MSE approximations. Hence, a conventional bandwidth selector, such as cross-validation, might even perform well in small samples, although it is not consistent for large samples since it does not lead to asymptotic undersmoothing as required for $\sqrt{n}$ consistency of the matching estimator.[16]

In the following, the performance of cross-validation as a practical data-driven bandwidth selector for small samples is analyzed. By comparing the MSE with bandwidth selected by cross-validation to the MSE at the optimal bandwidth (as given by the solid line in, e.g., Figs. 4 and 5), the precision loss and thus the potential for developing better bandwidth selectors can be assessed. If the precision losses are rather small, cross-validation may be a useful approach in small samples.

Leave-one-out cross-validation for nonparametric regression chooses the bandwidth h as

$$\arg\min_h \frac{1}{n_0}\sum_{i=1}^{n_0}\left(Y_i^0 - \hat{m}_{-i}\left(X_i^0; h\right)\right)^2,$$

where $\hat{m}_{-i}(x)$ is the leave-one-out estimate of $m(x)$ obtained from the source sample without observation i.[17] Notice that the cross-validation criterion depends only on the source sample observations; the target sample $\{X_j^1\}_{j=1}^{n_1}$ does not affect the bandwidth choice. This is another reason, besides the asymptotic undersmoothing, why cross-validation cannot lead to optimal bandwidth choices because the location of the target population is neglected.

With this data-driven cross-validation bandwidth selector, the MSE of kernel matching and local linear matching is simulated for all simulation designs of the previous section. In addition, the MSE of pair-matching and of ridge matching is simulated. The simulation results are summarized in Tables 1 and 2.[18] Table 1 shows the simulated MSE of the matching estimators relative to the benchmark pair-matching estimator, whereas Table 2 gives the MSE relative to the MSE at their *optimal*

Table 1. *MSE of matching with CV bandwidth selection, relative to pair-matching*

		$n_0 = n_1 = 40$			$n_0 = n_1 = 200$			$n_0 = n_1 = 1000$			$n_0 = 200, n_1 = 40$			$n_0 = 40, n_1 = 200$		
	Densities	Kernel	Loclin	Ridge	Kernel	Loclin	Ridge	Kernel	Loclin	Ridge	Kernel	Loclin	Ridge	Kernel	Loclin	Ridge
m_1	N_1-N_2	97	118	89	84	96	79	83	58	60	81	77	73	112	166	112
	N_1-N_3	91	266	87	71	154	65	71	119	83	80	131	73	92	319	91
	N_2-N_1	89	133	82	81	91	78	77	72	73	76	82	76	108	199	94
	N_2-N_3	79	134	73	75	86	77	76	64	83	76	79	76	83	199	79
	N_3-N_1	70	154	69	132	239	121	109	228	112	111	192	105	66	158	67
	N_3-N_2	72	98	68	83	94	82	79	111	82	81	83	78	62	108	61
m_2	N_1-N_2	71	138	68	66	111	61	65	76	67	56	84	51	81	150	77
	N_1-N_3	94	171	96	76	145	71	64	148	59	67	157	62	100	159	99
	N_2-N_1	91	100	80	88	97	68	82	93	65	61	77	50	108	108	95
	N_2-N_3	70	121	59	63	78	60	57	58	52	55	54	55	84	144	72
	N_3-N_1	75	147	123	103	75	98	98	129	120	89	95	88	74	146	128
	N_3-N_2	56	123	80	62	74	64	62	79	65	54	67	53	55	129	90
M_3	N_1-N_2	110	101	79	78	75	63	70	73	62	69	73	58	137	112	96
	N_1-N_3	141	133	113	73	103	64	52	157	59	67	106	61	140	132	124
	N_2-N_1	57	117	71	54	77	64	56	55	65	46	56	46	64	148	84
	N_2-N_3	70	93	62	64	75	63	66	60	49	58	62	62	78	108	70
	N_3-N_1	65	197	58	58	167	56	60	138	51	57	155	55	62	203	58
	N_3-N_2	80	105	67	75	86	62	61	78	77	67	71	61	90	119	72
m_4	N_1-N_2	60	130	59	56	81	57	65	98	61	51	67	48	65	155	67
	N_1-N_3	54	270	53	57	154	58	58	111	70	54	151	54	50	260	54
	N_2-N_1	72	127	75	64	75	63	66	59	61	53	55	58	85	160	92
	N_2-N_3	58	130	58	55	69	53	64	42	47	38	49	37	66	157	72
	N_3-N_1	47	173	54	75	148	79	68	275	74	65	137	65	48	179	54
	N_3-N_2	51	126	61	62	90	62	67	86	56	46	64	46	57	150	74
m_5	N_1-N_2	56	141	75	56	112	63	54	70	53	47	82	49	60	165	88
	N_1-N_3	50	263	101	56	174	51	62	153	54	49	196	44	51	273	101
	N_2-N_1	118	87	78	111	88	68	114	134	69	68	72	52	153	95	96
	N_2-N_3	50	92	59	51	77	52	65	102	49	31	44	32	59	113	69
	N_3-N_1	136	121	125	153	90	130	161	125	127	116	93	101	149	124	132
	N_3-N_2	79	96	75	100	88	57	119	117	57	59	62	38	94	104	88
m_6	N_1-N_2	101	88	90	93	83	81	87	84	95	81	80	82	123	90	102
	N_1-N_3	107	94	97	94	97	86	95	158	75	84	98	81	112	87	100
	N_2-N_1	92	123	83	84	89	76	87	91	67	81	82	79	103	157	94
	N_2-N_3	75	87	79	70	81	74	55	69	60	64	73	73	79	91	90
	N_3-N_1	176	194	171	135	247	106	106	116	96	107	157	97	196	214	196
	N_3-N_2	129	105	111	110	105	87	90	72	76	84	86	76	186	123	154
m_7	N_1-N_2	102	71	66	98	78	83	113	57	97	74	69	66	134	66	65
	N_1-N_3	126	58	55	124	94	97	109	75	106	98	68	80	137	53	53
	N_2-N_1	89	71	66	92	74	73	89	74	76	66	65	63	111	69	68
	N_2-N_3	85	71	64	89	69	68	88	76	56	69	68	63	105	69	68
	N_3-N_1	126	49	53	117	76	98	90	54	108	101	72	77	131	46	53
	N_3-N_2	102	70	66	100	76	80	120	59	85	76	75	63	128	66	66
m_8	N_1-N_2	114	84	91	123	106	87	125	117	68	82	83	67	138	84	107
	N_1-N_3	148	111	141	180	163	162	157	138	140	138	119	120	157	114	149
	N_2-N_1	60	92	61	61	90	57	57	112	51	48	66	47	69	110	66
	N_2-N_3	118	90	101	104	94	87	115	119	92	82	87	74	149	95	129
	N_3-N_1	54	155	120	61	115	59	71	116	69	59	113	56	50	157	131
	N_3-N_2	70	105	93	67	98	66	78	120	62	69	85	70	68	125	107
	Mean	87	123	81	85	104	75	83	102	74	71	90	65	98	137	91
	Median	80	117	75	77	90	68	77	92	67	68	79	63	91	127	89

Note: MSE of kernel matching (kernel), local linear matching (loclin) and ridge matching (ridge). MSE is given relative to the MSE of pair-matching (in%). Bandwidth values chosen by Akaike penalised cross-validation. The first column indicates the regression curve. The second column indicates the density combination. The rows 'Mean' and 'Median' give the mean and median, respectively, over the 48 different designs.

Table 2. *MSE of matching with CV bandwidth selection, relative to optimal bandwidth*

		$n_0 = n_1 = 40$			$n_0 = n_1 = 200$			$n_0 = n_1 = 1000$			$n_0 = 200, n_1 = 40$			$n_0 = 40, n_1 = 200$		
	Densities	Kernel	Loclin	Ridge	Kernel	Loclin	Ridge	Kernel	Loclin	Ridge	Kernel	Loclin	Ridge	Kernel	Loclin	Ridge
m_1	N_1-N_2	108	127	119	103	118	109	112	78	90	119	120	133	103	141	111
	N_1-N_3	104	165	114	106	124	107	113	127	126	115	156	147	105	170	109
	N_2-N_1	121	169	131	103	113	131	108	93	120	119	123	169	111	197	108
	N_2-N_3	108	186	121	104	122	123	100	79	126	134	200	199	105	182	99
	N_3-N_1	258	133	97	334	235	118	141	218	127	350	237	134	250	134	127
	N_3-N_2	109	114	105	116	124	114	100	142	133	149	175	166	108	111	107
m_2	N_1-N_2	169	139	126	101	127	104	107	92	107	111	107	112	140	117	118
	N_1-N_3	116	190	116	102	198	111	121	154	170	101	245	126	112	176	106
	N_2-N_1	126	113	120	120	122	111	114	138	109	98	120	118	114	113	105
	N_2-N_3	108	188	111	102	114	105	119	90	92	103	101	124	113	168	105
	N_3-N_1	792	203	164	105	107	103	101	150	153	112	141	104	751	239	164
	N_3-N_2	133	141	151	160	105	114	133	108	132	203	118	133	116	143	158
m_3	N_1-N_2	125	147	116	104	97	111	131	94	100	105	98	122	129	153	112
	N_1-N_3	137	183	121	100	141	122	80	208	106	100	122	123	131	208	127
	N_2-N_1	111	127	122	104	103	107	101	88	123	118	98	98	106	132	128
	N_2-N_3	105	131	103	101	104	101	113	78	79	100	112	119	111	140	103
	N_3-N_1	106	180	113	106	169	111	99	197	114	108	198	117	103	190	114
	N_3-N_2	108	139	104	107	114	103	98	104	116	102	119	122	114	142	99
m_4	N_1-N_2	129	205	111	126	107	114	141	136	132	262	154	166	101	173	107
	N_1-N_3	266	418	115	312	186	131	367	134	140	392	238	134	232	388	141
	N_2-N_1	167	149	105	104	114	105	114	114	92	133	126	174	137	131	120
	N_2-N_3	143	234	110	139	117	110	116	84	114	220	171	156	115	212	105
	N_3-N_1	476	426	158	297	102	102	93	190	104	364	112	96	465	416	183
	N_3-N_2	195	201	109	184	107	103	157	130	95	319	124	124	174	183	146
m_5	N_1-N_2	115	165	137	113	102	112	94	86	90	124	116	105	107	166	142
	N_1-N_3	185	232	255	258	257	121	300	291	91	246	342	118	185	233	281
	N_2-N_1	143	135	106	152	113	108	169	180	99	108	135	119	166	120	107
	N_2-N_3	149	141	121	159	116	104	121	238	98	189	109	108	138	142	117
	N_3-N_1	141	156	135	133	136	121	142	191	176	124	157	117	149	162	135
	N_3-N_2	110	141	126	133	117	101	218	179	90	110	138	113	113	124	122
m_6	N_1-N_2	102	100	105	103	102	101	115	101	132	101	113	119	104	110	107
	N_1-N_3	100	128	152	98	139	160	121	178	97	101	129	132	101	169	177
	N_2-N_1	105	146	103	106	115	105	108	103	85	107	116	122	116	171	106
	N_2-N_3	107	112	107	100	108	104	71	100	85	101	102	109	103	124	118
	N_3-N_1	109	239	114	104	169	100	106	138	148	106	145	113	108	277	116
	N_3-N_2	108	122	106	108	117	118	97	95	98	103	108	124	119	123	108
m_7	N_1-N_2	110	109	103	113	110	102	185	79	133	103	101	101	122	110	101
	N_1-N_3	108	135	109	113	164	100	111	99	118	107	107	100	111	149	107
	N_2-N_1	115	105	101	129	107	105	125	134	134	103	99	101	132	109	101
	N_2-N_3	111	108	101	129	101	99	120	108	77	110	102	102	128	113	104
	N_3-N_1	111	123	109	107	135	102	104	86	122	108	118	97	109	134	105
	N_3-N_2	108	107	103	118	109	102	155	93	113	103	105	92	116	110	98
m_8	N_1-N_2	106	105	118	108	116	137	122	132	97	102	115	142	114	105	111
	N_1-N_3	111	178	113	104	306	107	88	195	98	103	190	104	115	190	113
	N_2-N_1	151	118	117	155	122	104	99	174	84	227	113	104	128	134	106
	N_2-N_3	113	127	119	122	107	125	144	135	149	103	118	128	118	112	113
	N_3-N_1	155	202	290	221	139	118	229	168	134	192	157	119	147	221	351
	N_3-N_2	107	135	140	106	127	102	133	169	85	113	110	100	105	187	158
	Mean	150	162	124	134	131	111	130	135	113	146	139	123	146	166	127
	Median	112	141	115	108	116	107	115	131	111	110	119	119	115	146	111

Note: MSE of kernel matching (kernel), local linear matching (loclin) and ridge matching (ridge). MSE is given relative to the MSE at the optimal bandwidth value (in%). Bandwidth values chosen by Akaike penalised cross-validation.See note below Table 1.

bandwidth value. The relative MSE is given in percent. The first columns refer to the symmetric samples sizes $n_0 = n_1 = 40$ and $n_0 = n_1 = 200$ and $n_0 = n_1 = 1000$. The last columns refer to the asymmetric samples sizes: $n_0 = 200, n_1 = 40$ and vice versa. $n_0 = 200, n_1 = 40$ refers to a situation where the source sample is much larger than the target sample. This corresponds to the usual situation in treatment evaluation, where the number of control observations is much larger than the number of treated individuals. On the other hand, $n_0 = 40, n_1 = 200$ refers to a situation where only $n_0 = 40$ control observations are available, which are matched to $n_1 = 200$ treated observations. The rows refer to the different regression curves m_1 to m_8 and the 6 density combinations. The last two rows give the average over all simulation designs.

In Table 1, entries below 100% indicate that the respective matching estimator with cross-validation bandwidth selection performed better than pair-matching, whereas entries above indicate a worse result. As a general finding, kernel and ridge matching performed usually better than pair-matching. On the other hand, local linear matching is often worse than pair-matching, except for the sample size ($n_0 = 200, n_1 = 40$), where the source sample is much larger than the target sample. Another exception is the linear regression curve m_7, where local linear matching achieves substantial reductions in MSE by choosing large bandwidth values. Ridge matching has for all sample size configurations always the lowest MSE on average. For symmetric sample sizes ($n_0 = n_1$), its average efficiency gains vis-à-vis pair-matching increase from about 19% (median 25%) for sample size 40 to about 26% (median 33%) for sample size 1000. Kernel and local linear matching improve their relative position to pair-matching with growing sample size, too. These improvements, however, are much less pronounced for kernel matching: the MSE reductions are 13% in samples of size 40 and 17% for sample size 1000. Local linear matching performs significantly worse than pair-matching in small samples and seems to break even only at sample size 1000. For non-symmetric sample sizes, all local polynomial estimators become by another 10%-points more precise if the source sample is much larger than the target sample ($n_0 = 200, n_1 = 40$). The MSE of ridge matching is then about 35% below that of pair-matching. On the other hand, if the source sample is smaller than the target sample and thus the number of control observations is small ($n_0 = 40, n_1 = 200$), pair-matching becomes relatively more efficient because it uses the few control observations repeatedly. Only ridge matching still realizes significant reductions in MSE of about 10% vis-à-vis pair-matching.

In addition to these average precision gains, ridge matching further performs only rarely much worse than pair-matching. For sample sizes 40 and 200, its MSE is only in 2 of the 48 simulation designs more than 30% larger than the MSE of pair-matching, only once for sample size 1000, never in the sample size combination 200–40, and 5 times for sample size 40–200. For kernel matching these frequencies are more than twice as large, and they are furthermore much larger for local linear matching, which performs even for sample sizes 1000 and the favourable sample size combination $n = 200, n_1 = 40$ in more than 8 out of 48 simulation designs by more than 30% worse than pair-matching. This demonstrates that ridge matching not only performs better on average, but also that it is rather robust to the simulation design.

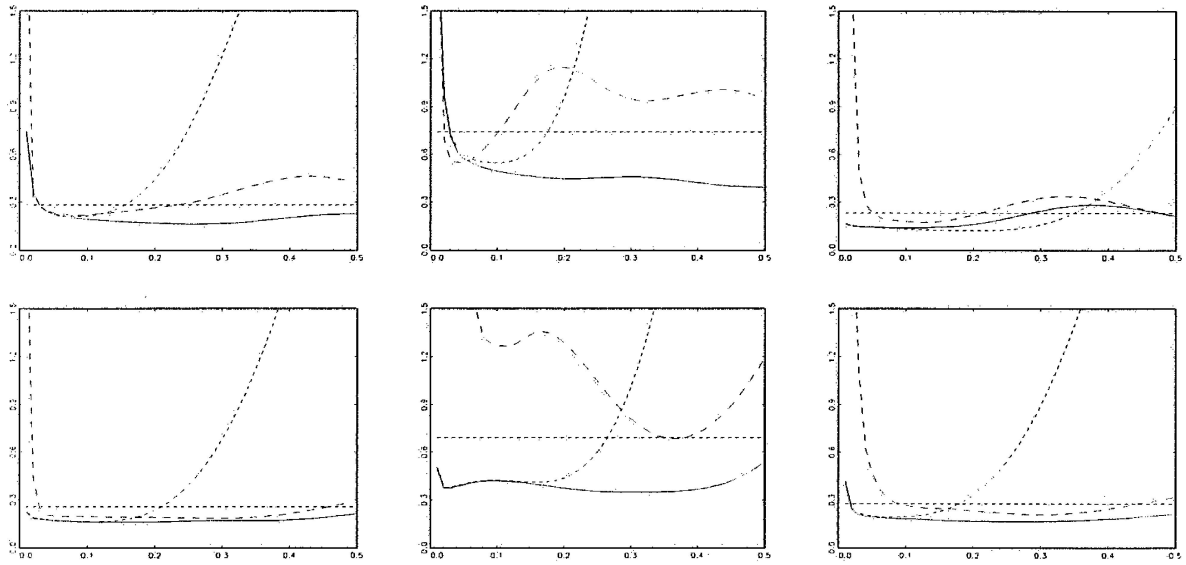
Table 1 showed that local polynomial matching estimators with cross-validation bandwidth choice can indeed yield more precise estimates than pair-matching. From the previous sections, however, we know that cross-validation does not lead to optimal bandwidth choices. To assess the efficiency loss due to the cross-validation bandwidth choice, Table 2 gives the MSE of the various matching estimators relative to the MSE at their *optimal* bandwidth values.[19] This indicates the potential for the development of improved bandwidth selectors. A value of 150%, for example, indicates that matching with cross-validation leads to a 50% higher MSE compared to a situation where the optimal bandwidth is known. Notice that several entries in Table 2 are smaller than 100%, indicating a lower MSE with cross-validation than at the simulated optimal bandwidth. Although this is largely due to noise in the simulations (particularly for sample size 1000 with only relatively few replications), ratios smaller than 100% could indeed occur, because the data-driven bandwidth selector chooses the bandwidth conditional on a given dataset, whereas the simulated optimal bandwidths are unconditional.

For small samples ($n_0 = 40$ or $n_1 = 40$), the relative MSE is around 150% for kernel matching, 160% for local linear matching and 125% for ridge matching. This efficiency ratio improves markedly for all three estimators when the sample size increases from 40 to 200, but does not improve further when the sample size is increased to 1000. For samples of size 200 or 1000, the MSE of kernel and local linear matching is about 30% higher than at the optimal bandwidth. For ridge matching it is about 15% higher.

Very large efficiency losses often occur for kernel and local linear matching with the more difficult density combinations (N_1, N_3) and (N_3, N_1), usually together with a bandwidth choice below the optimal bandwidth value. In general, the bandwidths chosen by cross-validation are on average smaller than the optimal bandwidths[20] for kernel and local linear matching in all sample size combinations, while no such clear pattern can be detected for ridge matching. However, this does not imply that selecting larger bandwidths would generally have been preferable for kernel or local linear matching. Hence, the development of better bandwidth selectors might be worthwhile for the kernel and local linear matching estimators, but its scope seems to be limited for ridge matching.

4.1. *Sensitivity to the bandwidth value*

The previous simulations with cross-validation bandwidth selection indicated a superior performance of kernel and ridge matching over local linear matching. Ridge matching, in particular, was most robust to the simulation design. The relatively weak performance of local linear matching, however, cannot



*Abscissa: bandwidth h from 0.01 to 0.50, Ordinate: MSE*1000; Pair-matching (horizontal line), kernel matching (short-dashed), local linear matching (long-dashed), and ridge matching (solid). Density combinations* (N_1, N_2), (N_1, N_3), (N_2, N_1) *from top left to top right picture, density combinations* (N_2, N_3), (N_3, N_1), (N_3, N_2) *from bottom left to bottom right picture.*

Fig. 6. *Simulated MSE for regression curve* m_3 *(sample size* $n_0 = n_1 = 200$*)*

be explained by first-order asymptotic theory, since local linear regression and ridge regression are asymptotically equivalent (see appendix). In addition, Section 3 already indicated that the asymptotic expressions may not be very reliable in samples of this size.

The reason for the better finite-sample performance of ridge matching seem to be the variance problems of local linear regression with small bandwidth values. Seifert and Gasser (1996) showed that the unconditional variance of the local linear regression estimator is infinite and that the conditional variance is unbounded in small samples. They demonstrated (page 269f) that the variance becomes extremely large for bandwidth values only slightly below the optimal bandwidth value (where optimal refers to mean integrated squared error). For the matching estimator, to achieve $\sqrt{n}$-consistency some undersmoothing with respect to the MISE-optimal bandwidth is required, which aggravates even further the problems with small bandwidth values. By adding a ridge term to the denominator, the conditional variance of the local linear ridge regression estimator becomes bounded. Therefore, the MSE of ridge matching will not increase as much for small bandwidth values as it does for the pure local linear matching estimator. Hence, the MSE of ridge matching should generally be flatter and consequently less sensitive to the bandwidth choice. This is further examined below.

The finite-sample differences between the matching estimators with cross-validation bandwidth selection could originate from two sources: Differences in the minimum value of the MSE and differences in the curvature of the MSE around its minimum. Whereas the minimum value shows the potential of the estimator, the curvature indicates the risk when cross-validation does not find the optimal bandwidth. This latter aspect may be particularly relevant in small samples, since cross-validation is known for rather variable bandwidth choices. To gain an insight into the relevance of these two sources, Fig. 6 and the corresponding figures in the supplementary appendix show the simulated true MSE for different bandwidth values for the designs of Section 3. Figure 6 displays the MSE of pair-matching (horizontal line), kernel matching (short-dashed), local linear matching (long-dashed) and ridge matching (solid) for the regression curve m_3 and sample size $n_0 = n_1 = 200$.[21]

Some general patterns emerge from Fig. 6, which are also found in many of the graphs for the other simulation designs. The problems of local linear matching with small bandwidths become obvious as its MSE becomes very large for bandwidth values approaching zero due to the variance problems of local linear regression (Seifert and Gasser 1996). Although local linear matching often performs better than pair-matching for some bandwidths, it does so usually only in rather narrow bandwidth regions. Furthermore, the MSE of local linear matching often has local minima in the more difficult density combinations (N_1, N_3), (N_3, N_1) even with large sample sizes. This corresponds to the trade-off between choosing a small bandwidth to estimate $m(x)$ with low bias in regions of dense data and choosing a large bandwidth to estimate $m(x)$ with less variability in regions of sparse data. It might be difficult for a data-driven bandwidth selector to distinguish between these minima to find the global minimum.

For kernel and ridge matching the results appear more favourable. Frequently kernel matching outperforms pair-matching in bandwidth regions from about $h = 0.05$ to 0.15.

However, at larger bandwidths its MSE increases steeply. Ridge matching appears much less sensitive to bandwidth choice and its MSE explodes less often at very small or large bandwidth values. Commonly its MSE is quite flat and lies below the MSE of pair-matching in bandwidth regions from $h = 0.03$ to 0.25.

Another way to assess robustness is to look at the average MSE in a region around the optimal bandwidth. Considering a neighbourhood of span 0.2 around h^{opt}, i.e. if h were drawn randomly from the interval $[h_{opt} - 0.1, h_{opt} + 0.1]$, the average efficiency gain of kernel matching relative to pair-matching is still 8% (median 16%) and for ridge matching 25% (median 34%) for sample size 200. In contrast, local linear matching would lead on average to a 71% *higher* MSE than pair-matching (median increase −2%). These figures are similar for samples of size 40 when allowing for a neighbourhood of span 0.4 around h^{opt} and for sample size 1000 with neighbourhood-span 0.15. For the sample size combination $(n_0 = 40, n_1 = 200)$ the robustness of ridge matching to bandwidth choice becomes even more apparent. If the bandwidth is randomly selected from the interval $h_{opt} \pm 0.2$ the MSE of ridge matching is on average still 13% (median 26%) lower than the MSE of pair-matching, whereas kernel matching on average would have a 12% (median increase 6%) larger MSE, and the MSE of local linear matching would be $2\frac{1}{2}$ times the MSE of pair-matching (median increase 15%).

Hence, the better small-sample performance of ridge matching seems largely be due to the lower curvature of its MSE.

5. Matching with estimated propensity score

Propensity score matching is a convenient and popular method when a large number of characteristics $Z \in \Re^k$ needs to be adjusted for. In practice, however, the propensity score (6) is usually unknown and needs to be estimated. In this section, the performance of pair, kernel, local linear and ridge matching with an estimated propensity score $\hat{X}$ is examined.

Samples of covariates Z_1, Z_2, Z_3 are drawn and the observations are assigned to the source and the target sample according to the selection rule

$$D_i = 1(\alpha + Z_i'\beta + \varepsilon \geq 0),$$

with ε a random error term. Observations with $D_i = 0$ belong to the source population and observations with $D_i = 1$ belong to the target population. The propensity score is given by

$$P(D = 1 \mid Z),$$

and is estimated by maximum likelihood probit.

Z_1 is a $\chi^2_{(1)}$ random variable, Z_2 is $\text{uniform}_{[0,1]}$ distributed, and Z_3 is either binary or normally distributed. Six different selection rules are examined, see Table 3. The ratio n_0/n_1 of source to target observations is random and around one.

Table 3. *Selection equation*

Model	Selection equation	Z_{3i}	ε_i
1	$D_i = 1(Z_{1i} + Z_{2i} + Z_{3i} - 0.5 < \varepsilon_i)$	$N(0,1)$	$N(0,4)$
2	$D_i = 1(Z_{1i} - Z_{2i} + 2.5Z_{3i} - 0.5 < \varepsilon_i)$	D	$U(0,12)$
3	$D_i = 1(Z_{1i} - Z_{2i} + Z_{3i} - 0.5 < \varepsilon_i)$	$N(0,1)$	$N(0,4)$
4	$D_i = 1(Z_{1i} + Z_{2i} - Z_{3i} - 0.5 < \varepsilon_i)$	$N(0,1)$	$N(0,4)$
5	$D_i = 1(-Z_{1i} - Z_{2i} + 2.5Z_{3i} < \varepsilon_i)$	D	$U(0,12)$
6	$D_i = 1(2Z_{1i} - Z_{2i} + 2.5Z_{3i} - 1.5 < \varepsilon_i)$	D	$U(0, \frac{49}{3})$

Note: Error $N(0, \sigma^2)$ stands for normal mean-zero random errors with variance σ^2, error $U(0, \sigma^2)$ denotes a uniform random error term with mean zero and variance σ^2. Variable Z_1 is $\chi^2_{(1)}$ divided by $\sqrt{2}$, variable Z_2 is uniform $U(0, 1)$, and variable Z_3 is normal in models 1, 3, and 4 and a dummy variable in the other models.

Table 4. *Outcome equations*

Model	Outcome equation with normal error
1	$Y_i = Z_{1i}Z_{2i} + Z_{3i}^2 + \sqrt{Z_{1i}} + u_i$
2	$Y_i = -Z_{1i} + Z_{2i} + u_i$
3	$Y_i = Z_{1i} \cdot 1(Z_{3i} > Z_{2i}) + u_i$
4	$Y_i = Z_{1i}Z_{3i} + Z_{2i}^2 + \sqrt{Z_{1i}} + u_i$
5	$Y_i = -Z_{2i} + Z_{3i} + u_i$
6	$Y_i = Z_{1i} + Z_{3i} \cdot 1(Z_{1i} > Z_{2i}) + Z_{2i} + u_i$

Note: The outcome variable Y_i is observed only for the non-participants, i.e. the observations with $D_i = 0$. The error term $u_i \sim N(0, 1)$.

The outcome variable is generated from one of the six regression curves of Table 4, disturbed by an additive normal error term.

For these 36 different simulation designs, the mean squared error of pair, kernel, local linear and ridge matching with estimated propensity score is simulated.[22] Samples of size $n = n_0 + n_1 = 200$, 500, and 2000, respectively, are considered. The simulation results relative to the MSE of pair-matching (in percent) are given in Table 5, where the columns one and two indicate the selection rule and the regression curve. More details can be found in Tables C1 to C3 in the supplementary appendix.

For all sample sizes, the local polynomial matching estimators are usually substantially superior to pair-matching. The relative efficiency of kernel and ridge matching vis-à-vis pair-matching seems to decrease somewhat with growing sample size from reductions in MSE of 32% (42%) at sample size 200 to 24% (35%) at sample size 2000 for kernel matching (ridge matching). On the other hand, the efficiency gains of local linear matching are stable at about 21%. Although kernel matching has on average a lower MSE than local linear matching, it seems to be less robust to the selection rule and regression curve, as it performs in about 7–9 out of the 36 designs worse than pair-matching. With local linear matching this occurs only in 3–4 designs. The MSE of ridge matching is never larger than that of pair-matching for

Table 5. *MSE of matching with estimated propensity score, relative to pair-matching*

		$n_0 = n_1 = 200$			$n_0 = n_1 = 5000$			$n_0 = n_1 = 2000$		
D_i	Y_i	Kernal	Locin	Ridge	Kernal	Locin	Ridge	Kernal	Locin	Ridge
1	1	93	108	71	79	129	73	78	143	76
	2	31	77	52	38	86	60	68	91	81
	3	45	75	46	55	76	43	80	74	50
	4	106	93	69	114	91	67	114	83	66
	5	22	64	42	22	64	41	20	52	40
	6	102	94	91	106	90	123	124	77	177
2	1	59	59	51	55	61	53	46	48	62
	2	67	72	67	64	74	63	50	73	61
	3	64	61	51	54	61	51	48	46	44
	4	57	76	56	55	66	53	49	90	50
	5	89	64	62	76	64	62	74	62	61
	6	67	64	54	64	64	56	57	48	51
3	1	37	99	42	43	130	43	52	188	48
	2	117	65	66	135	67	76	177	90	117
	3	103	83	63	132	101	69	177	76	85
	4	105	88	69	117	79	68	107	78	77
	5	60	85	47	52	90	45	51	94	44
	6	60	85	63	89	79	67	123	55	50
4	1	91	103	72	83	128	72	69	177	77
	2	31	80	52	38	93	63	86	83	82
	3	24	77	42	22	85	37	19	55	32
	4	30	78	41	29	88	39	32	81	40
	5	61	89	48	53	79	44	51	88	45
	6	103	74	54	120	73	54	123	74	67
5	1	66	96	62	60	79	60	56	72	75
	2	68	85	63	72	77	65	71	69	67
	3	39	66	52	43	69	49	43	63	49
	4	65	103	57	66	89	59	85	52	69
	5	98	83	78	96	76	89	113	74	97
	6	66	86	67	66	75	63	58	71	70
6	1	76	57	51	71	57	52	90	62	57
	2	67	68	63	65	64	60	65	65	63
	3	54	63	53	51	59	51	49	60	56
	4	57	70	56	56	70	56	61	85	48
	5	108	61	59	100	68	61	113	70	62
	6	63	71	56	63	66	54	65	61	60
	Mean	68	78	58	70	80	60	76	79	65
	Median	65	77	56	64	76	60	67	74	61

Note: MSE of kernel matching (kernal), local linear matching (loclin) and ridge matching (ridge). MSE is given relative to the MSE of pair-matching (in%). Bandwidth values chosen by Akaike penalised cross-validation. The first column indicates the selection rule D_i. The second column indicates the outcome equation Y_i. the rows 'Mean' and 'Median' give the mean and median, respectively, over the 36 different designs.

sample size 200, only once for sample size 500, and twice for sample size 2000.

Concerning variance and bias, the results are similar to the previous. Pair-matching and local linear matching are nearly unbiased, whereas about 25% of the MSE of kernel matching and about 20% of the MSE of ridge matching are due to squared bias. Given their lower average MSE, this indicates again that reducing local variance at the cost of incurring bias could be beneficial. For detecting systematic under- or oversmoothing of the cross-validation bandwidth selector, the local polynomial matching estimators are also evaluated at 0.7, 0.8, 0.9, 1.1, 1.2, and 1.3 times the bandwidth selected by cross-validation. For kernel matching smaller bandwidths would on average have been preferable, whereas local linear matching would have been slightly better off with larger bandwidths. Again, ridge matching is hardly affected by changes in the bandwidth value.

6. Conclusions

In this paper, the problem of optimal bandwidth choice has been analyzed and the finite-sample properties of various matching estimators been examined. An asymptotic approximation to their MSE has been derived and its accuracy in finite samples investigated. However, the approximations did not appear to be sufficiently reliable in small samples for being useful as the basis for a plug-in bandwidth selector.

On the other hand, conventional *cross-validation* turned out to be a quite fruitful method for bandwidth selection, albeit not being asymptotically optimal. Particularly, matching based on a modified local linear regression estimator of Seifert and Gasser (1996, 2000) appeared to be rather insensitive to bandwidth choice. Moreover, the potential for further efficiency gains through better bandwidth selection seems to be rather limited for ridge matching. As another main result of this paper, the relative ordering of the various estimators in terms of mean squared error remained remarkably stable across sample sizes and simulation schemes: *ridge matching* always turned out to be superior to all other estimators, followed by kernel matching. The relative ordering among pair-matching, local linear matching and weighting estimators is less clear-cut. Local linear matching is susceptible to regions of sparse data and sensitive to bandwidth choice. The weighting estimator is sensitive to trimming and its relative performance worsens with increasing sample size. Weighting without trimming, however, fails completely, and further usage of the weighting estimator would require the development of an appropriate method for estimating the optimal trimming level.

The MSE of ridge matching was on average about 25% smaller than the MSE of pair-matching when matching on an observed covariate. On the other hand, when matching on an estimated propensity score, the reduction in MSE is about 40%. Hence, pair-matching performs even worse with an estimated propensity score. Pair-matching becomes less precise (relative to all other estimators) when matching on estimated covariates, because it compares each target sample observation with only *one* source sample observation. Although the observations within each matched pair are supposed to have identical characteristics, they might be rather different if the propensity score is imprecisely estimated. Hence matching each target sample observation to many source sample observations (as in local polynomial matching) reduces not only the susceptibility of the estimate with respect to the variability in Y but also with respect to the variance of the estimated propensity scores. For kernel matching these precision gains are about 15 and 30%, respectively, but with scope for improvement through better bandwidth selection.

A reduction in MSE of about 40% means that pair-matching needs almost 70% more observations to achieve the same precision as ridge matching. If the source sample is larger than the target sample, which is often the case in treatment evaluation with a large control sample, the precision gains of local polynomial vis-a-vis pair-matching are even larger.

Appendix

Asymptotic MSE of ridge matching

In the following the mean squared error for ridge regression is examined. The ridge regression estimator (5)

$$\hat{m}_{\text{Ridge}}(x) = \frac{T_0}{S_0} + \frac{\delta T_1}{S_2 + \tau},$$

with $\delta = x - \bar{x}$, can be written as

$$= \frac{(s_2 + \tau)t_0 - s_1 t_1}{(s_2 + \tau)s_0 - s_1^2},$$

where $s_r = \sum K(\frac{X_i^0 - x}{h})(X_i^0 - x)^r$ and $t_r = \sum Y_i^0 K(\frac{X_i^0 - x}{h})(X_i^0 - x)^r$ are centered at x only.[23] Define further $\bar{s}_r = \sum K^2(\frac{X_i^0 - x}{h})(X_i^0 - x)^r$. For $h \to 0$ and with standard regularity conditions, see e.g. Fan and Gijbels (1996),

$$\begin{aligned} s_r &= E[s_r] + O_p(\sqrt{\text{Var}(s_r)}) \\ &= nh^{r+1}(f\mu_r + hf'\mu_{r+1} + O_p(h^2 + 1/\sqrt{nh})), \end{aligned}$$

and

$$\begin{aligned} \bar{s}_r &= E[\bar{s}_r] + O_p(\sqrt{\text{Var}(\bar{s}_r)}) \\ &= nh^{r+1}(f\bar{\mu}_r + hf'\bar{\mu}_{r+1} + O_p(h^2 + 1/\sqrt{nh})), \end{aligned}$$

where f is shorthand for $f_0(x)$ and m for $m(x)$. For a symmetric kernel μ_r is zero for r odd, and $\mu_0 = 1$ for a kernel integrating to one.

For the terms t_r, the conditional expectations and covariances are

$$\begin{aligned} E\big[t_r \mid X_1^0, \ldots, X_{n_0}^0\big] &= \sum K\left(\frac{X_i^0 - x}{h}\right)\big(X_i^0 - x\big)^r m\big(X_i^0\big) \\ &= \sum K\left(\frac{X_i^0 - x}{h}\right)\big(X_i^0 - x\big)^r \Big(m(x) + m'(x)\big(X_i^0 - x\big) \\ &\quad + \frac{m''(x)}{2}\big(X_i^0 - x\big)^2 + O_p\big(\big(X_i^0 - x\big)^3\big)\Big) \\ &= m(x)s_r + m'(x)s_{r+1} + \frac{m''(x)}{2}s_{r+2} + O_p(s_{r+3}), \end{aligned}$$

and

$$\begin{aligned} &\text{Cov}\big(t_k, t_l \mid X_1^0, \ldots, X_{n_0}^0\big) \\ &\quad = \sum K^2\left(\frac{X_i^0 - x}{h}\right)\big(X_i^0 - x\big)^{k+l}\sigma^2\big(X_i^0\big) = \sigma^2 \bar{s}_{k+l}, \end{aligned}$$

under the assumption of a constant variance $\sigma^2\big(X_i^0\big) = \sigma^2$.

With these preliminaries and after some algebra, the conditional bias of $\hat{m}_{\text{Ridge}}$ given the sample is given by

$$E\big[\hat{m}_{\text{Ridge}}(x) - m(x) \mid X_1^0, \ldots, X_{n_0}^0\big] = \frac{(s_2+\tau)E[t_0] - s_1 E[t_1]}{(s_2+\tau)s_0 - s_1^2} - m(x)$$

$$= \frac{(s_2+\tau)\big(ms_0 + m's_1 + \frac{m''}{2}s_2 + O_p(s_3)\big) - s_1\big(ms_1 + m's_2 + \frac{m''}{2}s_3 + O_p(s_4)\big)}{(s_2+\tau)s_0 - s_1^2} - m(x)$$

$$= h^2 \frac{h^2\frac{m''}{2}\big(f^2\mu_2^2 + O_p(h^2)\big) + \frac{\tau\mu_2}{nh}\big(m'f' + \frac{m''}{2}f + O_p(h)\big)}{h^2(f^2\mu_2 + O_p(h^2)) + \frac{\tau}{nh}(f + O_p(h^2))}$$

$$= h^2 \frac{h^2\frac{m''f^2\mu_2^2}{2} + O_p(h^4) + O_p\big(\frac{\tau}{nh}\big)}{h^2 f^2\mu_2 + O_p(h^4) + O_p\big(\frac{\tau}{nh}\big)},$$

and the conditional variance is

$$\text{Var}\big[\hat{m}_{\text{Ridge}}(x) \mid X_1^0, \ldots, X_{n_0}^0\big] = \frac{(s_2+\tau)^2\text{Var}(t_0) + s_1^2\text{Var}(t_1) - 2(s_2+\tau)s_1 Cov(t_0, t_1)}{\big((s_2+\tau)s_0 - s_1^2\big)^2}$$

$$= \sigma^2 \frac{(s_2+\tau)^2\bar{s}_0 + s_1^2\bar{s}_2 - 2(s_2+\tau)s_1\bar{s}_1}{\big((s_2+\tau)s_0 - s_1^2\big)^2}$$

$$= \frac{\sigma^2}{nh} \frac{h^4\mu_2^2 f^2(f\bar{\mu}_0 - hf'\bar{\mu}_1 + O_p(h^2)) + \frac{2\tau}{nh}h^2\mu_2(f^2\bar{\mu}_0 + O_p(h^2)) + \frac{\tau^2}{(nh)^2}(f\bar{\mu}_0 + hf'\bar{\mu}_1 + O_p(h^2))}{(h^2(f^2\mu_2 + O_p(h^2)) + \frac{\tau}{nh}(f + O_p(h^2)))^2}$$

$$= \frac{\sigma^2}{nh} \frac{h^4\mu_2^2 f^2(f\bar{\mu}_0 - hf'\bar{\mu}_1) + O_p(h^6) + O_p\big(\frac{\tau h^2}{nh}\big) + O_p\big(\frac{\tau^2}{n^2h^2}\big)}{\big(h^2 f^2\mu_2 + O_p(h^4) + O_p\big(\frac{\tau}{nh}\big)\big)^2}$$

If

$$\frac{\tau}{nh} = o_p(h^4), \tag{18}$$

the impact of the ridging term on the conditional bias and the conditional variance is overshadowed by the lower order terms $O_p(h^4)$ and $O_p(h^6)$ in the above expressions. In this case, ridging has no effect on the first-order asymptotic approximations to mean and variance.

For the particular choice of the ridge parameter

$$\tau = \frac{5}{16}h \cdot |\delta|$$

and using $\delta = -s_1/s_0$, it follows

$$\tau = \frac{5h}{16}\left|\frac{s_1}{s_0}\right| = \frac{5}{16}h^3\left|\frac{f'\mu_2 + O_p(h)}{f + O_p(h^2)}\right| = O_p(h^3).$$

Hence, for this ridge parameter, the condition (18) holds if

$$\frac{1}{nh} = o(h). \tag{19}$$

In nonparametric regression usually h is chosen such that $\frac{1}{nh} = O(h^4)$ to equilibrate variance and squared bias. For achieving $\sqrt{n}$-consistency of the matching estimator, the squared bias has to be of order n, such that $\frac{1}{n} = O(h^4)$. In both cases, the condition (19) is satisfied and ridge regression is asymptotically equivalent to local linear regression.

Acknowledgment

I am grateful for comments and suggestions to Yuanhua Feng, Bernd Fitzenberger, Michael Lechner, the editor and an anonymous referee. This research was supported by the Swiss National Science Foundation (project NSF 4043-058311) and the Grundlagenforschungsfonds HSG (project G02110112).

Notes

1. See e.g. Heckman, Ichimura and Todd (1997, 1998), Heckman *et al.* (1998) and the Symposium on the Econometrics of Matching of the Review of Economics and Statistics, 86 (2004).
2. X must contain all variables that affected D as well as Y^0. This is also known as *selection* on *observables* (Heckman and Robb 1985), *ignorable treatment assignment* (Rosenbaum and Rubin 1983), or *conditional independence assumption* (Lechner 1999).
3. In treatment evaluation, it is the probability of being assigned to treatment given the covariates x.
4. With respect to the semiparametric efficiency bound derived by Hahn (1998).
5. The results for the weighting estimators as well as additional simulation results are given in a supplementary appendix, available on the author's internet page.

6. X could be an index of several characteristics. In particular, X could be the propensity score, as discussed below. Hence, the restriction for X to be one-dimensional is not as restrictive as it appears.
7. The common support condition for identification requires that $f_0(x) > 0$ everywhere where $f_1(x) > 0$.
8. In applications pair-matching appears in two variants: matching with and without replacement. Matching without replacement reduces variance at the cost of a larger bias. However, matching without replacement is only possible if the source sample is larger than the target sample ($n_0 \geq n_1$), and it is likely to perform very poorly if $n_0 \approx n_1$. Matching without replacement is not further considered in this paper. For other pair-matching techniqes see e.g. Gu and Rosenbaum (1993).
9. The notation follows Seifert and Gasser (2000). Throughout the study always the Epanechnikov kernel $K(u) = \frac{3}{4}(1-u^2)1_{[-1,1]}(u)$ is used.
10. Local linear regression with an infinite bandwidth value corresponds to ordinary least squares regression. Therefore, least squares regression and pair-matching can be considered as the two extremes of matching estimators, where pair-matching uses the smallest local neighbourhood possible to estimate m, whereas least squares regression uses all observations equally.
11. In applications also other forms of matching on a one-dimensional index function, e.g. the Mahalanobis distance, are often observed.
12. The propensity score refers to the propensity of being in the target population. In treatment evaluation, the source population represents the non-treated and the target population the treated. Let D indicate whether a person got treated ($D = 1$) or not ($D = 0$). The treatment propensity given characteristics $\geq$ is $P(D = 1 \mid Z = z)$, which equals (6) by Bayes' theorem. As shown in Frölich (2002), the population size ratio $\frac{P_0}{P_1}$ in (6) is irrelevant for consistency of propensity score matching and can be set to any arbitrary number, e.g. $\frac{P_0}{P_1} = 1$.
13. In the supplementary appendix, also the small sample properties of Horvitz and Thompson (1952) weighting estimators are examined. Writing (1) as $E_1[Y] = \int m(x)\frac{f_1(x)}{f_0(x)} f_0(x)dx = E_0[Y\frac{f_1(X)}{f_0(X)}]$, then $E_1[Y]$ can be estimated by the *weighting estimator*

$$\widehat{E_1[Y]} = \frac{1}{n_0}\sum_{i=1}^{n_0} Y_i^0 \frac{\hat{f}_1(X_i^0)}{\hat{f}_0(X_i^0)}.$$

If covariate adjustment needs to account for multiple covariates $Z \in \Re^k$, then $E_1[Y] = \int E[Y \mid Z = z] \cdot f_{Z\mid 1}(z)\,dz = E_0[Y\frac{X}{1-X}]\frac{P_0}{P_1}$, and weighting by the propensity score X estimates $E_1[Y]$ as

$$\widehat{E_1[Y]} = \frac{1}{n_1}\sum_{i=1}^{n_0} Y_i^0 \frac{X_i^0}{1-X_i^0}.$$

14. Population regression curves m_1 to m_8: $m_1(x) = 0.4 + 0.25\sin(8x - 5) + 0.4\exp(-16(4x-2.5)^2)$, $m_2(x) = 0.5 - 4(x-0.2)^2 - 1.2\ln(1.1 - x)$, $m_3(x) = 0.2 + 2(x-0.9)^2 + 5(x-0.7)^3 + 100(x-0.6)^{10}$, $m_4(x) = 0.5 + 0.3e^{-2x}\sin(16x)$, $m_5(x) = 0.2 + \sqrt{x} - 0.6(x-0.1)^2$, $m_6(x) = -0.1 + 0.25(x+0.3)^{-1} + 0.4\exp(-24(x-0.25)^2) + 0.1\exp(-60(x - 0.75)^2)$, $m_7(x) = 0.15 + 0.7x$, $m_8(x) = 0.1 + 2(x-0.35)^2$.
15. For sample size $n_0 = n_1 = 40$, MSE is computed/simulated at the 50 bandwidth values $h = 0.02, 0.04, \ldots, 1.00$. For sample size 200, $h = 0.01, 0.02, \ldots, 0.50$. For sample size 1000, $h = 0.0075, \ldots, 0.375$.
16. For $\sqrt{n}$ consistency of the matching estimator, the squared bias has to be of order $O(\frac{1}{n})$. Cross-validation, however, chooses the bandwidth to balance squared bias and variance, such that bias is of order $O(\frac{1}{nh})$.
17. Also the performance of the penalized cross-validation bandwidth selectors of Akaike, Rice, and Shibata (see Pagan and Ullah (1999) p. 119) were compared and led to similar results. The Akaike penalised cross-validation selector chooses the bandwidth h as $\arg\min \exp(\frac{2}{n_0 h})\cdot\sum_{i=1}^{n_0}(Y_i^0 - \hat{m}(X_i^0; h))^2$.
18. More details can be found in Tables B1 to B5 in the supplementary appendix. As bandwidth search grid the same 50 bandwidth values as in the previous section are used. Number of replications is 10'000; for sample size 1000 only 100 replications. In the supplementary appendix, also additional results for the weighting estimator and for OLS are given.
19. I.e. relative to the minimum of their simulated MSE, as given for example in Figs. 4, 5 and 6. These minimum MSE values are given in Tables A1 to A5 in the supplementary appendix.
20. These are given in the Tables A1 to A5 in the supplementary appendix.
21. The MSE of kernel and of local linear matching are identical to the solid lines of Figs. 4 and 5.
22. Between 5'000 to 20'000 replications; for sample size 2000 only 200 replications. For kernel, local linear and ridge matching, the bandwidth is chosen by penalized cross-validation from the grid $h = 0.01, 0.02, \ldots, 0.80$.
23. This follows from $T_0 = t_0$ and $T_1 = t_1 + \delta t_0$ and $T_2 = t_2 + 2\delta t_1 + \delta^2 t_0$ and analogously for S_0, S_1 and S_2. Since furthermore, $S_1 = 0$ it follows with $S_1 = s_1 + \delta s_0$ that $\delta = -s_1/s_0$.

References

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

Angrist J. 1998. Estimating labour market impact of voluntary military service using social security data. Econometrica 66: 249–288.

Dehejia R. and Wahba S. 1999. Causal effects in non-experimental studies: Reevaluating the evaluation of training programmes. Journal of American Statistical Association 94: 1053–1062.

Fan J. 1993. Local linear regression smoothers and their minimax efficiency. Annals of Statistics 21: 196–216.

Fan J., Gasser T., Gijbels I., Brockmann M. and Engel J. 1997. Local polynomial regression: Optimal kernels and asymptotic minimax efficiency. Annals of the Institute of Mathematical Statistics 49: 79–99.

Fan J. and Gijbels I. 1996. Local Polynomial Modeling and its Applications. Chapman and Hall, London.

Frölich M. 2002. Propensity score matching without conditional independence assumption-with an application to the gender wage gap in the UK. mimeo, University of St. Gallen.

Gerfin M. and Lechner M. 2002. Microeconometric evaluation of the active labour market policy in switzerland. Economic Journal 112: 854–893.

Gu X. and Rosenbaum P. 1993. Comparison of multivariate matching methods: Structures, distance, and algorithms. Journal of Computational and Graphical Statistics 2: 405–420.

Hahn J. 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66: 315–331.

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

Heckman J., Ichimura H. and Todd P. 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. Review of Economic Studies 64: 605–654.

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

Heckman J. and Robb R. 1985. Alternative methods for evaluating the impact of interventions. In: Heckman J. and Singer B. (Eds.) Longitudinal Analysis of Labour Market Data, Cambridge University Press, Cambridge.

Hirano K., Imbens G. and Ridder G. 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.

Imbens G. 2000. The role of the propensity score in estimating dose-response functions. Biometrika 87: 706–710.

Lechner M. 1999. Earnings and employment effects of continuous off-the-job training in east germany after unification. Journal of Business and Economic Statistics 17: 74–90.

Little R. and Rubin D. 1987. Statistical Analysis with Missing Data. Wiley, New York.

Loader C. 1999. Bandwidth selection: Classical or plug-in?. Annals of Statistics 27: 415–438.

Pagan A. and A. Ullah 1999. Nonparametric Econometrics. Cambridge University Press. Cambridge.

Rosenbaum P. and Rubin D. 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70: 41–55.

Ruppert D. and Wand M. 1994. Multivariate locally weighted least squares regression. Annals of Statistics 22: 1346–1370.

Seifert B. and Gasser T. 1996. Finite-sample variance of local polynomials: Analysis and solutions. Journal of American Statistical Association 91: 267–275.

Seifert B. and Gasser T. 2000. Data adaptive ridging in local polynomial regression. Journal of Computational and Graphical Statistics 9: 338–360.