JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **24**: 163–185 (2009)
Published online 2 September 2008 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/jae.1026

RANDOM RECURSIVE PARTITIONING: A MATCHING METHOD FOR THE ESTIMATION OF THE AVERAGE TREATMENT EFFECT

GIUSEPPE PORRO[1]* AND STEFANO MARIA IACUS[2]

[1] *Department of Economics and Statistics, University of Trieste, P.le Europa 1, I-34127 Trieste, Italy*
[2] *Department of Economics, Business and Statistics, University of Milan, Via Conservatorio 7, I-20122 Milano, Italy*

SUMMARY

In this paper we introduce the Random Recursive Partitioning (RRP) matching method. RRP generates a proximity matrix which might be useful in econometric applications like average treatment effect estimation. RRP is a Monte Carlo method that randomly generates non-empty recursive partitions of the data and evaluates the proximity between two observations as the empirical frequency they fall in a same cell of these random partitions over all Monte Carlo replications. From the proximity matrix it is possible to derive both graphical and analytical tools to evaluate the extent of the common support between data sets. The RRP method is "honest" in that it does not match observations "at any cost": if data sets are separated, the method clearly states it.

The match obtained with RRP is invariant under monotonic transformation of the data. Average treatment effect estimators derived from the proximity matrix seem to be competitive compared to more commonly used estimators. RRP method does not require a particular structure of the data and for this reason it can be applied when distances like Mahalanobis or Euclidean are not suitable, in the presence of missing data or when the estimated propensity score is too sensitive to model specifications. Copyright © 2008 John Wiley & Sons, Ltd.

Received 1 October 2004; Revised 9 February 2007

1. INTRODUCTION

In social as well as in natural sciences researchers often need to estimate the effect of a "treatment" (like a social program, a medical therapy, etc) on a population of individuals. The estimation generally entails the comparison of an outcome variable between the group of the subjects exposed to the treatment and a "control" group which did not receive the treatment.

The evaluation of the differential treatment effect on the outcome for *treated versus control* individuals has to be done given the same pre-treatment conditions (see e.g. Heckman *et al.*, 1997, 1998). In experimental studies two similar groups of individuals are randomly selected from a population: one is then exposed to the treatment and the other is not. In observational studies the assignment to the treatment is usually non-randomized and, therefore, the distributions of pre-treatment covariates are different between the two groups. Finding control units similar to each treated individual represents the preliminary problem of the analysis. Therefore, a technique is required for *matching* observations coming from different data sets. If the match fails partially or

* Correspondence to: Professor Giuseppe Porro, Department of Economics and Statistics, University of Trieste, Piazzale Europa, 1 Trieste, 34127, Italy. E-mail: giuseppe.porro@econ.units.it

completely, it means that the distributions of the covariates in the two groups do not overlap. This is a case of (partial or total) lack of *common support* (see §2.2 and §2.4).

When there are many covariates the match might become an unfeasible task. Hence, since the seminal paper by Cochran and Rubin (1973), many authors have faced the matching problem and several matching techniques have been developed to overcome this dimensionality issue (see also Rubin 1973a,b). In their paper of 1973, Cochran and Rubin (among other contributions) proposed to solve the problem of multivariate matching using the Mahalanobis distance (Rubin, 1980) by matching the nearest available individuals. Later Rosenbaum and Rubin (1983) introduced the notion of *propensity score* (PS) as the probability that an individual receives the treatment, conditional on his/her covariates (see §3.1.3).

Rosenbaum (1989) introduced further the notion of *optimal matching* that is a matching strategy to group treated and control units in a way that minimizes the overall distance between observations (see §3.2). The drawback of all unconstrained methods based on distances or propensity scores is that, if two data sets are separated (i.e. they have no common support), a match can always be found among the "less distant" observations, but the achieved matching may be meaningless. In such cases, it is more effective the use of a *caliper* (a bound) on the distance or on the propensity score or a mix of the two. Rosenbaum and Rubin (1985a, b) and Gu and Rosenbaum (1993) show the performance of different methods under this setup.

Propensity score matching has been brought back to the attention of the statistical community after the work of Dehejia and Whaba (1999). In their paper the authors suggest that propensity score matching is a good way to reduce bias in the estimation of the average treatment effect in observational studies (no matter the data sets to be matched). The debate that followed (Dehejia and Whaba 2002, Smith and Todd 2005a, b and Dehejia, 2005) was mainly focused on the sensitivity of the match to the model used to estimate the propensity score. In this respect, different parametric as well as nonparametric procedures (see e.g. Stone *et al.*, 1995, Hirano *et al.*, 2003) to estimate the propensity score have been proposed.

In this paper we propose a new matching algorithm which is invariant under monotonic transformation of the data. This method, named *Random Recursive Partitioning* (RRP) does not rely on a particular distance or on a specific model to be estimated and it does not suffer of the problem of "matching at any cost". It works on the spatial distribution of the observations and tries to figure out, using Monte Carlo arguments, whether two observations can be considered equal.

This method generates a proximity (complementarily, a dissimilarity) measure between observations which can be easily interpreted as the belief of two observations to be equal in covariates in the sense that they lie in the same region of the space (whichever the nature of this space). Information coming from this dissimilarity can be used as both a graphical and numerical tool to examine the extent of the common support between data sets. Once the common support has been identified, the observations in the common support can be used to evaluate the portion of the average treatment effect that can be reliably estimated.

The RRP method can in fact be considered an alternative to the matching methods proposed so far in the literature when distributional hypotheses cannot be assumed or when distances, like Mahalanobis, are not appropriate because of the nature of the data. This method, like the Genetic Matching algorithm (Diamond and Sekhon, 2005), is computationally intensive but still fast enough to be a usable device in applications. A free software ready to use is available for the `R` statistical environment.

DOI: 10.1002/jae

The paper is organized as follows: in Section 2 we introduce the Random Recursive Partitioning method and discuss the tools for the identification of the common support between data sets. Section 3 describes the problem of average treatment effect estimation and presents several alternative estimators, including the ones derived from the RRP method. In Section 4 our methodology is applied to the NSW data set analyzed originally in Lalonde (1986) with the aim to compare the RRP-based estimators to their competitors currently available in the literature. Section 5 reports Monte Carlo evidence of the ability of RRP to reduce the bias in covariates and some results on the asymptotic bias and limiting distribution of ATT estimators based on RRP.

2. RANDOM RECURSIVE PARTITIONING ALGORITHM

The RRP method is based on *regression trees* (RT), which recently become quite popular in datamining applications. Briefly, a regression tree (Breiman *et al*., 1984) models the expected value of some response variable Y conditionally on a set of covariates $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ by partitioning the observations on the basis of their covariates. The resulting final partition is such that in each stratum the homogeneity of the outcome Y is the maximum achievable with respect to some criterion (e.g. deviance for continuous variates or Gini index for categorical response variable). Starting from the set of complete observations the algorithm explores every possible bipartite stratification of the observations generated by each covariate X_i and finally splits the observations in two subsets according to the stratification to the variable that leads to the maximum homogeneity of the outcome inside both groups: for example, suppose that the variable is X_j, then each observation is moved in one group if for this observation $X_j < x$ and to the other group if $X_j \geq x$ (where x is the some optimal splitting threshold identified by RT). Thus, after the first iteration two groups are formed. At the second step, for each of the two groups the same rule is applied and two new groups are generated from each of the former. The algorithm stops when enough homogeneity in the currently formed groups is reached or when the size of the groups is small enough. Intermediate subgroups are called *nodes* and the final nodes are called *leaves*. The set of all the leaves corresponds to the final stratification of the whole set of observations. At each step the algorithm generates non empty strata and these strata are generated recursively. Optimization is not global as the algorithm uses a one-step look-ahead strategy[2]. Regression trees are quite effective in describing the dependence of the response variable Y on the covariates X and their interactions when looking at their graphical representation.

Regression trees have some features that make them interesting in matching applications. In particular, the generated partition is invariant under monotonic transformations of the data $\mathbf{X}$ (see Breiman *et al*., 1984, pag. 57) because the algorithm only considers the order of the values of each covariate and not the values themselves (it is essentially the same argument of median versus arithmetic mean). Another interesting feature is that regression trees tends to overfit the data. Overfit is of course not a good feature if the tree is used to predict the response variable Y for new observations, but in matching applications this is exactly what we desire to have: the resulting partition tailors the structure of the data. RRP uses the regression trees algorithm only to partition the observations and then generates a proximity which turns out to be quite effective in solving the matching problem.

[2] Other versions of the algorithm have been developed since 1984 for multipartite stratification and global optimization but we are not going to use them here.

DOI: 10.1002/jae

2.1. The proximity and dissimilarity matrix generated by RRP

The regression trees algorithm needs a response variable, say Z, whose homogeneity has to be optimized inside the strata. We assign a fictitious response variable to the observations: we draw n random numbers z_i from the uniform distribution on [0,1] and associate them to the n observations of the sample.[3] So we are going to model $Z \sim (X_1, X_2, \ldots, X_d)$ where Z is a fictitious response variable[4] needed to feed the RT algorithm. We then let the RT algorithm grow a tree and obtain a *random*, *recursive* and *non empty* partition of the data. The proximity measure for this random partition is defined as follows: we set $\pi_{ij} = 1$ for all the observations with indexes i and j in the same leaf and set $\pi_{ij} = 0$ otherwise. So we obtain a matrix of 0's and 1's, where the 1's correspond to observations belonging to the same stratum. The dissimilarity measure is defined as $\delta_{ij} = 1 - \pi_{ij}$. This partition and the dissimilarity/proximity measure entirely depends on the variable Z: therefore we replicate this procedure R times and at each replication we draw n new random numbers z_i and grow a new tree. Denote by $\pi_{ij}^{(r)}$ the proximity measure for iteration $r = 1, \ldots, R$: the final proposed proximity measure is obtained as the average of the $\pi_{ij}^{(r)}$'s over the R replications, i.e.:

$$\Pi^{RRP} = \left[\pi_{ij} = \frac{1}{R}\sum_{r=1}^{R} \pi_{ij}^{(r)}\right] \quad \text{and} \quad \Delta^{RRP} = 1 - \Pi^{RRP}.$$

RRP can be seen as a Monte Carlo method on the space of non-empty and recursive partitions of the observations. We refer to Π^{RRP} as the RRP-*proximity matrix* and to Δ^{RRP} as the RRP-*dissimilarity matrix*. Please remark that RRP method does not rely on a particular distance but it rather generates a new dissimilarity measure and, due to the fact that RT algorithm is invariant under monotonic transformations of the data, so is the RRP proximity Π^{RRP}. Sometimes a preliminary discretization of continuous covariates is advisable to avoid too wide cells near the border of the support of the data. In our implementation we suggest to divide the range of each continuous covariate in 15–20 intervals. The discretization has also the nice side-effect to reduce the numerical complexity of our method.

A parameter that controls the size of the leaves generated by the RT algorithm, is the so called "minsplit" parameter. Setting minsplit = 10 means "split cells with at least 10 observations, otherwise stop". Empirical evidence and arguments of next section suggest to use very low minsplit values (either 5 or 10) for the RRP method.

Figure 1 shows some possible partitions generated by the RRP method. That picture clearly shows that the partitions generated by the RRP algorithm are different from simply slicing the the set of covariates $\mathbf{X}$ at random (for example it never generates empty cells). It should also be clear from Figure 1 that the "shape" of the data determines the partition.

The RRP method requires, at most, ordering of the variables involved, but works with any kind of data and also in the presence of missing data. So it could be used in matching problems whenever the use of other well known distances (e.g. Mahalanobis) are debatable.

[3] We should notice here that any random assignment which makes observations equally preferable with respect to this fictitious response variable can be chosen.

[4] It is to be stressed that this Z variable has nothing to do with the outcome variable of the observational study or with the treatment. It is just an artifact to run a regression tree.

DOI: 10.1002/jae

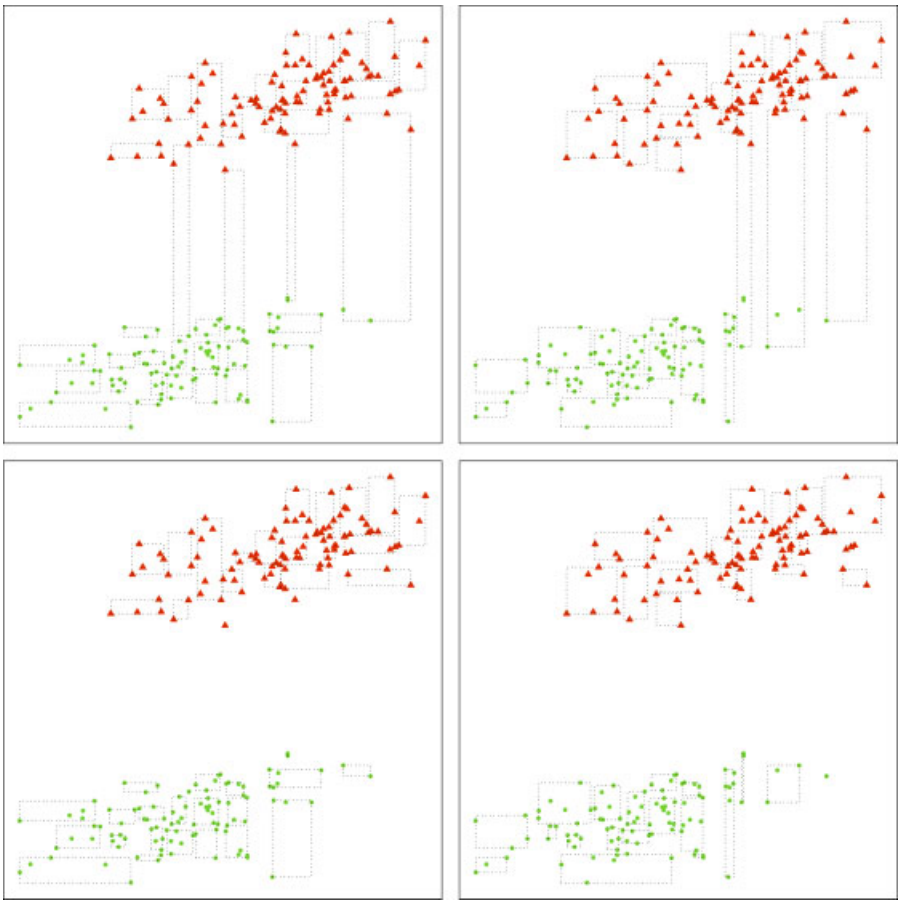

Figure 1. (UP): two random partitions generated by RRP. (DOWN): the same partitions in the above but with common support checking in each cell. Data are two multivariate normal samples with same spread and different means. Variables discretized in 15 intervals, minsplit = 8. This figure is available in color online at www.interscience.wiley.com/journal/jae

2.1.1 About the information coming from the proximity matrix

The RRP method is appreciable in that the elements of matrices Π^{RRP} and/or Δ^{RRP} can be easily interpreted. In particular, for some couple of units i and j, $\delta_{ij} = 0$ means that the two units *always* lie in the same cell. Conversely, $\delta_{ij} = 1$ means that units i and j *never* fall in the same cell. In our view it means that these two units should never be matched. It has to be stressed that Δ^{RRP} is not a distance matrix. Indeed, the proximity π_{ij} can be interpreted only as the belief of the event "observation i = observation j" to be true, but it gives no insight about the relative distance between observations. Observations with $\delta_{ij} = 1$ are simply different but we don't know how far they are in the space of covariates contrary to what a distance effectively measures. This means that a nearest neighbor approach on Δ^{RRP} it is not likely to be interesting without replacing $\delta_{ij} = 1$ with $\delta_{ij} = +\infty$. We will discuss this topic in §3.2.

DOI: 10.1002/jae

2.2. Preliminary reduction of the data

Prior to attempt a match between different data sets/groups, one should look at the distribution in covariates of the observations in the two groups to figure out if a match makes sense at all. Some geometric considerations can help to exclude a priori some control units which are likely to be out of the support of the treated units. We report in this section a couple of alternative devices. In the applications of §4.2 this preliminary reduction appears to be effective.

2.2.1 Selection via convex hull

Recently King and Zeng (2006a, b) proposed to identify the *convex hull* of one group (say the treated units) and to exclude from furhter analysis the units from the other group (the control units) which do not lie inside this convex hull. The convex hull of the treated units is the smallest subspace such that, for any two treated units, all points that are on the portion of hyperplane connecting them also belong to the subspace. This criterion is quite selective because it excludes from the common support all the control units that, although lying out of the convex hull, are *near* the treated units which define the boundaries of the convex hull (for further details see cited references).

2.2.2 Selection via hyper-rectangles

In this paper we propose another criterion that consists in constructing the smallest hyper-rectangle which includes all the treated units and to exclude from the analysis all control units not belonging to the hyper-rectangle. This method is less stringent than the convex hull criterion and more easy to implement. Moreover, it does not require any linear structure of the space: it only requires that, for each covariate, the minimum and maximum can be calculated[5]. Define $(m_i, M_i) = (\min_{j\in T} X_{ij}, \max_{j\in T} X_{ij})$, $i = 1, \ldots, p$, where T is the set of indexes for the treated units and X_{ij} is the value assumed by variable X_i on subject j. The hyper-rectangle is defined by the product $H = (m_1, M_1) \times \cdots \times (m_p, M_p)$. Then H corresponds to the region of the space $X_1 \times X_2 \times \cdots \times X_p$ which reasonably includes the common support of the two data sets.

2.2.3 About the balance check

The above methods do not assume any distributional hypotesis which are always hard to be verified in real world applications. One the contrary, it does not seem a good idea to use one of such criteria as a way to check for the *balancing property* inside the strata to further refine a match.

In fact, virtually any matching algorithm produces a stratification of the data in subgroups. Inside each stratum it might happen that control and treated units are not really homogeneous: to check for it, different parametric and nonparametric tests (such as t, Kolmogorov-Smirnov or Chi-Squared tests) are usually applied in the econometric literature. In situations where distributional hypothesis are hardly verified or when the strata consists of very few observations (like in our case) all these tests are likely to be very conservative (for an argument see Becker and Ichino, 2002). Conversely, both geometrical methods seem to be too severe and lead to unpleasant results when applied to small cells. We discuss now the drawbacks of the hyper-rectangle approach but similar considerations applies to the convex hull method.

[5] For non ordinal categorical and dichotomous variables one can take minimum and maximum of coding values without affecting the criterion.

DOI: 10.1002/jae

We make use of two-dimensional *ad hoc* examples for which (given a minsplit value) checking for homogeneity is not efficient whilst a preliminary finer discretization of the support of the covariates would have performed better. In Figure 2 top-left the hyper-rectangle H including the two treated units (full dots) does not contain other control units even if they are clearly "close" to the treated ones. The balance check based on the hyper-rectangle then creates two separate groups and no match between treated and control units can been obtained. Top-right: for the same data, if we choose a finer discretization of the covariates (the dashed line), four cells are created and matching without checking for the balancing property is more reasonable. Figure 2 middle: all controls lie in the hyper-rectagle and hence they are matched, but this is a case where one can see that treated and control units are not homogeneous. Again, a finer discretization gives more reasonable match between treated and control units. Figure 2 bottom shows a case where treated and control units are separate on one covariate. In this case the hyper-rectangle gives the correct answer leading to no possible match. However, if we use a finer discretization, the answer remains unchanged.

So, even if this are peculiar examples, our suggestion is not to use the hyper-rectangle check and to use instead a small minsplit along with a fine discretization of continuous variables. The hyper-rectangle as well as the convex hull criterion should probably be used before running any matching algorithm to roughly select a region which contains the common support between the treated and control units or, as any other balance check, in the RRP algorithm when strata are crowded enough. It should be pointed out, anyway, that it is not costless to suppress balance checking, particularly when a common support between treated and control units does not exists. Indeed, Figure 1 shows two sample partitions obtained with the RRP method. In the upper part the information on treatment variable is ignored and hence no balance check inside cells is applied. The lower part of the figure represents the same partition after applying the hyper-rectangle check: no cells contain observations of both groups, hence treated and control units have been separated.

2.3. EPBR property and the RRP method

Cochran and Rubin (1973) introduced (without naming it) the property of EPBR (Equal Percent Bias Reduction) with respect to the Mahalanobis metric saying that "*if* $\boldsymbol{x}$ *is spherical and symmetric (. . .) Mahalanobis distance implies the same percent reduction in bias for each* $x^{(k)}$". The EPBR property, formalized later in Rubin (1976a, b), excludes, for example, the unappealing case of a matching method which is able to reduce the bias on one covariate producing, at the same time, an increase in the bias on some others. Matching methods can also be *affinely invariant* in the sense that affine transformations on the covariates lead to the same match of the observations. Under the hypothesis of ellipsoidal distribution on the data—for example multinormal samples—Rubin and Thomas (1992a) show that any affinely invariant matching method is also EPBR. As already mentioned, RRP is invariant invariant under monotonic transformation of the data and not under general affine transforms. Still, what follows (see §5) appears to be competitive with Mahalanobis matching under conditions for EPBR and sometimes better when these conditions are not met.

DOI: 10.1002/jae

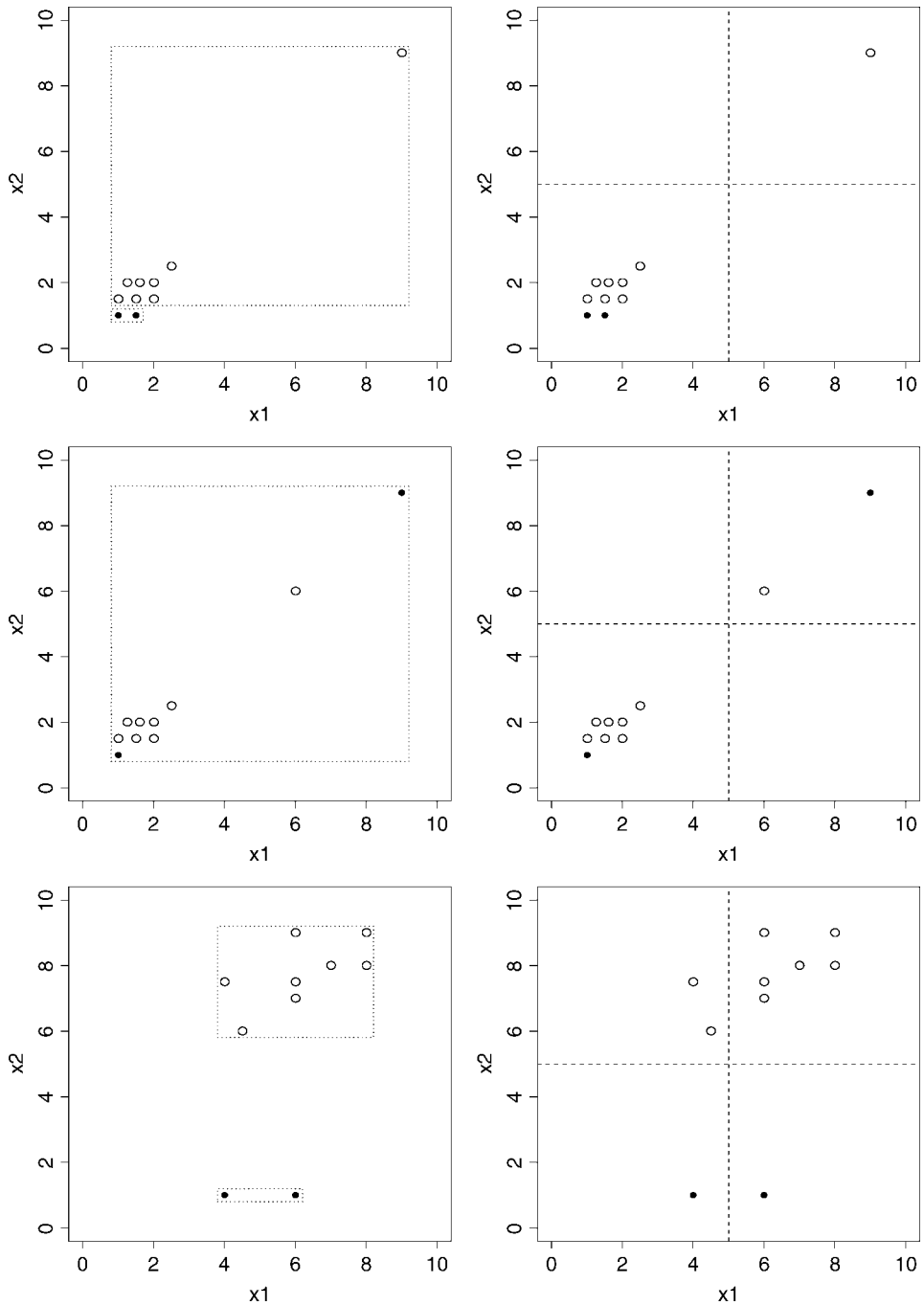

Figure 2. RRP with small minsplit. On the left: with balance check with hyper-rectangles, on the right a finer discretization on continuous variates. In some cases it is better to choose a finer discretization. (Filled dots = treated units, circles = control units)

DOI: 10.1002/jae

2.4. The proximity matrix and the identification of the common support

The proximity matrix generated by the RRP method contains insights for the determination of the extent of the common support between two data sets, for example experimental treated individuals T and non experimental controls C. The analysis can be made by a graphical inspection of Π^{RRP} and by calculating some related quantities. We illustrate it with a driving example. In Figure 3 two situations are shown corresponding to two data sets: one with partial common support between treated and control individuals (left column) and one with complete absence of a common support (right column). In the first row the two data sets are represented. These data have been randomly generated: they are made of 400 units (200 treated, 200 control units). In the second row we plotted the two Π^{RRP}'s. In particular, each plot represents the portion of Π^{RRP} for "treated" (rows) versus "treated and controls" (columns). It is evident that both matrices are dense in the part "treated" vs "treated". For the non overlapping data set (middle-right) the "treated vs controls" part of Π^{RRP} has all $\pi_{ij} = 0$ which means that no treated and control units ever fell in the same cell in the 250 RRP replications. This is of course the extreme case of complete absence of a common support and the graphical inspection of the proximity matrix helps in discovering it. This analysis becomes particularly effective in the case of $p \geq 3$ covariates where direct plotting of the data is essentially unavailable.

For the partial overlapping case (middle-left) only few π_{ij} are strictly positive indicating that only subsets of the control and treated groups have common support. In Figure 3 we also depicted the hyper-rectangle defined in §2.2.2. We remind that this rectangle corresponds to the region of the space $X_1 \times X_2$ which reasonably includes the common support. In this rectangle we counted 62 treated units. A priori these 62 treated units are the ones which can be potentially matched with some controls. If one counts the number of dark spots ($\pi_{ij} > 0$, i.e. units actually matched) in the plot of the Π^{RRP} (middle-left), she will find a total of 61 treated units.

2.4.1 The function $S(\lambda)$

Beside the graphical analysis of the RRP proximity matrix, an analytical tool can be derived from it. This is the curve $S(\lambda)$, $\lambda \in [0, 1]$, defined as follows: $S(\lambda) = \#\{i \in T : \max_{j \in C} \pi_{ij} \geq \lambda\}/n_T$, where n_T is the number of treated units. For fixed λ, the function $S(\lambda)$ is the proportion of treated units that matched some control units at least λ times over the R replications of the RRP algorithm. Indeed, $S(\lambda)$ is a non increasing function of λ and it rapidly goes to zero if no match is available. This behavior is well represented in Figure 3 where for the non overlapping case the curve just goes to zero[6] as soon as $\lambda > 0$. From the graph of the curve $S(\lambda)$ one should notice that the curve stabilizes around some level, say $S(\lambda^*)$, which is the proportion of treated units that have always been matched: these are the units belonging to the common support. We will refer to λ^* as the *empirical threshold* in the applications. The empirical threshold can be used to select only the treated units belonging to the common support.

3. AVERAGE TREATMENT EFFECT ESTIMATION

In the estimation of the average treatment effect (see Rubin, 1974, 1977, 1978), the differential effect of the treatment on some outcome variable between treated and control units has to be

[6] In the example $S(\lambda) = 0$ at point $\lambda = 0.05$ because we plotted $S(\lambda)$ on the grid $\lambda_i = i/20$, $i = 0, 1, \ldots, 20$.

DOI: 10.1002/jae

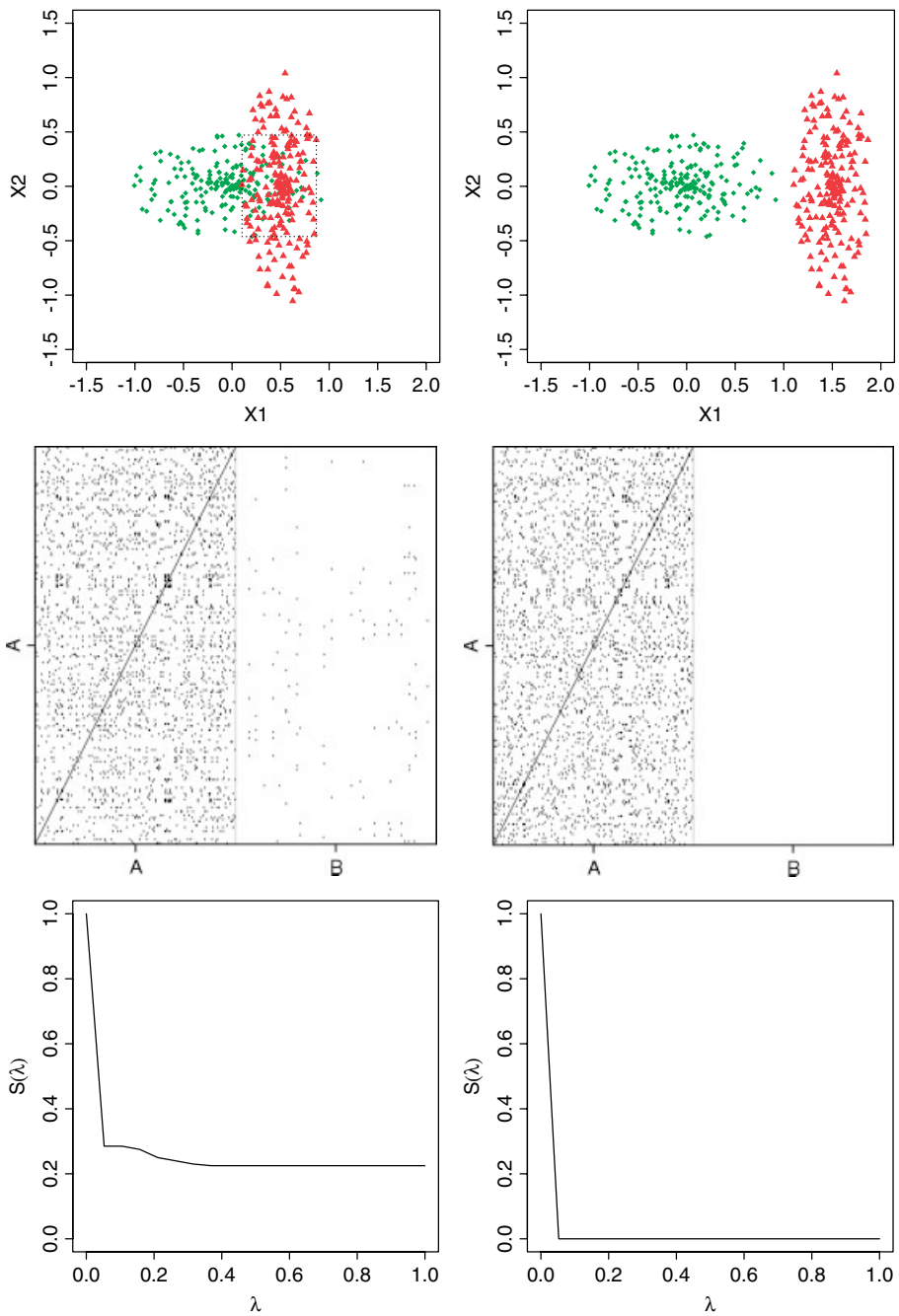

Figure 3. (UP): disjoint (right) and overlapping (left) data sets. (MIDDLE): the resulting proximity matrix (treated versus treated and controls) for the above data: variables discretized in 15 intervals, minsplit = 10, 250 RRP replications. (DOWN): the curve $S(\lambda)$ for the two data sets (see text). This figure is available in color online at www.interscience.wiley.com/journal/jae

DOI: 10.1002/jae

evaluated given the same pre-treatment conditions. Here is where the matching techniques are applied. Formally, for individual $i = 1, \ldots, N$, let (Y_i^T, Y_i^C) denote the two potential outcomes, Y_i^C being the outcome of individual i when she is not exposed to the treatment and Y_i^T the outcome of individual i when he is exposed to the treatment. If both Y_i^C and Y_i^T were observable, then the effect of the treatment on i would be simply $Y_i^T - Y_i^C$. The root of the problem is that only one of the two outcomes is observed whilst the *counterfactual* is to be estimated.

Often, the object of interest in applications is the average treatment effect on the subpopulation of the N_T treated subjects (ATT). Let τ be the ATT, then τ can be written as $\tau = \frac{1}{N_T}\sum_{i\in T}(Y_i^T - Y_i^C)$. As noticed, the first problem in practice is to estimate the unobserved outcome Y_i^C. The basic idea behind matching estimators is the following: for each treated unit i, matching estimators impute to Y_i^C the average outcome of control individuals similar to the treated i.

To ensure that the matching estimators identify and consistently estimate the treatment effects of interest, it is always assumed that: a) (*unconfoundedness*) assignment to treatment is independent of the outcome, conditional on the covariates; b) (*overlap*) the probability of assignment is bounded away from zero and one (see Rosenbaum and Rubin, 1983). These hypotheses imply the existence of a common support on pre-treatment covariates between treated and control units.

3.1. ATT estimators and the proximity matrix

It is possible to derive several estimators from Π^{RRP} or integrate the information coming from the matrix into other estimators. We review some estimators which will be used in the next sections.

3.1.1 The simple RRP-ATT estimator

A simple ATT estimator based on Π^{RRP} can be defined as follows:

$$\hat{\tau} = \frac{1}{n_T}\sum_{i\in T}\left(Y_i^T - \sum_{j\in C_i} f_{ij}Y_j^C\right) = \frac{1}{n_T}\sum_{i\in T}(Y_i^T - \hat{Y}_i^C) \tag{1}$$

where $C_i = \{j \in C : \pi_{ij} > 0\}$ and $f_{ij} = \pi_{ij}/\sum_{j\in C_i}\pi_{ij}$ is the relative frequency of match conditional to unit $i \in T$. This estimator will be denoted simply by RRP in the the tables.

3.1.2 The weighted RRP-ATT estimator

We can also integrate the information on the common support coming from the Π^{RRP} into the $\hat{\tau}$ of (1). Indeed, treated units which achieve high frequency of match provide a more reliable contribution to the estimation of the ATT. This reliability should be reflected by an ATT estimator in assigning different weights to each treated unit. For instance, we may define the following set of weights for $(Y_i^T - \hat{Y}_i^C) : \pi_i^{\max} = \max_{j\in C_i}\pi_{ij}$ and $q_i = \pi_i^{\max}/\sum_{i\in T}\pi_i^{\max}$, where the q_i's are the normalized version of these weights, i.e. $\sum_{i\in T} q_i = 1$. This set of weights, which takes into account the maximum number of times treated unit i has been matched to a control unit, is strictly related to our notion of measurement of the common support. We can now define a weighted ATT estimator $\tilde{\tau}$ as follows:

$$\tilde{\tau} = \sum_{i\in T}(Y_i^T - \hat{Y}_i^C)q_i \tag{2}$$

DOI: 10.1002/jae

Obviously, estimator in (2) reduces to (1) when $q_i = 1/n_T$. This estimator will be denoted by W.RRP in the tables. Other weighted estimators exists in the literature (see e.g. Hirano *et al.*, 2003) but we are not going to implement them here.

3.1.3 Propensity score and RRP

Rosenbaum and Rubin (1983) introduced the notion of propensity score (PS) as the probability that an individual i receives the treatment conditional on his/her covariates, sometimes denoted by $e_i(\mathbf{X})$. Of course, two observations with same values of covariates have the same propensity score: hence they propose to match on the basis of $e(\mathbf{X})$. Under the assumptions in Rubin and Thomas (1992a), the PS matching method is affinely invariant and hence EPBR. The erformance of the PS matching under different assumptions on the distribution of the covariates has been examined in Rubin and Thomas (1992b, 1996).

3.1.4 Nearest neighbor and the RRP

A widely used class of estimators is the one of the *nearest neighbor* estimators. They start by selecting one treated unit and matching it with the "nearest" k controls (NN(k), where k is usually 1, 4, 16) and *nearest* is with respect to some distance or score, for example the Mahalanobis distance or the propensity score. Then, another treated unit is chosen and matched using the same criterion. If control units are removed from the set of the controls after matching, the method is said *without replacement*. In this case, the order the treated units are chosen has an effect on the final match and hence on the ATT estimate. In the applications we will use a match *with replacement*. Even if it seems natural to define a NN(k) estimator on the RRP dissimilarity matrix, this is a reasonable strategy when $k = 1$ or when $\delta_{ij} = 1$ is set to $+\infty$. In the applications we calculate the NN(1) and NN(16) on the Mahalanobis distance (MAH(1) and MAH(16) in the tables) and on the RRP matrix (RRP(1) and RRP(16) respectively) without any modification of the δ_{ij} in order to show the difference in performance.

3.2. Optimal matching and RRP

We also show how to use Δ^{RRP} as an alternative dissimilarity matrix or to define a caliper on other distance matrices in the optimal matching algorithm. Rosenbaum (1989) introduced the notion of *optimal matching* which, for a given set of observations, a given distance and a set of weights, is a matching strategy that groups treated and control units in a way that minimizes the *overall distance* between observations. Given a stratification of α treated units and β controls in groups (i.e. a match of size (α, β)), under mild assumptions this match might be improved by transforming it into a *full match*[7] (for an extensive account on optimal matching see Rosenbaum, 2002 Ch. 10). Of course, nothing can be stated about which is the best match in general because this notion of optimality is strictly related to the distance adopted, as well as to the size (α, β) of the match. Nevertheless, this method is computationally more efficient then a greedy match (like the nearest neighbor) because it can be transformed into a problem of a "minimum cost flow in a network" (Bertsekas, 1991) for which highly efficient algorithms exist.

[7] A full match is a non overlapping stratification in which each treated is matched with at least one control in a stratum or viceversa a control is matched with one or more treated but does not allow for multiple treated versus multiple controls match.

DOI: 10.1002/jae

The drawback of this method, as well as of the nearest neighbor method, is that if two data sets have no common support (and without any restrictions) an optimal match can always be found but this does not guarantee any good property in the estimation of the average treatment effect. In such cases, it is more effective the use of some restriction. For example, an observation with propensity score $e_i(\mathbf{X})$ should not be matched with another one whose propensity score $e_j(\mathbf{X})$ is outside a prescribed radius/caliper r, i.e. if $|e_i(\mathbf{X}) - e_j(\mathbf{X})| > r$. Rosenbaum and Rubin (1985a) proposed the nearest available Mahalanobis metric matching within calipers defined by the propensity score[8] and showed its good performance. Also, Gu and Rosenbaum (1993) show the empirical performance of a full match using the same distance. A different caliper can be derived from Δ^{RRP}. We set $\delta_{ij} = +\infty$ when $\delta_{ij} \geq \delta^*$, for some value of $\delta^* \in [0, 1]$. The same restriction can be imposed on the corresponding terms of other distance matrices, like the Mahalanobis. We will use the following notation: MAH(F) for the the optimal full match ATT estimator on the Mahalanobis distance, MAH + P(F) for the same estimator with propensity score's caliper on the Mahalanobis distance, MAH + RRP(F) for the same estimator with calipers on the RRP-dissimilarities, RRP(F) for the ATT estimator based on the optimal full match on Δ^{RRP} and RRP + RRP(F) for the ATT estimator based on the optimal full match on Δ^{RRP} within calipers on the RRP-dissimilarities.

3.2.1 The "selected" RRP-ATT estimator

We define finally the "selected" RRP estimator (in a simple and a weighted version). This estimator is built solely on the treated units that have been matched at least λ^* % times with some other controls, where the empirical threshold λ^* is defined in §2.4.1. Thus the selected RRP estimator does not evaluate the *global* average treatment effect but only the effect restricted to the portion of treated individuals that can be reliably matched. In the tables the estimator will be denoted by SEL.ATT (and W.SEL.ATT for the weighted version).

3.3. Adjustment for difference-in-covariates

Provided we are given a consistent estimator $\hat{}_0(\mathbf{X})$ of $_0(\mathbf{X}) = E(Y^C|\mathbf{X})$, we can also adjust the residual bias for the difference-in-covariates (see e.g. Abadie and Imbens, 2005) obtaining, for example, the following adjusted version of the simple RRP-ATT estimator in equation (1):

$$\hat{\tau}' = \frac{1}{n_T} \sum_{i \in T} \left((Y_i^T - \hat{}_0(\mathbf{X}_i)) - \sum_{j \in C_i} f_{ij} (Y_j^C - \hat{}_0(\mathbf{X}_j)) \right) \tag{3}$$

When treated and control units are matched perfectly, i.e. both control and treated units have the same values in covariates, this bias correction has no effect. If the bias correction for the difference-in-covariates has a large impact on the ATT estimate, one can presume that the match has not been completely successful or the dependence of the outcome on the covariates is highly non-linear, since even small covariate imbalances lead to very different counterfactual estimates. This adjustment in covariates will be applied to all of the estimators presented so far.

[8] The distance between observations are set to $+\infty$ if their corresponding propensity score are as far as (or more then) 0.6 times the standard deviation of the distribution of the estimated propensity score.

DOI: 10.1002/jae

4. APPLICATION TO THE LALONDE (1986) DATA SET

In this section the performance of the RRP method is shown in a context of real data where the distributional assumptions for EPBR do not hold. We do not use propensity score based estimators as, in such a context, it is known that a good specification of the PS is crucial (the reader can refer to the thread[9] in the literature of propensity score matching which analyzes the same data set). We make use of the data from the National Supported Work (NSW) Demonstration, a job training program implemented during the Seventies in the United States. From April 1975 to August 1977 the program was carried out as a randomized experiment: some applicants were assigned to the program while some others, randomly chosen, were assigned to a control group and not allowed to participate to the program. The program provided training to the participants for 12–18 months and helped them in finding a job. As an effect, the program was supposed to yield an increase in the earnings of participants: therefore, real earnings in 1978 is the outcome variable of the analysis.

Several pre-treatment variables were registered about the applicants (both participants and controls): age (`age`), years of education (`education`), marital status (`married`), lack of an high school diploma (`nodegree`), race (`black`, `hispanic`) and real earnings in 1974 (`re74`) and 1975 (`re75`). Many of these variables are dicothomous, while the earnings variables are highly skewed and have point masses: therefore, the EPBR property conditions are hadly satisfied. As in the cited references, the two indicator variables `u74` and `u75` are included in the analysis to signal unemployement in 1974 and 1975.

The results of the program have been used by Lalonde (1986) to criticize the effectiveness of non-experimental matching in estimating the average treatment effect. Although a unique experimental target result cannot be easily defined, Lalonde adopted, as a target estimate of the average effect of the program, the simple difference in the means of the 1978 real earning across the participants and the control group: according to this definition, and using an experimental sample (LL sample) composed by 297 "treated" subjects (the participants) and 425 controls, Lalonde estimated an average effect of $886. In order to check the capability of non-experimental econometrics to estimate the effect of the treatment, Lalonde used several non-experimental control groups, coming from the Panel Study of Income Dynamics (PSID) and the Current Population Survey-Social Security Administration File (CPS) and tried to replicate the experimental target. He concluded that methods based on non-experimental data cannot correctly estimate the effect of the randomized experiment, consequently shedding some doubts on the reliability of these procedures. In more recent years, Dehejia and Wahba (1999, 2002) tried to show that the matching method based on propensity score can be successful in replicating the target estimated by Lalonde even in a non-experimental context. To this aim, they selected from the experimental sample used by Lalonde a subsample (DW sample) made of 185 treated and 260 control units. As a non-experimental control group, they used a sample of 2490 units coming from PSID[10]. Smith and Todd (2005a) questioned the results exposed by Dehejia and Wahba, arguing that the DW sample and their specification of the propensity score model are, to some extent, peculiar and that is the reason why they can replicate the experimental target. To support their argument, Smith and Todd suggest a more consistent (in their opinion) selection of the LL sample, restricting the number of treated individuals to 108 and to 142 for the controls (ST sample).

[9] See Dehejia and Whaba (1999, 2002), Smith and Todd (2005a, b) and Dehejia (2005).
[10] It is the PSID-1 sample used in Lalonde (1986).

DOI: 10.1002/jae

4.1. Results on the LL, DW and ST experimental samples

Tables I to III compare the performance of the RRP method to the estimates based on the Mahalanobis distance for the the experimental samples LL, DW and ST. The last row in the tables shows the estimated average effect after the adjustment for the difference-in-covariates. Tables I to III also report the number of treated and control units (n_T and n_C respectively) for which π is greater than the empirical threshold λ^* (defined in §2.4.1) used by SEL.RRP and W.SEL.RRP estimators.

For the LL sample, we adopt the simple difference in means between the treated and the control groups \$886 (RAW) as a comparison target. This is the only case (among LL, DW and ST) where the RAW estimate is a reasonable target provided the experimental design by Lalonde. Despite of the randomization of the experiment, some imbalance in covariates still exists (see Table VIII): the drawback of this imbalance on the ATT estimate will be discussed as well. For this data (see Table I), the RRP and RRP(F) estimates are quite close to the target \$886 and they seem to be better, compared to the alternative matching estimators.

The effect of the adjusment seems to be negligible for the RRP, RRP(F), SEL.RRP and W.RRP estimators, nevertheless only RRP and RRP(F) agree with the RAW estimate.

Table I. Comparison of performance of different estimators (see §3) on the LL experimental data. RRP based estimators with 1500 replications, minsplit = 5 and support of continuous variates divided in 20 subintervals. In this case the natural target is the value of the RAW estimator. For S.RRP and W.S.RRP $n_T = 113$, $n_C = 148$, $\lambda^* = 0.63$

Est.	RAW	MAH (F)	MAH (1)	MAH (16)	RRP (1)	RRP (16)	MAH + RRP(F)	RRP (F)	RRP	W.RRP	S.RRP	W.S.RRP
UnAdj.	886.3	574.7	−482.0	811.4	81.2	−394.0	820.3	918.5	802.3	628.0	361.0	412.3
Adj.	887.4	554.0	−465.2	716.5	81.5	103.3	796.2	929.8	802.2	647.3	361.0	412.3

Table II. Comparison of performance of different estimators (see §3) on the DW experimental data. RRP based estimators with 1500 replications, minsplit = 5 and support of continuous variates divided in 20 subintervals. For S.RRP and W.S.RRP $n_T = 73$, $n_C = 100$, $\lambda^* = 0.68$

Est.	RAW	MAH (F)	MAH (1)	MAH (16)	RRP (1)	RRP (16)	MAH + RRP(F)	RRP (F)	RRP	W.RRP	S.RRP	W.S.RRP
UnAdj.	1794.3	1430.2	778.4	1735.1	1277.6	1233.1	1683.7	1653.9	1778.2	1467.0	1112.5	1113.7
Adj.	1278.2	1071.2	396.9	1266.7	966.7	1204.8	1281.9	1351.5	1443.1	1302.0	989.8	990.6

Table III. Comparison of performance of different estimators (see §3) on the ST experimental data. RRP based estimators with 1500 replications, minsplit = 5 and support of continuous variates divided in 20 subintervals. For S.RRP and W.S.RRP $n_T = 17$, $n_C = 17$, $\lambda^* = 0.58$

Est.	RAW	MAH (F)	MAH (1)	MAH (16)	RRP (1)	RRP (16)	MAH + RRP(F)	RRP (F)	RRP	W.RRP	S.RRP	W.S.RRP
UnAdj.	2748.5	2349.7	1747.7	2719.9	3070.5	2111.0	2210.4	2326.8	2629.3	2525.8	1579.6	1489.4
Adj.	1883.0	1470.7	1112.3	1929.2	2015.6	1559.2	1499.9	1609.5	1856.1	2094.3	1579.6	1489.4

DOI: 10.1002/jae

In fact, the distributions of continuous variables in the original LL sample (see Table VIII) indicates that treated and control units have a common support (in terms of range) but the distribution of their continuous covariates are quite different in shape (see the corresponding quantiles). The SEL.RRP estimator (see §3.2.1) evaluates the average treatment effect restricted to the treated and control units (groups "a)" and "b)" in Table VII) which, besides belonging to the common support, have also quite similar distribution of continuous covariates. Table VII contains the summary statistics for the groups of treated units "a)" and control units "b)" such that the corresponding $\pi_{ij} \geq \lambda^* = 0.63$ and the statistics for less frequently matched treated units "c)" and controls "d)". A reduced dissimilarity can be noted on the distribution of earnings (`re74` and `re75`) of treated and control units frequently matched, compared to what happens on the complementary set and on the whole LL sample (see Table VIII)[11]. On one hand, this is an evidence of the reliability of the RRP method; on the other hand, it may lead to wonder about which is the real ATT of the Lalonde experiment. What we noticed, in fact, is that when the average effect is evaluated with the RRP estimator (therefore including all treated units in the estimation) we obtain results that are closer to the target adopted by Lalonde. On the contrary, when we take into account that the two samples do have different distribution of continuous covariates and restrict the estimation to the "closest" treated units (less than 40%) and controls (about 35%), we obtain different estimates (SEL.RRP) which should be considered reliable for the portion of the phenomenon included in the estimation.

The DW and ST samples are non-randomly selected subsamples of LL data: therefore the RAW estimate cannot be considered a good approximation of the real ATT. As a consequence, the RAW adjusted estimate is quite different from the unadjusted one in the DW and ST experimental sample, while the two values cannot be distinguished in the LL sample: this suggests that the DW and ST samples are even less balanced compared to the LL case. In the DW and ST samples, the effect of the adjustment is remarkable also for the RRP, RRP(F) and W.RRP but weaker (in the ST case, absent) for the SEL.RRP and W.SEL.RRP estimators. In these two cases, few can be said with our approach about the ATT on the whole set of treated units. Something more reliable, on the contrary, can be induced from the "selected" estimator. They in fact suggest stable estimates of the ATT for, respectively, 38% and 16% of the treated individuals[12].

Table IV. Comparison of performance of different estimators (see §3) on the DW (experimental treated) vs PSID (non experimental controls) whole data set. RRP based estimators with 1500 replications, minsplit = 5 and support of continuous variates divided in 20 subintervals. For S.RRP and W.S.RRP $n_T = 42$, $n_C = 33$, $\lambda^* = 0.53$

Est.	RAW	MAH (F)	MAH (1)	MAH (16)	RRP (1)	RRP (16)	MAH + RRP(F)	RRP (F)	RRP	W.RRP	S.RRP	W.S. RRP
UnAdj.	−15204.8	−1062.2	499.4	367.3	−5090.5	−10976.3	−3494.3	−4280.2	−896.9	−1148.7	−459.4	−464.0
Adj.	244.0	263.2	708.5	769.6	1111.4	172.5	1017.3	1283.7	1465.6	431.6	762.2	769.8

[11] A sharper evidence of this effect can be noted when less balanced samples are matched: see, e.g., Table IX about DWvsPSID.

[12] It is interesting to notice that in DW sample, even for the selected estimator, the adjustment for difference-in-covariates still shows to play a role. This seems to support the remarks in Smith and Todd (2005a, b) about the improper selection of units in the DW sample.

DOI: 10.1002/jae

4.2. Results on DW vs PSID non-experimental data

If a non experimental sample is used as a control group, generally the experimental target cannot be replicated: both Lalonde (1986) and Smith and Todd (2005a) show it by matching the DW experimental sample to the so called PSID-1 sample, composed by 2490 individuals drawn from the Panel Study of Income Dynamics. Smith and Todd (2005a) argued that, even if match is attained, the average treatment effect cannot be correctly estimated when treated and control samples come from different contexts. In this particulat case of DW versus PSID the different context is the local labour market, therefore, even if a common support between treated and control units were found, one cannot expect to replicate the raw target ($1794) of the DW experimental data. Table IV reports the ATT estimates obtained using the whole PSID-1 sample as the control group. Notice, first of all, that the adjusted estimates strongly differ from the unadjusted ones for all the applied estimators: this indicates the strong imbalance in covariates that survives the matching. The "raw" unadjusted ATT estimate is a negative quantity, reflecting the higher average earnings level of the non-experimental controls, compared to the earnings of the treated individuals. Almost all the estimators fail in replicating the "raw" DW experimental target and the estimates show a large variability across the estimators.

The selected (and weighted selected) RRP estimators consider only 23% of the treated units (and less than 2% of the controls): Table IX shows that the covariates of treated and control units with $\pi_{ij} \geq 0.53$ are much more balanced compared to the complementary groups. Despite of the selection, however, a large correction of the estimates is brought by the adjustment for the difference in covariates.

Previous results suggest a lack of common support between the two samples. The suspect can be verified evaluating how many controls are included in the portion of the space of covariates where the treated units lie. We consider the two preliminary data reductions presented in §2.2 based on the convex hull and hyper-rectangle. The hyper-rectangle criterion selects 1479 out of 2490 non experimental PSID controls. Table V shows the results of the ATT estimation: the remarkable difference between the adjusted and the unadjusted estimates is still observable and all the estimators fail to get the "raw" target. The convex hull criterion, which is even more selective, selects only 45 out of 2490 control units from the PSID data set. Nevertheless, this does not improve the performance of the ATT estimators very much (see Table VI).

It is worth noting what happens with the SEL.RRP and W.SEL.RRP estimators: only small subsamples of treated and untreated units are involved in both the hyper-rectangle and the convex hull case. The slight differences in the composition of the subsamples are enough to generate large differences in the estimated ATT: this can be explained by the non-linear effect of the covariates

Table V. Comparison of performance of different estimators (see §3) on the DW (experimental treated) vs PSID (non experimental controls) reduced to the hyper-rectangle common support (from 2490 to 1479 non experimental controls). RRP based estimators with 1500 replications, minsplit = 5 and support of continuous variates divided in 20 subintervals. For S.RRP and W.S.RRP $n_T = 19$, $n_C = 11$, $\lambda^* = 0.58$

Est.	RAW	MAH (F)	MAH (1)	MAH (16)	RRP (1)	RRP (16)	MAH + RRP(F)	RRP (F)	RRP	W.RRP	S.RRP	W.S.RRP
UnAdj.	−10295.1	−626.9	99.9	666.6	−7.0	−2933.1	−1781.8	−3099.8	−449.6	−563.7	3277.3	3277.3
Adj.	165.0	61.9	117.3	848.7	351.9	808.3	630.5	1514.9	859.0	1267.6	3277.3	3277.3

DOI: 10.1002/jae

Table VI. Comparison of performance of different estimators (see §3) on the DW (experimental treated) vs PSID (non experimental controls) reduced to the convex-hull common support (from 2490 to 45 non experimental controls). RRP based estimators with 1500 replications, minsplit = 5 and support of continuous variates divided in 20 subintervals. For S.RRP and W.S.RRP $n_T = 21$, $n_C = 12$, $\lambda^* = 0.47$

Est.	RAW	MAH (F)	MAH (1)	MAH (16)	RRP (1)	RRP (16)	MAH + RRP(F)	RRP (F)	RRP	W.RRP	S.RRP	W.S.RRP
UnAdj.	1495.3	3916.9	3926.1	1936.8	3117.0	1488.3	3546.6	2844.5	2647.7	2200.6	2417.9	2606.9
Adj.	738.7	2840.8	2868.4	1189.6	2619.4	636.5	2921.7	2273.8	2114.0	1893.3	2565.7	2702.9

on the outcome variable. At the same time, almost all the bias in covariates has been removed, as the high similarity between adjusted and unadjusted estimates indicates. Moreover, the ATT estimates on these subsamples includes a bias due to the different labour markets the units come from, therefore they cannot be considered reliable estimates.

5. MONTE CARLO RESULTS

Formal statistical properties of the RRP method and ATT estimators derived from it are not known at the moment of this writing but extensive Monte Carlo analysis has been done: the reader might want to refer to the unabridged version of this manuscript[13]. We concisely report the results in what follows: first we replicated the Monte Carlo experiment of Gu and Rosenbaum (1993). This experiment was intended to measure the ability of matching methods to reduce the bias in covariates as the number of continuous covariates increases along the sequence $p = 2, 5, 10, 15, 20$. Two samples are drawn from two p-dimensional multivariate Gaussian distributions with the same covariance matrix but with different mean vectors. The difference in mean represents the bias. In our experiment we compared the following matching methods (see §3 for the definition): MAH(F), MAH + P(F), MAH + RRP(F), RRP + RRP(F), MAH(1) and RRP(1). This is a setup in which EPBR property hold for Mahalanobis and propensity score matching. Indeed, as it can be expected (Rosenbaum and Rubin, 1985a), MAH + P showed one of the best performances. Moreover, MAH + P (with or without caliper) performs better than MAH (1) and RRP(1). What is interesting to observe is that, when $p \leq 5$, matching within a caliper induced by the Π^{RRP} yields a bias reduction which is comparable (in some cases, superior) to MAH + P. This is also true for MAH + RRP(F) and RRP + RRP(F). When the number of continuous covariates is 10 or more the curse of dimensionality seem to play a role for RRP.

We then replicated the experimental design proposed in Diamond and Sekhon (2005). This experiment was designed to test the ability of matching methods to reduce the bias in ATT estimation when the relationship between the outcome and the covariates is highly non-linear and when EPBR is unlikely to hold. What emerges from our experiment is that almost all estimators need to be adjusted in order to approach the target. Although ADJ.MAH(16) gets quite close to the target, its $\sqrt{MSE}$ is almost twice the variance of the SEL.RRP estimator. Moreover, the adjustment of SEL.RRP and W.SEL.RRP estimators has essentially no effect, which means that all the unbalance due to covariates has been removed by the correct selection of the treated and control units.

We also studied the asymptotic bias and limiting distribution of ATT estimators. To this end we generated a population of $N = 100\,000$ observations from the original LL data set using Latin

[13] available at http://services.bepress.com/unimi.

DOI: 10.1002/jae

Table VII. Summary statistics for the units (treated vs controls) such that their $\pi_{ij} \geq 0.63$, and for the group of less frequently matched units ($\pi_{ij} < 0.63$). It can bee seen that groups a) and b) have more homogeneous distributions in variables 're74' and 're75' then groups c) and d). LL data, split = 5, $R = 1500$

a) treated $\pi_{ij} \geq 0.63$ 113 units	age	education	re74	re75	black	married	nodegree	hispanic	u74	u75
Min	17.0	4.0	0.0	0.0						
1st Qu.	18.0	9.0	0.0	0.0						
Median	21.0	11.0	0.0	0.0						
Mean	22.2	10.3	706.4	661.1	0.05	0.97	0.21	0.96	0.35	0.39
3rd Qu.	25.0	11.0	934.4	851.6						
Max	48.0	13.0	9266.6	8497.4						
b) control $\pi_{ij} \geq 0.63$ 148 units	age	education	re74	re75	black	married	nodegree	hispanic	u74	u75
Min	17.0	3.0	0.0	0.0						
1st Qu.	18.9	9.0	0.0	0.0						
Median	20.0	10.0	0.0	0.0						
Mean	21.9	10.1	657.3	645.6	0.05	0.98	0.11	0.97	0.33	0.33
3rd Qu.	24.3	11.0	998.2	917.8						
Max	55.0	13.0	8784.2	8784.2						
c) treated $\pi_{ij} < 0.63$ 184 units	age	education	re74	re75	black	married	nodegree	hispanic	u74	u75
Min	17.0	4.0	0.0	0.0						
1st Qu.	21.0	9.0	0.0	481.1						
Median	25.0	11.0	3000.0	2636.4						
Mean	26.2	10.5	5330.0	4543.1	0.29	0.74	0.30	0.87	0.68	0.77
3rd Qu.	29.0	12.0	8549.0	6691.2						
Max	49.0	16.0	37432.0	37431.7						
d) control $\pi_{ij} < 0.63$ 277 units	age	education	re74	re75	black	married	nodegree	hispanic	u74	u75
Min	17.0	4.0	0.0	0.0						
1st Qu.	20.0	9.0	0.0	0.0						
Median	25.0	10.0	2432.0	2037.0						
Mean	25.8	10.2	5284.0	4299.0	0.28	0.78	0.22	0.84	0.65	0.70
3rd Qu.	29.0	11.0	7688.0	6540.0						
Max	54.0	14.0	39571.0	36941.0						

Hypercube Sampling (LHS) algorithm (see Iman and Conover 1982, Iman *et al.*, 1980). This population, assumed as the true population, is then resampled with sample sizes $n = 100$, 500 and 1000. Again, outcome variable is a highly non-linear function of the covariates. From the experiment, it emerges that all the estimators considered behave similarly with the exception of unadjusted MAH(16) and RRP(16), but both results are expected because a high fixed number of units are matched (see Abadie and Imbens, 2005). It is worth nothing that the asymptotic bias of

DOI: 10.1002/jae

Table VIII. Summary statistics for treated vs control units in the original LL data set

a) treated 297 units	age	education	re74	re75	black	married	nodegree	hispanic	u74	u75
Min	17.0	4.0	0.0	0.0						
1st Qu.	20.0	9.0	0.0	0.0						
Median	23.0	11.0	858.3	1117.0						
Mean	24.6	10.4	3571.0	3066.0	0.20	0.83	0.27	0.90	0.56	0.63
3rd Qu.	27.0	12.0	5491.5	4310.0						
Max	49.0	16.0	37431.7	37432.0						
b) control 425 units	age	education	re74	re75	black	married	nodegree	hispanic	u74	u75
Min	17.0	3.0	0.0	0.0						
1st Qu.	19.0	9.0	0.0	0.0						
Median	23.0	10.0	788.5	823.3						
Mean	24.5	10.2	3672.5	3026.7	0.20	0.84	0.19	0.89	0.54	0.58
3rd Qu.	28.0	11.0	4906.6	3649.8						
Max	55.0	14.0	39570.7	36941.3						

the unadjusted RRP estimators (with the exclusion of RRP(16)) is smaller than $\sqrt{n}$. This is not true for the adjusted case. The SEL.ATT estimator is one among the best estimators in this group because it always have lower bias and mean square error. This is again expected because, if the algorithm really isolates the portion of the sample that can be accurately matched, then the ATT estimate is reliable. It should be also noticed that this estimator has quite low mean square error. All the estimators tend to have a asymptotically Gaussian distribution with few exceptions.

6. CONCLUDING REMARKS ON THE USE OF RRP

The proximity matrix Π^{RRP} and the function $S(\lambda)$ are useful tools to analyze the extent of the common support between two data sets. It is always advisable to apply a preliminary data reduction through the geometric methods proposed (hyper-rectangle or convex hull), in order to ease the identification of the common support.

If the RRP method is considered reasonable—as we believe—the key to interpret the results of an application of RRP on some data set should be the following: if RRP and SEL.RRP estimates agree, then one can rely on the fact that all treated units did found a "close enough" control (i.e. their counterfactual). If, in addition, the adjusted estimates ADJ.RRP and ADJ.SEL.RRP are also concordant with RRP and SEL.RRP, this indicates that large part of the bias has been removed and/or there is only a linear effect of covariates on the outcome: hence, the ATT estimate can be considered reliable.

On the contrary, if RRP and SEL.RRP do not agree, it means that[14] we are not able to reliably estimate the ATT on the whole set of treated units, but we might be able to provide a good estimate only on a subset of the treated units. In fact, if SEL.RRP and

[14] In the absence of further information on the experimental conditions, as in the LL randomized experiment, where a reasonable target is available.

DOI: 10.1002/jae

Table IX. Summary statistics for the units (treated vs controls) such that their $\pi_{ij} \geq 0.53$, and for the group of less frequently matched units ($\pi_{ij} < 0.53$). It can bee seen that groups a) and b) have more homogeneous distributions then groups c) and d). DWvsPSID data, split = 5, $R = 1500$

a) treated $\pi_{ij} \geq 0.53$ 42 units	age	education	re74	re75	black	married	nodegree	hispanic	u74	u75
Min	17.0	6.0	0.0	0.0						
1st Qu.	21.0	10.0	0.0	0.0						
Median	23.0	11.0	1140.0	334.0						
Mean	23.2	11.0	3580.0	1927.3	0.10	0.86	0.48	1.00	0.57	0.57
3rd Qu.	25.0	12.0	5381.0	2826.4						
Max	35.0	12.0	20280.0	13830.6						
b) control $\pi_{ij} \geq 0.53$ 33 units	age	education	re74	re75	black	married	nodegree	hispanic	u74	u75
Min	18.0	6.0	0.0	0.0						
1st Qu.	21.0	10.0	0.0	0.0						
Median	23.0	11.0	3135.0	3652.0						
Mean	24.1	10.7	4258.0	3800.0	0.15	0.73	0.45	1.00	0.73	0.70
3rd Qu.	25.0	12.0	6583.0	5460.0						
Max	34.0	12.0	18613.0	16113.0						
c) treated $\pi_{ij} < 0.53$ 143 units	age	education	re74	re75	black	married	nodegree	hispanic	u74	u75
Min	17.0	4.0	0.0	0.0						
1st Qu.	20.0	9.0	0.0	0.0						
Median	26.0	10.0	0.0	0.0						
Mean	26.6	10.2	1660.0	1416.0	0.17	0.80	0.24	0.92	0.21	0.35
3rd Qu.	29.5	11.0	0.0	1338.0						
Max	48.0	16.0	35040.0	25142.0						
d) control $\pi_{ij} < 0.53$ 2457 units	age	education	re74	re75	black	married	nodegree	hispanic	u74	u75
Min	18.0	0.0	0.0	0.0						
1st Qu.	26.0	11.0	10776.0	10742.0						
Median	33.0	12.0	18613.0	17903.0						
Mean	35.0	12.1	19633.0	19268.0	0.75	0.13	0.70	0.97	0.92	0.90
3rd Qu.	44.0	14.0	26450.0	26696.0						
Max	55.0	17.0	137149.0	156653.0						

ADJ.SEL.RRP do agree, then we can state that this is a reliable estimate[15] of the "restricted" ATT.

If none of the above is true, i.e. RRP = SEL.RRP and also the adjustment has large effects, no reliable ATT estimate can be drawn from RRP method.

[15] Unless there is a confounding factor (e.g. the "labour market" factor in the DW vs PSID application).

DOI: 10.1002/jae

The RRP procedure is computationally intensive but, being a Monte Carlo method, it can be easily parallelized and a ready-to-use software has been written for the `R` statistical environment (see R Development Core Team, 2005) in the form of a package named `rrp` and available at http://CRAN.R-project.org.

ACKNOWLEDGMENTS

We are grateful to two anonimous referees and the A.E. John Rust for careful reading of the manuscript. Their criticism and suggestions have been very elucidating and ended up in a deeply revised version of the first manuscript.

REFERENCES

Abadie A, Imbens G. 2005. Large sample properties of matching estimators for average treatment effects, *Econometrica*, **74**(1): 235–267.

Becker S, Ichino A. 2002. Estimation of Average Treatment Effects Based on Propensity Scores, *The Stata Journal*, **2**(4): 358–377.

Bertsekas D. 1991. *Linear network optimization: algorithm and codes*, Cambridge, MA: MIT Press.

Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. *Classification and Regression Trees*, Monterey, Wadsworth and Brooks-Cole.

Cochran W, Rubin DB. 1973. Controlling Bias in Observational Studies: A Review, *Sankhya A*, **35**: 417–446.

Dehejia R. 2005. Practical propensity score matching: a reply to Smith and Todd, *Journal of Econometrics*, **125**(1–2): 355–364.

Dehejia R, Wahba S. 1999. Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs, *Journal of the American Statistical Association*, **94**: 1053–1062.

Dehejia R, Wahba S. 2002. Propensity score matching methods for Non-experimental causal studies, *Review of Economics and Statistics*, **84**(1): 151–161.

Diamond A, Sekhon JS. 2005. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. Mimeo, available at http://sekhon.polisci.berkeley.edu/papers/GenMatch.pdf.

Gu XS, Rosenbaum PR. 1993. Comparison of multivariates matching methods: structures, distances and algorithms, *Journal of Computational and Graphical Statistics*, **2**: 405–420.

Heckman J, Ichimura H, Todd P. 1997. Matching as econometric evaluation estimator: evidence from evaluating a job training programme, *Review of Economic Studies*, **64**(4): 605–654.

Heckman J, Ichimura H, Smith J, Todd P. 1998. Characterizing selection bias using experimental data, *Econometrica*, **66**(5): 1017–1098.

Hirano K, Imbens G, Ridder G. 2003. Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score, *Econometrica*, **71**: 1161–1189.

Iman RL, Conover WJ. 1982. A distribution-free approach to inducing rank correlation among input variables, *Communications in Statistics*, **B11**: 311–334.

Iman RL, Davenport JM, Zeigler DK. 1980. Latin Hypercube Sampling (Program User's Guide), SAND79 –1473.

King G, Zeng L. 2006a. The Dangers of Extreme Counterfactuals, *Political Analysis*, **14**(2): 131–159.

King G, Zeng L. 2006b. When Can History Be Our Guide? The Pitfalls of Counterfactual Inference, *International Studies Quarterly*, forthcoming. Available at http://gking.harvard.edu.

Lalonde R. 1986. Evaluating the Econometric Evaluations of Training Programs, *American Economic Review*, **76**: 604–620.

R Development Core Team 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Software available at http://www.R-project.org.

Rosenbaum PR. 1989. Optimal matching in observational studies, *Journal of the American Statistical Association*, **84**: 1024–1032.

DOI: 10.1002/jae

Rosenbaum PR. 2002. *Observational studies*, Second Edition, Springer-Verlag: New York.
Rosenbaum PR, Rubin DB. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, **70**: 41–55.
Rosenbaum PR, Rubin DB. 1985a. Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score, *The American Statistician*, **39**: 33–38.
Rosenbaum PR, Rubin DB. 1985b. The Bias Due to Incomplete Matching, *Biometrics*, **41**: 103–116.
Rubin DB. 1973a. The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies, *Biometrics*, **29**: 185–203.
Rubin DB. 1973b. Matching to Remove Bias in Observational Studies, *Biometrics*, **29**: 159–183.
Rubin DB. 1974. Estimating Causal Effects of Treatments in Randomized and non-randomized Studies, *Journal of Educational Psychology*, **66**: 688–701.
Rubin DB. 1976a. Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples, *Biometrics*, **32**: 109–120.
Rubin DB. 1976b. Multivariate Matching Methods That are Equal Percent Bias Reducing, II: Maximums on Bias Reduction for Fixed Sample Sizes, *Biometrics*, **32**: 121–132.
Rubin DB. 1977. Assignment to Treatment Group on the Basis of a Covariate, *Journal of Educational Statistics*, **2**: 1–26.
Rubin DB. 1978. Bayesian inference for causal effects: The Role of Randomization, *Annals of Statistics*, **6**: 34–58.
Rubin DB. 1980. Bias Reduction Using Mahalanobis-Metric Matching, *Biometrics*, **36**: 293–298.
Rubin DB, Thomas N. 1992a. Affinely Invariant Matching Methods with Ellipsoidal Distributions, *The Annals of Statistics*, **20**: 1079–1093.
Rubin DB, Thomas N. 1992b. Characterizing the Effect of Matching Using Linear Propensity Score Methods with Normal Distributions, *Biometrika*, **79**: 797–809.
Rubin DB, Thomas N. 1996. Matching Using Estimated Propensity Scores: Relating Theory to Practice, *Biometrics*, **52**: 249–264.
Smith J, Todd P. 2005a. Does Matching Overcome Lalondes Critique of Nonexperimental Estimators?, *Journal of Econometrics*, **125**(1–2): 305–353.
Smith J, Todd P. 2005b. Rejoinder (to Dehejia, 2005), *Journal of Econometrics*, **125**(1–2): 365–375.
Stone RA, Obrosky DS, Singer DE, Kapoor WN, Fine MJ. 1995. Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia, *Medical Care*, **33**: AS56–AS66.

DOI: 10.1002/jae