AUTHOR QUERY FORM

ELSEVIER	**Journal:** Statistics and Probability Letters **Article Number:** 8013	**Please e-mail your responses and any corrections to:** **E-mail: corrections.esch@elsevier.river-valley.com**

Dear Author,

Please check your proof carefully and mark all corrections at the appropriate place in the proof (e.g., by using on-screen annotation in the PDF file) or compile them in a separate list. Note: if you opt to annotate the file with software other than Adobe Reader then please also highlight the appropriate place in the PDF file. To ensure fast publication of your paper please return your corrections within 48 hours.

For correction or revision of any artwork, please consult http://www.elsevier.com/artworkinstructions.

Location in article	Query / Remark click on the Q link to go Please insert your reply or correction at the corresponding line in the proof
Q1	Your article is registered as a regular item and is being processed for inclusion in a regular issue of the journal. If this is NOT correct and your article belongs to a Special Issue/Collection please contact s.sekar@elsevier.com immediately prior to returning your corrections.
Q2	Please confirm that given names and surnames have been identified correctly.
	Please check this box or indicate your approval if you have no corrections to make to the PDF file ☐

Thank you for your assistance.

ARTICLE IN PRESS

Statistics and Probability Letters xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Q1 On the support of matching algorithms

Q2 Massimo Cannas [a,*], Gavino Puggioni [b]

[a] Department of Business and Economic Sciences, University of Cagliari, Viale Sant'Ignazio 83, 09123 Cagliari, Italy
[b] Department of Computer Science and Statistics, University of Rhode Island, 9 Greenhouse Road, Suite 2, Kingston, RI, 02881, USA

ARTICLE INFO

Article history:
Received 6 October 2016
Received in revised form 16 May 2017
Accepted 20 August 2017
Available online xxxx

MSC:
62K99
05A99

Keywords:
Causal inference
Quasi experimental design
Enumerative combinatorics
Rook polynomials
Radius matching

ABSTRACT

In causal inference a matching algorithm assigns a subset of control units to each treated unit. We explore the support of matching algorithms by using combinatorial techniques to provide counting results, and investigate the role of the dimension of covariates.

1. Introduction

In observational studies where the treatment is not randomized the group of treated and control units may have systematic differences in covariates which would not appear in an experimental design. Matching methods are a common tool for obtaining a subcollection of the data where the distribution of observed covariates looks only randomly different across the two groups (see Imbens and Rubin (2016) for a presentation of the classic causal inference framework including matchings and other quasi and non-experimental methods, and Stuart (2010) for a comprehensive review of matching methods specifically). Usually matching methods are defined rather informally as procedures for pre-treatment of observational data (Rosenbaum, 2002) even if they can be identified with the embedded matching algorithm

Definition 1.1. Let T be the set of treated units and C the set of controls. A *matching algorithm* is any function

$$\theta : T \to \mathcal{P}(C).$$

where $\mathcal{P}(\cdot)$ denotes the power set. The subset S of $T \times \mathcal{P}(C)$ given by $S \equiv \{(a, b) \mid b = \theta(a)\}$ (i.e. the graph of θ) is called *matching*. The pair (T, B), where B is the multiset of the images of θ, where at least one of the images should be nonempty, is called *matched dataset*.

The use of the word *matching* to indicate the output is consistent with the graph theory literature since S is equivalent to the edge set of a matching in the bipartite graph with vertex bipartition $T \cup C$. The expression *matched subset* can be connected to the relevant output of causal inference estimation. In fact, common statistics for evaluation of causal effects

* Corresponding author.
E-mail addresses: massimo.cannas@unica.it (M. Cannas), gpuggioni@uri.edu (G. Puggioni).

http://dx.doi.org/10.1016/j.spl.2017.08.009
0167-7152/

Please cite this article in press as: Cannas M., Puggioni G., On the support of matching algorithms. Statistics and Probability Letters (2017), http://dx.doi.org/10.1016/j.spl.2017.08.009.

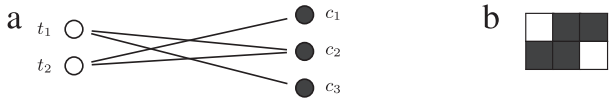

Fig. 1. Two representations of the matching: $t_1 \to c_2, c_3$; $t_2 \to c_1, c_2$. The associated matched dataset is the pair $(T, B) \equiv (\{t_1, t_2\}, \{c_1, c_2, c_2, c_3\})$.

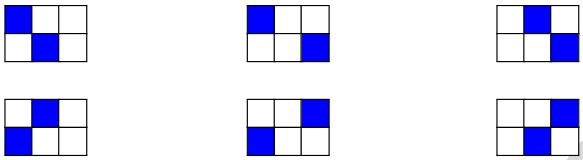

Fig. 2. All possible 1-1 matchings without replacement for the case $m = 2,\ n = 3$. Each vertical pair corresponds to a matched dataset.

are a function of the difference in means between the treated and controls. Thus, for inferential purposes, it is irrelevant how treated units are paired to subset of control units, provided that the same treated and control units are selected and each control is repeated the same number of times. Definition 1.1 does not require all subsets to be non-empty. In fact, it is not always possible to find suitable controls for all treated units, which are then eliminated. A matching where all treated units are assigned at least one control unit is a *full* matching, otherwise it is a *partial matching*.

Remark 1.1. The matching can be visualized using common "point-edge" representation for bipartite graphs (Fig. 1, left). An alternative representation is given by a board of dimensions $m \times n$ where m is the number of treated, n is the number of controls and matches are represented by darker squares (Fig. 1, right).

Remark 1.2. It follows from Definition 1.1 that the same matched dataset arises from several matchings (see Fig. 2), so the number of matched datasets is always less than or equal the number of matchings, but they are not necessarily in fixed proportion.

It is apparent that a matching algorithm works on a finite matching space (i.e. its *support*) given by the set of all possible matchings and matched samples that the algorithm can reach given its specific characteristics. Following Rosenbaum (2002) we distinguish some major characteristics of matching algorithms affecting the support size. A matching (algorithm) is *with replacement* if a control unit can be assigned to more than one treated unit, otherwise it is *without replacement*. Furthermore, if the number of control units assigned to each treated unit is the same for all treated units we have a *fixed size* matching, the main case being that of size one (1-to-1 matching), otherwise we have a *variable* matching. Finally, although Definition 1.1 does not mention the need of a specific metric, virtually all matching algorithms use a (pseudo) metric to evaluate the similarity between pair of observations in the covariates space.

In this paper, we give enumeration results for both matched datasets and matchings as a function of the different specifications of the matching algorithm: with replacement, one to one; with replacement, one to variable; without replacement, one to one; without replacement, one to variable. Sections 2 and 3 cover the full and partial cases, respectively. In Section 4 we show how the number of covariates can shrink the size of the matching space. Section 5 discusses aspects of matching space size and optimality.

2. Full matching

Let us consider the case where each treated unit has been associated with at least one control unit. A full matching is highly desirable in causal inference studies focusing on the Average Treatment effect on the Treated (ATT), which is a common target parameter. A nice feature of the counting problem being considered here is that all results can be derived using basic combinatorics, such as the addition, multiplication and inclusion–exclusion principles. In order to prove the result we recall a classic result and derive a corollary.

Lemma 2.1. *The number of solutions of*

$$x_1 + \cdots + x_n = m \tag{1}$$

in non-negative integers is $\binom{m+n-1}{m}$.

Corollary. *The number of solutions of Eq. (1) in non-negative integers, subject to the constraint* $x_i \geq c > 0$ *for a fixed* i*, is* $\binom{m-c+n-1}{m-c}$.

Please cite this article in press as: Cannas M., Puggioni G., On the support of matching algorithms. Statistics and Probability Letters (2017), http://dx.doi.org/10.1016/j.spl.2017.08.009.

Table 1
Number of matchings and matched datasets from m treated and n controls.

Replacement		1 to 1	1 to variable
Yes	Matchings	n^m	$(2^n-1)^m$
	Matched	$\binom{n+m-1}{m}$	$\sum_{k=m}^{mn}\sum_{j=0}^{n}(-1)^j\binom{n}{j}\binom{n+k-1-j(m+1)}{n-1}$
No	Matchings	$\frac{n!}{(n-m)!}$	$\sum_{k=m}^{n}\binom{n}{k}S(k,m)m!$
	Matched	$\binom{n}{m}$	$\sum_{k=m}^{n}\binom{n}{k}$

Proof. If we add the restriction that one of the x_i must be greater or equal than c, the problem can be solved by setting $x_i = y_i + c$ and counting the number of non-negative integer solutions of the resulting equation using Lemma 2.1. □

We can now state the two main propositions, concerning the cases with and without replacement, respectively. All results are summarized in Table 1.

Proposition 2.1. *Given m treated units and n control units the size of the matching space when matching with replacement is given in the upper part of Table 1*

Proof.

(i) *matchings 1 to 1* For a 1 to 1 matching exactly one out of n controls must be chosen for each of the m treated unit so by multiplication principle there are n^m total possibilities.

(ii) *matchings 1 to v* One out of $2^n - 1$ non empty elements of $\mathcal{P}(C)$ must be chosen for each of the m treated and the result again follows from multiplication principle.

(iii) *matched 1 to 1* According to Definition 1.1 this is the number of multisets of size m which can be extracted from the set of n controls. Let x_i be the number of times that control i matched, $i = 1, \ldots, n$. Then the number required is $\binom{m+n-1}{m}$ by Lemma 2.1.

(iv) *matched 1 to v* This case is analogous to the previous one but now the number of matched controls (counted with their multiplicity) can be any k in $\{m, \ldots, mn\}$. Moreover, we further require that $x_i \leq m$ for each i since no control can be matched more than m times. Thus, for fixed k and i, the number of matched datasets can be found from Corollary by putting $c = m$ and excluding solutions satisfying the constraint. Let a solution have property P_i if $x_i \geq m + 1$ holds for exactly i indices, $i = 1 \cdots n$. We are looking for the number of solutions having none of the properties, which can be found by subtracting from the total number of solutions the number of solutions having at least one P_i. On using the inclusion–exclusion principle to avoid overcounting when summing over i we find:

$$\binom{k+n-1}{k} - \left[\binom{n}{1}\binom{k+n-1-(m+1)}{k-(m+1)} - \binom{n}{2}\binom{k+n-1-2(m+1)}{k-2(m+1)} + \cdots + (-1)^{n-1}\binom{n}{n}\binom{k+n-1-n(m+1)}{k-n(m+1)}\right].$$

Summing the previous expression over $k = m, \ldots, mn$ and rearranging the terms gives the result. □

Proposition 2.2. *Given m treated and n control units the size of the matching space when matching without replacement is as in the lower part of Table 1.*

Proof.

(i) *matchings 1-1* There are n possibilities for matching the first treated unit, $n - 1$ for the second and so on until the mth unit, which can be matched in $n - m + 1$ ways. The result follows by multiplication principle.

(ii) *matchings 1-variable* Consider a k-element subset of the controls. Any partition of the subset into m parts gives $m!$ different matchings by considering all possible ways of assigning its parts to the treated. Thus, the total number of variable matchings using k controls is $S(k, m)m!$, where $S(k, m)$ is the number of partitions of $\{1, \ldots, k\}$ into m parts (i.e. the Stirling number of the second kind). Note that $k \geq m$ in order to have a full match and there are $\binom{n}{k}$ ways of choosing these k distinct controls. The result follows by multiplication principle.

(iii) *matched 1-1* To build a 1-1 matched dataset m distinct controls must be selected from the n available ones. The total number of matched datasets is then the number of selections of m distinct objects from n objects, i.e. $\binom{n}{m}$.

(iv) *matched 1-variable* The previous argument easily extends to the variable case. Here the number of controls to be selected is at least m and no more than n and the result follows from the addition principle. □

Please cite this article in press as: Cannas M., Puggioni G., On the support of matching algorithms. Statistics and Probability Letters (2017), http://dx.doi.org/10.1016/j.spl.2017.08.009.

3. Partial matching

If some of the treated units could not be paired with any control units we have a partial matching (and thus a partial matched dataset) that can be considered a particular case of the counting problems introduced so far, if only s treated units can be paired, $s = 0, \ldots, m$. Hence, it is possible to use the following three step procedure: first, find the number of ways of selecting s treated units, that is $\binom{m}{s}$; then, find the number of *full* matchings (matched datasets) with s units; finally, multiply the numbers above and sum over s.

Example 3.1. Let $a(n, m)$ be the number of 1-1 matched dataset without replacement when drops are allowed. Since there are $\binom{m}{s}$ ways of choosing s treated units and $\binom{n}{s}$ matched datasets with s units, then $a(n, m) = \sum_{s=0}^{m} \binom{m}{s}\binom{n}{s}$.

Although the above procedure is general easier formulae can sometimes be obtained by *ad hoc* reasoning. In the example above first add m empty controls to the n original ones to allow for the possibility of drops and then observe that in this new setting a matched dataset can be obtained in $\binom{m+n}{m}$ ways. Similarly, it turns out that compact formulae for the number of partial matchings can be obtained using rook polynomials.

3.1. Rook polynomials

The board representation in Fig. 1 is particularly amenable to be used in conjunction with a nice counting tool, known as the rook polynomial, which was first introduced by John Riordan (see Riordan, 2002, Ch 7) to count permutations with restricted positions.

Definition 3.1. The rook polynomial of a board C is

$$r(x, C) = r_0(C) + r_1(C)x + r_2(C)x^2 + \cdots$$

where $r_k(C)$ is the number of ways of placing k *non-attacking* rooks on C.

Here non-attacking means that no two rooks can be on the same row or column and so they cannot attack each other under classic chess rules. For example, if C is 2×3 then $r(x, C) = 1 + 6x + 6x^2$ and the coefficient of x^2 is the number of ways of putting two non attacking rooks on C (see Fig. 2). For a general $m \times n$ board it can be shown that $r_k(C) = \binom{n}{k}\frac{m!}{(m-k)!}$. For our purposes we can slightly extend the notion of rook polynomial to prove the following

Lemma 3.1. *Let $r^{nh}(C, x)$ and $r^{nv}(C, x)$ be the rook polynomials for non-horizontally and non-vertically attacking rooks, respectively. The generic coefficients of these polynomials are*

$$r_k^{nv}(C) = \binom{n}{k} m^k \qquad \text{and} \qquad r_k^{no}(C) = \binom{m}{k} n^k.$$

Proof. Note that it is enough to prove the first since $r_k^{no}(C_{m\times n}) = r_k^{nv}(C_{n\times m})$ by reversing the table. To prove the first consider that the rooks must be on different columns and k different columns can be selected in $\binom{n}{k}$ ways. Then the rooks must be placed in any of the m rows and this can be done in m^k ways. The result now follows from multiplication principle. □

We thus have the following proposition:

Proposition 3.1. *If drops are allowed the number of*: (i) *1-1 matchings without replacement is* $\sum_{k=1}^{m}\binom{n}{k}\binom{m}{k}m!$ (ii) *1-variable matchings without replacement is* $\sum_{k=1}^{n}\binom{n}{k}m^k$ (iii) *1-1 matchings with replacement is* $\sum_{k=1}^{m}\binom{m}{k}n^k$ (iv) *1-variable matchings with replacement is* $\sum_{k=1}^{mn}\binom{mn}{k}$.

Proof. The properties of non-row and non-column attacking correspond to the 1-1 and the non replacement features of the matching, respectively. Thus: (i) a 1-1 matching without replacement can be obtained putting $k = 1, \ldots, m-1$ non attacking rooks on C; (ii) a 1-variable matching without replacement can be obtained putting $k = 1, \ldots, n$ non column attacking rooks on C; (iii) a 1-1 matching with replacement can be obtained by placing $k = 1, \ldots, m$ non-row attacking rooks on C; and (iv) for a 1-variable matching with replacement just place $k = 1, \ldots, mn$ rooks without requiring the properties. Summing over the suitable rook coefficients gives the results. □

4. The radius matching case

The enumeration results obtained so far concerned a general matching space. However, the effective size of the matching support usually depends also on other aspects, such as the number of covariates available. We illustrate this point considering the *radius matching* procedure first proposed by Rosenbaum and Rubin (1985) and recently reviewed by Huber et al. (2015). In radius matching, euclidean distances between all pairs of treated and control units are calculated. Then, for each treated unit, all controls within a fixed radius $\epsilon > 0$ are used as matches. The maximal support of this algorithm is the set of one

Please cite this article in press as: Cannas M., Puggioni G., On the support of matching algorithms. Statistics and Probability Letters (2017), http://dx.doi.org/10.1016/j.spl.2017.08.009.

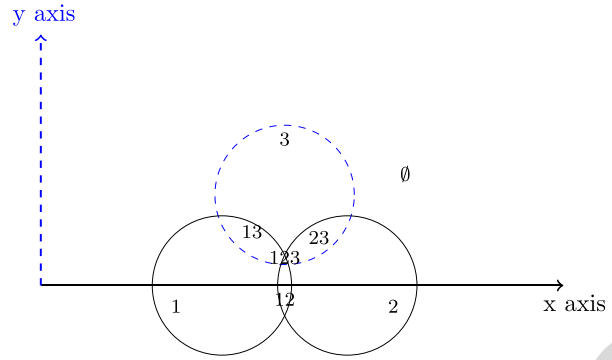

Fig. 3. When $n = 3$ feasible matchings can have no more than (i) $r(n = 3, p = 2) = 7$ distinct elements in $\mathcal{R}^2$; (ii) $r(n = 3, p = 1) = 5$ distinct elements in $\mathcal{R}$.

to variable matchings. However, some matchings are unattainable if distances are calculated in a "low" dimensional space. In fact, the maximum possible number of distinct subsets of the controls that can appear in a matching depends on the dimension of the space where the units are paired and so on the number of covariates p. To clarify this point let us consider treated units t_i and controls c_j as points in the p-dimensional space and define $A_i \equiv \{j| \quad dist(t_i, c_j) < \epsilon\}$ with $i = 1, \ldots, m$ as the set of the labels of the controls matched to the ith treated unit.The family A_i univocally identifies a matching. We call the family *feasible* if there exists such an $\epsilon > 0$.

The feasibility can be interpreted in terms of Venn diagrams, in which the centers of circles around the control units play the role of the c_j and points t_i corresponding to treated are placed in regions of overlap/non-overlap of circles according to the desired subset A_i. It is well known (since Venn, 1880) that in the Euclidean plane the maximum number of circles of equal size (ϵ-neighborhoods) that can be arranged to give all possible subsets is three (see Fig. 3). Additional circles might be added to such a diagram, but not all possibilities of overlap/non-overlap will be achieved. The generalization to p-dimensional Euclidean spaces has been considered in the literature. A result that at most $p+1$ "independent" (hyper)spheres are possible in a p-dimensional space was reported by Rényi et al. (1951) with some corrections given in Grünbam (1975) (this paper may be read online, Venn diagrams and independent families of sets). It follows that $n - 1$ is a sufficiently high dimension to allow feasibility of all 2^n possible subsets A_i of $\{1, \ldots, n\}$. We find that the size of the matching space is smaller than the maximal one whenever the number of covariates p is smaller than $(n - 1)$. In fact, we can prove the following

Proposition 4.1. *Let m and n be the number of treated and control units and p be the number of covariates. If (Euclidean) radius matching is performed the size of the matching space is not greater than:*

$$n_{m,n,p} = \sum_{k=1}^{2^{p+1}-1+2^p(n-p-1)} \binom{2^n - 1}{k} a_k$$

where $a_1 = 1$ and $a_j = j^m - \binom{j}{1}a_1 - \binom{j}{2}a_2 - \cdots - \binom{j}{j-1}a_{j-1}$.

Proof. Let r be the maximum number of distinct elements in a matching. We say that a matching has property P_i if it has *exactly* i distinct elements. If we denote by N_{P_i} the number of matchings satisfying the ith property we have

$$n_{m,n,p} = \sum_{i=1}^{r} N_{P_i} = \binom{2^n - 1}{1} a_1 + \binom{2^n - 1}{2} a_2 + \cdots + \binom{2^n - 1}{r} a_r$$

where a_k denotes the number of matchings that can be obtained with k *distinct* elements. Clearly, $a_1 = 1$. Also, the value of a_j can be found by subtracting from the total number of matchings with j elements the number of matchings with less than j distinct elements, so $a_j = j^m - \sum_{s<j}\binom{j}{s}a_s$.

It remains to determine $r = r(n, p)$. In a p-dimensional space there are no more than $p+1$ independent circles and then no more than $(2^{p+1}-1)$ *non empty* distinct regions A_i (see Fig. 3). These regions can be generated if there are at least $p+1$ controls and any additional control generates no more than 2^p further regions, so $r = (2^{p+1}-1)+2^p(n-p-1) = 2^p(n-p+1)-1$. □

When $p \geq n - 1$ we have $r = 2^n - 1$ and thus $n_{m,n,p} - (2^n - 1)^m = 0$, as expected from Proposition 2.1; otherwise the difference is positive and increases as p goes to one. This implies that it is advisable to calculate distances at least on a $(p - 1)$ dimensional space to avoid undesired shrinking of the matching space, i.e there must be at least $(p - 1)$ *matching covariates*.

5. Discussion

In this work we used elementary combinatorics to find the maximal size of the matching space as function of the number of treated and controls and we showed that the size of the matching space can shrink when metric matching is performed.

Please cite this article in press as: Cannas M., Puggioni G., On the support of matching algorithms. Statistics and Probability Letters (2017), http://dx.doi.org/10.1016/j.spl.2017.08.009.

Causal inference using observational data requires careful selection of a matched dataset in order to find reliable estimates of causal effects. Usually a "good" matched dataset is found when some measure of covariate imbalance is minimized and validated through reasonable diagnostics. Finding the optimum by enumerating all datasets is almost always impossible when the size of the matching space is too large. In order to mitigate this problem two routes have been chosen. The first is based on classic combinatorial optimization algorithms. An optimal matching procedure based on the Hungarian algorithm was proposed by Rosenbaum (1989). While moving on the whole matching space the method can be computationally intensive. More recently Ming and Rosenbaum (2001) reformulated the optimal search as an assignment problem. The second, perhaps more popular, route involves the basic idea of moving on a subset of the matching space, sacrificing optimality to gain computational feasibility. One or both of the following principles have been followed: moves on a random portion of the matching space; moves on the restricted matching space induced by some kind of covariate compression, for instance the propensity score (PS) (Rosenbaum and Rubin, 1983) or through calipers.

The use of "greedy" matching algorithms which do not allow to change an already matched pair is an example of random moving and may imply the loss of the optimal matching. Similarly, the PS forces the matching algorithm to work in a one-dimensional space which is generally smaller than the original one. This shrinking is desired and it is justified by demonstrating the large samples balancing property of the PS. However exact knowledge of the PS is rare in practice and both estimation and replacement increase the variance of causal estimates. Intuitively, the larger the number of covariates pooled into the PS the higher the risk of losing good matchings. Rosenbaum recommends an hybrid strategy consisting in the joint use of optimal matching and PS calipers "to speed the algorithm that finds the optimal match" (see section 3.4 ibid.). However, this optimality is generally on the induced *restricted* matching space. In fact, if all the units inside the caliper are taken as matches this procedure is the optimal version of the radius matching cited in the example of Section 4.

Acknowledgments

We are grateful to the contributors of math.stackexchange.com (in particular to hardmath) for helping with the derivation of the dimensionality bound and for pointing us to relevant references in Section 4 and to an anonymous referee for helpful comments.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.spl.2017.08.009.

References

Grünbam, B., 1975. Venn diagrams and independent families of sets. Amer. Math. Monthly 82, 12–23.
Huber, M., Lechner, M., Steinmayr, A., 2015. Radius matching on the propensity score with bias adjustment: tuning parameters and finite sample behaviour. Empir. Econ. 49, 1–31.
Imbens, G.W., Rubin, D., 2016. Causal Inference for Statistics, Social and Biomedical Sciences. Cambridge University Press.
Ming, K., Rosenbaum, P., 2001. A note on optimal matching with variable controls using the assignment algorithm. J. Comput. Graph. Statist. 10, 455–463.
Rényi, A., Rényi, V., Surányi, J., 1951. Sur l'indépendance des domaines simples dans l'Espace Euclidien a n dimensions. Colloq. Math. 2, 130–135.
Riordan, J., 2002. Introduction to Combinatorial Analysis. Dover Publications, Inc., New York. Unabridged republication of the edition published by John Wiley and Sons, Inc., New York, 1958.
Rosenbaum, P., 1989. Optimal matching for observational studies. J. Amer. Statist. Assoc. 84, 1024–1029.
Rosenbaum, P., 2002. Observational Studies, Springer Series in Statistics, second ed.
Rosenbaum, P., Rubin, D., 1983. The central role of the propensity score for the estimation of causal effects in observational studies. Biometrika 70, 41–55.
Rosenbaum, P., Rubin, D., 1985. Constructing a control group using multivariate matched sampling options that incorporate the propensity score. Amer. Statist. 39, 33–38.
Stuart, E., 2010. Matching methods for causal inference: a review and a look forward. Statist. Sci. 25, 1–21.