



**“AN OPTIMIZATION-BASED MATCHING
ESTIMATOR: LARGE AND SMALL SAMPLE
PROPERTIES”**

**TESIS PARA OPTAR AL GRADO DE
DOCTOR EN ECONOMÍA**

Alumno: Juan Díaz Maureira

Profesor Guía: Jorge Rivera

Santiago, Diciembre 2014

AN OPTIMIZATION-BASED MATCHING
ESTIMATOR: LARGE AND SMALL SAMPLE
PROPERTIES

Juan Díaz Maureira

A thesis submitted for the degree of Doctor of
Philosophy at the University of Chile

December 2014

Thesis Advisor:

Professor Jorge Rivera

Committee members:

Professor Eduardo Engel

Professor Francisco Gallego

Professor Alejandro Jofré

ABSTRACT

This work proposes a novel matching estimator where weights and the choice of neighbors used are endogenously determined by solving an optimization problem. The estimator is non-parametric and is based on finding, for each unit that needs to be matched, sets of observations such that a convex combination of their covariates has the same value of the covariates as the unit to be matched, or with minimized distance between them. Since there is generally more than one set per each unit, the method chooses the one with the closest covariate values.

In this work we contribute to the matching literature by linking the choice of matches and weights to the improvement of post-matching covariate balance in a simple way: an optimization problem that optimizes individual covariate balance. It is worth mentioning that the developed method is not an algorithm that iteratively checks covariate balance until convergence. Instead, it incorporates an individual balance criterion in the objective function that determines the weights used in each match. It can be written as a linear program that allows us to use standard optimization techniques to solve the problem quickly. To aid research, we provide a new R library called `blopmatching`.

Regarding asymptotic properties, we show that our estimator of the ATE attains standard limit properties (consistency and normality), and it has a conditional bias that is $O_p(N^{-2/k})$. It is worth mentioning that this order improves the order $N^{1/k}$ attained by the NN -matching estimator. In fact, $O_p(N^{-2/k})$ could be attained by the NN -matching estimator in the only case in which the conditional expectation of the outcome variable is a linear expression in covariates, a condition under which the conditional bias of our estimators is as good as we want. Besides, even though the proposed estimator of the ATE is not \sqrt{N} -consistent in general, we show that if the number of control units increases faster than the number of treated units, then our matching estimator of the ATT attains the \sqrt{N} -consistency, as its bias rate is better than the one attained by the NN -matching estimator.

Finally, as regards finite sample properties, we implement the proposed estimator to data from the National Supported Work Demonstration finding an outstanding performance even though when using alternative control groups from a non-experimental sample. In addition, by performing Monte Carlo experiments with designs based on the related literature that includes misspecification of the selection equation, we study its performance in finite samples. We find that our estimator provides good post-matching balance and performs well in terms of bias and variance when compared to nearest neighbor matching estimators (for both covariates and propensity score) and the normalized inverse probability weighting estimator. Major improvements are observed when there is underspecification of the selection equation for estimating the propensity score. Hence, our method gives researchers a new alternative matching estimator that prevents the selection of an arbitrary number of neighbors or the estimation of the propensity score.

DEDICATORIA

A mi compañera Carla.

A mi hija Emilia y a los/las que estén por llegar.

A mis padres Ana Luisa y Juan Eduardo.

A mis hermanos Ivonne y Rodolfo.

ACKNOWLEDGEMENTS

I would like to thank the people that have contributed to the initiation and completion of my graduate studies.

I am indebted to my thesis supervisor and friend: Professor Jorge Rivera. To him I owe the completion of this thesis. Jorge may never fully realize the impact he has had on my academic career and life.

I would like to thank Professor Tomás Rau for his ideas and active participation in this work. His contribution played a key role in completion of the small sample properties of the proposed estimator. In fact, Professor Rau, my advisor, and I wrote an article called “*A matching estimator based on a bi-level optimization problem*”, which has already been accepted for publishing in “*The Review of Economics and Statistics*”. Some parts of the Introduction, §2, and §4 are based on that paper, and thus on Professor Rau’s contributions.

I would also like to thank Professor Roberto Cominetti for his ideas and contribution in the part related to asymptotic properties of the method. Without him, we would not have shown the improvement in the bias rate of the proposed estimator. Indeed, Professor Cominetti, my advisor, and I are finishing an article called “*Large sample properties of an optimization-based matching estimator*”, which will soon be sent to a journal for its review. Some parts of the Introduction, §2, and §3 are based on that work, and thus on Professor Cominetti’s contributions.

I would like to thank George Vega and Daniel Espinoza for helping us with the R library and Justin McCrary for providing us copies of the code used to generate the results in Busso, DiNardo and McCrary (forthcoming).

I gratefully acknowledge the financial support of Conicyt fellowship. Additional funding was provided by Fondecyt, Projects No. 11095181 (Jorge Rivera and Tomás Rau) and No. 1130468 (Jorge Rivera and Roberto Cominetti).

Last but not least, I would like to thank the committee members of this thesis: Professor Eduardo Engel, Professor Francisco Gallego, and Professor Alejandro Jofré.

CONTENTS

1. <i>Introduction</i>	6
2. <i>The optimization-based matching estimator</i>	11
2.1 Basic concepts and notation	11
2.2 Formal aspects of the optimization-based matching estimator	12
2.3 Example	14
3. <i>Large sample properties</i>	17
3.1 Standing assumptions	18
3.2 Some results in geometric probability theory	19
3.2.1 The probability of not being in the convex hull of the nearest neighbors	19
3.2.2 The number of times that a unit is used as a match	22
3.3 Asymptotic properties	25
3.3.1 The order of the conditional bias	25
3.3.2 Variance, consistency, and normality properties	30
3.3.3 Estimating the variance	32
4. <i>Small sample properties</i>	34
4.1 NSW Demonstration	34
4.2 Empirical Monte Carlo	35
4.3 Finite sample properties and misspecification	38
5. <i>Appendix: Figures and tables</i>	40

1. INTRODUCTION

Matching estimators¹ have been widely used in the impact evaluation literature during the past decades. These methods essentially rest on the imputation of a *potential outcome* to an individual, built as a weighted average of the observed outcomes of his closest neighbors from the corresponding counterfactual set. Given that, the individual treatment effect is usually defined as the difference between the imputed and actual outcome. One popular estimator is the simple matching estimator (also known as nearest neighbor matching estimator) studied by Abadie & Imbens (2006), where the outcome to be imputed is defined as the average of the outcomes from a certain number of closest neighbors, with closeness defined by a distance induced by a norm. The choice of the number of neighbors is up to the researcher and the weights are simply the reciprocal of this number. As Imbens & Wooldridge (2009) state, “little is known about the optimal number of matches, or about data-dependent ways of choosing it”.

In this thesis we propose a matching estimator where the number of units used in each match and the weighting scheme are endogenously determined from the solution of an optimization problem. The first task deals with finding sets of observations such that a convex combination of their covariates has exactly the same covariate values as the unit to be matched, when possible, or otherwise where their distance is minimized. Since this problem may have more than one solution, the second task consists on implementing a refinement criterion that looks for the set with the closest covariate values to the unit to be matched. These problems (involving the two tasks described) can be written as one optimization problem whose admissible points belong to the solution set of another optimization problem. In the optimization literature this is called a *bi-level optimization problem* (BLOP).²

To fix ideas, assume we are interested in finding a match for individual i with a unique real-valued characteristic $X_i \in \mathbb{R}$ and outcome variable $Y_i \in \mathbb{R}$. Also assume that individuals in the counterfactual group are indexed by $j = 1, \dots, n$, $n \geq 2$, with characteristics and outcomes given by $X_j \in \mathbb{R}$ and $Y_j \in \mathbb{R}$ respectively. Without loss of generality, let us assume that $X_1 < \dots < X_s < X_i < X_{s+1} < \dots < X_n$. The first task of this optimization problem is

¹ This Introduction is a convex combination of the Introductions of two papers. One belongs to the work named “*A matching estimator based on a bi-level optimization problem*”, which has been written by Professor Tomás Rau, Professor Jorge Rivera, and me. The second one is in the article called “Large sample properties of an optimization-based matching estimator”, which has been written by Professor Jorge Rivera, Professor Roberto Cominetti, and me. Thus, some parts of this Introduction are based on their contributions.

² For details see Colson, Marcotte & Savard (2007).

to look for individuals in the counterfactual group such that the convex combination of their characteristics is as close as possible to X_i . This problem has many solutions. For instance, combining X_s and X_{s+1} will match X_i but the combination of X_1 with X_2 and X_n will also get an exact match. Hence, the second task is another optimization problem that identifies the solution that minimizes the sum of distances to the power of two between X_i and the covariates from individuals used from the counterfactual group. In this example the solution of the BLOP is the weighting scheme given by

$$\lambda_s = \frac{X_{s+1} - X_i}{X_{s+1} - X_s}, \quad \lambda_{s+1} = \frac{X_i - X_s}{X_{s+1} - X_s}, \quad \lambda_j = 0, \quad j \neq s, s + 1,$$

which exactly matches X_i .³ Hence, our approach uses only units s and $s + 1$ to perform the match and, according to this approach, the potential outcome to be imputed to individual i is

$$\hat{Y}_i = \sum_{j=1}^n \lambda_j Y_j = \lambda_s Y_s + \lambda_{s+1} Y_{s+1}.$$

When $X \in \mathbb{R}^k$ is a k -dimensional vector of characteristics ($k > 1$), solving the BLOP could be difficult due to the great number of alternatives that any optimization algorithm has to evaluate in order to achieve a global solution. To circumvent this complexity, we present an equivalent formulation of the BLOP as a *linear programming* problem, for which there is a vast literature in optimization that allows us to solve the BLOP efficiently.

Regarding the existing literature, this estimator is related to recent methods that use covariate balance as a metric for selecting either the propensity score model or tuning parameters. Diamond & Sekhon (2013) propose a search algorithm to iteratively check and improve covariate balance. This is done by using a generalized version of the Mahalanobis distance that places different weights on the covariates used in the matching algorithm. The algorithm then iteratively chooses the weighting matrix that provides the best covariate balance according to a loss function depending on measures of imbalance such as the Kolmogorov-Smirnov test statistic. These weights are common for all the units to be matched, generating a distance metric to perform the matching optimizing post-matching covariate balance. Our approach differs from theirs since it looks for different weights that minimizes the distance between the match (built with these weights) and the unit. This optimization is done for each observation that needs to be matched so the weights are different observation per observation. Thus, the method we develop is an optimization problem (not an iterative algorithm) that jointly finds the weights and determines the number of neighbors used per each observation. The choice of neighbors is determined by the fact that units with zero weight are then not used in the match. Meanwhile, Graham, Pinto & Egel (2012) propose a modified inverse probability

³ It is straightforward to check that $X_i = \lambda_s X_s + \lambda_{s+1} X_{s+1}$. Additionally, as it will be discussed later, the number of matches will be determined by a refinement criterion that minimizes the sum of distances between X_i and the covariates from the units used from the counterfactual group.

weighting estimator that also maximizes covariate balance and has the property of being doubly robust (see Robins, Rotnitzky & Zhao (1994), Robins, Rotnitzky & Zhao (1995)). These weights are optimally estimated to balance the covariates and inversely depend on estimates propensity score estimates. Finally, Imai & Ratkovic (2014) propose an inverse propensity score weighting estimator that simultaneously optimizes the covariate balance and the prediction of treatment assignment. The last two approaches are inverse probability weighting methods that require the estimation of the propensity score. Our method, however, is a nonparametric matching method in characteristics with an optimal choice of weights that by construction improves the covariate balance.

This thesis is also related to the literature that tries to determine the number of neighbors to use when doing matching. Even though there is not a theoretical foundation, many researchers use cross-validation to select the number of neighbors. For instance, Frölich (2004) uses that in his investigations of finite sample properties of matching estimators. Also, Galdo, Smith & Black (2008) discuss how to select the bandwidth when doing matching following a weighted cross-validation strategy. The BLOP strategy solves the issue of the number of neighbors selection since the optimal weights that are different from zero define the units from the counterfactual group, which participate in the optimal convex combination. Indeed, from Caratheodory's Theorem (see Rockafellar (1972)), the number of units should be at most the pretreatment characteristic vector dimension plus one. This is relevant since when the estimation is implemented employing the entire sample (as we propose), there is no need to fix the number of nearest neighbors to be used for estimation.

After presenting the notation, intuition, and the formal aspects of the proposed method in §2, we show in §3 the large sample properties of the matching estimator we introduce. It is worth mentioning that asymptotic properties of nonparametric matching estimators have been scarcely studied in the program evaluation literature. As far as we know, the most general results were provided by Abadie & Imbens (2006) when presenting, under general conditions, the limit properties of the well-known *NN-matching estimator*. In particular, they show that this estimator for the ATE has a conditional bias whose stochastic order is $N^{1/k}$, with k the dimension of continuous covariates. Indeed, by directly applying their results, and after adapting some notations, it can be shown that the same asymptotic properties hold true for any matching estimator in which the missing potential outcome –see Rosenbaum & Rubin (1983)– to be imputed to any unit that needs to be matched is defined as a weighted average of observed outcomes of a *fixed number* of its first nearest neighbors having the opposite treatment, with weighting schemes for doing so depending on units or not.

The main results of §3 are two. First, under practically the same conditions as those used by Abadie & Imbens (2006), we show that the BLOP matching estimator of the ATE attains standard limit properties, and has a conditional bias that is $O_p(N^{-2/k})$, where k is the number of continuous pretreatment variables and N is the sample size. It is important

to say that this order improves the order $N^{1/k}$ attained by the NN-matching estimator. In fact, $O_p(N^{-2/k})$ could be attained by the *NN-matching estimator* in the only case in which the conditional expectation of the outcome variable is a linear expression in covariates, a condition under which the conditional bias of BLOP estimators is as good as one wants –see Theorem 3.3.1 in §3.3–. Second, even though the BLOP matching estimator of the ATE is not \sqrt{N} -consistent, we show that if the number of control units increases faster than the number of treated units, then the BLOP matching estimator of the ATT attains the \sqrt{N} -consistency, and its bias rate is better than the one attained by the *NN-matching estimator*.

As we have mentioned, an important aspect concerning the BLOP approach is that although the optimization problems involved in its definition use the entire sample of counterfactuals to perform the missing potential outcome, from Caratheodory’s Theorem –see Rockafellar (1972)– it follows that the number of them that actively participate in the convex combination performing the covariates of the unit that is being matched is, at most, $k + 1$. I have, however, that these units are not necessarily the first $k + 1$ nearest neighbors to it. This “*lack of control*” with respect to the closeness of units employed is precisely one of the fundamental difficulties we face when developing our results. Nevertheless, we overcome this challenge by using some novel contributions in geometric probability theory we develop.

Due to the nature of the BLOP matching estimator, an initial result we will need is the probability that a random vector does not belong to the convex hull of a certain number of nearest neighbors. Using a property in Cover & Efron (1967), which extends a result in Wendel (1962), Theorem 3.2.1 in §3.2.1 states that this probability can be bounded above by an expression that converges exponentially to zero with the number of matches employed. This fact, along with the nature of BLOP’s solution, will allow us to overcome the aforementioned *lack of control*.⁴ Note that when the number of counterfactuals employed is fixed exogenously as in most standard matching approaches –see Imbens & Wooldridge (2009)–, this probability is a constant value, hence the norm of matching discrepancies, and therefore the balance in terms of covariates reached by the method, becomes the relevant expressions defining the order of the conditional bias.

On the other hand, once the BLOP is solved, let us say, for a control unit, it is then defined a *random polytope*⁵ that is given by the convex hull of covariates of treated units actively participating in the BLOP’s solution for this control unit. Of course a treated unit participates in such realization whenever it is a vertex of such polytope. In §3.2.2 we investigate the number of times that, in expected value, a treated (control) unit is a vertex

⁴ Under the standing assumptions here assumed, Theorem 5.4 in Evans, Jones & Schmidt (2002) states that the α -moments of the norm of the M -matching discrepancy can be bounded above by an expression that is *polynomial* in M (with degree equal to α). Therefore, even for the “worst case” when the BLOP uses the farthest counterfactual in the sample, the “exponential decreasing” of the probability overcomes the “polynomial increasing” of distances between covariates, a result that leads us to conclude that the expected value of the balance of covariates reached by the BLOP converges to zero exponentially with the size of the sample. Hence, the BLOP approach restores relevance to the weighting scheme for the order obtained.

⁵ See Majumdar, Comtet & Randon-Furling (2010) for concepts and main properties of random polytopes.

of the polytopes arising when the BLOP is solved for all control (treated) units. Proposition 3.2.2 in that section states that, under general conditions, this number can be bounded above by a constant that does not depend on the sample size. By using this result, we are able to show the asymptotic normality and variance properties of the BLOP estimators.

Finally, to assess the performance of the BLOP estimator in finite samples, in Section §4 we implement different empirical designs and Monte Carlo exercises. We use data from National Supported Work Demonstration (NSW) to see its performance compared to its natural competitor: the nearest neighbor matching estimator. Then, using the data generating processes from Busso, DiNardo & McCrary (forthcoming), we implement Monte Carlo experiments in order to assess its performance in terms of absolute bias, variance and covariate balance. we find significant improvements in comparison to other matching estimators employed in the literature, especially when the propensity score is under specified. In this part of the work we use a consistent estimator of its marginal variance that we provide in §2. A new **R** package called **blopmatching** implements the method.

This thesis is organized as follows. After this Introduction, §2 presents the methodology and the assumptions employed. In addition to that, in §2 we give some intuition behind the estimator and exemplify the method by using a simple setup. Then, §3 shows the large sample properties of the proposed method. To end, §4 presents the performance of the introduced method in finite samples.

2. THE OPTIMIZATION-BASED MATCHING ESTIMATOR

In this section¹ we illustrate and formalize the proposed matching estimator beginning with an introduction of some basic notations concerning binary program evaluation and certain mathematical concepts needed for properly setting up the method. In addition to that, this part also gives some intuition behind the estimator we propose.

2.1 Basic concepts and notation

Following Imbens & Wooldridge (2009), the binary program to be evaluated is represented by a random variable $\Omega = (W, Y, X)$, with $W \in \{0, 1\}$ indicating whether a treatment was received ($W = 1$) or not ($W = 0$) by the individual whose covariates or pretreatment characteristics is the vector $X \in \mathbb{X} \subseteq \mathbb{R}^k$. The observed outcome is $Y = WY(1) + (1 - W)Y(0) \in \mathbb{R}$, with $Y(1)$ and $Y(0)$ being the potential outcomes –see Rosenbaum & Rubin (1983) and Rubin (1973)–. Given the above, the average treatment effect, ATE, of the program is² $\tau = \mathbb{E}(Y(1) - Y(0))$, while the average treatment effect on the treated, ATT, is $\tau_{tre} = \mathbb{E}(Y(1) - Y(0) | W = 1)$. For $x \in \mathbb{X}$ and $w \in \{0, 1\}$, the conditional expectation and conditional variance of Y are, respectively, $\mu(x, w) = \mathbb{E}(Y | X = x, W = w)$ and $\sigma^2(x, w) = \mathbb{V}(Y | X = x, W = w)$.

A sample of size $N \in \mathbb{N}$ of Ω is denoted by $\Omega_N = \{(W_i, Y_i, X_i), i = 1, \dots, N\}$, and for this, N_0 and N_1 are the number of control and treated units, respectively. Of course, $N = N_0 + N_1$. We make the convention here and through the rest of the paper that the control units are indexed by $1, \dots, N_0$, so the treated ones are labeled as $N_0 + 1, \dots, N_0 + N_1$.

Without loss of generality, in this paper we use the Euclidean norm, $\|\cdot\|$, as the matching metric, and we also assume that the matching is performed with replacement. Given that, borrowed from Abadie & Imbens (2006), for $i \in \{1, \dots, N\}$ and $m \in \mathbb{N}$, $m \leq N_{1-W_i}$, we set

$$j_m(i) \in \begin{cases} \{1, \dots, N_0\} & \text{if } W_i = 1, \\ \{N_0 + 1, \dots, N\} & \text{if } W_i = 0, \end{cases}$$

¹ This Section contains some parts of two papers. On one hand, it is based on the work named “*A matching estimator based on a bi-level optimization problem*”, which has been written by Professor Tomás Rau, Professor Jorge Rivera, and me. On the other hand, it is based on the article called “*Large sample properties of an optimization-based matching estimator*”, which has been written by Professor Jorge Rivera, Professor Roberto Cominetti, and me. Thus, some parts of this Section are contributions of theirs.

² Throughout this paper, we denote the underlying probability by \mathbb{P} and mathematical expectation by \mathbb{E} .

as the index of the unit that is the m th nearest neighbor to unit i in the opposite treatment group.

As far as mathematical concepts are concerned, a vector sum $\lambda_1 X_1 + \dots + \lambda_N X_N$ is called a *convex combination* of vectors $X_1, \dots, X_N \in \mathbb{R}^k$ if the coefficients λ_j are all non-negative and $\lambda_1 + \dots + \lambda_N = 1$. The set of these weights is the *simplex* of dimension N , hereafter denoted by $\Delta_N = \left\{ (\lambda_1, \dots, \lambda_N) \in \mathbb{R}_+^N, \sum_{j=1}^N \lambda_j = 1 \right\}$. The *convex hull* of $\{X_1, \dots, X_N\}$ is the set of all convex combinations of these vectors, which throughout this thesis is denoted as $\text{co}\{X_1, \dots, X_N\} = \left\{ \sum_{j=1}^N \lambda_j X_j, (\lambda_1, \dots, \lambda_N) \in \Delta_N \right\}$.

We recall that the *projection* of $X \in \mathbb{R}^k$ onto $\text{co}\{X_1, \dots, X_N\}$ is, by definition, the nearest vector to X belonging to that set. Denoting it by $\text{Proj}(X)$, it follows that³

$$\|X - \text{Proj}(X)\| = \min_{Z \in \text{co}\{X_1, \dots, X_N\}} \|X - Z\| = \min_{(\lambda_1, \dots, \lambda_N) \in \Delta_N} \left\| X - \sum_{j=1}^N \lambda_j X_j \right\|. \quad (2.1)$$

Finally, for integer M , the convex hull of the first M matches to unit i will play a quite relevant role in this work. This subset is denoted by

$$\text{co}\{X_{j_1(i)}, \dots, X_{j_M(i)}\} = \left\{ \sum_{m=1}^M \lambda_m X_{j_m(i)}, (\lambda_1, \dots, \lambda_M) \in \Delta_M \right\}.$$

2.2 Formal aspects of the optimization-based matching estimator

We begin this part with a simple reasoning that will serve to present the main contributions of the large sample properties presented in this work. All the formal aspects and the proofs are postponed to Sections 3.2 and 3.3 below.

From Imbens & Wooldridge (2009) we already know that most matching methods approximate the unobserved (potential) outcome of a treated unit $i \in \{N_0 + 1, \dots, N\}$, namely $Y_i(0) = \mu_0(X_i)$ ignoring error terms, by an expression of the form

$$\widehat{Y}_i(0) = \sum_{m=1}^M \xi_m Y_{j_m(i)},$$

where M is an exogenous number of matches employed, and $(\xi_1, \dots, \xi_M) \in \Delta_M$ the weighting scheme used to perform the method.⁴ After performing a second order Taylor expansion of $Y_{j_m(i)} = \mu_1(X_{j_m(i)})$, $m = 1, \dots, M$, around X_i , it is not difficult to realize that, assuming

³ The uniqueness of this point comes from the fact that $\text{co}\{X_1, \dots, X_N\}$ is a *convex* and *compact* set. See Rockafellar (1972) for details on properties of convex sets.

⁴ The value of M is usually left to the researchers' criterion, and the manner that one defines the weighting schemes leading to the different matching methods currently available. For instance, the NN -matching estimator considers $\xi_m = 1/M$, $m = 1, \dots, M$, while for Kernel-based methods it is assumed that $\xi_m = K(1/\|X_i - X_{j_m(i)}\|)$, with $K(\cdot)$ a given kernel function –see Heckman, Ichimura & Todd (1998)–.

that there are constants $L_1, L_2 > 0$, the upper bounds of derivatives of μ_1 onto \mathbb{X} , that the *unit-level bias* can be approximated as

$$\left| \widehat{Y}_i(0) - Y_i(0) \right| \sim L_1 \left\| X_i - \sum_{m=1}^M \xi_m X_{j_m(i)} \right\| + L_2 \sum_{m=1}^M \xi_m \|X_i - X_{j_m(i)}\|^2, \quad (2.2)$$

its stochastic order thus depending on the order of the norm of matching discrepancies, and particularly from the *balance* in terms of covariates reached by the method.

In order to minimize the individual conditional bias, we could initially be naturally tempted to employ weighting schemes that minimize the right-hand side of the expression in (2.2), an approach that becomes pointless due to the fact that the constants involved in that expression are unknown. Given that, instead of attempting the minimization of that expression as a whole, we propose an approximated solution which, first of all, seeks the weighting schemes that minimize the covariates balance reached by the method and, once this optimization problem has been solved, we propose a second optimization problem to find the solution that minimizes the quadratic part of the approximation in (2.2). In order to achieve the best possible balance, we propose that the first problem should be solved using the entire sample of counterfactuals instead of a fixed number of matches as for standard approaches. Formally, for a treated unit $i \in \{N_0 + 1, \dots, N\}$, the first optimization problem our propose is

$$\mathcal{F}_i \quad : \quad \min_{(\xi_1, \dots, \xi_{N_0}) \in \Delta_{N_0}} \left\| X_i - \sum_{m=1}^{N_0} \xi_m X_m \right\|,$$

whose solution set is denoted by $\operatorname{argmin}\{\mathcal{F}_i\}$. Since that set contains several solutions, the second optimization problem we propose is

$$\mathcal{S}_i \quad : \quad \min_{(\lambda_1, \dots, \lambda_{N_0}) \in \operatorname{argmin}\{\mathcal{F}_i\}} \sum_{m=1}^{N_0} \lambda_m \|X_i - X_m\|^2. \quad (2.3)$$

Given the two aforementioned problems, note that they can be stated as a linear optimization problem. Indeed, it is straightforward to notice that the proposed weighting scheme

solves the following linear optimization problem

$$\begin{aligned}
& \min_{(\lambda_1, \dots, \lambda_{N_0})} \sum_{j=1}^{N_0} \lambda_j \|X_i - X_j\|^2 \\
& \text{s.t.} \\
& \sum_{j=1}^{N_0} \lambda_j X_j = \text{Proj}(X_i), \\
& \sum_{j=1}^{N_0} \lambda_j = 1, \quad \lambda_j \geq 0, \quad j = 1, \dots, N_0.
\end{aligned} \tag{2.4}$$

Thus the estimation can be efficiently implemented using optimization routines available in the literature. Due to the set of admissible points of problem (2.3) are the solutions of another optimization problem, namely (2.3), it is called a *bi-level optimization problem*, BLOP, (see Colson et al. (2007)).

The main difficulty in solving linear optimization problem (2.4) is determining the orthogonal projection of X_i onto the convex hull of opposite units covariates, $\text{Proj}(X_i)$, which can be solved efficiently using methods currently available in optimization literature (see Botkin & Stoer (2005)).

After properly configuring the problems above in terms of involving covariates for control units, the weighting scheme that solves problem \mathcal{S}_i for any unit i is denoted as⁵

$$\lambda^i = (\lambda_1^i, \dots, \lambda_{N_1 - W_i}^i) \in \Delta_{N_1 - W_i},$$

and therefore, the potential outcome imputed to this unit according to that approach is

$$\hat{Y}_i^b(0) = (1 - W_i)Y_i + W_i \sum_{m=1}^{N_0} \lambda_m^i Y_m, \quad \hat{Y}_i^b(1) = W_i Y_i + (1 - W_i) \sum_{m=1}^{N_1} \lambda_m^i Y_{m+N_0}.$$

Hence, the BLOP matching estimator of the ATE and ATT, denoted $\hat{\tau}^b$ and $\hat{\tau}_{tre}^b$ respectively, are given by

$$\hat{\tau}^b = \frac{1}{N} \sum_{i=1}^N \left(\hat{Y}_i^b(1) - \hat{Y}_i^b(0) \right), \quad \hat{\tau}_{tre}^b = \frac{1}{N_1} \sum_{i=1}^N W_i \left(\hat{Y}_i^b(1) - \hat{Y}_i^b(0) \right). \tag{2.5}$$

2.3 Example

Let the number of covariates be $k = 2$, and take a *treated* unit i , with observed outcome $Y_i(1) = Y_i \in \mathbb{R}$, and covariates $X_i = (1, a) \in \mathbb{R}^2$, with $0 \leq a \leq 1$. There are $N_0 \geq 3$

⁵ From well-known convexity properties –see Rockafellar (1972)–, $\text{argmin}\{\mathcal{F}_i\}$ is a nonempty, convex and compact subset, thus the optimization problem \mathcal{S}_i has always a solution. In fact, since we consider only continuous covariates, without loss of generality we may assume that this problem has a unique solution.

control units indexed by $j = 1, \dots, N_0$, each one having observed outcome $Y_j(0) = Y_j \in \mathbb{R}$, and covariates $X_1 = (0, 0)$, $X_2 = (2, 0)$ and $X_j = (1, 1 + j/N_0)$ for $j = 3, \dots, N_0$. Figure 2.1 illustrates the configuration of these covariates. There, the convex hull of $\{X_1, \dots, X_{N_0}\}$ is the shaded triangle with vertices X_1 , X_2 and X_{N_0} .

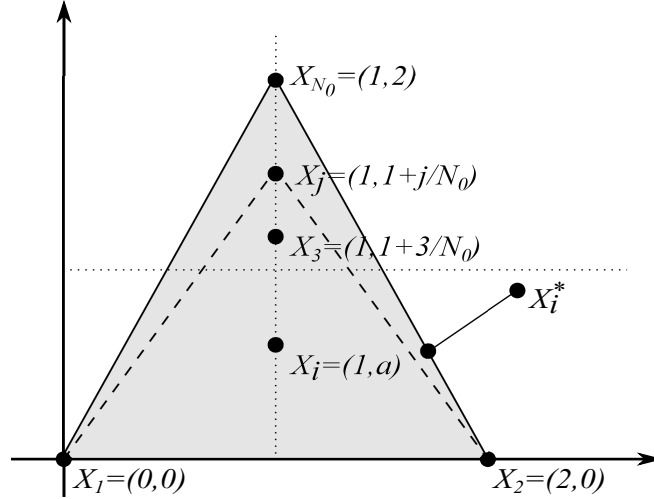


Fig. 2.1: Spatial configuration of covariates $\{X_1, \dots, X_{N_0}\}$ and the convex hull of them.

For a given $j \in \{3, \dots, N_0\}$, highlighted by dashed lines in Figure 2.1 we have that X_i belongs to the interior of the triangle with vertices X_1 , X_2 and X_j , which is equivalent to state that there is a vector of weights $(\lambda_1, \dots, \lambda_{N_0}) \in \Delta_{N_0}$ such that $X_i = \sum_{s=1}^{N_0} \lambda_s X_s$, with $\lambda_s > 0$ for $s = 1, 2, j$, and 0 otherwise. In fact, for a given j as before, it is easy to see the non-null components of this vector of weights are given by

$$\lambda_1 = \frac{j + (1 - a) N_0}{2(j + N_0)}, \quad \lambda_2 = \frac{j + (1 - a) N_0}{2(j + N_0)}, \quad \lambda_j = \frac{a N_0}{j + N_0}. \quad (2.6)$$

Consequently, varying j from 3 to N_0 , we have at least, $N_0 - 2$ solutions of problem (2.3). Leading from this is the question of which of these convex combinations should be chosen to match X_i . First note that any of these solutions generate convex combinations that exactly match X_i . At this stage, it is important to note that when the matching is performed using the nearest neighbor approach a perfect match is not guaranteed. For instance, when the parameter a is equal to 1 and the number of neighbors chosen is equal to 3, we have that the nearest neighbors to unit i are units 3, 4, and 5. It is easy to check that convex combinations of these unit's covariate will not achieve a perfect match for X_i . This occurs since the nearest neighbor approach does not incorporate an explicit covariate balance criterion when choosing the neighbors.

Consider now problem (2.3). When evaluated in the convex combination with covariates

X_1, X_2, X_j and weights from relation (2.6), its objective function value is

$$\lambda_1 \|X_i - X_1\|^2 + \lambda_2 \|X_i - X_2\|^2 + \lambda_j \|X_i - X_j\|^2 = 1 - a^2 + \frac{2aj}{j + N_0} + \frac{aj^2}{N_0(j + N_0)}, \quad (2.7)$$

which clearly attains a minimum value when $j = 3$. This corresponds to the triangle with the closest covariates to X_i . Thus, the weighting scheme is given by relation (2.6) evaluated at $j = 3$, implying that the missing potential outcome to impute to unit i according to this approach is

$$\widehat{Y}_i(0) = \left(\frac{3 + (1 - a)N_0}{2(3 + N_0)} \right) Y_1 + \left(\frac{3 + (1 - a)N_0}{2(3 + N_0)} \right) Y_2 + \left(\frac{aN_0}{3 + N_0} \right) Y_3,$$

where the estimated individual treatment effect for this unit is given by $Y_i - \widehat{Y}_i(0)$.

When the unit needing matching does not belong to the counterfactual group's convex hull of covariates, we find the convex combination that exactly balances the projection of its covariates onto that convex hull. Illustrated in Figure 2.1, this case occurs when X_i^* , instead of X_i , is the vector of covariates of unit i . Denoting by $\text{Proj}(X_i^*)$ the projection of X_i^* onto $\text{co}\{X_1, \dots, X_{N_0}\}$, our solution uses covariates X_2 and X_{N_0} to perform that projection, with optimal weights proportional to the distance from $\text{Proj}(X_i^*)$ to X_2 and $\text{Proj}(X_i^*)$ to X_{N_0} .

3. LARGE SAMPLE PROPERTIES

The goal of this part of the work¹ is to calculate the large sample properties of the matching estimator we have presented. However, before showing the asymptotic results, we state some needed results on geometric probability theory that will be used later.

From a geometric point of view, the problem \mathcal{F}_i for a treated unit $i \in \{N_0 + 1, \dots, N\}$ concerns the weighting schemes that serve to perform the *projection*² of X_i onto the convex hull of covariates of control units, $\mathbf{co}\{X_1, \dots, X_{N_0}\}$. The fact that the BLOP approach considers the entire sample of control units to perform that projection does not imply that all of them are finally employed to build that point. Indeed, a vector X_m , $m \in \{1, \dots, N_0\}$, participates actively in the construction of that projection when $\lambda_m^i > 0$, and from Caratheodory's Theorem –see Rockafellar (1972)–, we already know that their number should be, at most, $k + 1$. It is worth mentioning that these units are not necessarily the $k + 1$ first nearest neighbors to unit i , otherwise the limit properties of the BLOP matching estimators can be readily obtained from arguments employed in Abadie & Imbens (2006). These units are determined endogenously by the BLOP approach.

The “lack of control” regarding the closeness of units that participate in the realization of X_i by the BLOP approach is precisely the main difficulty we face for showing our results. In order to overcome this fundamental difficulty, denoting the value of problem \mathcal{F}_i for a treated unit i by

$$\nu\{\mathcal{F}_i\} = \left\| X_i - \sum_{m=1}^{N_0} \lambda_m^i X_m \right\|, \quad (3.1)$$

conditional on the sample, we have that the expected value of the balance reached by the BLOP method for this unit complies with

$$\mathbb{E}(\nu\{\mathcal{F}_i\} | W_i = 1, \{X_\iota, W_\iota\}_{\iota=1}^N) = \nu\{\mathcal{F}_i\} \mathbb{P}(\nu\{\mathcal{F}_i\} \neq 0). \quad (3.2)$$

When the sample size increases, it is clear that $\nu\{\mathcal{F}_i\}$ converges to zero. In fact, since that expression can be bounded above by $\|X_i - X_{j_1(i)}\|$, we have that, in the extreme, it can be

¹ This Section contains some parts of the work called “*Large sample properties of an optimization-based matching estimator*”, which has been written by Professor Jorge Rivera, Professor Roberto Cominetti, and me. In particular, Professor Cominetti played a key role in proving several results. Thus, this Section are based on Professor Cominetti's contributions.

² We recall that the projection of $X \in \mathbb{R}^k$ onto a convex set C is the vector that solves $\min_{c \in C} \|X - c\|$.

assumed to be $O(N^{-1/k})$. This result therefore gives relevance to the above probability for the order of that expectation. In this regard, we first note that $\nu\{\mathcal{F}_i\} \neq 0$ is equivalent to saying that $X_i \notin \text{co}\{X_1, \dots, X_{N_0}\}$. Hence, under the standing assumptions, a fundamental result for our purposes, which we show in §3.2.1, states that this probability can be bounded above by an expression that goes to zero exponentially in the number of control units. Therefore, regardless of the units that are used when solving the optimization problem \mathcal{F}_i , the conditional expected value in (3.2) attains an arbitrary order of convergence, implying that the order of the conditional bias of the BLOP matching estimator is dominated by the order of expectation of the *value* of the optimization problem \mathcal{S}_i , namely $\nu\{\mathcal{S}_i\}$, which for the treated unit i is given by

$$\nu\{\mathcal{S}_i\} = \sum_{m=1}^{N_0} \lambda_m^i \|X_i - X_m\|^2. \quad (3.3)$$

Of course, the aforementioned lack of control once again implies that we cannot use Lemma 2 in Abadie & Imbens (2006) to obtain the order of the conditional bias of $\hat{\tau}^b$. The techniques we use to conclude the proofs are developed in §3.3.

Finally, with the aim of studying the variance and asymptotic normality properties of the BLOP matching estimators, the aforementioned lack of control is, of course, the main issue we must overcome in order to obtain its standard limit properties. In this regard, and similarly to Abadie & Imbens (2006), the fundamental question to be addressed concerns the number of times, on average, that a certain unit was used as a match by the entire sample of its opposites after solving the optimization problem (2.3). The main result in §3.2.2 states that this number can be bounded above by a constant. Given that result, the limit properties of the BLOP matching estimator that remain to be completed follow directly from corresponding results in Abadie & Imbens (2006) for the NN -matching estimator.

3.1 Standing assumptions

Regarding aforementioned concepts, the following hypotheses are quite standard in the program evaluation literature, and they will be part of our standing assumptions.³

Assumption 1. Regularity conditions: \mathbb{X} is compact and convex, with unitary Lebesgue measure in \mathbb{R}^k ; the density of X is bounded away from zero, with bounded partial derivatives at each point of \mathbb{X} .

Assumption 2. Unconfoundedness: $W \perp\!\!\!\perp ((Y(0), Y(1)) \mid X)$.

Assumption 3. Overlap: there is $c \in]0, 1[$ such that $0 < \mathbb{P}(W = 1 \mid X) < 1 - c$.

³ See Heckman et al. (1998), Imbens & Wooldridge (2009) and Rosenbaum & Rubin (1983) for a detailed discussion on them.

By Assumptions **2** and **3**, $\mu(x, w)$ coincides with $\mu_w(x) = \mathbb{E}(Y(w) \mid X = x)$. These mappings are relevant for the purposes of this thesis, and the following regularity conditions will be assumed throughout this work.

Assumption 4. Regularity of conditional mappings: for $w \in \{0, 1\}$, $\mu(\cdot, w)$, is twice continuously differentiable on \mathbb{X} , and $\sigma^2(\cdot, \cdot)$ is uniformly bounded in $\mathbb{X} \times \{0, 1\}$.

The following condition will be part of our standing assumptions as well.

Assumption 5. For each $N \in \mathbb{N}$, (W_i, X_i, Y_i) , $i = 1, \dots, N$, are independent draws from the distribution of Ω .

Remark 3.1.1. Instead of Assumption **1**, Abadie & Imbens (2006) assume that \mathbb{X} is compact and convex, and that the density of X is bounded and bounded away from zero. The remaining conditions we need are the same as those considered for studying the asymptotic properties of the NN-matching estimator.

3.2 Some results in geometric probability theory

3.2.1 The probability of not being in the convex hull of the nearest neighbors

The purpose of this part is to obtain a proper upper bound for the probability that X_i does not belong to the convex hull of covariates of its first M nearest neighbors in the opposite treatment group,

$$\mathbb{P}(X_i \notin \mathbf{co}\{X_{j_1(i)}, \dots, X_{j_M(i)}\}) = \mathbb{P}(0_k \notin \mathbf{co}\{U_{1,i}, \dots, U_{M,i}\}), \quad (3.4)$$

where $U_{m,i} = X_i - X_{j_m(i)}$ is the m th matching discrepancy, $m = 1, \dots, M$.

Following Cover & Efron (1967) we say that a set of random vectors $\{\xi_1, \dots, \xi_M\}$ in \mathbb{R}^k , with $M > k$, is in *general position* if, with probability one, every k -elements subset is linearly independent. From that work (see page 218), we have that this property holds when these vectors are “*selected independently according to a distribution absolutely continuous with respect to natural Lebesgue measure*”. Hence, from Assumptions **1**, **2**, **3** and **5**, it is not difficult to show that the subset of covariates $\{X_1, \dots, X_N\}$ is in general position. Besides, it is also clear that any M -subset of $\{X_1, \dots, X_N\}$, with $M > k$, is in general position as well, and that this property remains valid under *translation*. All of these facts imply that for a large enough N , $i \in \{1, \dots, N\}$ and $M > k$, $\{U_{1,i}, \dots, U_{M,i}\}$ is in general position.

A remarkable result in Wendel (1962), slightly extended by Cover & Efron (1967), says that if the set of random vectors $\{\xi_1, \dots, \xi_M\}$ of \mathbb{R}^k , with $M > k$, is in general position,

and the joint distribution of them is invariant under reflections through the origin,⁴ then the probability of a half-space containing that set of vectors existing is equal to

$$\frac{1}{2^{M-1}} \sum_{s=0}^{k-1} \binom{M-1}{s} = \frac{1}{2^{M-1}} \sum_{s=0}^{k-1} \frac{(M-1)!}{s!(M-1-s)!}.$$

From the fact that the convex hull of $\{\xi_1, \dots, \xi_M\}$ is the intersection of all half-spaces containing that set of vectors, and after bounding the factorial terms in last relation, it is not difficult to conclude that there is a constant θ such that

$$\mathbb{P}(0_k \notin \mathbf{co}\{\xi_1, \dots, \xi_M\}) \leq \theta \frac{M^k}{2^M}.$$

The last inequality cannot be directly applied for upper bounding the probability in (3.4), since even though the set of matching discrepancies $\{U_{1,i}, \dots, U_{M,i}\}$ is in general position, the joint distribution of them could be far from being invariant under reflections through the origin. However, the following technical result helps us to overcome this drawback (see the proof in the §Appendix).

Proposition 3.2.1. *If Assumptions 1 holds, then the joint distribution of the first M matching discrepancies can be bounded above by a strictly positive mapping, which properly re-scaled by a constant yields a distribution function that is invariant under reflections through the origin.*

Proof. Following Abadie & Imbens (2006), from a sample of $\{X_j\}_{j=1}^N \subset \mathbb{R}^k$, we have the probability that $X_i = x$ is the m th closest match of z is given by

$$f_{j_m}(x) = N \binom{N-1}{m-1} f(x) (1 - \mathbb{P}(\|X - z\| \leq \|x - z\|))^{N-m} (\mathbb{P}(\|X - z\| \leq \|x - z\|))^{m-1},$$

where $f(\cdot)$ is the density function of covariates. Denoting $F(x) = \mathbb{P}(\|X - z\| \leq \|x - z\|)$, the conditional distribution of $X_s = \tilde{x}$ being the r th closest match of z , given that $X_{j_m} = x$ for $r > m$, is the same as the distribution of the $(r - m)$ th closest match of z obtained from a sample of size $N - m$ from a population whose distribution is simply $F(\cdot)$ truncated on the left at x , the latter given by the following expression:

$$\begin{aligned} f_{j_r}^{j_m}(\tilde{x} | x) &= \frac{f_{j_m, j_r}(x, \tilde{x})}{f_{j_m}(x)} \\ &= (N - m) \binom{N - m - 1}{r - m - 1} \frac{f(\tilde{x})}{(1 - F(x))} \left(\frac{F(\tilde{x}) - F(x)}{1 - F(x)} \right)^{r - m - 1} \left(\frac{1 - F(\tilde{x})}{1 - F(x)} \right)^{N - r}. \end{aligned}$$

Thus, the joint distribution of probability that $X_i = x$ and $X_s = \tilde{x}$ are the m th and r th

⁴ That is, for any subset A_1, \dots, A_M of \mathbb{R}^k , $\mathbb{P}(\delta_1 Z_1 \in A_1, \dots, \delta_M Z_M \in A_M)$ has the same value for all 2^M choices of $\delta_i = \pm 1$, $i = 1, \dots, M$.

($r > m$) nearest neighbors of z respectively is:

$$f_{j_m, j_r}(x, \tilde{x}) = \frac{N!}{(m-1)!(r-m-1)!(N-r)!} f(x) f(\tilde{x}) (F(\tilde{x}) - F(x))^{r-m-1} (1 - F(\tilde{x}))^{N-r}.$$

Hence, by following the above arguments and performing some calculus, denoting $x = (x_{j_1}, \dots, x_{j_M})$ we can show that the joint distribution of the first M closest matches is:

$$f_{j_1, \dots, j_M}(x) = \frac{N!}{(N-M)!} \left(\prod_{s=1}^M f(x_{j_s}) \right) (1 - F(x_{j_M}))^{N-M},$$

which after transforming to the matching discrepancy, $U_m = X_{j_m} - z$, and denoting $u = (u_{j_1}, \dots, u_{j_M})$, we can conclude the following relation:

$$f_{j_1, \dots, j_M}(u) = \frac{N!}{(N-M)!} \left(\prod_{s=1}^M f(z + u_{j_s}) \right) (1 - \mathbb{P}(\|X - z\| \leq \|u_{j_M}\|))^{N-M}.$$

Finally, denoting $V_m = N^{1/k} U_m$, and $v = (v_{j_1}, \dots, v_{j_M})$, we have that

$$f_{j_1, \dots, j_M}(v) = \frac{N! N^{-M}}{(N-M)!} \left(\prod_{s=1}^M f\left(z + \frac{v_{j_s}}{N^{1/k}}\right) \right) \left(1 - \mathbb{P}\left(\|X - z\| \leq \frac{\|v_{j_M}\|}{N^{1/k}}\right) \right)^{N-M}, \quad (3.5)$$

from which we can readily conclude the following inequality⁵

$$f_{j_1, \dots, j_M}(v) \leq \bar{f}^M \exp\left(-\underline{f} \frac{\|v_{j_M}\|^k}{(M+1) \Gamma(1+k/2)} \frac{\pi^{k/2}}{\Gamma(1+k/2)}\right), \quad (3.6)$$

where $0 < \underline{f} < \bar{f} < \infty$ are the lower and upper bounds of the distribution $f(\cdot)$, respectively. Using the right term in (3.6) we can define the distribution as stated.

Remark 3.2.1. Using (3.5) it can be shown that the joint distribution of the first M nearest neighbors converges to the following distribution, which indeed is invariant under reflections through the origin:

$$\lim_{N \rightarrow \infty} f_{j_1, \dots, j_M}(v) = f(z)^M \exp\left(-\|v_{j_M}\|^k f(z) \frac{\pi^{k/2}}{\Gamma(1+k/2)}\right).$$

□

The following result comes directly from the properties presented above.

Theorem 3.2.1. If Assumptions 1, 2, 3, and 5 hold, for a large enough N , $i \in \{1, \dots, N\}$

⁵ Here we use the fact that for $N > M$, $\frac{N-M}{N} \geq \frac{1}{M+1}$.

and $M > k$, there is a constant $\gamma > 0$ such that

$$\mathbb{P}(X_i \notin \mathbf{co}\{X_{j_1(i)}, \dots, X_{j_M(i)}\}) \leq \gamma \frac{M^k}{2^M}.$$

Remark 3.2.2. What Theorem 3.2.1 states is that the probability of not being in the convex hull of the first M nearest neighbors is bounded above by a term that goes to zero exponentially in M . The constant γ in that relation comes from the unknown distribution function in Proposition 3.2.1. From this theorem, it is also clear that, given the sample, for each treated unit i (and similar for controls) we have that:

$$\mathbb{P}(X_i \notin \mathbf{co}\{X_1, \dots, X_{N_0}\}) \leq \gamma \frac{N_0^k}{2^{N_0}}.$$

3.2.2 The number of times that a unit is used as a match

For a control unit $j \in \{1, \dots, N_0\}$, after solving the optimization problem \mathcal{S}_j , the vector of covariates of a treated unit $i \in \{N_0 + 1, \dots, N\}$ participates in the convex combination performing X_j (or its projection onto $\mathbf{co}\{X_{N_0+1}, \dots, X_N\}$) whenever $\lambda_{i-N_0}^j > 0$.⁶ From a geometric point of view, this means that vector X_i is a vertex of the polytope defined by the convex hull of covariates of treated units associated with the solution of problem \mathcal{S}_j , whose set of indexes is $M_j = \{i' \in \{N_0 + 1, \dots, N\}, \lambda_{i'-N_0}^j > 0\}$. For the case that j is a treated unit, $M_j = \{i' \in \{1, \dots, N_0\}, \lambda_{i'}^j > 0\}$, and given that, the number of times that a unit i , either control or treated, is a vertex of such polytopes is

$$T(i) = W_i \sum_{j=1}^{N_0} \mathbb{1}\{i \in M_j\} + (1 - W_i) \sum_{j=1}^{N_1} \mathbb{1}\{i \in M_j\},$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, which is equal to 1 if the argument is true and 0 otherwise. The corresponding sum of weights associated with unit i is given by

$$K(i) = W_i \sum_{j=1}^{N_0} \lambda_{i-N_0}^j + (1 - W_i) \sum_{j=1}^{N_1} \lambda_i^{N_0+j}.$$

The following result will be quite relevant when we study the properties of the variance of the BLOP matching estimator in §3.3.2. Roughly speaking, it states that the sum (to the power of any integer) of the weights associated with the unit i when it was used as a counterfactual individual when performing the BLOP matching estimator is a constant, on average.

⁶ We have that the components of vector λ^j are $\lambda_1^j, \dots, \lambda_{N_1}^j$, thus a treated unit $i \in \{N_0 + 1, \dots, N_0 + N_1\}$ is associated with $\lambda_{N_0-i}^j$.

Proposition 3.2.2. *If Assumptions 1, 2, 3, and 5 hold, then for each unit i and integer α , $\mathbb{E}((K(i))^\alpha)$ is bounded uniformly in N .*

Proof. Assume for a while that $\alpha = 1$, and without loss of generality, the proof is performed for a treated unit $i_1 \in \{N_0 + 1, \dots, N\}$. In that case, for the sake of simplicity regarding notation, for a control unit j and $i \in \{N_0 + 1, \dots, N\}$, we set $\lambda_i^j \equiv \lambda_{i-N_0}^j$. Since $K(i_1) \leq T(i_1)$, it is clear that the following inequality holds:

$$\mathbb{E}(K(i_1)) \leq \mathbb{E}\left(\sum_{j=1}^{N_0} \mathbf{1}\{i_1 \in M_j\}\right),$$

and from the fact $\{\mathbf{1}\{i \in M_j\}, j \in \{1, \dots, N_0\}, i \in \{N_0 + 1, \dots, N\}\}$ are identically distributed, we can readily conclude that for any treated unit i and a control j ,

$$\mathbb{E}(K(i_1)) \leq N_0 \mathbb{P}(i \in M_j). \quad (3.7)$$

Denoting by $\mathcal{C}_1 = \mathbf{co}\{X_{N_0+1}, \dots, X_N\}$, the convex hull of covariates of the entire sample of treated units, from standard decomposition of $\mathbb{P}(i \in M_j)$ using the “belonging to set \mathcal{C}_1 ” as the conditional event, we have that

$$\mathbb{P}(i \in M_j) = \mathbb{P}(i \in M_j | X_j \in \mathcal{C}_1) \mathbb{P}(X_j \in \mathcal{C}_1) + \mathbb{P}(i \in M_j | X_j \notin \mathcal{C}_1) \mathbb{P}(X_j \notin \mathcal{C}_1),$$

and by Theorem 3.2.1,

$$\mathbb{P}(i \in M_j) \leq \mathbb{P}(i \in M_j | X_j \in \mathcal{C}_1) \mathbb{P}(X_j \in \mathcal{C}_1) + \gamma \frac{N_1^k}{2^{N_1}}. \quad (3.8)$$

Conditional on $\{X_j \in \mathcal{C}_1\}$, let \mathcal{M}_j the subset of indexes of treated units that are associated with the minimum number of nearest neighbors to unit j that are necessary to build X_j (or its projection as the case may be) as a convex combination of their covariates.⁷ Hence, since $\mathbb{P}(i \in M_j | X_j \in \mathcal{C}_1) = \mathbb{P}(i \in \mathcal{M}_j | X_j \in \mathcal{C}_1)$, and partitioning the event $\{X_j \in \mathcal{C}_1\}$ into subevents⁸ $\{X_j \in \Delta\mathcal{C}_j(m)\} = \{X_j \in \mathcal{C}_j(m) \setminus \mathcal{C}_j(m-1)\}$, $m = 2, \dots, N_1$, it follows that⁹

$$\mathbb{P}(i \in M_j | X_j \in \mathcal{C}_1) = \sum_{m=2}^{N_1} \mathbb{P}(i \in \mathcal{M}_j | X_j \in \Delta\mathcal{C}_j(m)) \mathbb{P}(X_j \in \Delta\mathcal{C}_j(m)).$$

⁷ If this number is \mathbf{m} , then $\mathcal{M}_j = \{j_1(j), \dots, j_{\mathbf{m}}(j)\}$, and for each $m < \mathbf{m}$, $X_j \notin \mathcal{C}_j(m)$ and $X_j \in \mathcal{C}_j(\mathbf{m})$.

⁸ The set-difference between A and B is denoted by $A \setminus B = \{c \in A, c \notin B\}$.

⁹ $X_j \in \Delta\mathcal{C}_j(m)$ corresponds to say that this vector belongs to the convex hull of its m nearest neighbors and does not belong to the convex hull of its $m-1$ nearest neighbors.

Using the identical distribution of the aforementioned random variables,

$$\sum_{i=1}^{N_1} \mathbb{P}(i \in \mathcal{M}_j | X_j \in \Delta \mathcal{C}_j(m)) = N_1 \mathbb{P}(i \in \mathcal{M}_j | X_j \in \Delta \mathcal{C}_j(m)), \quad (3.9)$$

and the fact that $\mathbb{P}(i \in \mathcal{M}_j | X_j \in \Delta \mathcal{C}_j(m)) = \mathbb{E}(\mathbb{1}\{i \in \mathcal{M}_j\} | X_j \in \Delta \mathcal{C}_j(m))$, implies¹⁰

$$\mathbb{P}(i \in \mathcal{M}_j | X_j \in \Delta \mathcal{C}_j(m)) = \frac{m}{N_1}. \quad (3.10)$$

Thus, the combination of (3.9) and (3.10) yields

$$\mathbb{P}(i \in M_j | X_j \in \mathcal{C}_1) = \frac{1}{N_1} \sum_{m=2}^{N_1} m \mathbb{P}(X_j \in \Delta \mathcal{C}_j(m)),$$

and by Theorem 3.2.1,

$$\mathbb{P}(i \in M_j | X_j \in \mathcal{C}_1) \leq \sum_{m=2}^k \frac{m}{N_1} + \sum_{m=k+1}^{N_1} \frac{\gamma m^{k+1}}{2^m N_1} \leq \frac{\gamma_2}{N_1},$$

for some constant $\gamma_2 > 0$. This last inequality along with relations (3.7) and (3.8) give

$$\mathbb{E}(K(i_1)) \leq N_0 \left(\frac{\gamma_2}{N_1} + \gamma \frac{N_1^k}{2^{N_1}} \right),$$

and therefore, using the well known Chernoff's inequality, we obtain the result for the case $\alpha = 1$, i.e., there is a constant κ_1 such that $\mathbb{E}(K(i_1)) \leq \kappa_1$. For the case $\alpha = 2$, we first notice that for $j, j' \in \{1, \dots, N_0\}$, $j \neq j'$, using the convention above regarding the weighting scheme, Assumption 5 implies that $\lambda_{i_1}^j$ and $\lambda_{i_1}^{j'}$ are independent random variables. Given that, after doing some simple algebra,

$$(K(i_1))^2 \leq K(i_1) + 2 \sum_{j'=1, j' \neq j}^{N_0} \lambda_{i_1}^{j'} \sum_{j=1}^{N_0} \lambda_{i_1}^j,$$

and then, by taking expectation and using the independence condition mentioned above, it follows that

$$\mathbb{E}(K(i_1))^2 \leq \kappa_1 + 2\kappa_1 \mathbb{E} \left(\sum_{j'=1, j' \neq j}^{N_0} \lambda_{i_1}^{j'} \right) \leq \kappa_1 + 2\kappa_1^2.$$

The proof for any $\alpha > 2$ comes readily using an inductive argument. \square

¹⁰ Here we use the fact that $\sum_{i=1}^{N_1} \mathbb{E}(\mathbb{1}\{i \in \mathcal{M}_j\} | X_j \in \Delta \mathcal{C}_j(m)) = m$.

3.3 Asymptotic properties

Before going into details regarding limit properties, it is worth presenting a breakdown of the bias of the BLOP matching estimator, which is a useful tool to understand what variables play a relevant role in determining the results in this work. By following Abadie & Imbens (2006), and doing some algebra, it can be shown that $\hat{\tau}^b - \tau = A^b + E^b + B^b$, where

$$A^b = \frac{1}{N} \sum_{i=1}^N (\mu_1(X_i) - \mu_0(X_i)) - \tau, \quad E^b = \frac{1}{N} \sum_{i=1}^N (2W_i - 1) (1 + K(i)) \epsilon_i, \quad (3.11)$$

with $\epsilon_i = Y_i - \mu_{W_i}(X_i)$, $i = 1, \dots, N$, and B^b is the bias of $\hat{\tau}^b$, conditional on $\{(W_i, X_i)\}_{i=1}^N$, which after some calculus is given by:

$$B^b = \frac{1}{N} \left(\sum_{i=1}^{N_0} \sum_{m=1}^{N_1} \lambda_m^i (\mu_1(X_{m+N_0}) - \mu_1(X_i)) + \sum_{i=1+N_0}^N \sum_{m=1}^{N_0} \lambda_m^i (\mu_0(X_i) - \mu_0(X_m)) \right). \quad (3.12)$$

Note that after taking expectation to $\hat{\tau}^b - \tau$, the only term that survives is B^b , so the order of the bias is dominated by the order of this term, which, as we shall see, depends on the order of the unit-level conditional bias.

In a similar manner to the approximation performed in (2.2), for a treated unit $i \in \{N_0 + 1, \dots, N\}$, after performing a second order Taylor expansion of μ_0 around X_i –which is possible under Assumption 4–, in view of Assumption 1 we have that the absolute value of its unit-level conditional bias attains the next inequality –see (3.1) and (3.3)–:

$$\left| \sum_{m=1}^{N_0} \lambda_m^i (\mu_0(X_i) - \mu_0(X_m)) \right| \leq L_1 \nu\{\mathcal{F}_i\} + L_2 \nu\{\mathcal{S}_i\} + O \left(\sum_{m=1}^{N_0} \lambda_m^i \|X_i - X_m\|^3 \right) \quad (3.13)$$

where constants L_1 and L_2 are the upper bounds, over \mathbb{X} , of the first and second derivatives of that mapping. It is clear that inequality (3.13) can be built to the unit-level conditional bias of any control unit, where μ_0 has to be replaced by μ_1 , and the covariates values of optimization problems need to be well configured (the constants can be assumed to be the same).

3.3.1 The order of the conditional bias

From inequality (3.13), it is not difficult to realize that the order of the unit-level conditional bias of any unit i (and therefore of the whole conditional bias as we see later) depends on the properties of both $\nu\{\mathcal{F}_i\}$ and $\nu\{\mathcal{S}_i\}$. In what follows we show that the order of the former term is as good as one wants, this because $\nu\{\mathcal{F}_i\}$ is zero with a probability tending to one as N goes to infinity. Meanwhile, the asymptotic behavior of the latter term depends on the properties of the norm of matching discrepancies to the power of two, which are associated

with counterfactual units that are not necessarily the nearest neighbors to the individual under analysis.

The following property is the key result in this part. It states that the order of the expectation of the unit-level conditional bias of individual i is $N_{1-W_i}^{2/k}$.

Proposition 3.3.1. *If Assumptions 1 – 5 hold, then*

$$\mathbb{E} \left(\left| \sum_{m=1}^{N_0} \lambda_m^i (\mu_0(X_i) - \mu_0(X_m)) \right| \middle| W_i = 1, X_i, \{W_j, X_j\}_{j=1}^N \right) = O \left(N_0^{-2/k} \right). \quad (3.14)$$

and

$$\mathbb{E} \left(\left| \sum_{m=1}^{N_1} \lambda_m^i (\mu_1(X_{m+N_0}) - \mu_1(X_i)) \right| \middle| W_i = 0, X_i, \{W_j, X_j\}_{j=1}^N \right) = O \left(N_1^{-2/k} \right). \quad (3.15)$$

Proof. Without loss of generality, the proof is performed for the relation in (3.14). For a treated unit i , the conditionals in (3.14) is denoted by $\theta_i = \{W_i = 1, X_i, \{W_j, X_j\}_{j=1}^N\}$, and for $m \leq N_0$, we set

$$\mathcal{C}_i(m) = \mathbf{co}\{X_{j_1(i)}, \dots, X_{j_m(i)}\}.$$

For the case $m = N_0$ we have that $\mathcal{C}_i(N_0) = \mathbf{co}\{X_1, \dots, X_{N_0}\}$, which does not depend on the unit i , thus this set denoted as \mathcal{C}_0 . We also denote $\mathbf{q}_i(m) = \mathbb{P}(X_i \notin \mathcal{C}_i(m))$, and let $\mathbf{p}_i(m) = 1 - \mathbf{q}_i(m)$. Hence, ignoring the order term in right-hand side of (3.13) for a while, we are now concerned with the study of $\psi_i = \mathbb{E} \left(L_1 \nu\{\mathcal{F}_i\} + L_2 \nu\{\mathcal{S}_i\} \middle| \theta_i \right)$, which obviously can be written as

$$\psi_i = \mathbb{E} \left(\psi_i \middle| X_i \notin \mathcal{C}_0 \right) \mathbf{q}_i(N_0) + \mathbb{E} \left(\psi_i \middle| X_i \in \mathcal{C}_0 \right) \mathbf{p}_i(N_0).$$

Denoting the diameter of \mathbb{X} by $\delta > 0$, pursuant to Theorem 3.2.1,

$$\mathbb{E} \left(\psi_i \middle| X_i \notin \mathcal{C}_0 \right) \mathbf{q}_i(N_0) \leq (L_1 \delta + L_2 (k+1) \delta^2) \gamma \frac{N_0^k}{2^{N_0}},$$

and then,

$$\mathbb{E} \left(N_0^{2/k} \psi_i \middle| X_i \notin \mathcal{C}_0 \right) \mathbf{q}_i(N_0) = o(1). \quad (3.16)$$

On the other hand, it is clear that $\mathbb{E} \left(\psi_i \middle| X_i \in \mathcal{C}_0 \right) = \mathbb{E} \left(L_2 \nu\{\mathcal{S}_i\} \middle| \theta_i, X_i \in \mathcal{C}_0 \right)$. Hence, after partitioning the event $\{X_i \in \mathcal{C}_0\}$ into the events

$$\{X_i \in \Delta \mathcal{C}_i(m)\} = \{X_i \in \mathcal{C}_i(m) \setminus \mathcal{C}_i(m-1)\}, \quad m = 2, \dots, N_0,$$

each one of them having a probability of occurrence of $\mathbf{p}_i(m) \mathbf{q}_i(m-1)$, which indeed is less

than or equal to $\mathbf{q}_i(m-1)$, it follows that

$$\mathbb{E} \left(\psi_i \mid X_i \in \mathcal{C}_0 \right) \mathbf{p}_i(N_0) \leq \sum_{m=2}^{N_0} \mathbb{E} \left(L_2 \nu\{\mathcal{S}_i\} \mid \theta_i, X_i \in \Delta\mathcal{C}_i(m) \right) \mathbf{q}_i(m-1). \quad (3.17)$$

For $m \leq N_0$, notice that when $X_i \in \Delta\mathcal{C}_i(m)$, then there is a vector $(\xi_1, \dots, \xi_m) \in \Delta_m$ such that $X_i = \sum_{s=1}^m \xi_s X_{j_s(i)}$, which, by definition of problem \mathcal{S}_i , implies that

$$\nu\{\mathcal{S}_i\} = \sum_{s=1}^{N_0} \lambda_s^i \|X_i - X_s\|^2 \leq \sum_{s=1}^m \xi_s \|X_i - X_{j_s(i)}\|^2 \leq \|X_i - X_{j_m(i)}\|^2.$$

On the other hand, when $m > k$, Theorem 3.2.1 implies that $\mathbf{q}_i(m-1) \leq \gamma \frac{m^k}{2^m}$. All of this in (3.17) give

$$\begin{aligned} \mathbb{E} \left(\psi_i \mid X_i \in \mathcal{C}_0 \right) \mathbf{p}_i(N_0) &\leq \sum_{m=2}^k \mathbb{E} \left(L_2 \|X_i - X_{j_m(i)}\|^2 \mid \theta_i, X_i \in \Delta\mathcal{C}_i(m) \right) + \\ &\quad \sum_{m=k+1}^{N_0} \mathbb{E} \left(L_2 \|X_i - X_{j_m(i)}\|^2 \mid \theta_i, X_i \in \Delta\mathcal{C}_i(m) \right) 2\gamma \frac{m^k}{2^m} \end{aligned} \quad (3.18)$$

Now, in view of standing assumptions, Theorem 5.4 in Evans et al. (2002) implies that for a large enough N , and therefore N_0 from Assumptions **2** and **3**,

$$\mathbb{E} \left(N_0^{2/k} \|X_i - X_{j_m(i)}\|^2 \right) = \eta_1 \frac{\Gamma(m+2/k)}{\Gamma(m)} + O \left(\frac{1}{N_0^{1/k-\rho}} \right), \quad (3.19)$$

for some constant $\eta_1 > 0$, $\rho \in]0, 1/k[$, and $m \leq N_0$. Moreover, from relations (5.36) and (5.44) in Evans et al. (2002) –see pag. 2848 –, the order expression in the right-hand side of (3.19) does not depend on m , which implies that it can be bounded above by some constant. Hence, using the following straightforward inequalities

$$\frac{\Gamma(m+2/k)}{\Gamma(m)} \leq \frac{\Gamma(m+2)}{\Gamma(m)} \leq \eta_2 m^2,$$

for some $\eta_2 > 0$, it follows that for unit i , a large enough N and $m \leq N_0$,

$$\mathbb{E} \left(N_0^{2/k} \|X_i - X_{j_m(i)}\|^2 \right) \leq \eta_3 m^2 \quad (3.20)$$

for some constant $\eta_3 > 0$. Applying (3.20) to the right-hand side in (3.18) yields

$$\mathbb{E} \left(\psi_i \mid X_i \in \mathcal{C}_0 \right) \mathbf{p}_i(N_0) \leq \frac{\eta_3 L_2}{N_0^{2/k}} \left(\sum_{m=2}^k m^2 + 2\gamma \sum_{m=k+1}^{N_0} \frac{m^{2+k}}{2^m} \right) \leq \frac{C}{N_0^{2/k}}$$

for some constant C . This last inequality along with (3.16) implies the result for the order of ψ_i . The remaining to conclude is straightforward from the results just presented. \square

Corollary 3.3.1. *If Assumptions 1 – 5 hold and $\mu_w(\cdot)$, $w = 0, 1$, is flat over \mathbb{X} , then for any integer $\beta > 0$*

$$\mathbb{E} \left(\left| \sum_{m=1}^{N_0} \lambda_m^i (\mu_0(X_i) - \mu_0(X_m)) \right| \middle| W_i = 1, X_i, \{W_j, X_j\}_{j=1}^N \right) = o(N_0^{-\beta}).$$

and

$$\mathbb{E} \left(\left| \sum_{m=1}^{N_1} \lambda_m^i (\mu_1(X_{m+N_0}) - \mu_1(X_i)) \right| \middle| W_i = 0, X_i, \{W_j, X_j\}_{j=1}^N \right) = o(N_1^{-\beta}).$$

Proof. This result is straightforward from Proposition 3.3.1, since in this case it holds that only the linear term of Taylor's expansion of $\mu_w(\cdot)$, expression (3.13), is different from zero. \square

The following result is directly from Proposition 3.3.1 and Corollary 3.3.1.

Theorem 3.3.1. *If Assumptions 1 – 5 hold, then $B^b = O_{\mathbf{p}}(N^{-2/k})$. In addition, if $\mu_w(\cdot)$, $w = 0, 1$, is flat over \mathbb{X} , then $B^b = o_{\mathbf{p}}(N^{-\beta})$ for any integer $\beta > 0$.*

Proof. For the first part of the Theorem we have that after developing (3.12), we have

$$\begin{aligned} \mathbb{E} \left(N^{2/k} |B^b| \right) &\leq \mathbb{E} \left(\frac{N^{2/k}}{N} \sum_{i=1}^{N_0} \mathbb{E} \left(\left| \sum_{m=1}^{N_1} \lambda_m^i (\mu_1(X_{m+N_0}) - \mu_1(X_i)) \right| \middle| X_i, \{W_j, X_j\}_{j=1}^N \right) \right) + \\ &\quad \mathbb{E} \left(\frac{N^{2/k}}{N} \sum_{i=N_0+1}^N \mathbb{E} \left(\left| \sum_{m=1}^{N_0} \lambda_m^i (\mu_0(X_i) - \mu_0(X_m)) \right| \middle| X_i, \{W_j, X_j\}_{j=1}^N \right) \right), \end{aligned}$$

and by Proposition 3.3.1, and doing some algebra, there is a constant ϱ such that

$$\mathbb{E} \left(N^{2/k} |B^b| \right) \leq \varrho \mathbb{E} \left(\left(\frac{N}{N_1} \right)^{2/k} \left(\frac{N_0}{N} \right) + \left(\frac{N}{N_0} \right)^{2/k} \left(\frac{N_1}{N} \right) \right).$$

The proof concludes after using Chernoff and Markov's inequalities in the last relation. The second part of this Theorem is direct by using Corollary 3.3.1. \square

We end this part studying some additional properties concerning the BLOP matching estimator of the ATT, which from (2.5) is

$$\hat{\tau}_{tre}^b = \frac{1}{N_1} \sum_{i=N_0+1}^N \left(\hat{Y}_i^b(1) - \hat{Y}_i^b(0) \right).$$

After performing some simple algebra, we can show that the conditional bias of this estimator is given by

$$B_{tre}^b = \frac{1}{N_1} \left(\sum_{i=N_0+1}^N \sum_{m=1}^{N_0} \lambda_m^i (\mu_0(X_i) - \mu_0(X_m)) \right).$$

For this estimator, the following assumption will replace Assumption 5 we have used for studying the limit properties of the BLOP matching estimator of the ATE.

Assumption 6. *Conditional on $W_i = w$, the sample consists of independent draws from $Y, X|W = w$ for $w = 0, 1$, and for some $r > 1$,*

$$\frac{N_1^r}{N_0} \rightarrow \theta < \infty. \quad (3.21)$$

The following result is straightforward using the properties above.

Corollary 3.3.2. *If Assumptions 1 – 4 and 6 hold, then $B_{tre}^b = O_{\mathbf{p}}(N_1^{-2r/k})$. In addition, if $\mu_0(\cdot)$ is flat over \mathbb{X} , then $B_{tre}^b = o_{\mathbf{p}}(N_1^{-\beta})$ for any integer $\beta > 0$.*

Proof. In view of Assumption 6 we can apply Proposition 3.3.1 to conclude $B_{tre}^b = O_{\mathbf{p}}(N_0^{-2/k})$. Hence, using (3.21) we can readily obtain the result. The remainder of this property is direct after using Corollary 3.3.1. \square

Four comments before ending this section. First, what Theorem 3.3.1 states is that the BLOP matching estimator of the ATE is not \sqrt{N} -consistent, in general, since the bias rate depends on the number of employed pretreatment variables. In fact, the order is worse as the number of pretreatment variables increases. However, it is worth mentioning that the analysis above is based on continuous covariates, thus when covariates are discrete variables, they do not generate bias as the sample size tends to infinity. Second, Theorem 3.3.1 also says that BLOP matching estimator of the ATE attains an order of its bias that is better than the $N^{1/k}$ attained by the NN -matching estimator. Informally speaking, the bias of the BLOP matching estimator vanishes with a smaller sample size (the square root of the sample needed by the NN matching). Third, under certain conditions it is certainly possible to remove the bias, so that the estimator becomes \sqrt{N} -consistent, with no asymptotic bias. This happens, for instance, when we have at most two continuous covariates to build the ATE ($k = 1, 2$), or when the conditional expectations of the outcomes are flat. Finally, when estimating BLOP matching for ATE, a bias correction is feasible as in Abadie & Imbens (2011a) in order to remove the bias.

3.3.2 Variance, consistency, and normality properties

Since we have already obtained the rate of convergence of the conditional bias, Proposition 3.2.2 is the most relevant result we need for obtaining the variance and asymptotic normality properties of both $\hat{\tau}^b$ and $\hat{\tau}_{tre}^b$. Hence, the proofs and partial results we show below basically follow the arguments provided by Abadie & Imbens (2006) when studying such properties for the NN -matching estimator.

After performing some simple calculus, we can show that the variance of $\hat{\tau}^b$, conditional on $\{X_i, W_i\}_{i=1}^N$, is given by

$$\mathbb{V}\left(\hat{\tau}^b \mid \{X_i, W_i\}_{i=1}^N\right) = \frac{1}{N^2} \sum_{i=1}^N (1 + K(i))^2 \sigma^2(X_i, W_i), \quad (3.22)$$

while for $\hat{\tau}_{tre}^b$ it is

$$\mathbb{V}\left(\hat{\tau}_{tre}^b \mid \{X_i, W_i\}_{i=1}^N\right) = \frac{1}{N_1^2} \sum_{i=1}^N (W_i - (1 - W_i)K(i))^2 \sigma^2(X_i, W_i). \quad (3.23)$$

In the following, the *normalized conditional variance* of $\hat{\tau}^b$ and the *variance of the conditional mean* are denoted, respectively, by

$$V^{CV} = N \mathbb{V}\left(\hat{\tau}^b \mid \{(W_i, X_i)\}_{i=1}^N\right), \quad V^{CM} = \mathbb{E}\left((\mu_1(X) - \mu_0(X) - \tau)^2\right),$$

and for $\hat{\tau}_{tre}^b$ these concepts are denoted by

$$V_{tre}^{CV} = N_1 \mathbb{V}\left(\hat{\tau}_{tre}^b \mid \{(W_i, X_i)\}_{i=1}^N\right), \quad V_{tre}^{CM} = \mathbb{E}\left((\mu_1(X) - \mu_0(X) - \tau_{tre})^2 \mid W = 1\right).$$

Lemma 3.3.1. *If Assumptions 1 – 5 hold, then $\mathbb{E}(V^{CV}) = O(1)$. If Assumptions 1 – 4 and 6 hold, then*

$$\frac{N_0}{N_1} \mathbb{E}(V_{tre}^{CV}) = O(1).$$

Proof. For the first part, using (3.22), the proof is direct from Assumption 4 and Proposition 3.2.2. For the second part, the argument is the same, but using (3.23). \square

The following technical condition is needed for the result below.

Assumption 7. *For $w \in \{0, 1\}$, $\sigma^2(\cdot, w)$ is Lipschitz in \mathbb{X} and bounded away from zero, the fourth-moment of $Y(w)$ are uniformly bounded in \mathbb{X} .*

Proposition 3.3.2. *Suppose Assumptions 1 – 5 and 7 hold, then $\hat{\tau}^b \xrightarrow{\mathbb{P}} \tau$ and*

$$\frac{\sqrt{N}(\hat{\tau}^b - \tau - B^b)}{\sqrt{V^{CV} + V^{CM}}} \xrightarrow{\mathbb{D}} \mathcal{N}(0, 1).$$

Suppose Assumptions **1** – **4**, **6**, and **7** hold, then $\widehat{\tau}_{tre}^b \xrightarrow{\mathbb{P}} \tau_{tre}$ and

$$\frac{\sqrt{N_1} (\widehat{\tau}_{tre}^b - \tau_{tre} - B_{tre}^b)}{\sqrt{V_{tre}^{CV} + V_{tre}^{CM}}} \xrightarrow{\mathbb{D}} \mathcal{N}(0, 1).$$

Proof. We show the proof for the ATE, because it is direct for the ATT after that result. From the standard law of large numbers, we already know

$$\left(\frac{1}{N} \left(\sum_{i=1}^N (\mu_1(X_i) - \mu_0(X_i)) \right) - \tau \right) \xrightarrow{\mathbb{P}} 0,$$

and from definition of E^b in (3.11), we have

$$\mathbb{E} \left(N (E^b)^2 \right) = \frac{1}{N} \sum_{i=1}^N \mathbb{E} \left((1 + K(i))^2 \epsilon_i^2 \right) = \mathbb{E} \left((1 + K(i))^2 \sigma^2(X_i, W_i) \right),$$

and by Lemma 3.3.1, $\mathbb{E}(N (E^b)^2) = O(1)$. Thus, using Markov's inequality, and the order of convergence of B^b , we can readily conclude the proof of consistency. In order to show the normality property, from Lemma 3.3.1 we have that V^{CV} is bounded in N and from Assumptions **1** and **4** the same holds for V^{CM} . From the fact that $\sqrt{N} (\widehat{\tau}^b - \tau - B^b) = \sqrt{N} A^b + \sqrt{N} E^b$, the *Standard Central Limit Theorem* and properties of E^b give

$$\sqrt{N} A^b \xrightarrow{\mathbb{D}} \mathcal{N}(0, V^{CM}). \quad (3.24)$$

Finally, using the *Linderberg-Feller Central Limit Theorem*¹¹, Proposition 3.2.2 and following the same argumentation provided by Abadie & Imbens (2006) when showing their Theorem 4 (here the necessity of Assumption **7**), it can be shown that

$$\frac{\sqrt{N} E^b}{\sqrt{V^{CV}}} \xrightarrow{\mathbb{D}} \mathcal{N}(0, 1). \quad (3.25)$$

Because (3.24) and (3.25) are asymptotically independent, we conclude the proof. \square

We conclude this part by recalling conditions under which our estimators are \sqrt{N} -consistent. Of course a trivial case holds when the conditional expectations, $\mu_w(\cdot)$, are flat on the supporting set. In addition, due to the order of conditional bias we have obtained, this property for the BLOP matching estimator of the ATE holds true as well when $k = 1$ or $k = 2$, this being in fact the only case, besides the trivial one, when this estimator attains the \sqrt{N} -consistency. For the estimator of the ATT we have, however, that this property is also obtained when the number of control units increases faster than the number of treated units as stated by Assumption **6**. Summing up, these results are presented in the following

¹¹ This theorem remains valid conditional on $\{(W_i, X_i)\}_{i=1}^N$, which is relevant for our case.

corollary.

Corollary 3.3.3. *Suppose that Assumptions 1 – 5 and 7 hold, and that $k = 1$ or $k = 2$, (and/or $\mu_w(\cdot)$, $w = 0, 1$, is flat over \mathbb{X}), then $\hat{\tau}^b \xrightarrow{\mathbb{P}} \tau$ and*

$$\frac{\sqrt{N} (\hat{\tau}^b - \tau)}{\sqrt{V^{\text{CV}} + V^{\text{CM}}}} \xrightarrow{\mathbb{D}} \mathcal{N}(0, 1).$$

Suppose Assumptions 1 – 4, 6, and 7 hold, and $r > \frac{k}{4}$ (and/or $k = 1$ or $k = 2$, and/or $\mu_0(\cdot)$ is flat over \mathbb{X}), then $\hat{\tau}_{tre}^b \xrightarrow{\mathbb{P}} \tau_{tre}$ and

$$\frac{\sqrt{N_1} (\hat{\tau}_{tre}^b - \tau_{tre})}{\sqrt{V_{tre}^{\text{CV}} + V_{tre}^{\text{CM}}}} \xrightarrow{\mathbb{D}} \mathcal{N}(0, 1).$$

3.3.3 Estimating the variance

In this part we present marginal variance estimators for the average treatment effect and average treatment effect on the treated. Following Abadie & Imbens (2006), a variance estimator of the *conditional mean* of $\hat{\tau}^b$ is given by

$$\hat{\mathbb{V}} \left(\mathbb{E}(\hat{\tau}^b | \{X_i, W_i\}_{i=1}^N) \right) = \frac{1}{N^2} \sum_{i=1}^N \left(\left(\hat{Y}_i(1) - \hat{Y}_i(0) - \hat{\tau}^b \right)^2 - (1 + K(i)^2) \hat{\sigma}_i^2 \right), \quad (3.26)$$

while the estimator for the *expected value* of the conditional variance is

$$\hat{\mathbb{E}} \left(\mathbb{V}(\hat{\tau}^b | \{X_i, W_i\}_{i=1}^N) \right) = \frac{1}{N^2} \sum_{i=1}^N (1 + K(i)^2) \hat{\sigma}_i^2, \quad (3.27)$$

where $\hat{\sigma}_i^2$ is a BLOP matching estimator of the conditional variance $\sigma^2(X_i, W_i) = \mathbb{V}(Y | X = X_i, W = W_i)$. That is, the BLOP matching estimator of the conditional variances requires solving problem (2.3) for each unit that needs to be matched, using covariates from units in the same treatment group and leaving the i -th unit out, instead of using covariates from units in the the opposite treatment group. Using (3.26) and (3.27) we have a consistent estimator of the variance of $\hat{\tau}^b$

$$\hat{V}(\hat{\tau}^b) = \frac{1}{N^2} \sum_{i=1}^N \left(\left(\hat{Y}_i(1) - \hat{Y}_i(0) - \hat{\tau}^b \right)^2 + \left[(1 + K(i)^2) - (1 + K(i)^2) \right] \hat{\sigma}_i^2 \right),$$

and restricting this estimator to the sub-sample of treated units, after some simple manipulation we have the estimator of the variance of $\hat{\tau}_{tre}^b$

$$\hat{V}(\hat{\tau}_{tre}^b) = \frac{1}{N_1^2} \sum_{i=1}^N \left(W_i \left(\hat{Y}_i(1) - \hat{Y}_i(0) - \hat{\tau}_{tre}^b \right)^2 + (1 - W_i) \left[(K(i)^2) - K(i)^2 \right] \hat{\sigma}_i^2 \right).$$

Given Proposition 3.2.2, the proof of consistency of the variance estimator for $\hat{\tau}^b$ (and $\hat{\tau}_{tre}^b$) is a straightforward extensions of those from Abadie & Imbens (2006) for the simple matching estimator.

4. SMALL SAMPLE PROPERTIES

This part¹ shows results based on an empirical application and Monte Carlo evidence. In particular, we implement the proposed estimator to data from the National Supported Work (NSW) Demonstration and evaluate its performance in an Empirical Monte Carlo design based on this data. In addition to that, in this section we also carry out Monte Carlo simulations with different designs taken from Busso et al. (forthcoming) and evaluate the BLOP's performance comparing it to other matching estimators in cases of correct and misspecification of the outcome and selection equations.

4.1 NSW Demonstration

The National Supported Work (NSW) Demonstration data is a well-known test to assess the finite sample properties of matching estimator. In particular, it has been analyzed by Lalonde (1986), Dehejia & Wahba (1999), Smith & Todd (2005), and Abadie & Imbens (2011*b*) among other authors.

In order to assess the performance of the BLOP matching estimator we provide estimates of the average treatment on the treated (ATT) effect using two control groups: the control group from the experimental sample from Lalonde (1986) and a control group from the Panel Study of Income Dynamics (PSID) used by Dehejia & Wahba (1999), Smith & Todd (2005), and Abadie & Imbens (2011*b*) among others.

Table 5.1 presents some summary statistics of the data. As can be seen from the experimental data, treated and control units are well balanced in terms of sample means. Indeed we fail to reject the null hypothesis of the differences being equal to zero for the 9 covariates considered. However, when comparing the treated units from the experimental data with those from the control group in the non-experimental data (PSID), we can see that the samples differ significantly in terms of first moment for 8 of the 9 covariates. The only case in which we fail to reject the null hypothesis of means equality is for the Hispanic dummy variable. Thus, using the non-experimental data is an interesting scenario to check the balancing properties of the proposed estimator.

¹ This part contains one Section of the paper named “*A matching estimator based on a bi-level optimization problem*”, which has been written by Professor Jorge Rivera, Professor Tomás Rau, and me. Indeed, Professor Rau played a key role in completion of the numerical exercises. Thus, some parts of this Section are Professor Rau's contributions.

In Table 5.2 we present the ATT results for both control groups. In addition, we compare our estimator with those from the nearest neighbor matching estimator with differing number of matches (from 1 to all). As can be observed, with the experimental data (control group) the estimate is very close to the benchmark (US\$1,794) and is comparable to those obtained with nearest neighbor using 16 neighbors. However, for the BLOP estimator only an average of 2.66 neighbors per observation were needed (ranging from 1 to 8). With the BLOP matching estimator we do not need to worry about the choice of number of matches.

When analyzing the results for the PSID control group we can see that the nearest neighbor estimator performs relatively well with 1 and 4 neighbors but very badly when the number of neighbors increases. The BLOP estimator gives a slightly higher estimate than the nearest neighbor and it does not explode (in terms of bias) as the nearest neighbor since it chooses optimally the number of neighbors. In this case, for each unit to match, only 2.44 neighbors were needed on average. In Figure 5.1 we present the histograms for the number of units used for the experimental and the PSID sample. In the left panel we show the histogram for the experimental data. As can be observed, it is highly left skewed with a median value equal to 2. The right panel shows the histogram for the number of units used with the PSID control group. As with the experimental data, the histogram is left skewed but less so. Additionally, the median value of the number of units used in this case is equal to 2 as well.

As we mentioned before, the non-experimental data from the PSID is quite different from experimental data's control group. Thus, it is interesting to analyze the post matching balance of our estimator. To do so, we compare the sample mean and Kolmogorov-Smirnov distances among the treated units and their counterfactuals, constructed with the units chosen for each match and the computed optimal weights. Table 5.3 presents the results of the post-matching balance. The BLOP estimator is able to balance the 9 covariates for sample means (failing to reject the null of means equality at 1%). For the continuous covariates we perform a Kolmogorov-Smirnov test and compute the p-values implementing bootstrapping with 500 repetitions. For 3 of the 4 continuous covariates we cannot reject the null of equality of treated units distribution and their counterfactuals.

4.2 *Empirical Monte Carlo*

In this section we implement an empirical Monte Carlo experiment to evaluate the performance of the BLOP matching estimator for the average treatment effect on the treated. We compare its performance in terms of bias, variance, and post-matching covariate balance to the nearest neighbor matching estimator (on covariates and propensity score) and the normalized inverse probability weighting estimators (IPW).²

The empirical Monte Carlo design is taken from Busso et al. (forthcoming). They focus on the African American subsample in the experimental group (156 individuals); the control

² See Imbens (2004) for a discussion of this estimator.

group is taken from the PSID (624). The covariates considered are similar to the previous subsection (age, education, marital status, earnings in 1974 and 1975, and unemployment in 1974 and 1975) plus a dummy for high school dropouts. Also, an interaction between the 1974 and 1975 unemployment dummies and an interaction between 1974 and 1975 earnings are included. Last, earnings in 1974 and 1975 squared complete the full set of covariates. Let X_i be the list of covariates excluding squared terms and interactions and Z_i the full set of covariates (including squared terms and interactions).

The data generation process (DGP) is explained in detail in Busso et al. (forthcoming) but we want to touch on it briefly. The main framework is given by the following equations

$$Y_i(0) = m(Z_i) + \sigma\epsilon \quad (4.1)$$

$$T_i^* = \alpha + \beta Z_i - U_i \quad (4.2)$$

where Z_i is a function of the covariates (described previously) and U_i follows a standard logistic distribution. Each sample is constructed such that covariates (X_i) are drawn from a population model (that uses first and second moments from the original sample). Then, U_i are drawn from a standard logistic distribution and T^* is generated using equation (4.2) where we use, instead of α and β , the coefficients from a logistic regression estimated on the original NSW sample. Hence, a latent treatment variable T^* is generated using equation (4.2). Using a linear function for $m(\cdot)$ and sampling ϵ from a standard normal distribution (independent of X_i), we use equation (4.1) to generate $Y_i(0) = \delta_0'Z_i + \epsilon_{0i}$. Instead of δ_0 we use the coefficients from a regression of $Y_i(0)$ on Z_i using control observations in the NSW sample. The root mean squared error of the regression is assigned to σ_0^2 . $Y_i(1)$ is constructed analogously regressing $Y_i(1)$ on Z_i using the treated units from the NSW sample. Last, we construct $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$.

We draw 5,000 random samples of size $N=400$. The population treatment effect on the treated for this benchmark case is \$2,334. Since the benchmark case has poor/bad overlap, Busso et al. (forthcoming) also consider a case with good overlap to evaluate the performance of different estimators under the two cases.³

In Table 5.4, we present the estimates of different matching estimators to assess the BLOP's performance in comparison to commonly used estimators such as nearest neighbor in characteristics, nearest neighbor in propensity score, and the normalized inverse probability weighting (IPW) estimator (Hirano, Imbens & Ridder 2003). In this exercise, the propensity score is correctly specified so nearest neighbor matching estimators on the propensity score and IPW are well specified. As can be seen, for the design with poor overlap (Panel A) the BLOP estimator (on characteristics) performs well in terms of bias when compared to

³ In this case, the parameters of the selection to treatment equation are divided by 5 so the random component in the assignment to treatment (U_i) is relatively more important, implying a better overlap of the data.

the nearest neighbor estimator in characteristics. As noted, when the number of neighbors increase the bias increases by a factor of 11 and its variance decreases to a third, which remark the importance of the choice of the number of matches.⁴ One plausible explanation for the dramatic increase in the bias is related to the overlap being poor. Thus, increasing the number of neighbors necessarily implies to include units far away to the units that need to be matched. When compared to the IPW, the BLOP matching estimator in covariates performs slightly better than the IPW in terms of bias and variance. We also compute the BLOP estimator on the (correctly specified) propensity score. As shown in Table 5.4, the BLOP matching estimator on the propensity score performs similarly to the IPW in terms of bias and variance. Nearest neighbor matching estimators on the propensity score performs well with few neighbors (between 1 and 4) but when they increase to 16 the performance is poor in terms of bias.

When analyzing the results with good overlap (Table 5.4, Panel B) the results are qualitatively similar to those with poor overlap. The BLOP matching estimator in covariates and the IPW performs similarly well in terms of bias and variance. Nearest neighbor matching estimator on covariates also performs well with 4 neighbors in terms of bias that increases by a factor of 10 with 16 neighbors. Nearest neighbor matching on propensity score achieves its best performance in term of bias with 4 neighbors, which increases by a factor of 24 with 16 neighbors. Thus, the BLOP matching estimator has an excellent performance given that solves an important issue: it determines the number of neighbors used (by finding the weights that optimize covariate balance), an open question for nearest neighbor estimators.

Next we check the balance performance of the estimators analyzed in Table 5.4. In Table 5.5 we show the p-values of the difference in means test for all the covariates for treated individuals and their predicted covariates using the different matching estimators. The BLOP matching estimator has the higher average p-value in both designs (panels A and B). In general, matching estimators on the (correctly specified) propensity score and IPW performs well. However, when overlap is poor, nearest neighbor matching estimator on covariates performs poorly relative to the BLOP or IPW matching estimators when the number of neighbors increase (both designs).

Hence, the empirical Monte Carlo evidence suggests that when the propensity score is correctly specified, matching on propensity score with few neighbors (between 1 and 4) and reweighting (IPW) performs well in terms of bias, variance, and overlap. The BLOP matching estimator also performs well without needing to estimate the propensity score. The next section analyzes the case of misspecification in the selection equation and, hence, the propensity score estimation.

⁴ See Härdle (1990) for details about the trade-off bias-variance in nearest neighbor estimators.

4.3 Finite sample properties and misspecification

In this section we implement Monte Carlo simulations with different designs taken from Busso et al. (forthcoming) and evaluate the BLOP's performance comparing it to other matching estimators in cases of correct and misspecification of the outcome and selection equations.

Design 1 considers a linear model for both the outcome and the selection equation. The number of observations is $N = 200$, and there are 4 covariates (X_1, X_2, X_3, X_4) normally distributed with zero mean and a block diagonal matrix given by Σ .⁵ This structure permits correlation between only X_1 and X_2 and only X_3 and X_4 . The main framework follows equations (4.1) and (4.2). As in the empirical Monte Carlo of Section 3, Z_i is a function of the covariates specified below and U_i follows a standard logistic distribution. Hence, a latent treatment variable T^* is generated using equation (4.2). Using a linear function for $m(\cdot)$ and sampling ϵ from a standard normal distribution (independent of X_i) and setting $\sigma = 1$, we use equation (4.1) to generate $Y_i(0)$. We assume a constant treatment effect equal to 1 to generate $Y_i(1) = T_i + Y_i(0)$ where $T = \mathbf{1}(T^* > 0)$ where $\mathbf{1}(\cdot)$ is an indicator function that is equal to one when the argument is true and equal to zero otherwise.

Design 2 considers a linear model with interactions for both the outcome and the selection equation. Hence, Z_i includes the four linear terms (X_1, X_2, X_3, X_4) plus the six interactions $(X_1X_2, X_1X_3, \text{ and so on})$.

For both designs, we consider cases of correct and misspecification for the propensity score. Table 5.6 shows the results of this Monte Carlo exercise after 5,000 replications. Focusing on absolute bias and empirical variance of the estimators we see that the BLOP in characteristics performs well in terms of absolute bias compared to the nearest neighbor matching estimator (NN) in Design 1 (columns 1 to 3, Table 5.6). The BLOP estimator based on the correctly specified propensity score (linear, column 1 in Table 5.6) is very competitive with nearest neighbor estimates on the propensity score and IPW. When the propensity score is misspecified (interactions only, column 2 in Table 5.6) the performance of the BLOP on characteristics is outstanding relative to the other estimators. When over specified (linear plus interactions, column 3), the comparison is similar to the case in which the propensity score is correctly specified.

In Design 2 (columns 4 to 6, Table 5.6), the true specification of the propensity score (and $m(Z_i)$) is a linear plus interaction equation (column 6, Table 5.6). When the propensity score is underspecified (linear, column 4 in Table 5.6), the nearest neighbor matching estimator (on covariates) performs well with 1 and 4 neighbors. With 16 neighbors, nearest neighbor matching on covariates performs similarly to the BLOP but with 64 neighbors its bias greatly increases. Estimates based on the propensity score perform poorly in terms of bias with 1 and 4 neighbors but improve significantly with 64 neighbors.

⁵ The lower right and upper left blocks of Σ are given by $\frac{1}{3} \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$.

The second case of underspecification is when the selection equation considers only interactions but no linear terms (interactions, column 5 in Table 5.6) and the results are similar to the previous case. When the selection equation is correctly specified (linear plus interactions, column 6 in Table 5.6), estimates based on the propensity score perform very well. Specifically nearest neighbor on propensity score with 1 and 4 neighbors, IPW and the BLOP on the propensity score are very close in absolute bias and variance. Thus, under misspecification, the BLOP (on covariates) performs very well in comparison to other estimators based on the propensity score, especially when this is underspecified.

5. APPENDIX: FIGURES AND TABLES

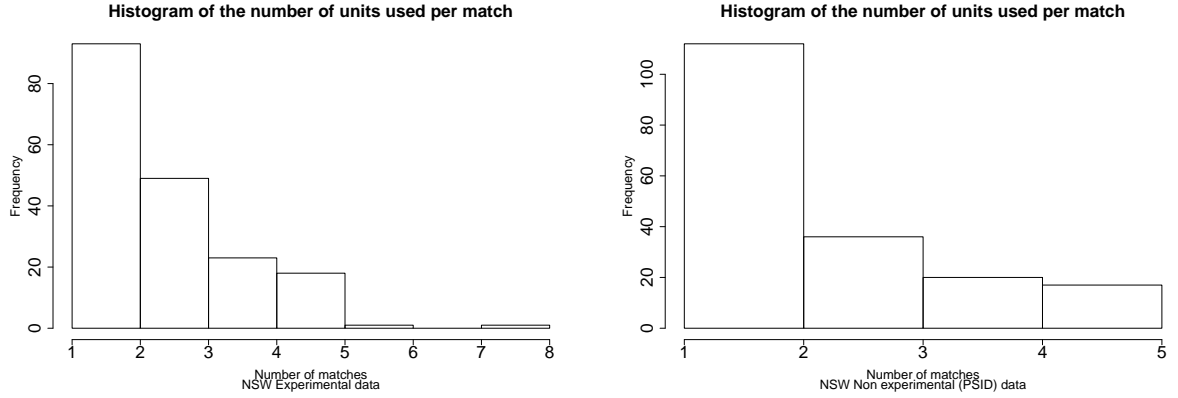


Fig. 5.1: Histograms of the number of units used

Tab. 5.1: Summary statistics

Variable	Experimental data				Non-experimental PSID		p-value	
	Treated (185)		Control (260)		Control (2490)		Treat/	Treat/
	Mean	(SD)	Mean	(SD)	Mean	(SD)	Control Exp.	Control PSID
Age	25.8	(7.2)	25.1	(7.1)	34.9	(10.4)	0.27	0.00
Education	10.3	(2.0)	10.1	(1.6)	12.1	(3.1)	0.15	0.00
Black	0.84	(0.36)	0.83	(0.38)	0.25	(0.43)	0.65	0.00
Hispanic	0.06	(0.24)	0.11	(0.31)	0.03	(0.18)	0.06	0.13
Married	0.19	(0.39)	0.15	(0.36)	0.87	(0.34)	0.33	0.00
Earnings '74	2.10	(4.89)	2.11	(5.69)	19.4	(13.41)	0.98	0.00
Earnings '75	1.53	(3.22)	1.27	(3.10)	19.1	(13.60)	0.39	0.00
Unemployed '74	0.71	(0.46)	0.75	(0.43)	0.10	(0.28)	0.33	0.00
Unemployed '75	0.60	(0.49)	0.68	(0.47)	0.09	(0.30)	0.07	0.00

Notes: Earnings are expressed in thousands of 1978 dollars. The last two columns show the p-values for the differences in means test between treated and controls from the two samples.

Tab. 5.2: Estimates for the NSW data

	Neighbors					
	Optimal	$k = 1$	$k = 4$	$k = 16$	$k = 64$	All
<i>Panel A: Experimental control group</i>						
BLOP	1728.9 (637.5)	-	-	-	-	-
NN Covariates	-	1223.2 (852.9)	1994.6 (712.7)	1753.3 (714.2)	2204.9 (704.9)	1794.3 (650.6)
NN P-Score	-	1608.2 (804.6)	1970.9 (723.2)	1863.8 (680.5)	1847.1 (674.9)	1794.3 (650.6)
<i>Panel B: Non-Experimental control group (PSID)</i>						
BLOP	2338.9 (697.3)	-	-	-	-	-
NN Covariate	-	2073.5 (1673.6)	1618.7 (1511.8)	469.2 (1101.6)	-111.6 (820.7)	-15204.8 (938.6)
NN P-Score	-	2141.7 (1557.8)	2061.8 (1461.0)	1014.4 (1364.9)	-182.1 (977.9)	-15204.8 (938.6)

Notes: The estimator presented is for the ATT. BLOP refers to our proposed estimator, NN refers to Nearest Neighbor matching estimator and IPW to the normalized Inverse Probability Weighting estimator. For the experimental sample the optimal number of matches was 2.66 (ranging from 1 to 8) and for the non-experimental was 2.44 (ranging from 1 to 5). Standard errors follow the approach of Abadie and Imbens (2006).

Tab. 5.3: Post Matching Balance: NSW data with PSID controls

Variable	Mean	Mean	t-statistic	KS
	Treated	Control	p-value	p-value
Age	25.82	27.06	0.092	0.004
Education	10.35	10.33	0.934	0.800
Black	0.84	0.84	1.000	-
Hispanic	0.06	0.06	1.000	-
Married	0.19	0.19	0.995	-
Earnings 74'	2095.57	2211.63	0.818	0.590
Earnings 75'	1532.06	2563.57	0.013	0.062
Unemployed 74'	0.71	0.60	0.953	-
Unemployed 75'	0.60	0.71	0.968	-

Notes: The post-matching balance corresponds to the comparison of treated units and their counterfactuals constructed with the units chosen for each match and the optimal weights computed. The average treatment effect is 2,338.9 and the average number of units used in the matching was 2.44 with a minimum of 1 and a maximum of 5. The KS p-value corresponds to the bootstrapped p-value for the Kolmogorov-Smirnov test for the distribution of the non-binary covariates.

Tab. 5.4: Empirical Monte Carlo, $N = 400$

	BLOP		NN Covariates				NN P-Score		
	Covariates	P-Score	$k = 1$	$k = 4$	$k = 16$	$k = 1$	$k = 4$	$k = 16$	IPW
<i>Panel A: Bad Overlap</i>									
Abs. Bias $\times 1000$	97.89	28.84	116.11	353.11	1369.62	25.15	213.79	1194.95	147.70
Variance $\times n$	3379.62	5139.79	3442.77	1908.39	1014.64	5341.96	3090.08	1686.45	3610.69
<i>Panel B: Good Overlap</i>									
Abs. Bias $\times 1000$	10.69	34.56	71.19	24.10	233.15	27.43	2.16	51.62	13.82
Variance $\times n$	795.36	990.36	859.22	678.79	633.60	1039.88	787.57	720.98	829.45

Notes: BLOP corresponds to our estimator performing the matching on characteristics and BLOP P-Score to our estimator doing the matching on the estimated propensity score from a correctly specified model. NN Covariates corresponds to the k -nearest neighbor matching estimator on characteristics and NN P-Score to the k -nearest neighbor matching estimator on the propensity score. IPW corresponds to the normalized Inverse Probability Weighting estimator. Bad overlap corresponds to the NSW DGP that mimics the overlap of the original NSW sample. In Panel B, Good overlap corresponds to the NSW DGP in which the coefficients of the selection equation are divided by 5. See Figure 3 of Busso, DiNardo and McCrary (forthcoming) for more details.

Tab. 5.5: P-values, post-matching Balance Empirical Monte Carlo, $N = 400$

Variable	BLOP		NN Covariates				NN P-Score				IPW
	Covariates	P-Score	$k = 1$	$k = 4$	$k = 16$	$k = 1$	$k = 4$	$k = 16$	$k = 16$		
<i>Panel A: Bad Overlap</i>											
Age	0.22	0.32	0.11	0.04	0.00	0.30	0.28	0.20	0.24		
Education	0.43	0.24	0.24	0.28	0.20	0.25	0.31	0.43	0.27		
Dropout	0.82	0.23	0.51	0.24	0.40	0.23	0.30	0.42	0.25		
Married	0.74	0.45	0.32	0.13	0.00	0.39	0.39	0.14	0.34		
Unemployed 74'	0.47	0.49	0.07	0.01	0.00	0.42	0.45	0.18	0.31		
Unemployed 75'	0.99	0.30	0.41	0.28	0.02	0.44	0.46	0.14	0.26		
Earnings 74'	0.76	0.52	0.49	0.04	0.00	0.38	0.34	0.04	0.38		
Earnings 75'	0.33	0.54	0.99	0.86	0.18	0.26	0.31	0.12	0.39		
Average p-value	0.59	0.39	0.39	0.23	0.10	0.33	0.36	0.21	0.31		
Min p-value	0.22	0.23	0.07	0.01	0.00	0.23	0.28	0.04	0.24		
<i>Panel B: Good Overlap</i>											
Age	0.78	0.54	0.44	0.49	0.43	0.48	0.56	0.57	0.57		
Education	0.79	0.50	0.44	0.50	0.48	0.50	0.57	0.60	0.58		
Dropout	0.95	0.49	0.64	0.48	0.33	0.48	0.57	0.58	0.57		
Married	0.92	0.58	0.50	0.50	0.38	0.48	0.57	0.59	0.57		
Unemployed 74'	0.89	0.63	0.42	0.40	0.16	0.51	0.60	0.63	0.55		
Unemployed 75'	0.95	0.57	0.44	0.53	0.41	0.51	0.59	0.62	0.55		
Earnings 74'	0.90	0.66	0.90	0.34	0.04	0.45	0.53	0.54	0.59		
Earnings 75'	0.91	0.69	0.93	0.49	0.12	0.44	0.51	0.52	0.59		
Average p-value	0.89	0.58	0.59	0.46	0.29	0.48	0.56	0.58	0.57		
Min p-value	0.78	0.49	0.42	0.34	0.04	0.44	0.51	0.52	0.55		

Notes: BLOP corresponds to our estimator performing the matching on characteristics and BLOP (P-Score) to our estimator doing the matching on the estimated propensity score from a correctly specified model. NN Covariates corresponds to the k -nearest neighbor matching estimator on characteristics and NN P-Score to the k -nearest neighbor matching estimator on the propensity score. IPW corresponds to the normalized Inverse Probability Weighting estimator. Bad overlap corresponds to the NSW DGP that mimics the overlap of the original NSW sample. In Panel B, Good overlap corresponds to the NSW DGP in which the coefficients of the selection equation are divided by 5. See Figure 3 of Busso, DiNardo and McCrary (forthcoming) for more details.

Tab. 5.6: Monte Carlo Evidence with Misspecification: Busso, DiNardo and McCrary (forthcoming)
 $N = 200, 5000$ repetitions

	Design 1			Design 2		
	Linear (1)	Interactions (2)	Linear + Interactions (3)	Linear (4)	Interactions (5)	Linear + Interactions (6)
<i>Panel A: Absolute Bias $\times 1000$</i>						
BLOP Covariates	110.6	110.6	110.6	86.9	86.9	86.9
BLOP P-Score	6.1	595.6	11.9	101.8	175.5	1.7
NN Covariates						
$k = 1$	183.5	183.5	183.5	26.7	26.7	26.7
$k = 4$	274.7	274.7	274.7	45.4	45.4	45.4
$k = 16$	428.5	428.5	428.5	97.9	97.9	97.9
$k = 64$	560.7	560.7	560.7	138.0	138.0	138.0
NN P-Score						
$k = 1$	6.6	595.8	12.6	101.3	174.3	2.0
$k = 4$	26.6	592.6	31.5	101.3	169.5	6.6
$k = 16$	107.1	584.1	108.4	91.7	156.7	29.9
$k = 64$	389.4	571.6	395.6	34.2	122.9	83.1
IPW	3.4	595.3	10.0	108.7	177.9	3.3
<i>Panel B: Variance $\times n$</i>						
BLOP Covariates	8.3	8.3	8.3	7.1	7.1	7.1
BLOP P-Score	9.6	11.2	10.5	8.2	8.3	9.2
NN Covariates						
$k = 1$	7.9	7.9	7.9	7.1	7.1	7.1
$k = 4$	6.1	6.1	6.1	5.3	5.3	5.3
$k = 16$	6.0	6.0	6.0	5.1	5.1	5.1
$k = 64$	6.6	6.6	6.6	5.4	5.4	5.4
NN P-Score						
$k = 1$	10.2	12.1	11.2	8.9	9.1	9.9
$k = 4$	7.1	8.1	7.6	6.1	6.2	6.6
$k = 16$	5.9	7.2	6.4	5.3	5.6	5.5
$k = 64$	6.0	7.1	6.5	5.4	5.6	5.3
IPW	7.1	7.2	7.8	5.2	5.6	5.8

Notes: Design 1 corresponds to a linear equation for the outcome equation and linear equation for the selection equation. Design 2 corresponds to a linear plus interaction equation for the outcome and selection equations. Thus, in Design 1, the column (1) corresponds to a correctly specified model for the propensity score, column (2) to a misspecified model and column (3) to an over specified model for the propensity score. For Design 2, columns (4) and (5) correspond to a miss-specified model for the propensity score and column (6) to a correctly specified model for the propensity score.

BIBLIOGRAPHY

- Abadie, A. & Imbens, G. W. (2006), ‘Large sample properties of matching estimator for average treatment effect’, *Econometrica* **74**(1), 235–267.
- Abadie, A. & Imbens, G. W. (2011a), ‘Bias-corrected matching estimators for average treatment effects’, *Journal of Business & Economic Statistics* (29), 1–11.
- Abadie, A. & Imbens, G. W. (2011b), ‘Bias-corrected matching estimators for average treatment effects’, *Journal of Business & Economic Statistics* **29**(1), 1–11.
- Botkin, N. & Stoer, J. (2005), ‘Minimization of convex functions on the convex hull of a point set’, *Math. Meth. Oper. Res.* **62**, 167–185.
- Busso, M., DiNardo, J. & McCrary, J. (forthcoming), ‘New evidence on the finite sample properties of propensity score reweighting and matching estimators’, *Review of Economics and Statistics* .
- Colson, B., Marcotte, P. & Savard, G. (2007), ‘An overview of bilevel optimization’, *Annals of Operations Research* **153**, 235–256.
- Cover, T. & Efron, B. (1967), ‘Geometrical probability and random points on a hypersphere’, *The Annals of Mathematical Statistics*. **38**, 213–220.
- Dehejia, R. H. & Wahba, S. (1999), ‘Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs’, *Journal of the American Statistical Association* **94**(448), 1053–1062.
- Diamond, A. & Sekhon, J. S. (2013), ‘Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies’, *Review of Economics and Statistics* **95**(3), 932–945.
- Evans, D., Jones, A. & Schmidt, W. (2002), ‘Asymptotic moments of near-neighbour distance distributions’, *Proc. R. Soc. Lond.* **458**, 2839–2849.
- Frölich, M. (2004), ‘Finite-sample properties of propensity-score matching and weighting estimators’, *Review of Economics and Statistics* **86**(1), 77–90.
- Galdo, J. C., Smith, J. & Black, D. (2008), ‘Bandwidth selection and the estimation of treatment effects with unbalanced data’, *Annals of Economics and Statistics / Annales d’Économie et de Statistique* (91/92), 189–216.
- Graham, B. S., Pinto, C. C. D. X. & Egel, D. (2012), ‘Inverse probability tilting for moment condition models with missing data’, *Review of Economic Studies* **79**(3), 1053–1079.

-
- Härdle, W. (1990), *Smoothing Techniques: With Implementation in S*, Springer-Verlag, New York.
- Heckman, J., Ichimura, H. & Todd, P. (1998), ‘Matching as an econometric evaluation estimator’, *Review of Economic Studies* **65**, 261–294.
- Hirano, K., Imbens, G. W. & Ridder, G. (2003), ‘Efficient estimation of average treatment effects using the estimated propensity score’, *Econometrica* **71**(4), 1161–1189.
- Imai, K. & Ratkovic, M. (2014), ‘Covariate balancing propensity score’, *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **76**(1), 243–263.
- Imbens, G. W. (2004), ‘Nonparametric estimation of average treatment effects under exogeneity: A review’, *The Review of Economics and Statistics* **86**(1), 4–29.
- Imbens, G. W. & Wooldridge, J. M. (2009), ‘Recent developments in the econometrics of program evaluation’, *Journal of Economic Literature* (47), 5–86.
- Lalonde, R. J. (1986), ‘Evaluating econometric evaluations of training programs with experimental data’, *The American Economic Review* **76**(4), 604–620.
- Majumdar, S., Comtet, A. & Randon-Furling, J. (2010), ‘Random convex hulls and extreme value statistics’, *Journal of Statistical Physics* **138**, 995–1009.
- Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994), ‘Estimation of regression coefficients when some regressors are not always observed’, *Journal of the American Statistical Association* **89**(427), 846–866.
- Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1995), ‘Analysis of semiparametric regression models for repeated outcomes in the presence of missing data’, *Journal of the American Statistical Association* **90**(429), 106–121.
- Rockafellar, R. (1972), *Convex Analysis*, Princeton University Press, New Jersey.
- Rosenbaum, P. & Rubin, D. (1983), ‘The central role of the propensity score in observational studies for causal effects’, *Biometrics* (70), 41–55.
- Rubin, D. (1973), ‘Matching to remove bias in observational studies’, *Biometrika* **29**, 159–183.
- Smith, J. A. & Todd, P. E. (2005), ‘Does matching overcome lalonde’s critique of nonexperimental estimators?’, *Journal of Econometrics* **125**(1-2), 305 – 353.
- Wendel, J. (1962), ‘A problem in geometric probability’, *Math. Scand.* **11**, 109–111.