Large sample properties of an optimization-based matching estimator

Roberto Cominetti* Juan Díaz† Jorge Rivera‡

November 26, 2014

DRAFT. Not for distribution.

Abstract

This paper mainly concerns the the asymptotic properties of the *BLOP matching estimator* introduced by Díaz, Rau & Rivera (Forthcoming), showing that this estimator of the ATE attains the standard limit properties, and that its conditional bias is $O_p(N^{-2/k})$, with k the dimension of continuous covariates. Even though this estimator is not $\sqrt{N}$-consistent in general, when the order of magnitude of the numbers of control units is bigger than the one of treated units, we show that the BLOP matching estimator of ATT is $\sqrt{N}$-*consistent*. Finally, for a general nonparametric setting, the conditional bias of matching estimators that use a constant number of matches to perform the potential outcomes cannot attain the aforementioned stochastic orders, regardless of the weighting schemes used to perform the potential outcomes. The proof of these results uses novel contributions in the field of geometric probability theory we provide in this work. Our results improve the obtained by Abadie & Imbens (2006) when studying the limit properties of the well known NN-matching estimator.

Keywords: Matching estimator, treatment effect, nonparametric methods, asymptotic properties.

JEL Classification: C01, C14, C61.

*Department of Industrial Engineering, Universidad de Chile. Santiago, Chile, *email*: rccc@dii.uchile.cl

†Department of Economics, Universidad de Chile. Santiago, Chile, *email*: juadiaz@fen.uchile.cl

‡Department of Economics, Universidad de Chile. Santiago, Chile, *email*: jrivera@econ.uchile.cl

1 Introduction

Asymptotic properties of nonparametric matching estimators have been scarcely studied in the program evaluation literature.[1] 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 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.

This paper addresses the asymptotic properties of the *BLOP matching estimators* introduced by Díaz et al. (Forthcoming). These estimators are based on finding, for each unit that needs to be matched, sets of observations in the opposite treatment group such that a *convex combination* of them has the same covariate values as it, or minimizing the distance between them. Since this optimization problem may have more than one solution in terms of weighting schemes, the procedure that such scheme determine establishes a refinement criterion that looks for the set with the closest covariates values to the unit that is being matched, which is done by choosing, from the solutions, the one that minimizes the weighted sum of the norm of matching discrepancies to the power of two.[2] Therefore, the weighting schemes used to perform this estimator are dependent on units, and the number of counterfactuals employed for this purpose is endogenously determined by the method.

The main results of this paper are two. First, under practically the same conditions as those used by Abadie & Imbens (2006), we show that the BLOP matching estimator of ATE attains standard limit properties, and has a conditional bias that is $O_p(N^{-2/k})$. It is worth mentioning that this

[1] From Imbens & Wooldridge (2009), the matching estimators we are concerned with in this work perform the potential outcome imputed to any unit that needs to be matched as a weighted sum of observed outcomes of a fixed number of its nearest neighbors having the opposite treatment. Without loss of generality, weights for doing so can be assumed to belong to the simplex of dimension of that number, a condition that avoids the potential outcomes being out of range regarding the observed ones. This approach is applied for estimating the average treatment effect, ATE, the average treatment effect on the treated, ATT, and/or any other estimator defined on sub-samples of units.

[2] In the presentation of their approach, Díaz et al. (Forthcoming) consider the minimization of the weighted sum of the norm of matching discrepancies instead of the squares of these distances. Hence, the refinement criterion here assumed is directly derived from theirs, and can be implemented using the codes that they provide in that work. See details in §2.2 below.

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 reaches an arbitrary order –see Theorem 4.1 in §4–. Second, even though the BLOP matching estimator of 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, as its bias rate is better than the one attained by the NN-matching estimator.

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, $k+1$. We 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.[3] 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 3.1 in §3.1 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

[3]The limit properties presented by Díaz et al. (Forthcoming) are for a restricted version of the BLOP matching estimator, performed using a finite number of matches instead of the entire sample. Given that, it is not difficult to prove that this version of the BLOP estimator attains the same asymptotic properties as the NN-matching estimator.

[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.

of matching discrepancies, and therefore the balance in terms of covariates reached by the method, become 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 we investigate the number of times that, in expected value, a treated (control) unit is a vertex of the polytopes arising when the BLOP is solved for all control (treated) units. Proposition 3.2 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.

2 Preliminaries

2.1 Basic concepts, notation and standing assumptions

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 $Y = W\,Y(1) + (1 - W)\,Y(0) \in \mathbb{R}$, 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 is[6] $\tau = \mathbb{E}(Y(1) - Y(0))$, while the average treatment effect on the treated, ATT, is $\tau_{tre} = \mathbb{E}(Y(1) - Y(0) \,|\, W = 1)$. Regarding these concepts, the following hypotheses are quite standard in the program evaluation literature, and they will be part of our standing assumptions.[7]

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:*** $W \perp\!\!\!\perp ((Y(0), Y(1)) \mid X)$.

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

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

[7] See Heckman, Ichimura & Todd (1998), Imbens & Wooldridge (2009) and Rosenbaum & Rubin (1983) for a detailed discussion on them.

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

For $x \in \mathbb{X}$ and $w \in \{0,1\}$, the conditional expectation and conditional variance of Y are, respectively, $\mu(x,w) = \mathbb{E}(Y \mid X = x, W = w)$ and $\sigma^2(x,w) = \mathbb{V}(Y \mid X = x, W = w)$. 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 paper, 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* $\mathbb{X} \times \{0,1\}$.

A sample of size $N \in \mathbb{N}$ of Ω is denoted by $\Omega_N = \{(W_i, Y_i, X_i),\ i = 1, \ldots, N\}$, 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 $1, ..., N_0$, so the treated ones are labeled as $N_0 + 1, ..., N_0 + N_1$.

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), $i = 1, \ldots, N$, *are independent draws from the distribution of* Ω.

Remark 2.1. *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.*

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 $i \in \{1, \ldots, N\}$ and $m \in \mathbb{N}$, $m \leq N_{1-W_i}$, we set

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

as the index of the unit that is the mth nearest neighbor to unit i in the opposite treatment group. For integer M, the convex hull of the first M matches to unit i will play a quite relevant role in

this paper. This subset is denoted by

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

with $\Delta_M=\{(\lambda_1,\ldots,\lambda_M)\in\mathbb{R}_+^M,\,\lambda_1+\ldots+\lambda_M=1\}$, the *Simplex* of dimension M.

2.2 The BLOP matching estimator and some intuition behind the results

We begin this part with a simple reasoning that will serve to present the main contributions of this work. All the formal aspects and the proofs are postponed to Sections 3 and 4 below.

From Imbens & Wooldridge (2009) we already know that most matching methods approximate the unobserved (potential) outcome of a treated unit $i\in\{N_0+1,\ldots,N\}$, 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 $(\xi_1,\ldots,\xi_M)\in\Delta_M$ the weighting scheme used to perform the method.[8] In view of standing assumptions, after performing a second order Taylor expansion of $Y_{j_m(i)}=\mu_1(X_{j_m(i)})$, $m=1,\ldots,M$, around X_i, it is not difficult to realize that there are constants L_1, $L_2>0$, the upper bounds of derivatives of μ_1 onto $\mathbb{X}$, such that the *unit-level bias* can be approximated as

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

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. From Lemma 2 in Abadie & Imbens (2006),[9] it follows that the stochastic order of the unit-level conditional bias is $N_0^{1/k}$, since the order of the balance term in (1) dominates the order of the quadratic part of it, which is $N_0^{2/k}$ by that Lemma. This property is the key ingredient used by Abadie & Imbens (2006) to

[8] 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 $\xi_m=1/M$, $m=1,\ldots,M$, 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 et al. (1998)–.

[9] Under the assumptions mentioned in Remark 2.1, this result states that the α-moment of $\|X_i-X_{j_m(i)}\|$ is $O(N^{-\alpha/k})$, $m=1,\ldots,M$. To this result holds, a key condition is that M is a constant.

prove that the stochastic order of the conditional bias of the NN-matching estimator is $N^{1/k}$. In fact, by following their proofs, it is an easy matter to verify that this result is also attained by any matching estimator for which the number of matches employed is fixed exogenously, and the weighting schemes belong to the simplex of dimension equals to that number. This result certainly enhances their contributions.

Another consequence of approximation in (1) is that, for general nonparametric settings, the conditional bias of any matching method that uses a fixed number of matches should be, at most, $O(N^{-2/k})$. Moreover, even if this number increases with the sample size, say $M = N_0$ in the case of a treated unit being matched, it could occur that the order obtained by the conditional bias could be worse than the one obtained using a constant number of matches. For instance, in case $M = N_0$, since the NN-matching estimator, with probability one, uses all the counterfactuals to perform the potential outcomes, the fact that $\|X_i - X_{N_0}\|$ tends to the diameter of the supporting set of covariates when the sample size tends to infinity, is certainly a drawback for the purposes of improving the order of its conditional bias by using that number of matches.

With the aim of reaching the order $N^{2/k}$ for the conditional bias, we could initially be naturally tempted to employ weighting schemes that minimize the right-hand side of the expression in (1), 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, Díaz et al. (Forthcoming) 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, they propose a second optimization problem to find the solution that minimizes the quadratic part of the approximation in (1). In order to achieve the best possible balance, they 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 $i \in \{N_0 + 1, \ldots, N\}$, the first optimization problem they propose is

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

whose solution set is denoted by $argmin\{\mathcal{F}_i\}$. Given that set, the second optimization problem

that yields the weighting scheme to be employed by the BLOP approach they introduce is

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

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 as[10]

$$\lambda^i \;\; = \;\; (\lambda^i_1, \ldots, \lambda^i_{N_{1-W_i}}) \in \Delta_{N_{1-W_i}},$$

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

$$\widehat{Y}^b_i(0) \, = \, (1-W_i)Y_i + W_i \sum_{m=1}^{N_0} \lambda^i_m \, Y_m, \qquad \widehat{Y}^b_i(1) \, = \, W_i Y_i + (1-W_i) \sum_{m=1}^{N_1} \lambda^i_m \, Y_{m+N_0}.$$

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

$$\widehat{\tau}^b \, = \, \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{Y}^b_i(1) - \widehat{Y}^b_i(0) \right), \qquad \widehat{\tau}^b_{tre} \, = \, \frac{1}{N_1} \sum_{i=1}^{N} W_i \, \left(\widehat{Y}^b_i(1) - \widehat{Y}^b_i(0) \right). \tag{3}$$

From a geometric point of view, the problem $\mathcal{F}_i$ for a treated unit $i \in \{N_0+1, \ldots, N\}$ concerns the weighting schemes that serve to perform the *projection*[11] of X_i onto the convex hull of covariates of control units, $\mathbf{co}\{X_1, \ldots, X_{N_0}\}$. 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, $m \in \{1, \ldots, N_0\}$, participates actively in the construction of that projection when $\lambda^i_m > 0$, 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

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

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

by the BLOP approach is precisely the main difficulty we face for showing our results. In fact, that issue implies that we cannot use aforementioned Lemma 2 to obtain the order of the conditional bias of this estimator. In order to overcome this fundamental difficulty, denoting the value of problem $\mathcal{F}_i$ for a treated unit i by

$$\nu\{\mathcal{F}_i\} \quad = \quad \left\| X_i - \sum_{m=1}^{N_0} \lambda_m^i X_m \right\|, \tag{4}$$

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

$$\mathbb{E}\left(\nu\{\mathcal{F}_i\} \,|\, W_i = 1, \{X_\iota, W_\iota\}_{\iota=1}^N\right) \quad = \quad \nu\{\mathcal{F}_i\}\ \mathbb{P}(\nu\{\mathcal{F}_i\} \neq 0). \tag{5}$$

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 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 $X_i \notin \mathbf{co}\{X_1, \ldots, X_{N_0}\}$. Hence, under the standing assumption, a fundamental result for our purposes, which we show in §3.1, 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 (5) 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

$$\nu\{\mathcal{S}_i\} \quad = \quad \sum_{m=1}^{N_0} \lambda_m^i \, \|X_i - X_m\|^2. \tag{6}$$

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 $\widehat{\tau}^b$. The techniques we use to conclude the proofs are developed in §4.

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). The main result in §3.2 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 Some results in geometric probability theory

3.1 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,

$$\mathbb{P}(X_i \notin \mathbf{co}\{X_{j_1(i)}, \ldots, X_{j_M(i)}\}) \quad = \quad \mathbb{P}(0_k \notin \mathbf{co}\,\{U_{1,i}, \ldots, U_{M,i}\}), \tag{7}$$

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

Following Cover & Efron (1967) we say that a set of random vectors $\{\xi_1, \ldots, \xi_M\}$ 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 $\{X_1, \ldots, X_N\}$ is in general position. Besides, it is also clear that any M-subset of $\{X_1, \ldots, X_N\}$, 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, $i \in \{1, \ldots, N\}$ and $M > k$, $\{U_{1,i}, \ldots, U_{M,i}\}$ is in general position.

A remarkable result in Wendel (1962), slightly extended by Cover & Efron (1967), says that if the set of random vectors $\{\xi_1, \ldots, \xi_M\}$ of $\mathbb{R}^k$, with $M > k$, is in general position, and the joint distribution of them is invariant under reflections through the origin,[12] then the probability of a

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

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 $\{\xi_1,\ldots,\xi_M\}$ 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

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

The last inequality cannot be directly applied for upper bounding the probability in (7), since even though the set of matching discrepancies $\{U_{1,i},\ldots,U_{M,i}\}$ 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.

Proposition 3.1. *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 mth closest match of z is given by

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

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 rth 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 mth and rth $(r > m)$

nearest neighbors of z respectively is:

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

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

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

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

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

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

$$f_{j_1,\ldots,j_M}(v) = \frac{N!N^{-M}}{(N-M)!}\left(\prod_{s=1}^{M} f\left(z+\frac{v_{j_s}}{N^{1/k}}\right)\right)\left(1-\mathbb{P}\left(||X-z||\leq\frac{||v_{j_M}||}{N^{1/k}}\right)\right)^{N-M}, \qquad (8)$$

from which we can readily conclude the following inequality[13]

$$f_{j_1,\ldots,j_M}(v) \;\leq\; \bar{f}^{M} exp\left(-\underline{f}\,\frac{||v_{j_M}||^k}{(M+1)}\,\frac{\pi^{k/2}}{\Gamma(1+k/2)}\right), \qquad (9)$$

where $0<\underline{f}<\bar{f}<\infty$ are the lower and upper bounds of the distribution $f(\cdot)$, respectively. Using the right term in (9) we can define the distribution as stated. □

Remark 3.1. *Using (8) 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:*

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

The following result comes directly from the properties presented above.

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

Theorem 3.1. *If Assumptions* **1**, **2**, **3**, *and* **5** *hold, for a large enough* N, $i \in \{1, \ldots, N\}$ *and* $M > k$, *there is a constant* $\gamma > 0$ *such that*

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

Remark 3.2. *What Theorem 3.1 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 3.1. From this theorem, it is also clear that, given the sample, for each treated unit* i *(and similar for controls) we have that:*

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

3.2 The number of times that a unit is used as a match by the BLOPs

For a control unit $j \in \{1, \ldots, N_0\}$, after solving the optimization problem $\mathcal{S}_j$, the vector of covariates of a treated unit $i \in \{N_0 + 1, \ldots, N\}$ participates in the convex combination performing X_j (or its projection onto $\mathbf{co}\{X_{N_0+1}, \ldots, X_N\}$) whenever $\lambda^j_{i-N_0} > 0$.[14] 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 $M_j = \left\{ i' \in \{N_0 + 1, \ldots, N\}, \, \lambda^j_{i'-N_0} > 0 \right\}$. For the case that j is a treated unit, $M_j = \left\{ i' \in \{1, \ldots, N_0\}, \, \lambda^j_{i'} > 0 \right\}$, and given that, the number of times that a unit i, either control or treated, is a vertex of such polytopes is

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

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

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

[14] We have that the components of vector λ^j are $\lambda^j_1, \ldots, \lambda^j_{N_1}$, thus a treated unit $i \in \{N_0 + 1, \ldots, N_0 + N_1\}$ is associated with $\lambda^j_{N_0-i}$.

The following result will be quite relevant when we study the properties of the variance of the BLOP matching estimator in §4.2. 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.

Proposition 3.2. *If Assumptions* **1**, **2**, **3**, *and* **5** *hold, then for each unit i and integer α, $\mathbb{E}\left((K(i))^{\alpha}\right)$ 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 $i_1 \in \{N_0 + 1, \ldots, N\}$. In that case, for the sake of simplicity regarding notation, for a control unit j and $i \in \{N_0 + 1, \ldots, N\}$, 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:

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

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

$$\mathbb{E}\left(K(i_1)\right) \quad \leq \quad N_0\, \mathbb{P}(i \in M_j). \tag{10}$$

Denoting by $\mathcal{C}_1 = \mathbf{co}\{X_{N_0+1}, \ldots, X_N\}$, 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 3.1,

$$\mathbb{P}(i \in M_j) \quad \leq \quad \mathbb{P}(i \in M_j \,|\, X_j \in \mathcal{C}_1)\, \mathbb{P}(X_j \in \mathcal{C}_1) + \gamma\, \frac{N_1^k}{2^{N_1}}. \tag{11}$$

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.[15] Hence, since

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

$\mathbb{P}\left(i \in M_j \,|\, X_j \in \mathcal{C}_1\right) = \mathbb{P}\left(i \in \mathcal{M}_j \,|\, X_j \in \mathcal{C}_1\right)$, and partitioning the event $\{X_j \in \mathcal{C}_1\}$ into subevents[16] $\{X_j \in \Delta\mathcal{C}_j(m)\} = \{X_j \in \mathcal{C}_j(m) \setminus \mathcal{C}_j(m-1)\}$, $m = 2, \ldots, N_1$, it follows that[17]

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

Using the identical distribution of the aforementioned random variables,

$$\sum_{i=1}^{N_1} \mathbb{P}(i \in \mathcal{M}_j \,|\, X_j \in \Delta\mathcal{C}_j(m)) = N_1\, \mathbb{P}(i \in \mathcal{M}_j \,|\, X_j \in \Delta\mathcal{C}_j(m)), \tag{12}$$

and the fact that $\mathbb{P}(i \in \mathcal{M}_j \,|\, X_j \in \Delta\mathcal{C}_j(m)) = \mathbb{E}\left(\mathbb{1}\{i \in \mathcal{M}_j\} \,|\, X_j \in \Delta\mathcal{C}_j(m)\right)$, implies[18]

$$\mathbb{P}(i \in \mathcal{M}_j \,|\, X_j \in \Delta\mathcal{C}_j(m)) = \frac{m}{N_1}. \tag{13}$$

Thus, the combination of (12) and (13) 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 3.1,

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

for some constant $\gamma_2 > 0$. This last inequality along with relations (10) and (11) give

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

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 $\mathbb{E}\left(K(i_1)\right) \leq \kappa_1$. For the case $\alpha = 2$, we first notice that for $j, j' \in \{1, \ldots, N_0\}$, $j \neq j'$, 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

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

[17] $X_j \in \Delta\mathcal{C}_j(m)$ 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.

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

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. □

4 Asymptotic properties of the BLOP matching estimator

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 $\widehat{\tau}^b - \tau = A^b + E^b + B^b$, where

$$A^b = \frac{1}{N} \sum_{i=1}^{N} (\mu_1(X_i) - \mu_0(X_i)) - \tau, \quad E^b = \frac{1}{N} \sum_{i=1}^{N} (2W_i - 1)\,(1 + K(i))\,\epsilon_i, \tag{14}$$

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

$$B^b = \frac{1}{N} \left(\sum_{i=1}^{N_0} \sum_{m=1}^{N_1} \lambda_m^i (\mu_1(X_{m+N_0}) - \mu_1(X_i)) + \sum_{i=1+N_0}^{N} \sum_{m=1}^{N_0} \lambda_m^i (\mu_0(X_i) - \mu_0(X_m)) \right). \tag{15}$$

Note that after taking expectation to $\widehat{\tau}^b - \tau$, 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 manner similar to the approximation performed in (1), for a treated unit $i \in \{N_0 + 1, \ldots, N\}$, 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 (4) and (6)–:

$$\left|\sum_{m=1}^{N_0}\lambda_m^i(\mu_0(X_i)-\mu_0(X_m))\right| \;\leq\; L_1\,\nu\{\mathcal{F}_i\}+L_2\,\nu\{\mathcal{S}_i\}+O\left(\sum_{m=1}^{N_0}\lambda_m^i\|X_i-X_m\|^3\right), \qquad (16)$$

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 (16) 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).

4.1 The order of the conditional bias

The following property is the key result in this part.

Proposition 4.1. *If Assumptions* **1** – **5** *hold, then*

$$\mathbb{E}\left(\left|\sum_{m=1}^{N_0}\lambda_m^i(\mu_0(X_i)-\mu_0(X_m))\right| \,\Big|\, W_i=1,X_i,\{W_j,X_j\}_{j=1}^N\right)=O\left(N_0^{-2/k}\right). \qquad (17)$$

and

$$\mathbb{E}\left(\left|\sum_{m=1}^{N_1}\lambda_m^i(\mu_1(X_{m+N_0})-\mu_1(X_i))\right| \,\Big|\, W_i=0,X_i,\{W_j,X_j\}_{j=1}^N\right)=O\left(N_1^{-2/k}\right). \qquad (18)$$

Proof. Without loss of generality, the proof is performed for the relation in (17). For a treated unit i, the conditionals in (17) 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

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

For the case $m=N_0$ we have that $\mathcal{C}_i(N_0)=\mathbf{co}\{X_1,\ldots,X_{N_0}\}$, 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 (16) 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 3.1,

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

and then,

$$\mathbb{E}\left(N_0^{2/k}\,\psi_i \,\middle|\, X_i \notin \mathcal{C}_0\right)\,\mathbf{q}_i(N_0) = o(1). \tag{19}$$

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

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

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

$$\mathbb{E}\left(\psi_i \,\middle|\, X_i \in \mathcal{C}_0\right)\,\mathbf{p}_i(N_0) \;\leq\; \sum_{m=2}^{N_0} \mathbb{E}\left(L_2\,\nu\{\mathcal{S}_i\} \,\middle|\, \theta_i,\, X_i \in \Delta\mathcal{C}_i(m)\right)\,\mathbf{q}_i(m-1). \tag{20}$$

For $m \leq N_0$, notice that when $X_i \in \Delta\mathcal{C}_i(m)$, then there is a vector $(\xi_1, \ldots, \xi_m) \in \Delta_m$ 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 3.1 implies that $\mathbf{q}_i(m-1) \leq \gamma\,\frac{m^k}{2^m}$. All of this in (20) give

$$\begin{aligned}\mathbb{E}\left(\psi_i \middle| X_i \in \mathcal{C}_0\right)\mathbf{p}_i(N_0) \;\leq\; & \sum_{m=2}^{k} \mathbb{E}\left(L_2\,\|X_i - X_{j_m(i)}\|^2 \,|\, \theta_i, X_i \in \Delta\mathcal{C}_i(m)\right) + \\ & \sum_{m=k+1}^{N_0} \mathbb{E}\left(L_2\,\|X_i - X_{j_m(i)}\|^2 \,|\, \theta_i, X_i \in \Delta\mathcal{C}_i(m)\right) 2\gamma\,\frac{m^k}{2^m}. \end{aligned} \tag{21}$$

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**,

$$\mathbb{E}\left(N_0^{2/k}\,\|X_i - X_{j_m(i)}\|^2\right) \;=\; \eta_1\,\frac{\Gamma(m+2/k)}{\Gamma(m)} + O\left(\frac{1}{N_0^{1/k-\rho}}\right), \tag{22}$$

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 (22) 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 $m \leq N_0$,

$$\mathbb{E}\left(N_0^{2/k}\, \|X_i - X_{j_m(i)}\|^2\right) \leq \eta_3\, m^2 \tag{23}$$

for some constant $\eta_3 > 0$. Applying (23) to the right-hand side in (21) yields

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

for some constant C. This last inequality along with (19) implies the result for the order of ψ_i. The remaining to conclude is straightforward from the results just presented. □

Corollary 4.1. *If Assumptions* **1** – **5** *hold and* $\mu_w(\cdot)$, $w = 0, 1$, *is flat over* $\mathbb{X}$, *then* $B^b = o_{\mathbf{p}}(N^{-\beta})$ *for any integer* $\beta > 0$.

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

and

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

Proof. This result is straightforward from Proposition 4.1, since in this case it holds that only the linear term of Taylor's expansion of $\mu_w(\cdot)$, expression (16), is different from zero. □

The following result is directly from Proposition 4.1 and Corollary 4.1.

Theorem 4.1. *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 (15), we have

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

and by Proposition 4.1, 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 4.1. □

We end this part studying some additional properties concerning the BLOP matching estimator of the ATT, which from (3) is

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

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 $r > 1$,*

$$\frac{N_1^r}{N_0}\to\theta<\infty. \tag{24}$$

The following result is straightforward using the properties above.

Corollary 4.2. *If Assumptions* **1** – **4** *and* **6** *hold, then* $B^b_{tre} = O_{\mathbf{p}}\left(N_1^{-2r/k}\right)$. *In addition, if* $\mu_0(\cdot)$ *is flat over* $\mathbb{X}$, *then* $B^b_{tre} = o_{\mathbf{p}}(N_1^{-\beta})$ *for any integer* $\beta > 0$.

Proof. In view of Assumption **6** we can apply Proposition 4.1 to conclude $B^b_{tre} = O_{\mathbf{p}}(N_0^{-2/k})$. Hence, using (24) we can readily obtain the result. The remainder of this property is direct after using Corollary 4.1. □

4.2 Variance, consistency, and normality properties

Proposition 3.2 is the most relevant result we need for obtaining the variance and asymptotic normality properties of both $\widehat{\tau}^b$ and $\widehat{\tau}^b_{tre}$. 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 $\widehat{\tau}^b$, conditional on $\{X_i, W_i\}_{i=1}^N$, is given by

$$\mathbb{V}\left(\widehat{\tau}^b \middle| \{X_i, W_i\}_{i=1}^N\right) = \frac{1}{N^2}\sum_{i=1}^{N}(1+K(i))^2\,\sigma^2(X_i, W_i), \tag{25}$$

while for $\widehat{\tau}^b_{tre}$ it is

$$\mathbb{V}\left(\widehat{\tau}^b_{tre} \middle| \{X_i, W_i\}_{i=1}^N\right) = \frac{1}{N_1^2}\sum_{i=1}^{N}(W_i - (1-W_i)K(i))^2\,\sigma^2(X_i, W_i). \tag{26}$$

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

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

and for $\widehat{\tau}^b_{tre}$ these concepts are denoted by

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

Lemma 4.1. *If Assumptions* **1** – **5** *hold , then* $\mathbb{E}\left(V^{\mathrm{CV}}\right) = O(1)$. *If Assumptions* **1** – **4** *and* **6** *hold,*

then

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

Proof. For the first part, using (25), the proof is direct from Assumption **4** and Proposition 3.2. For the second part, the argument is the same, but using (26). □

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

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

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

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

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 (14), we have

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

and by Lemma 4.1, $\mathbb{E}(N\left(E^b\right)^2) = O(1)$. 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 4.1 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 $\sqrt{N}\,\left(\widehat{\tau}^b - \tau - B^b\right) = \sqrt{N}\,A^b + \sqrt{N}\,E^b$, the *Standard*

Central Limit Theorem and properties of E^b give

$$\sqrt{N}\, A^b \quad \xrightarrow{\mathbb{D}} \quad \mathcal{N}(0, V^{\mathrm{CM}}). \tag{27}$$

Finally, using the *Linderberg-Feller Central Limit Theorem*[19], Proposition 3.2 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

$$\frac{\sqrt{N}\, E^b}{\sqrt{V^{\mathrm{CV}}}} \quad \xrightarrow{\mathbb{D}} \quad \mathcal{N}(0, 1). \tag{28}$$

Because (27) and (28) are asymptotically independent, we conclude the proof. □

We conclude this part by presenting 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 corollary.

Corollary 4.3. *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* $\widehat{\tau}^b \quad \xrightarrow{\mathbb{P}} \quad \tau$ *and*

$$\frac{\sqrt{N}\, \left(\widehat{\tau}^b - \tau\right)}{\sqrt{V^{\mathrm{CV}} + V^{\mathrm{CM}}}} \quad \xrightarrow{\mathbb{D}} \quad \mathcal{N}(0, 1).$$

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* $\widehat{\tau}^b_{tre} \quad \xrightarrow{\mathbb{P}} \quad \tau_{tre}$ *and*

$$\frac{\sqrt{N_1}\, \left(\widehat{\tau}^b_{tre} - \tau_{tre}\right)}{\sqrt{V^{\mathrm{CV}}_{tre} + V^{\mathrm{CM}}_{tre}}} \quad \xrightarrow{\mathbb{D}} \quad \mathcal{N}(0, 1).$$

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

References

Abadie, A. & Imbens, G. W. (2006), 'Large sample properties of matching estimator for average treatment effect', *Econometrica* **74**, 235–267.

Cover, T. & Efron, B. (1967), 'Geometrical probability and random points on a hypersphere', *The Annals of Mathematical Statistics.* **38**, 213–220.

Díaz, J., Rau, T. & Rivera, J. (Forthcoming), 'A matching estimator based on a bi-level optimization problem', *The Review of Economics and Statistics* .

Evans, D., Jones, A. & Schmidt, W. (2002), 'Asymptotic moments of near-neighbour distance distributions', *Proc. R. Soc. Lond.* **458**, 2839–2849.

Heckman, J., Ichimura, H. & Todd, P. (1998), 'Matching as an econometric evaluation estimator', *Review of Economic Studies* **65**, 261–294.

Imbens, G. W. & Wooldridge, J. M. (2009), 'Recent developments in the econometrics of program evaluation', *Journal of Economic Literature* **47**, 5–86.

Majumdar, S., Comtet, A. & Randon-Furling, J. (2010), 'Random convex hulls and extreme value statistics', *Journal of Statistical Physics* **138**, 995–1009.

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.

Wendel, J. (1962), 'A problem in geometric probability', *Math. Scand.* **11**, 109–111.