SERIE DOCUMENTOS DE TRABAJO

FACULTAD ECONOMÍA Y NEGOCIOS
UNIVERSIDAD DE CHILE

DEPARTAMENTO DE ECONOMÍA

SDT 351

A MATCHING ESTIMATOR BASED ON A BI-LEVEL OPTIMIZATION PROBLEM

Autores: Juan Díaz, Tomás Rau y Jorge Rivera

Santiago, Marzo de 2012

A matching estimator based on a bi-level optimization problem*

Juan Díaz† Tomás Rau‡ Jorge Rivera§

January 17, 2012

Abstract

This paper proposes a matching estimator where the size of the weights and the number of neighbors are endogenously determined from the solution of a *bi-level optimization problem*. The *first level* problem minimizes the distance between the characteristics of an individual and a convex combination of characteristics of individuals belonging to the corresponding counterfactual set, and with the *second level* we choose a solution point of the first level that minimizes the sum of the distances between the characteristics of the individual under analysis and those from the counterfactuals employed in the optimal convex combination. We show that this estimator is consistent and asymptotically normal. Finally we study its behavior in finite samples by performing Monte Carlo experiments with designs based on the related literature. In terms of bias, standard deviation and mean square error, we find significant improvements using our estimator in comparison to the *simple matching estimator*, widely employed in the literature.

Keywords: Matching estimator, treatment effect estimator, non-parametric methods, bi-level optimization.

JEL Classification: C01, C14, C61.

*We are indebted to the useful comments of Roberto Cominetti, Guido Imbens, and seminar participants at Universidad de Chile, and the annual meeting of the Chilean Economic Association (SECHI). The usual disclaimer applies. Authors thank the funding provided by Fondecyt, Project Number 1095181, and ISCI. Tomás Rau thanks the funding provided by the Millenium Science Initiative from the Ministry of Economy, Development and Tourism to Microdata Center, Project NS100041

†Department of Economics, Universidad de Chile. Diagonal Paraguay 257, Santiago, Chile, *email*: juadiaz@fen.uchile.cl

‡Instituto de Economía, Pontificia Universidad Católica de Chile. Vicuña Mackenna 4860, Santiago, Chile, *email*: trau@uc.cl

§Department of Economics, Universidad de Chile. Diagonal Paraguay 257, Santiago, Chile, *email*: jrivera@econ.uchile.cl

1 Introduction

Matching estimators have been widely used in the impact evaluation literature during the past decades. Without any assumption about the distribution of involved variables, 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 by means of this matching estimator is usually defined as the difference between the imputed and actual outcome. One of the most popular estimators is the *simple 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 according to 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) point out "little is known about the optimal number of matches, or about data-dependent ways of choosing it".

In this paper we propose a matching estimator where the number of neighbors and the weighting scheme are endogenously determined from the solution of a *bi-level optimization problem* (from now on, BLOP), that is, an optimization problem where the restrictions are defined by another optimization problem (see Colson, Marcotte & Savard (2007)). The *first level* of the problem deals with finding weights belonging to the simplex associated with the control (treated) units such that the convex combination of their characteristics employing those weights, approximates as closely as possible the characteristics of a treated (control) unit. Since this problem may have more than one solution, the *second level* of the optimization problem is a refinement criterion which looks for a solution point of the first level that minimizes the sum of the distances between the characteristics of each individual used in the convex combination and the one under analysis by weighting each *individual distance* with weights obtained from the first level.

We would like to remark that in this setup, optimal weights could depend on characteristics, and the number of them, that are different from zero, may vary depending on individual characteristics as well. In fact, the weights that are non-null define the counterfactuals participating in the optimal convex combination, which are, not necessarily, the nearest neighbors. From Caratheodory's Theorem (see Rockafellar (1972)), the number of them should be at most the number of pre-treatment characteristics plus one. This is relevant for practical issues, since when the estimation is implemented with the whole sample (as we propose), there is no need to fix the number of nearest neighbors to be employed for estimation as in Abadie & Imbens (2006), which is one of the contributions of our proposed methodology.

In order to illustrate the method, suppose that there is a unique real-valued characteristic, which for the individual under analysis, labeled $i = 0$, is noted by $X_0 \in \mathbb{R}_+$. Suppose additionally that the potential counterfactuals for this individual are indexed by $i = 1, \ldots, n,\ n \geq 2$, with characteristics and outcomes given by $X_i \in \mathbb{R}_+$ and $Y_i \in \mathbb{R}$ respectively. Without loss

of generality, let us assume that $X_1 < ... < X_k < X_0 < X_{k+1} < \ldots < X_n$. The *first level* of the optimization problem mentioned above looks for the counterfactuals such that the convex combination of their characteristics are as close as possible to X_0. Obviously since this problem has multiple solutions, we consider a refinement criterion in order to select the *best* solution. As we mentioned, for the *second level* we propose an optimization problem that looks for *the solution* that minimizes the sum of weighted distances between the corresponding counterfactual X's and X_0, with weights from the convex combination coming from the first level. Intuitively, we try to perform the first level using the *closest possible* neighbors to individual $i = 0$. In the example, and with the exception of pathological cases, the solution of the BLOP should be

$$w_k^* = \frac{X_{k+1} - X_0}{X_{k+1} - X_k}, \quad w_{k+1}^* = \frac{X_0 - X_k}{X_{k+1} - X_k}, \; w_i^* = 0, \; i \neq k, \; k+1,$$

and according to our approach, the potential outcome to be imputed to individual $i = 0$ is

$$\widehat{Y_0} = \sum_{i=1}^{n} w_i^* \; Y_i = w_k^* \; Y_k + w_{k+1}^* \; Y_{k+1}.$$

When X is a vector of characteristics, solving the BLOP is certainly a difficult task due to the great number of alternatives that any optimization algorithm has to evaluate in order to achieve a global solution. In order to circumvent this complexity, we follow an approach based on Cominetti & San Martín (1994) and Cominetti (1999). According to this approach, the BLOP is approximated by a *parametrized* unrestricted optimization problem, where the objective function is a weighted sum of the first and second level, plus a penalty term that determines that the weights should belong to the Simplex. The penalty term depends on a positive parameter which also participates in the weighting of the first and second level as mentioned. Given this, it can be shown that when the parameter approaches to zero, the parametrized solution of the unrestricted problem converges to the solution of the bi-level problem we are dealing with. Employing these optimal weights we define the *bi-level matching estimators*, which are formally introduced in Section 2.

Regarding large sample properties of the bi-level matching estimator, in Section 3 we show that under the same conditions assumed by Abadie & Imbens (2006), this estimator is consistent and asymptotically normal. Indeed, these limit properties hold for any *generic matching estimator* as we define in Section 2.1. Finally, in Section 4 we study finite sample properties of the bi-level matching estimators. For that purpose, based on Frölich (2004) and Busso, DiNardo & McCrary (2009) we set up the experiments, implementing Monte Carlo simulations in order to assess its performance in terms of bias, variance, and mean square error, finding significant improvements in comparison with the *simple matching estimator*.

2 Generic and bi-level matching estimators

In this section we introduce, for a nonparametric framework, the concept of *generic matching estimator*, i.e., a matching estimator for which the weighting scheme is general, with only the conditions that they are non-negative and add up to one. Immediately after that, we define the *bi-level matching estimator*, which is a particular case of a generic estimator. The limit properties we show in Section 3 are valid for generic matching estimators, and then for the bi-level matching ones.

2.1 Generic matching estimators

Following standard notation (see Imbens & Wooldridge (2009)), set $k \in \mathbb{N},\ k \geq 1$, and $\mathbb{X} \subseteq \mathbb{R}^k$ a non-empty set, and consider a random variable $\Omega = (T, X, Y)$ whose values belong to $\{0,1\} \times \mathbb{X} \times \mathbb{R}$, with T indicating whether a *treatment* is received ($T = 1$) or not ($T = 0$) by an individual with a vector of *characteristics* X. The observed *outcome* is $Y = T\,Y_1 + (1-T)\,Y_0$, with Y_1 and Y_0 being the outcome that the agent would have obtained under treatment or the absence of treatment, respectively. For $x \in \mathbb{X}$ and $\delta \in \{0,1\}$, define the *conditional expectation* and *conditional variance* of Y as, respectively,

$$\mu_\delta(x) = \mathbb{E}[Y_\delta \mid X = x], \quad \sigma^2_\delta(x) = \mathbb{V}[Y_\delta \mid X = x]. \tag{1}$$

$$\mu(x,\delta) = \mathbb{E}[Y \mid X = x, T = \delta],\ \sigma^2(x,\delta) = \mathbb{V}[Y \mid X = x, T = \delta]. \tag{2}$$

The assumptions below are needed for the definitions. They are quite standard in the literature (see Rosenbaum & Rubin (1983), Heckman, Ichimura & Todd (1998), Imbens (2004) and Imbens & Wooldridge (2009) for a detailed discussion on them).

Assumption 1. *$\mathbb{X}$ is a compact and convex subset of $\mathbb{R}^k$ and the distribution of X is bounded away from zero.*

Assumption 2. *Unconfounded Treatment Assignment: $T \perp [(Y_0, Y_1) \mid X]$.*

Assumption 3. *Overlap: $0 < \mathbb{P}[T = 1 \mid X] < 1$.*

Note that under Assumption 2, corresponding terms in (1) are coincident with those in (2).

Definition 2.1. *The average treatment effect is denoted by τ, and the average treatment effect on the treated by τ_{tre}, i.e.,*

$$\tau = \mathbb{E}\left[\mathbb{E}[Y|X = x, T = 1] - \mathbb{E}[Y|X = x, T = 0]\right],$$

$$\tau_{tre} = \mathbb{E}[\mathbb{E}[Y|X = x, T = 1] - \mathbb{E}[Y|X = x, T = 0]|T = 1].$$

In what follows, for $n \in \mathbb{N}$ let us define $I^n = \{1, \ldots, n\}$ and assume as known a sample of size $n \in \mathbb{N},\ n \geq 2$, of Ω, say

$$\Omega^n = \{(T_i, X_i, Y_i), i \in I^n\}.$$

The set of indexes for treated and control individuals are noted, respectively, by

$$I_1^n = \{i \in I^n,\ T_i = 1\}\ ,\ I_0^n = \{i \in I^n,\ T_i = 0\}, \tag{3}$$

whose respective cardinals are $n_1 \geq 1$ and $n_0 \geq 1$[1] .

Definition 2.2. *For $M \in \mathbb{N},\ M \geq 1$, the set of M-closest neighbors to individual $i \in I^n$ belonging to its counterfactual set is noted by $\mathcal{N}_i^{n,M}$, that is,*

$$\mathcal{N}_i^{n,M} = \left\{ X_j \text{ s.t } j \in I_{1-T_i}^n,\ \sum_{s \in I_{1-T_i}^n} \mathbb{1}_{\{\|X_s - X_i\| \leq \|X_j - X_i\|\}} \leq M \right\},$$

and the set of indexes for which $X_j,\ j \in I^n$, belongs to this set is

$$\mathcal{I}_i^{n,M} = \{j \in I^n,\ X_j \in \mathcal{N}_i^{n,M}\} \subseteq I_{1-T_i}^n.$$

Next condition will be assumed through the paper: for each $M \in \mathbb{N}$ and $i \in I^n$,

$$\#\mathcal{N}_i^{n,M} = \#\mathcal{I}_i^{n,M} = \min\{n, M\}. \tag{4}$$

In the remaining part of this section, we fix $M_0,\ M_1 \in \mathbb{N}$. They are the number of closest neighbors to be employed for estimate τ and τ_{tre}. We assume

$$1 \leq M_0 \leq n_0, \quad 1 \leq M_1 \leq n_1. \tag{5}$$

Under (4) and (5), for $i \in I^n$ consider arbitrary weights

$$\lambda_i^{n,M_{1-T_i}} = (\lambda_{ij}^{n,M_{1-T_i}}) \in \Delta_{M_{1-T_i}},$$

and let us define *generic weights* as

$$\mathcal{L}(n, M_0) = \{\lambda_i^{n,M_0} \in \Delta_{M_0},\ i \in I_1^n\}, \quad \mathcal{L}(n, M_1) = \{\lambda_i^{n,M_1} \in \Delta_{M_1},\ i \in I_0^n\}. \tag{6}$$

For $\delta \in \{0, 1\}$ and $i \in I_{1-\delta}^n$, by reordering the indexes in $\mathcal{I}_i^{n,M_\delta}$ in order to properly evaluate the summation below, we define the *generic outcome* based on (6) as

$$\widehat{Y}_{\delta i}(\lambda_i^{n,M_\delta}) = \sum_{j \in \mathcal{I}_i^{n,M_\delta}} \lambda_{ij}^{n,M_\delta}\ Y_j. \tag{7}$$

[1]The cardinal of a finite set A will be denoted as $\#A$, $\|\cdot\|$ is the Euclidean norm in the corresponding space and $\mathbb{1}_{\{\cdot\}}$ is the *indicator function*, that is, $\mathbb{1}_{\{\cdot\}} = 1$ if the argument is true and 0 otherwise.

It is important to note that weights (6) (and then the subsequent generic outcome (7)) may depend on individual characteristics. Also, the size and the number of them that are different from zero varies with $i \in I^n$. The only condition we assume is that weights should belong to the Simplex of the corresponding space.

Based on standard definitions for matching estimators of the *average treatment effect* and *average treatment effect on the treated*, the *generic matching estimators* we define below are simply straightforward extension of them, based on generic weights (6). For our purposes, the only relevance of this concept is that the asymptotic properties we present in Section 3 are valid for this type of estimators, being the bi-level matching estimators a particular case. Certainly the literature has provided us several alternatives to define weighting schemes for matching estimators (see Imbens & Wooldridge (2009)) which are particular cases of the generic estimators.

Definition 2.3. Generic matching estimator
Based on (6) and (7), the "generic matching estimator for the average treatment effect" is defined by

$$\widehat{\tau}(\mathcal{L}(n,M_0),\mathcal{L}(n,M_1)) = \frac{1}{n}\left[\sum_{i\in I_1^n}\left[Y_i - \widehat{Y}_{0i}(\lambda_i^{n,M_0})\right] + \sum_{i\in I_0^n}\left[\widehat{Y}_{1i}(\lambda_i^{n,M_1}) - Y_i\right]\right], \tag{8}$$

whereas the "generic matching estimator for the treatment effect on the treated" as

$$\widehat{\tau}_{tre}(\mathcal{L}(n,M_0)) = \frac{1}{n_1}\sum_{i\in I_1^n}\left[Y_i - \widehat{Y}_{0i}(\lambda_i^{n,M_0})\right]. \tag{9}$$

2.2 Bi-level matching estimators

In what follows, we define a type of generic matching estimator, where the weighting scheme is endogenously determined from the solution of a bi-level optimization problem (BLOP). For that purpose, let us consider a sample of size $n \in \mathbb{N},\ n \geq 2$, of Ω, namely $\Omega^n = \{(T_i, X_i, Y_i), i \in I^n\}$, and assume as given $M_0,\ M_1 \in \mathbb{N}$ complying with (5). Since the bias of any generic treatment effect estimator depends on the distance among characteristics of an individual and his closest neighbors on the counterfactual set (see Appendix section), as a first attempt to define weights for our approach we consider a *solution* of the optimization problem $\mathcal{P}_i^{n,M_{1-T_i}},\ i \in I^n$, defined by[2]

$$\mathcal{P}_i^{n,M_{1-T_i}}: \quad \min_{\lambda_i=(\lambda_{ij})\in\Delta_{M_{1-T_i}}} \|X_i - \sum_{j\in\mathcal{I}_i^{n,M_{1-T_i}}} \lambda_{ij}X_j\|^2. \tag{10}$$

[2]In the following, if necessary, we assume a reordering of the indexes belonging to $\mathcal{I}_i^{n,M_\delta},\ \delta \in \{0,1\}$, in order to perform adequately the summation.

Problem (10) is the *first level* we are interested to solve. However, the solution set of this first level, denoted *argmin* $[\mathcal{P}_i^{n,M_{1-T_i}}]$, may contain -usually does- several points. In order to avoid an indetermination for the value of the estimator based on weights from the solution set of (10), we consider a refinement criterion in order to determine the weights to be employed in the *bi-level matching estimator*. Based on a continuity argument, for the *second level* problem we propose to seek for a solution point that complementarily minimizes the sum of a *weighted* distance between X_i and X_j participating in the convex combination (i.e., $X_j,\ j \in \mathcal{I}_i^{n,M_{1-T_i}}$, for which $\lambda_{ij} > 0$ at the optimum). Formally, for $i \in I^n$ the BLOP is posed in the following form:

$$\min_{(\lambda_{ij})\ \in\ argmin\ \left[\mathcal{P}_i^{n,M_{1-T_i}}\right]} \sum_{j \in \mathcal{I}_i^{n,M_{1-T_i}}} \lambda_{ij}\ \|X_i - X_j\|. \tag{11}$$

Certainly introducing this optimization problem makes the complexity of any algorithm searching for a numerical solution of the BLOP increases in a combinatorial manner. However, we circumvent this difficulty by transforming the bi-level problem on a penalized and unrestricted optimization problem following Cominetti & San Martín (1994), and Cominetti (1999). For that purpose, for $i \in I^n$ and $\lambda = (\lambda_j) \in \mathbb{R}^{M_{1-T_i}}$, let us begin defining the auxiliary mappings $f_{is} : \mathbb{R}^{M_{1-T_i}} \to \mathbb{R},\ s = 1, \ldots, M_{1-T_i} + 2$, such that

$$f_{is}(\lambda) = -\lambda_s,\ 1 \leq s \leq M_{1-T_i},\quad f_{iM_{1-T_i}+1}(\lambda) = -f_{iM_{1-T_i}+2}(\lambda) = \sum_{k=1}^{M_{1-T_i}} \lambda_k - 1. \tag{12}$$

Based on (12), it is easy to check that

$$\Delta_{M_{1-T_i}} = \left\{\lambda \in \mathbb{R}^{M_{1-T_i}},\ f_{is}(\lambda) \leq 0,\ s = 1, \ldots, M_{1-T_i} + 2\right\}.$$

In order to internalize the fact that optimal weights solving the BLOP should belong to the Simplex, for $r > 0$ and $i \in I^n$, consider the following *exponential penalty function*

$$\psi_i : \mathbb{R}_{++} \times \mathbb{R}^{M_{1-T_i}} \to \mathbb{R} \mid \psi_i(r, \lambda) = \sum_{s=1}^{M_{1-T_i}+2} \exp\left(\frac{f_{is}(\lambda)}{r}\right). \tag{13}$$

Given that, the next unrestricted penalized optimization problem, namely $\mathcal{P}_i^{n,M_{1-T_i}}(r)$, is a key ingredient for empirical implementation

$$\min_{\lambda_i = (\lambda_{ij}) \in \mathbb{R}^{M_{1-T_i}}} \left\|X_i - \sum_{j \in \mathcal{I}_i^{n,M_{1-T_i}}} \lambda_{ij} X_j\right\|^2 + \sqrt{r} \sum_{j \in \mathcal{I}_i^{n,M_{1-T_i}}} \lambda_{ij} \|X_i - X_j\| + r\ \psi_i(r, \lambda_i). \tag{14}$$

For a correct definition of problem (14) the disclaimer regarding the reordering of indexes of $\mathcal{I}_i^{n,M_{1-T_i}}$ applies. Problem $\mathcal{P}_i^{n,M_{1-T_i}}(r)$ is a *hierarchical* nested version of the BLOP we are interested to solve (see Cominetti (1999)).

Proposition 2.1. *For $i \in I^n$ and $r > 0$, problem $\mathcal{P}_i^{n,M_{1-T_i}}(r)$ admits a unique solution that converges to a point belonging to argmin $[\mathcal{P}_i^{n,M_{1-T_i}}]$ when r approaches 0^+. Furthermore, this limit point solves the BLOP ((11)).*

Proof. From well known convexity properties, for each $r > 0$ it holds that $\mathcal{P}_i^{n,M_{1-T_i}}(r)$ admits a unique solution, namely $\lambda_i^{n,M_{1-T_i}}(r) \in \mathbb{R}^{M_{1-T_i}}$. From Theorems $3.3 - 3.4$ in Cominetti (1999), it holds that $\lambda_i^{n,M_{1-T_i}}(r)$ converges when r goes to 0^+, complying with

$$\lambda_i^{*n,M_{1-T_i}} \equiv \lim_{r \to 0^+} \lambda_i^{n,M_{1-T_i}}(r) \in argmin\ [\mathcal{P}_i^{n,M_{1-T_i}}] \subseteq \Delta_{M_{1-T_i}}. \tag{15}$$

Finally, the fact that $\lambda_i^{*n,M_{1-T_i}} \in \Delta_{M_{1-T_i}}$ solves the bi-level problem (i.e., along with (15), minimizes the middle term in (14)) comes directly from Theorem 5.2 in Cominetti (1999). □

Setting

$$\mathcal{L}^*(n, M_0) = \{\lambda_i^{*n,M_0} \in \Delta_{M_0},\ i \in I_1^n\}, \quad \mathcal{L}^*(n, M_1) = \{\lambda_i^{*n,M_1} \in \Delta_{M_1},\ i \in I_0^n\}, \tag{16}$$

definition below comes directly from Definition 2.3.

Definition 2.4. Bi-level matching estimators

Given $\Omega^n = \{(T_i, X_i, Y_i), i \in I^n\}$ a sample of size $n \in \mathbb{N},\ n \geq 2$, of Ω, and $M_0,\ M_1$ complying with (5), the "bi-level matching estimator" for the average treatment effect and for the average treatment effect on the treated are respectively defined as

$$\widehat{\tau}^*(n, M_0, M_1) \equiv \widehat{\tau}(\mathcal{L}^*(n, M_0), \mathcal{L}^*(n, M_1)) \tag{17}$$

$$\widehat{\tau}^*_{tre}(n, M_0) \equiv \widehat{\tau}_{tre}(\mathcal{L}^*(n, M_0)) \tag{18}$$

For numerical estimation of $\widehat{\tau}^*(n, M_0, M_1)$ and $\widehat{\tau}^*_{tre}(n, M_0)$ based on a certain sample it is necessary to solve problem (14) for each individual $i \in I^n$, which will undoubtedly add complexity to the method we propose compared, for instance, with the "simple nearest neighbor estimator" analyzed by Abadie & Imbens (2006). Moreover, since the exponential penalty term in (14) is "out of range" for r close enough to zero, we have that the standard methods of optimization do not necessarily apply for determining the limit solution of our problem. In spite of this, based on Cominetti & Dussault (1994) we can efficiently implement a numerical algorithm for evaluating the limit solution of problem (14), which attains super-linear convergence when the penalty parameter approaches to zero.

In section 4 we implement numerical exercises where we compare its performance in terms of bias, standard deviation, and mean square error with those from the simple nearest neighbor estimator. For this purpose, we solve the BLOP employing to whole sample, thus avoiding the choice of the number of nearest neighbors to be employed for the estimation: they will be determined endogenously as part of the solution of the optimization problem. indeed, they are those for which the optimal weights from the BLOP are strictly positive.

3 Limit properties of generic matching estimators

In this section we study the asymptotic properties of generic treatment effect estimators from Definition 2.3, which consequently will be valid for bi-level matching estimators.

Assumption $4-6$ below were employed by Abadie & Imbens (2006) with similar purposes than here.

Assumption 4. *For each $\delta \in \{0,1\}$, both $\mu(\cdot,\delta)$ and $\sigma^2(\cdot,\delta)$ are Lipschitz on $\mathbb{X}$.*

Assumption 5. *For each $\delta \in \{0,1\}$, the fourth-moment of Y_δ are uniformly bounded on $\mathbb{X}$, and $\sigma^2_\delta(\cdot,\cdot)$ is bounded away from zero.*

Assumption 6. *$\{\Omega^n\}_{n\in\mathbb{N}}$ are independent draws from the distribution of (T,X,Y).*

Given M_0 and M_1 satisfying (5), for any weighting scheme (6) we denote the *normalized conditional variance*, the *variance of the conditional mean* and the *conditional bias* of the generic matching estimator for the average treatment effect (8) as, respectively,

$$V^n = n \cdot \mathbb{V}\left[\widehat{\tau}(\mathcal{L}(n,M_0),\mathcal{L}(n,M_1)) \mid \Omega^n_Y\right], \quad V = \mathbb{E}[(\mu_1(X)-\mu_0(X)-\tau)^2],$$

$$B^n = \frac{1}{n}\left[\sum_{i\in I^n}\sum_{j\in\mathcal{I}_i^{n,M_{1-T_i}}} \lambda_{ij}^{n,M_{1-T_i}} \cdot [2T_i-1]\cdot[\mu_{1-T_i}(X_i)-\mu_{1-T_i}(X_j)]\right] \tag{19}$$

Proposition below is an straightforward extension of Theorems 3 and 4 in Abadie & Imbens (2006), where they assume fixed weights. The proof is given in the Appendix section. It is easy to show that, restricting the variance and bias to the treated sub-sample, results below hold.

Proposition 3.1. *Under Assumptions 1, 2, 3, 4 and 6 we have that*[3]

$$\widehat{\tau}(\mathcal{L}(n,M_0),\mathcal{L}(n,M_1)) \xrightarrow{\mathbb{P}} \tau.$$

Furthermore, if additionally Assumption 5 is satisfied, then

[3] In the following, for a sequence of random variables, convergence in probability will be noted by means of $\xrightarrow{\mathbb{P}}$, whereas convergence in distribution by $\xrightarrow{\mathbb{D}}$.

$$\frac{\sqrt{n}\cdot[\widehat{\tau}(\mathcal{L}(n,M_0),\mathcal{L}(n,M_1))-\tau-B^n]}{\sqrt{V^n+V}} \xrightarrow{\mathbb{D}} \mathcal{N}(0,1).$$

4 Finite sample properties of the bi-level matching estimators

In this section we implement Monte Carlo simulations to evaluate the small sample properties of the bi-level matching estimator for the average treatment effect on the treated (18). We compare its performance in terms of bias, standard deviation and mean square error to the simple nearest-neighbor treatment effect on the treated estimator studied by Abadie & Imbens (2006). The latter will be noted as NN estimator.

For the data generating process we follow Frölich (2004) and Busso et al. (2009). Thus, considering a sample of size $n \in \mathbb{N}$, for each $i \in I = \{1,\dots,n\}$ we define θ_i as a latent variable given by

$$\theta_i = \nu + \kappa\cdot\xi_i - u_i \in \mathbb{R},$$

where $\xi_i \sim \mathcal{N}(0,1)$, u_i is distributed standard Cauchy and $\nu,\ \kappa$ defining the *control-treated ratio*, i.e., the values of $n_0,\ n_1$. The observed treatment indicator is given by

$$T_i = \mathbb{1}_{\{\theta_i>0\}} \in \{0,1\},$$

and the outcome variable $Y_i \in \mathbb{R},\ i \in I$, is generated according to

$$Y_i = \beta\cdot T_i + m(X_i) + \epsilon_i \in \mathbb{R},$$

where $X_i = (x_{ij}) \in \mathbb{R}^k$ is a vector of *characteristics* such that $x_{ij} \sim \mathcal{N}(0,1),\ j = 1,\dots,k$ and $\epsilon_i \sim \mathcal{N}(0,0.1)$. This last assumption implies homocedasticity. Regarding $m : \mathbb{R}^k \to \mathbb{R}$ we evaluate for the following expressions

$$\begin{array}{lcl} \textit{Linear} & : & m(X_i) = \sum_{j=1}^k x_{ij} \\ \textit{Quadratic} & : & m(X_i) = \sum_{j=1}^k x_{ij} + \sum_{j=1}^k x_{ij}^2 \end{array}$$

Note that we are considering constant characteristic-specific treatment effects. This means that the individual treatment effect does not depend on the characteristics, implying that both the average treatment effect and the average treatment effect on the treated should be equal to β. For the exercise we report later on, we assume $\beta = 5$. Finally, we assume that $\xi_i = x_{i1}$.

From Abadie & Imbens (2008) we already know that standard resampling techniques fail to provide reliable estimators of the variance for the bi-level matching estimator for the average treatment effect on the treated. In order to treat this problem, we implement the resampling schemes developed by de Luna, Johansson & Sjöstedt-de Luna (2010), who show that a (circular) block bootstrap of estimated individual causal effects (EICE in their

notation) ordered respect to the matching covariates, overcomes the common difficulty of resampling under autocorrelated individual treatment effects. In this report the estimator of the variance is computed with a circular block bootstrap of size 10.

For the experiment we consider $n = 100$ and a control-treated ratio $1 : 1$, i.e, $n_0 = n_1 = 50$, and for the bi-level matching estimator for the average treatment effect on the treated we assume $M_0 = n_0$. The evaluation of the *NN* estimator considers diverse scenarios according with the number of nearest neighbors employed, ranging from $M = 1$ to $M = 16$.

The exercise is replicated for different numbers of characteristics, evaluating for $k = 1$, $k = 5$ and $k = 10$.

For the numerical solution of problem (14), we consider a stopping rule of $r = 10^{-7}$, which yields an adequate (indeed, a generous) level of precision for the results.

Simulations are based on $1,000$ replications, and for each one we consider $10,000$ bootstrap replications for determining the variance of the estimator. For each estimator, based on the results from the replications we report the bias (BIAS), the standard deviation (SD) and the mean-squared-error (MSE). The results are summarized in Table 1 below.

Table 1: *Simulation results for the bi-level and NN estimator of the treatment effect on the treated.* $n = 100$ *and treated:control=1:1. All characteristics are distributed* $\sim \mathcal{N}(0, 1)$. *Stopping criterion:* $r = 10^{-7}$

Length of vector	**Estimator**	$m(X)$ **is Linear**			$m(X)$ **is Quadratic**		
		BIAS	**SD**	**MSE**	**BIAS**	**SD**	**MSE**
$k = 1$	**Bi-level** $\widehat{\tau}^{*}_{tre}(100, 50)$	**0.042**	**0.042**	**0.008**	**0.188**	**0.161**	**0.132**
	NN with M=1	0.039	0.042	0.007	0.178	0.156	0.106
	NN with M=2	0.056	0.047	0.009	0.244	0.188	0.150
	NN with M=4	0.090	0.062	0.017	0.363	0.254	0.262
	NN with M=6	0.122	0.076	0.027	0.460	0.305	0.378
	NN with M=8	0.153	0.089	0.039	0.544	0.348	0.495
	NN with M=16	0.265	0.131	0.099	0.789	0.460	0.926
$k = 5$	**Bi-level** $\widehat{\tau}^{*}_{tre}(100, 50)$	**0.234**	**0.142**	**0.115**	**-0.648**	**0.597**	**1.460**
	NN with M=1	0.264	0.199	0.172	1.718	0.554	3.555
	NN with M=2	0.302	0.187	0.185	1.862	0.555	4.054
	NN with M=4	0.351	0.194	0.226	2.008	0.580	4.672
	NN with M=6	0.386	0.204	0.264	2.084	0.599	5.027
	NN with M=8	0.416	0.215	0.300	2.122	0.615	5.221
	NN with M=16	0.499	0.251	0.421	2.117	0.663	5.319
$k = 10$	**Bi-level** $\widehat{\tau}^{*}_{tre}(100, 50)$	**0.356**	**0.250**	**0.324**	**-0.714**	**0.842**	**2.600**
	NN with M=1	0.364	0.349	0.494	3.326	0.793	12.641
	NN with M=2	0.397	0.317	0.477	3.395	0.755	12.978
	NN with M=4	0.448	0.306	0.519	3.411	0.748	13.093
	NN with M=6	0.481	0.311	0.565	3.378	0.756	12.891
	NN with M=8	0.499	0.319	0.596	3.320	0.764	12.527
	NN with M=16	0.558	0.345	0.713	3.025	0.790	10.770

From the resulting values of the experiments regarding BIAS, SD and MSE we have that

the bi-level matching estimator for the average treatment effect on the treated outperforms the NN estimator in practically all designs, except for the *linear case* with $k = 1$ and $M = 1$, where all the indicators are practically the same for both estimators. Additionally, we note that independently of the number of characteristics, the gain in all indicators increases significantly as the number of used neighbors, M, increases. This gain is more significant for the quadratic than for the linear representation. Finally, the methodology we propose uses, in average, a number of neighbors no greater than the length of the vector of the characteristics plus one, that is, $k+1$. As shown in Table 2, based on the simulations results for $k = 1$ our method will not use more than two individuals to build the optimal convex combination (indeed, it uses 1.9 units on average). This fact is consistent with the well known Caratheodory's Theorem (see Rockafellar (1972)), since the number of vectors of any solution of problem $\mathcal{P}_i^n$ in (10) should be, at most, $k+1$.

Table 2: *Average of the number of individuals used by the bi-level matching estimator for the average treatment effect on the treated according with the value of the optimal weight* λ.

k	$\lambda > 10^{-5}$	$\lambda > 10^{-2}$
1	1.9	1.9
5	4.5	4.3
10	6.2	5.9

5 Conclusions

In this paper we propose a matching estimator in which the weighting scheme comes from the solution of the *bi-level optimization problem* (BLOP). This is an interesting feature from the applied work since, as Imbens and Woolridge (2009) points out, "little is known about the optimal number of matches, or about data-dependent ways of choosing it".

We provide a data-dependent way of choosing the number of matches and the weights of each match. These weights can be characterized as the limit of the solution of the penalized unrestricted optimization problem when the penalty parameter, r, approaches zero. By means of the weighting scheme arising from the BLOP, we define the *bi-level matching estimator* for the average treatment effect and for the treatment effect on the treated.

We perform Monte Carlo simulations in order to evaluate its performance in finite samples. We find that, for different scenarios used in the literature, the bi-level matching estimator for the treatment effect on the treated outperforms the corresponding simple matching estimator in terms of bias, standard deviation and mean square error.

Finally, we would like to mention that for this last exercise, the BLOP is solved considering the whole sample, thus avoiding the ex-ante choice of the number of nearest neighbors and weighting scheme. This advantage applies consequently for practical aspects regarding the estimation with real data.

References

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

Abadie, A. & Imbens, G. W. (2008), 'On the failure of the bootstrap for matching estimators', *Econometrica* **76**, 1537–1557.

Busso, M., DiNardo, J. & McCrary, J. (2009), 'Finite sample properties of semiparametric estimators of average treatment effects', *Manuscript, University of California-Berkeley*.

Colson, B., Marcotte, P. & Savard, G. (2007), 'An overview of bilevel optimization', *Annals of Operations Research* **153**, 235–256.

Cominetti, R. (1999), 'Nonlinear averages and convergence of penalty trajectories in convex programming', *Lectures Notes in Economics and Mathematical Systems* **477**, 65–78.

Cominetti, R. & Dussault, J.-P. (1994), 'A stable exponential-penalty algorithm with superlinear convergence', *Journal of Optimization Theory and Applications* **83**(2), 285–309.

Cominetti, R. & San Martín, J. (1994), 'Asymptotic analysis of the exponential penalty trajectory in linear programming', *Mathematical Programming* **67**, 169–187.

de Luna, X., Johansson, P. & Sjöstedt-de Luna, S. (2010), 'Bootstrap inference for k-nearest neighbour matching estimators', (5361).

Frölich, M. (2004), 'Finite-sample properties of propensity-score matching and weighting estimators', *Review of Economics and Statistics* **86**(1), 77–90.

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

Imbens, G. W. (2004), 'Nonparametric estimation of average treatment effects under exogenety: A review', *The Review of Economics and Statistics* **86**, 4–29.

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

Rockafellar, R.T. (1972), *Convex Analysis*. New Jersey, U.S.: Princeton University Press.

Rosenbaum, P. & Rubin, D. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrics* **70**, 41–55.

Appendix: proof of Proposition 3.1

The proof of Proposition 3.1 is based on Abadie & Imbens (2006) modifying and/or generalizing their arguments to get our results. We introduce some lemmata and auxiliary concepts which help us to simplify the expressions.

In the following, given $\Omega^n = \{(T_i, X_i, Y_i), i \in I^n\}$ a sample of size $n \in \mathbb{N},\ n \geq 2$, of Ω, denote $\Omega^n_Y = \{(T_i, X_i), i \in I^n\}$, and for fix $M_0,\ M_1$ complying with (5), from any generic weights $\mathcal{L}(n, M_0) = \{\lambda_i^{n,M_0} \in \Delta_{M_0},\ i \in I_1^n\},\ \ \mathcal{L}(n, M_1) = \{\lambda_i^{n,M_1} \in \Delta_{M_1},\ i \in I_0^n\}$, define

$$c_i^n = \{\lambda_{ji}^{n,M_{T_i}},\ j \in I_{1-T_i}^n\},$$

that is, *the set of weights for which individual $i \in I^n$ is used as one of the M_{1-T_i} nearest neighbors of any other individual on his counterfactual set.* Note that c_i^n could be an empty set. Finally let us define

$$\mathcal{C}_i^n = \begin{cases} 0 & \text{if } c_i^n = \emptyset, \\ \\ \displaystyle\sum_{j \in I_{1-T_i}^n} \lambda_{ji}^{n,M_{T_i}} & \text{if } c_i^n \neq \emptyset. \end{cases}$$

Lemma 5.1. *Under Assumptions 1, 2, 3 and 6, we have that $\mathcal{C}_i^n = O_p(1)$ and for each $\alpha > 0$, $\mathbb{E}[(\mathcal{C}_i^n)^\alpha]$ is uniformly bounded in $n \in \mathbb{N}$.*

Proof. Let $\mathcal{K}_i^n$ be the times that individual $i \in I^n$ is used as one of the M_{T_i} nearest neighbors for some individual in his counterfactual set. By definition, note that $\mathcal{C}_i^n \leq \mathcal{K}_i^n$, which implies that for each $\alpha > 0$, $\mathbb{E}[(\mathcal{C}_i^n)^\alpha] \leq \mathbb{E}[(\mathcal{K}_i^n)^\alpha]$. From Lemma 3 in Abadie & Imbens (2006) we already know that $\mathbb{E}[(\mathcal{K}_i^n)^\alpha]$ is uniformly bounded in $n \in \mathbb{N}$, which allows us to end the proof. □

Denoting $\widehat{\tau} = \widehat{\tau}(\mathcal{L}(n, M_0), \mathcal{L}(n, M_1))$ and following Abadie & Imbens (2006) we can readily show that

$$\widehat{\tau} - \tau = A + E + B, \tag{20}$$

where

$$A = \frac{1}{n} \cdot \sum_{i \in I^n} (\mu_1(X_i) - \mu_0(X_i)) - \tau, \tag{21}$$

$$E = \frac{1}{n} \cdot \sum_{i \in I^n} (2T_i - 1)(1 + \mathcal{C}_i^n)\epsilon_i, \tag{22}$$

with $\epsilon_i = Y_i - \mu_{T_i}(X_i)$ and $B = B^n$ from (19).

Lemma 5.2. *Under Assumptions 1, 2, 3, 4 and 6, follows that* $B = O_p(n^{-1/k})$.

Proof. Consider a sample $\Omega^n,\ n \in \mathbb{N}$, $i \in I^n$ and $j \in \mathcal{I}_i^{n,M_{1-T_i}}$. Under Assumption 1, from Lemma 2 in Abadie & Imbens (2006) holds for each $\alpha > 0$,

$$\mathbb{E}\left[n_{1-T_i}^{\alpha/k} \cdot \|X_i - X_j\|^\alpha\right]$$

is uniformly bounded in n_{1-T_i} (constant $\gamma > 0$, which applies for both, n_0 and n_1). On the other hand, from Assumption 4, there exists $L > 0$ (maximum of respective Lipschitz constants) such that

$$\|\mu_1(X_i) - \mu_1(X_j)\| \leq L\ \|X_i - X_j\|, \quad \|\mu_0(X_i) - \mu_0(X_j)\| \leq L\ \|X_i - X_j\|. \tag{23}$$

Denoting $B_{ij} = T_i\ [\mu_0(X_i) - \mu_0(X_j)] - (1 - T_i)\ [\mu_1(X_i) - \mu_1(X_j)]$, from (19) follows directly

$$\mathbb{E}[n^{2/k}\ B^2] = n^{2/k-2}\ \mathbb{E}\left[\left(\sum_{i \in I^n} \sum_{j \in \mathcal{I}_i^{n,M_{1-T_i}}} \lambda_{ij}^n\ B_{ij}\right)^2\right],$$

and then, applying Cauchy-Schwarz inequality and considering that

$$\sum_{j \in \mathcal{I}_i^{n,M_{1-T_i}}} (\lambda_{ij}^n)^2 \leq 1,$$

we can readily obtain the inequality below

$$\mathbb{E}[n^{2/k}\ B^2] \leq n^{2/k-1}\ \sum_{i \in I^n} \mathbb{E}\left[\sum_{j \in \mathcal{I}_i^{n,M_{1-T_i}}} {B_{ij}}^2\right].$$

Thus, expanding terms from last inequality and considering (23), we have that

$$\begin{aligned}
\mathbb{E}[n^{2/k}\ B^2] \ \leq\ & L^2\ n^{2/k-1}\ \mathbb{E}\left[\frac{1}{n_0^{2/k}} \sum_{i \in I_1^n} \sum_{j \in \mathcal{I}_i^{n,M_0}} \mathbb{E}\left[n_0^{2/k} \|X_i - X_j\|^2 | \{T_i\}_{i=1}^n, X_i\right]\right] \\
& + L^2\ n^{2/k-1} \cdot \mathbb{E}\left[\frac{1}{n_1^{2/k}} \sum_{i \in I_0^n} \sum_{j \in \mathcal{I}_i^{n,M_1}} \mathbb{E}\left[n_1^{2/k} \|X_i - X_j\|^2 | \{T_i\}_{i=1}^n, X_i\right]\right]
\end{aligned}$$

which leads us to conclude

$$\mathbb{E}[n^{2/k}\ B^2] \leq L_2 M_0 \cdot \mathbb{E}\left[\left(\frac{n}{n_0}\right)^{2/k} \frac{n_1}{n}\right] + L_2 M_1 \cdot \mathbb{E}\left[\left(\frac{n}{n_1}\right)^{2/k} \frac{n_0}{n}\right],$$

with $L_2 = \gamma \cdot L^2$. Thus, from the well known Chernoff's and Markov's inequalities we end the proof. □

Lemma 5.3. *Under Assumptions 1, 2, 3 and 6,*

$$\mathbb{E}\left[n\ \mathbb{V}[\widehat{\tau} \mid \Omega_Y^n]\right] = O(1).$$

Proof. From a direct calculation, we have that

$$\mathbb{E}[n\ \mathbb{V}[\widehat{\tau} \mid \Omega_Y^n]] = \frac{1}{n} \sum_{i \in I^n} \mathbb{E}\left[(1+\mathcal{C}_i^n)^2\ \sigma^2(X_i, T_i)\right] = \mathbb{E}\left[(1+\mathcal{C}_i^n)^2\ \sigma^2(x,t)\right].$$

From assumptions and properties of continuous mapping on compact sets, we have that there exists $\overline{\sigma}$ such that $\sigma(x,\delta) \leq \overline{\sigma}$ for all $(x,\delta) \in \mathbb{X} \times \{0,1\}$, which implies that

$$\mathbb{E}[n\ \mathbb{V}[\widehat{\tau} \mid \Omega_Y^n]] \leq \overline{\sigma}^2\ \mathbb{E}[(1+\mathcal{C}_i^n)^2].$$

Finally, from Lemma 5.1 we know that $\mathbb{E}[(1+\mathcal{C}_i^n)^2]$ is uniformly bounded in $n \in \mathbb{N}$, from which we conclude the proof. □

Proof. **Proof of Proposition 3.1**

From the *standard law of large numbers* follows immediately that

$$\left[\frac{1}{n}\ \left(\sum_{i \in I^n} (\mu_1(X_i) - \mu_0(X_i))\right) - \tau\right] \xrightarrow{\mathbb{P}} 0,$$

which along with Lemma 5.2 allows us to conclude

$$B = O_p(n^{1/k}) = o_p(1). \tag{24}$$

On the other hand, since

$$\mathbb{E}[n\ E^2] = \frac{1}{n} \sum_{i \in I^n} \mathbb{E}\left[(1+\mathcal{C}_i^n)^2\ \epsilon_i^2\right] = \mathbb{E}\left[(1+\mathcal{C}_i^n)^2\ \sigma^2(X_i, T_i)\right],$$

employing Lemma 5.1 we have that $\mathbb{E}[n\ E^2] = O(1)$, and then, from well known Markov's inequality,

$$E = O_p(n^{1/2}) = o_p(1). \tag{25}$$

From (24) and (25) the proof of consistency is done.

In order to show the normality property, from Lemma 5.3 follows that $V^n,\ n \in \mathbb{N}$, is bounded, and from Assumptions 1 and 4, it is so for V. On the other hand, from (20) we have

$$\sqrt{n}\ (\widehat{\tau} - \tau - B) = \sqrt{n}\ \ (A) + \sqrt{n}\ \ (E)\,,$$

and then, from the *Standard Central Limit Theorem* we can conclude

$$\sqrt{n}\ A \xrightarrow{\mathbb{D}} \mathcal{N}(0, V). \tag{26}$$

Finally, from the *Linderberg-Feller Central Limit Theorem* and Lemma 5.1 above, following the same argumentation provided by Abadie & Imbens (2006) to show their Theorem 4, we can deduce

$$\frac{\sqrt{n}\ E}{\sqrt{V^n}} \xrightarrow{\mathbb{D}} \mathcal{N}(0, 1). \tag{27}$$

We remark that the Linderberg-Feller Central Limit Theorem is valid conditional on Ω_Y^n, which is relevant for our case. Finally, considering that estimators in (26) and (27) are asymptotically independent, we end the proof.

□