FACULTAD ECONOMÍA Y NEGOCIOS
UNIVERSIDAD DE CHILE
DEPARTAMENTO DE ECONOMÍA

SERIE DE DOCUMENTOS DE TRABAJO

SDT 389

A geometric probability property and consequences on asymptotic properties of a matching estimator

Autores:
Roberto Cominetti
Juan Díaz
Jorge Rivera

Santiago, Septiembre de 2014

sdt@econ.uchile.cl
econ.uchile.cl/publicaciones

A geometric probability property and consequences on asymptotic properties of a matching estimator

Roberto Cominetti[*] Juan Díaz[†] Jorge Rivera[‡]

September, 2014

Working Paper. Not for quotation.

Abstract

This article concerns the limit properties of the "bi-level matching" estimator introduced by Díaz, Rau & Rivera (Forthcoming). Under usual conditions widely employed in the program evaluation literature, we show the conditional bias of that estimator is $O(N^{-2/k})$, with $k \in \mathbb{N}$ the dimension of covariates.

Keywords: Matching estimator, non-parametric methods, bi-level optimization, limits properties.

JEL Classification: C01, C14, C61.

1 Preliminaries

1.1 Notation and basic concepts

The binary program to be evaluated is represented by a random variable $\Omega = (W, Y, X)$, where $W \in \{0, 1\}$ indicates whether a treatment was received ($W = 1$) or not ($W = 0$) by the individual. The outcome for this treatment is $Y = W\,Y(1) + (1-W)\,Y(0) \in \mathbb{R}$, with $Y(1)$ and $Y(0)$ the potential outcomes –see Rosenbaum & Rubin (1983)–, and X is the vector of pretreatment variables or covariates, whose supporting set is $\mathbb{X} \subseteq \mathbb{R}^k$. The number $k \in \mathbb{N}$ is the dimension of covariates, which are assumed to be continuous variables.

This paper mainly concerns the limit properties of the population average treatment effect (ATE) of the program, denoted $\tau = \mathbb{E}(Y(1) - Y(0))$.

In the following, for $x \in \mathbb{X}$ and $w \in \{0, 1\}$, the conditional mean and conditional variance of Y is denoted by $\mu(x, w) = \mathbb{E}(Y \mid X = x, W = w)$ and $\sigma^2(x, w) = \mathbb{V}(Y \mid X = x,\, W = w)$, respectively. It is clear that $\tau = \mathbb{E}\left(\mu(x, 1) - \mu(x, 0)\right)$.

Given $\Omega_N = \{(W_i, Y_i, X_i),\, i = 1, \ldots, N\}$ a sample of size $N \in \mathbb{N}$ of Ω, we denote by N_0 and N_1 the number of control and treated units, respectively. We will assume that control units are indexed by $1, ..., N_0$, thus the treated ones are labeled by $N_0 + 1, ..., N_0 + N_1$ $(= N)$.

In this article we use the Euclidean norm, denoted $\|\cdot\|$, as the matching metric, which is not a restrictive condition for the purposes of this work. We also assume that the matching is performed with replacement, so each unit could be employed as match more than once.

[*] 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

By following Abadie & Imbens (2006), for $i \in \{1, \ldots, N\}$ and $m \in \mathbb{N}$, $1 \leq m \leq N_{1-W_i}$, we denote by

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

the index of the unit that is the mth nearest neighbour to unit i in terms of covariate values, among the units having the treatment opposite to that of unit i (namely, the counterfactual set of unit i). The corresponding matching discrepancy is denoted by

$$U_{m,i} = X_i - X_{j_m(i)}.$$

The convex hull of the subset[1] of covariates of the first m nearest neighbours to unit i is denoted as $\mathcal{K}_i(m) = \mathbf{co}\left\{X_{j_1(i)}, \ldots, X_{j_m(i)}\right\}$, while the convex hull of covariates of controls and treated units is denoted, respectively, by $\mathcal{K}_0 = \mathbf{co}\{X_1, \ldots, X_{N_0}\}$ and $\mathcal{K}_1 = \mathbf{co}\{X_{N_0+1}, \ldots, X_{N_0+N_1}\}$. These subsets will play a relevant role in this work.

1.2 The bi-level matching estimator

The *bi-level matching estimator* was introduced by Díaz et al. (Forthcoming). According to this approach, the vector of weights used to perform the potential outcome imputed to a treated unit $i \in \{N_0 + 1, \ldots, N\}$ solves the next optimization problem:

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

where *argmin*$\{\mathcal{F}_i\}$ is the solution set of the optimization problem

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

After configuring the problems above in terms of covariates to consider the case i is a control unit, the weighting scheme we are looking for is denoted by

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

Using weights (1), the potential outcome imputed to unit $i \in \{1, \ldots, N\}$ is

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

thus the bi-level matching estimator for the ATE is given by:

$$\widehat{\tau}^b \quad = \quad \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{Y}^b_i(1) - \widehat{Y}^b_i(0) \right). \tag{2}$$

[1] The convex hull of a subset $\{X_1, \ldots, X_n\}$ of $\mathbb{R}^k$, denoted $\mathbf{co}\{X_1, \ldots, X_n\}$, consists of all the vectors of the form $\lambda_1 X_1 + \ldots + \lambda_n X_n$, with $\lambda_1 \geq 0, \ldots, \lambda_n \geq 0$, $\lambda_1 + \ldots + \lambda_n = 1$. The set of such weights is called the *Simplex* of $\mathbb{R}^n$, hereafter denoted by Δ_n.

Finally, the next auxiliary concepts will be useful later. For $i \in \{1, \ldots, N\}$, the value of optimization problems $\mathcal{F}_i$ and $\mathcal{S}_i$ is given, respectively, by

$$\min\{\mathcal{F}_i\} \quad = \quad W_i \left\| X_i - \sum_{m=1}^{N_0} \lambda_m^i X_m \right\| + (1 - W_i) \left\| X_i - \sum_{m=1}^{N_1} \lambda_m^i X_{N_0+m} \right\|,$$

$$\min\{\mathcal{S}_i\} \quad = \quad W_i \sum_{m=1}^{N_0} \lambda_m^i \, \|X_i - X_m\|^2 + (1 - W_i) \sum_{m=1}^{N_1} \lambda_m^i \, \|X_i - X_{m+N_0}\|^2.$$

1.3 Standing assumptions and some direct consequences

The standing assumption we present below are quite standard in the program evaluation literature. We refer to Abadie & Imbens (2006), Heckman, Ichimura & Todd (1998), Imbens & Wooldridge (2009) and Rosenbaum & Rubin (1983) for a detailed discussion on these conditions.

Assumption 1. *$\mathbb{X}$ is compact and convex, with unitary Lebesgue measure in $\mathbb{R}^k$.*

Assumption 2. *The density of X is bounded away from zero, continuos on $\mathbb{X}$ and has bounded partial derivatives at each point of $\mathbb{X}$.*

Assumption 3. *$W \perp\!\!\!\perp ((Y(0), Y(1)) \mid X)$.*

Assumption 4. *There is $0 < c < 1$ such that $0 < \mathbb{P}(W = 1 \mid X) < (1 - c)$.*

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

Assumption 6. *For $w \in \{0, 1\}$, $\mu(\cdot, w)$ is twice continuously differentiable on $\mathbb{X}$.*

Some direct consequences of standing assumptions are presented by claims below. For that, we need to introduce the following conditional mean of Y, which will be useful later: for $x \in \mathbb{X}$ and $w \in \{0, 1\}$, we set $\mu_w(x) = \mathbb{E}(Y \mid X = x)$.

Claim 1.1. *N_0 and N_1 tend to infinity (with probability one) when N goes to infinity.*

This comes directly from Assumptions **3**, **4** and **5**.

Claim 1.2. *For each $x \in \mathbb{X}$ and $w \in \{0, 1\}$, $\mu_w(x) = \mu(x, w)$.*

This result is a straightforward consequence of Assumptions **3** and **4** –see Abadie & Imbens (2006)–.

Claim 1.3. *For i a treated unit, there are constants L_1 and L_2 such that*

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

Using Assumptions **1** and **6**, and Claim 1.2, this inequality can be readily obtained after some calculus involving the second order Taylor expansion of μ_0. There, constants L_1 and L_2 are the upper bounds, over the supporting set, of the first and second derivatives of that mapping. Properly configured in terms of covariates and the conditional bias mapping, it is clear that a similar inequality than above holds for the case i is a control unit, this using the same constants as stated.

Claim 1.4. *For integer α, there is a constant $c_\alpha > 0$ such that for N large enough and $m \leq N_{1-W_i}$,*

$$\mathbb{E}\left(N_{1-W_i}^{\alpha/k}\,\|U_{m,i}\|^\alpha\right) \quad \leq \quad c_\alpha\, m^\alpha. \tag{3}$$

This property is one of the key results we need to prove our main contributions, which is a straightforward consequence of Theorem 5.4 in Evans, Jones & Schmidt (2002).

Theorem 1.1. Evans et al. (2002)

Under Assumptions **1** *–* **5***, for $i \in \{1,\ldots,N\}$, $\alpha \in \mathbb{N}$ and $0 < \rho < 1/k$, there is constant $c(\alpha,k) \in \mathbb{R}_+$ such that for all N_{1-W_i} large enough and $m \leq N_{1-W_i}$,*

$$\mathbb{E}\left(N_{1-W_i}^{\alpha/k}\,\|U_{m,i}\|^\alpha\right) \quad = \quad c(\alpha,k)\,\frac{\Gamma(m+\alpha/k)}{\Gamma(m)} + O\left(\frac{1}{N_{1-W_i}^{1/k-\rho}}\right).$$

From relations (5.36) and (5.44) in Evans et al. (2002), we can appreciate that the order expression in last equation does not depend on m, which implies it can be bounded above by some constant. Given that, inequality (3) comes from the fact that $\Gamma(m+\alpha/k)/\Gamma(m) \leq (1+\alpha)^\alpha\, m^\alpha$.

2 The probability of the convex hull of nearest neighbours

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

$$\mathbf{Pr}(X_i \notin \mathcal{K}_i(M)) \quad = \quad \mathbf{Pr}(0_k \notin \mathbf{co}\,\{U_{1,i},\ldots,U_{M,i}\}).$$

With the aim as stated, 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 pag. 218), we have this property holds under the case these vectors are *“selected independently according to a distribution absolutely continuous with respect to natural Lebesgue measure”*. Given that, in view of our standing assumptions, it is clear that the subset of covariates $\{X_1,\ldots,X_N\}$ is in general position.[2] Moreover, 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*. Next result is straightforward (the proof is omitted).

Lemma 2.1. *Under Assumptions* **1** *–* **5***, for N large enough, $i \in \{1,\ldots,N\}$ and $M > k$, we have $\{U_{1,i},\ldots,U_{M,i}\}$ is in general position.*

A remarkable result in Wendel (1962), slightly extended by Cover & Efron (1967), states 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,[3] then the probability of existing a half-space containing that set of vectors is given by

$$C(M,k) \quad = \quad \frac{1}{2^{M-1}}\sum_{s=0}^{k-1}\binom{M-1}{s}. \tag{4}$$

[2] In fact, only Assumptions **1**, **2**, **3** and **5** are needed to obtain this property.

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

By developing the summation in (4), and properly bounding the binomial coefficient in last relation, it is easy to show that there is a constant $\theta > 0$ such that $C(M,k) \leq \theta\,\frac{M^k}{2^M}$. Now, taking into account the convex hull of $\{\xi_1,\ldots,\xi_M\}$ is the intersection of all half-spaces containing that vectors, using last approximation we can readily conclude the following inequality:

$$\mathbf{Pr}\left(0_k \notin \mathbf{co}\{\xi_1,\ldots,\xi_M\}\right) \quad \leq \quad \theta\,\frac{M^k}{2^M}. \tag{5}$$

Of course this last result cannot be directly applied to approximate $\mathbf{Pr}(X_i \notin \mathcal{K}_i(M))$: in spite of $\{U_{1,i},\ldots,U_{M,i}\}$ is in general position –Lemma 2.1–, the joint distribution of these vectors could be far from being invariant under reflections through the origin. However, next technical result helps us to overcome this drawback (see the proof in §Appendix).

Proposition 2.1. *If Assumptions* **1** *and* **2** *hold, 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.*

Finally, combining Proposition 2.1, Lemma 2.1 and inequality in (5), we can readily conclude the following property.

Theorem 2.1. *Under Assumptions* **1** – **5**, *for N large enough, $i \in \{1,\ldots,N\}$ and $M > k$, there is constant $\gamma > 0$ such that*

$$\mathbf{Pr}(X_i \notin \mathcal{K}_i(M)) \quad \leq \quad \gamma\,\frac{M^k}{2^M}.$$

3 Large sample properties

Following Abadie & Imbens (2006), and performing some simple calculus, it can be readily shown the conditional bias of the bi-level matching estimator, denoted B_N^b, is given by the following expression:

$$B_N^b \quad = \quad \frac{1}{N}\left(\sum_{i=1}^{N_0}\sum_{m=1}^{N_1}\lambda_m^i(\mu_1(X_{m+N_0}) - \mu_1(X_i)) + \sum_{i=1+N_0}^{N_1+N_0}\sum_{m=1}^{N_0}\lambda_m^i(\mu_0(X_i) - \mu_0(X_m))\right). \tag{6}$$

Next property is one the main results of this paper. The proof is given in §Appendix.

Proposition 3.1. *If Assumptions* **1** – **6** *hold, then*

$$\mathbb{E}\left(\left|\sum_{m=1}^{N_0}\lambda_m^i(\mu_0(X_i) - \mu_0(X_m))\right|\middle|W_i = 1, X_i, \{W_j, X_j\}_{j=1}^N\right) = O\left(N_0^{-2/k}\right). \tag{7}$$

and

$$\mathbb{E}\left(\left|\sum_{m=1}^{N_1}\lambda_m^i(\mu_1(X_{m+N_0}) - \mu_1(X_i))\right|\middle|W_i = 0, X_i, \{W_j, X_j\}_{j=1}^N\right) = O\left(N_1^{-2/k}\right). \tag{8}$$

Using Proposition 3.1, next result is straightforward.

Theorem 3.1. *If Assumptions* **1** – **6** *hold, then*

$$B_N^b \quad = \quad O_{\mathbf{p}}(N^{-2/k}).$$

Proof. After developing characterization in (6), we have

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

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

From Proposition 3.1, there is a constant ψ such that

$$\mathbb{E}\left(N^{2/k}|B_N^b|\right) \leq \psi\,\mathbb{E}\left(\frac{N^{2/k}}{N}\left(\frac{N_0}{N_1^{2/k}}+\frac{N_1}{N_0^{2/k}}\right)\right)=\psi\,\mathbb{E}\left(\left(\frac{N}{N_1}\right)^{2/k}\left(\frac{N_0}{N}\right)+\left(\frac{N}{N_0}\right)^{2/k}\left(\frac{N_1}{N}\right)\right).$$

Using the well known Chernoff's and Markov's inequalities in last relation, we can readily conclude the proof. □

4 Appendix

4.1 Proof of Proposition 2.1

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

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

where $f(\cdot)$ is the density function of covariates. Denoting $F(x)=\mathbf{Pr}(||X-z||\leq||x-z||)$, 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, this last 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:

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

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

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

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

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

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

$$f_{j_1,\dots,j_M}(v) = \frac{N!N^{-M}}{(N-M)!}\left(\prod_{s=1}^{M} f\left(z+\frac{v_{j_s}}{N^{1/k}}\right)\right)\left(1-\mathbf{Pr}\left(||X-z|| \leq \frac{||v_{j_M}||}{N^{1/k}}\right)\right)^{N-M}, \tag{9}$$

from which we can readily conclude the following inequality[4]

$$f_{j_1,\dots,j_M}(v) \quad \leq \quad \bar{f}^M exp\left(-\underline{f}\,\frac{||v_{j_M}||^k}{(M+1)}\,\frac{\pi^{k/2}}{\Gamma(1+k/2)}\right), \tag{10}$$

where $0 < \underline{f} < \bar{f} < \infty$ are the lower and upper bounds of the distribution $f(\cdot)$, respectively. Using the right term in (10) we can define the distribution as stated. □

Remark 4.1. *Using (9) 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:*

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

4.2 Proof of Proposition 3.1

Proof. Without loss of generality, the proof is done for relation in (7). For a treated unit $i \in \{N_0+1,\dots,N_0+N_1\}$ and $M \leq N_0$, we set

$$\mathbf{q}_i(M) = \mathbf{Pr}(X_i \notin \mathcal{K}_i(M)), \quad \mathbf{p}_i(M) = 1-\mathbf{q}_i(M) = \mathbf{Pr}(X_i \in \mathcal{K}_i(M)),$$

and the conditionals in (7) is denoted as $\theta_i = \{W_i = 1, X_i, \{W_j, X_j\}_{j=1}^N\}$.

The proof begins with the study of the stochastic order of the following part of the approximation in Claim 1.3:

$$\psi_i = \mathbb{E}\left(L_1 \min\{\mathcal{F}_i\} + L_2 \min\{\mathcal{S}_i\}\,\middle|\,\theta_i\right) = \mathbb{E}\left(\psi_i\,\middle|\,X_i \notin \mathcal{K}_0\right)\mathbf{q}_i(N_0) + \mathbb{E}\left(\psi_i\,\middle|\,X_i \in \mathcal{K}_0\right)\mathbf{p}_i(N_0).$$

Denoting by $\delta > 0$ the diameter of $\mathbb{X}$, Theorem 2.1 implies

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

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

this leading to the following result:

$$\mathbb{E}\left(N_0^{2/k}\,\psi_i\,\middle|\,X_i \notin \mathcal{K}_0\right)\,\mathbf{q}_i(N_0) = o(1). \tag{11}$$

Because $\min\{\mathcal{F}_i\} = 0$ when $X_i \in \mathcal{K}_0$, it follows that

$$\mathbb{E}\left(\psi_i\,\middle|\,X_i \in \mathcal{K}_0\right) \quad = \quad \mathbb{E}\left(L_2\,\min\{\mathcal{S}_i\}\,\middle|\,\theta_i,\, X_i \in \mathcal{K}_0\right),$$

and then, decomposing the event $X_i \in \mathcal{K}_0$ into the disjoint events[5]

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

each having probability $\mathbf{p}_i(m)\,\mathbf{q}_i(m-1)\,(\leq \mathbf{q}_i(m-1))$, we can conclude

$$\mathbb{E}\left(\psi_i\,\middle|\,X_i \in \mathcal{K}_0\right)\,\mathbf{p}_i(N_0) \quad \leq \quad \sum_{m=2}^{N_0}\mathbb{E}\left(L_2\,\min\{\mathcal{S}_i\}\,\middle|\,\theta_i,\, X_i \in \Delta\mathcal{K}_i(m)\right)\,\mathbf{q}_i(m-1). \tag{12}$$

For $m \leq N_0$, it is clear the condition $X_i \in \Delta\mathcal{K}_i(m)$ implies $\min\{\mathcal{S}_i\} \leq \|U_{m,i}\|^2$. On the other hand, when $m > k$, Theorem 2.1 implies $\mathbf{q}_i(m-1) \leq \gamma\,\frac{m^k}{2^m}$. Using all of these facts in (12) gives the following inequality:

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

Applying Claim 1.4 to last inequality implies that –there constant c_2 comes from (3)–

$$\mathbb{E}\left(\psi_i\,\middle|\,X_i \in \mathcal{K}_0\right)\,\mathbf{p}_i(N_0) \quad \leq \quad \frac{c_2\,L_2}{N_0^{2/k}}\left(\sum_{m=2}^{k} m^2 + 2\,\gamma\sum_{m=k+1}^{N_0}\frac{m^{2+k}}{2^m}\right). \tag{13}$$

Finally, because both summations in the right side of (13) are uniformly bounded, there is a constant $C > 0$ such that

$$\mathbb{E}\left(\psi_i\,\middle|\,X_i \in \mathcal{K}_0\right)\,\mathbf{p}_i(N_0) \quad \leq \quad \frac{C}{N_0^{2/k}},$$

which along with (11) lead us to conclude that

$$\mathbb{E}\left(L_1\,\min\{\mathcal{F}_i\} + L_2\,\min\{\mathcal{S}_i\}\,\middle|\,W_i = 1, X_i, \{W_j, X_j\}_{j=1}^{N}\right) \quad = \quad O\left(N_0^{-2/k}\right).$$

Using similar arguments as before, it is straightforward to show the order term of the approximation in Claim 1.3 is also $O\left(N_0^{-2/k}\right)$, which ends the proof. □

[5] In the following, the set-difference between A and B is denoted by $A \setminus B = \{c \in A,\, c \notin B\}$.

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.

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.

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