Biometrika (XXXX), **XXX**, X, *p*. 1
Printed in Great Britain Advance Access publication on XXXX

On Rosenbaum's Rank-based Matching Estimator

By Matias D. Cattaneo

Department of Operations Research and Financial Engineering, Princeton University, Princeton, New Jersey 08544, U.S.A.

cattaneo@princeton.edu

Fang Han

Department of Statistics, University of Washington, Seattle, Washington 98195, U.S.A.

fanghan@uw.edu

and Zhexiao Lin

Department of Statistics, University of California, Berkeley, Berkeley, California 94720, U.S.A.

zhexiaolin@berkeley.edu

Summary

In two influential contributions, Rosenbaum (2005, 2020a) advocated for using the distances between component-wise ranks, instead of the original data values, to measure covariate similarity when constructing matching estimators of average treatment effects. While the intuitive benefits of using covariate ranks for matching estimation are apparent, there is no theoretical understanding of such procedures in the literature. We fill this gap by demonstrating that Rosenbaum's rank-based matching estimator, when coupled with a regression adjustment, enjoys the properties of double robustness and semiparametric efficiency without the need to enforce restrictive covariate moment assumptions. Our theoretical findings further emphasize the statistical virtues of employing ranks for estimation and inference, more broadly aligning with the insights put forth by Peter Bickel in his 2004 Rietz lecture (Bickel, 2004).

Some key words: Average treatment effect; Matching estimator; Rank-based statistic; Regression adjustment; Semiparametric efficiency.

1. Introduction

Consider the problem of estimating the average treatment effect (ATE),

$$\tau \equiv E(Y(1) - Y(0)),$$

based on an observational study encompassing n observations of a binary treatment $D \in \{0, 1\}$, some measured pre-treatment covariates $X \in \mathbb{R}^d$, and an outcome $Y \equiv Y(D) \in \mathbb{R}$ that is realized from the two potential outcomes $(Y(0), Y(1))$. Among the techniques employed to estimate τ, nearest neighbor (NN) matching stands as one of the most widely adopted and comprehensible approaches; see Stuart (2010) and references therein. These estimators aim to impute the missing potential outcome of each unit in one treatment group by finding units from the opposite treatment group whose covariate profile closely resembles that of the unit with the missing potential

outcome. The quantification of covariate similarity relies on the user-specified *distance metric* between the (distributions of the) covariates of different units (from different treatment groups).

Abadie & Imbens (2006) laid out the mathematical groundwork to study NN matching estimators employing the Euclidean distance metric, while Abadie & Imbens (2011) established a root-n central limit theorem for a bias-corrected version of those NN matching estimators. More recently, Lin et al. (2023) established connections between NN estimators and augmented inverse probability weighted (AIPW) methods (Robins et al., 1994; Scharfstein et al., 1999), thereby establishing double robustness and semiparametric efficiency theory for NN matching estimators when the number of matches diverges to infinity with the sample size.

The aforementioned theoretical work, however, centers around the Euclidean distance metric for determining NN matches. This approach may exhibit sensitivity to alterations in scale and to the presence of extreme outliers or heavy-tailed distributions. Indeed, all the existing theoretical results on matching assume covariates with compact support, which is theoretically hard to alleviate, if not impossible. On the other hand, in practice, distance metrics are often derived from a "standardized" representation of the data, and the selection of a distance metric is an important factor in causal inference because various metrics can lead to different conclusions (Rosenbaum, 2020a, Chapter 9).

This paper focuses on a particular standardization approach that identifies NNs by measuring the Euclidean distance between the component-wise ranks of the covariates X, as proposed for the celebrated Rosenbaum's rank-based matching estimator (Rosenbaum, 2020a, Chapter 9.3). The concept of rank-based standardization is straightforward to interpret, easy to implement, and computationally efficient, while also being scale-invariant and insensitive to heavy-tailed distributions. Furthermore, due to their data adaptivity, rank-based methods are often used in treatment effect settings such as for analysis of experiments (Rosenbaum, 2020a), regression discontinuity plots (Calonico et al., 2015), and binscatter regressions (Cattaneo et al., 2024).

The challenge in formally studying rank-based methods lies on the theoretical side, as the transformation of the covariates into their ranks disrupts the independence structure of the original data, and thus complicates the subsequent statistical analysis. Our main theorem, Theorem 1, offers the *first* theoretical analysis of Rosenbaum's rank-based matching estimator, elucidating its appealing properties when combined with regression adjustments. In particular, our theory not only confirms Rosenbaum's intuition that the rank-based distance can limit the influence of outliers and heavy-tailed distributions (Rosenbaum, 2020a, page 210), but also demonstrates that the rank-based matching estimator can be doubly robust and semiparametrically efficient, particularly without imposing restrictive moment assumptions on the distribution of the covariates. More broadly, our results align with Peter Bickel's 2004 Rietz lecture advocating for "standardization by ranks" when performing statistical and machine learning related tasks (Bickel, 2004).

Our paper also offers two technical contributions, which may be of independent interest. First, Theorem 1 establishes consistency, asymptotic linearity, and semiparametric efficiency of Rosenbaum's Rank-based Matching Estimator under generic high-level conditions on the regression adjustment. The proof of that theorem relies on a careful combination of empirical process theory for rank-based statistics and tools established in Lin & Han (2022) and Lin et al. (2023) for matching estimators involving a growing number of nearest neighbors, which are generalized herein to accommodate standardization/transformation functions, of which component-wise ranking is one particular example. Second, Theorem 2 presents novel mean square and uniform convergence rates for series estimators when the covariates are generated via possibly unknown functions of the original independent variables, of which component-wise ranking is one particular example. Those results are proven for general series estimators (Newey, 1997; Belloni et al., 2015) with covariate-generated conditioning variables, and thus lead to suboptimal uniform approximation

results, but we also discuss how they may be upgraded to deliver optimal uniform convergence rates for partition-based series estimators (Huang, 2003; Cattaneo & Farrell, 2013; Cattaneo et al., 2020, 2024).

2. Setup

We adopt the standard potential outcomes causal model for a binary treatment, where it is assumed that there are n independent and identically (i.i.d.) distributed realizations $\{X_i, D_i, Y_i(0), Y_i(1)\}_{i=1}^n$, of a quadruple $(X, D, Y(0), Y(1))$. In practice, we are only able to observe a part of the data, i.e., $\{X_i, D_i, Y_i \equiv Y_i(D_i)\}_{i=1}^n$. The goal is to conduct estimation and inference for the population ATE,

$$\tau = E(Y(1) - Y(0)),$$

based only on the observed data.

Rosenbaum's rank-based matching approach estimates τ by plugging the component-wise ranks of X_i's, instead of the original values, into the NN matching mechanism, with each unit matched to M units in the opposite treatment group with replacement. It proceeds in three steps as follows.

Step 1. Given a sample $\{(X_i, D_i, Y_i)\}_{i=1}^n$ with $X_i = (X_{i,1}, \ldots, X_{i,d})^\top$, introduce the vector of component-wise (scaled) ranks such that for any $i \in \{1, \ldots, n\}$,

$$\widehat{U}_i \equiv (\widehat{U}_{i,1}, \ldots, \widehat{U}_{i,d})^\top, \text{ with } \widehat{U}_{i,k} \equiv \frac{1}{n}\sum_{j=1}^n \mathbb{1}(X_{j,k} \le X_{i,k}),\ k \in \{1, \ldots, d\}.$$

Here $\mathbb{1}(\cdot)$ stands for the indicator function. The vector of marginal population CDFs is $F(\cdot) : \mathbb{R}^d \to [0,1]^d$, such that for any input $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$F(x) \equiv (F_1(x_1), \ldots, F_d(x_d))^\top, \text{ with } F_k(x_k) \equiv \mathrm{pr}(X_{1,k} \le x_k),\ k \in \{1, \ldots, d\}.$$

Define $U \equiv F(X) \in [0,1]^d$ and for each $i \in \{1, \ldots, n\}$, $U_i \equiv F(X_i) \in [0,1]^d$.

Step 2. Employ regression adjustment to correct for the estimation bias from matching. Let $\widehat{\mu}_0(\cdot)$ and $\widehat{\mu}_1(\cdot)$ be mappings from $[0,1]^d$ to $\mathbb{R}$ such that they separately estimate the conditional means of the outcomes,

$$\mu_0(u) \equiv E(Y \mid U = u, D = 0) \qquad \text{and} \qquad \mu_1(u) \equiv E(Y \mid U = u, D = 1).$$

We obtain $\widehat{\mu}_0$ and $\widehat{\mu}_1$ by regressing Y_i's in either the treatment or control group on the corresponding $\{\widehat{U}_i\}$'s, respectively.

Step 3. Implement bias-corrected nearest neighbor matching on $(\widehat{U}_i, D_i, Y_i)$'s. Specifically, let $\mathcal{J}(i)$ represent the index set of the M-NNs of $\widehat{U}_i$ in $\{\widehat{U}_j : D_j = 1 - D_i\}_{j=1}^n$, measured using the Euclidean metric $\|\cdot\|$ with *ties broken in arbitrary way*. The resulting rank-based matching estimator is

$$\widehat{\tau} \equiv \frac{1}{n}\sum_{i=1}^n \big(\widehat{Y}_i(1) - \widehat{Y}_i(0)\big), \tag{1}$$

where, for $\omega \in \{0, 1\}$,

$$\widehat{Y}_i(\omega) \equiv \begin{cases} Y_i, & \text{if } D_i = \omega, \\ \frac{1}{M}\sum_{j \in \mathcal{J}(i)} \big(Y_j + \widehat{\mu}_\omega(\widehat{U}_i) - \widehat{\mu}_\omega(\widehat{U}_j)\big), & \text{if } D_i = 1 - \omega \end{cases}.$$

3. MAIN RESULT

Following the notation system of Abadie & Imbens (2006, 2011), let $K(i)$ represent the number of matched times for each unit i, i.e.,

$$K(i) \equiv \sum_{j=1, D_j=1-D_i}^{n} \mathbb{1}\big(i \in \mathcal{J}(j)\big).$$

In this paper, however, $K(i)$ denotes the matched times according to *the rank-based distance*, not the original Euclidean distance. Following Lin & Han (2022) and Lin et al. (2023), the rank-based bias-corrected matching estimator in (1) can be represented as an AIPW estimator:

$$\widehat{\tau} = \widehat{\tau}^{\text{reg}} + \frac{1}{n}\sum_{i=1}^{n}(2D_i - 1)\Big(1 + \frac{K(i)}{M}\Big)\widehat{R}_i, \tag{2}$$

where

$$\widehat{\tau}^{\text{reg}} \equiv \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{\mu}_1(\widehat{U}_i) - \widehat{\mu}_0(\widehat{U}_i)\big) \text{ and } \widehat{R}_i \equiv Y_i - \widehat{\mu}_{D_i}(\widehat{U}_i), \text{ for } i \in \{1, \dots, n\}.$$

We employ this insight throughout our large sample distributional analysis of $\widehat{\tau}$.

To establish our main theorem we impose some assumptions. The first assumption is posed to regulate basic features of the data generating distribution, in particular making the estimation problem identifiable. Conceptually, the main difference with prior literature is that the assumption concerns the triple (U_i, D_i, Y_i)'s, that is, X_i is replaced by U_i, the scaled population rank.

Assumption 1. (i) For almost all $u \in [0,1]^d$, D is independent of $(Y(0), Y(1))$ conditional on $U = u$, and there exists a fixed constant $c > 0$ such that $c < e(u) \equiv \text{pr}(D = 1 \mid U = u) < 1 - c$.

(ii) $[(X_i, D_i, Y_i)]_{i=1}^n$ are i.i.d. following the joint distribution of (X, D, Y).

(iii) $E\{(Y(\omega) - \mu_\omega(U))^2 \mid U = u\}$ is uniformly bounded for almost all $u \in [0,1]^d$ and $\omega = 0, 1$.

(iv) $E(\mu_\omega^2(U))$ is bounded for $\omega = 0, 1$.

The second assumption ensures that the regression adjustment procedure is, at least, well-posited in the sense that the estimator $\widehat{\mu}_\omega(x)$ is uniformly consistent for some well-behaved (possibly misspecified) conditional expectation. Let $\|\cdot\|_\infty$ be the L^∞ function norm.

Assumption 2. For $\omega = 0, 1$, there exists a deterministic, possibly changing with n, continuous function $\bar{\mu}_\omega(\cdot) : [0,1]^d \to \mathbb{R}$ such that $E(\bar{\mu}_\omega^2(U))$ is uniformly bounded and the estimator $\widehat{\mu}_\omega(x)$ satisfies $\|\widehat{\mu}_\omega - \bar{\mu}_\omega\|_\infty = o_{\text{P}}(1)$.

The next assumption regulates the population rank-transformed random vector U, which identifies the copula distribution for X (Joe, 2014). This assumption requires X to be continuous, but discrete components of X can be easily handled by conditioning (Stuart, 2010).

Assumption 3. The Lebesgue density of U exists and is continuous over its support.

The next three assumptions concern the case of consistent population regression functions $\mu_0(U)$ and $\mu_1(U)$, and will be used in contrast to Assumption 2, where the regression adjustment procedure is allowed to be inconsistent for the population ranked-based regression functions. More precisely, Assumption 2 vis-á-vis Assumptions 4–6 are used for establishing the double robustness and semiparametric efficiency of $\widehat{\tau}$, respectively. Using standard multi-index notation, let Λ_k be the set of all d-dimensional vectors of nonnegative integers $t = (t_1, \dots, t_d)$ such

that $|t| = \sum_{i=1}^{d} t_i = k$ with k any positive integer, and $\partial^t \mu_\omega$ denotes the corresponding partial derivative of μ_ω.

Assumption 4. For $\omega = 0, 1$, μ_ω is continuous and the estimator $\widehat{\mu}_\omega(x)$ satisfies $\|\widehat{\mu}_\omega - \mu_\omega\|_\infty = o_{\mathrm{P}}(1)$.

Assumption 5. (i) $E\{(Y(\omega) - \mu_\omega(U))^2 \mid U = u\}$ is uniformly bounded away from zero for almost all $u \in [0,1]^d$ and $\omega = 0, 1$.

(ii) There exists a constant $c > 0$ such that $E(|Y(\omega) - \mu_\omega(U)|^{2+c} \mid U = u)$ is uniformly bounded for almost all $u \in [0,1]^d$ and $\omega = 0, 1$.

(iii) $\max_{t \in \Lambda_{\max\{\lfloor d/2 \rfloor, 1\}+1}} \|\partial^t \mu_\omega\|_\infty$ is bounded, where $\lfloor \cdot \rfloor$ stands for the floor function.

Assumption 6. For $\omega = 0, 1$, the estimator $\widehat{\mu}_\omega(x)$ satisfies

$$\max_{t \in \Lambda_{\max\{\lfloor d/2 \rfloor, 1\}+1}} \|\partial^t \widehat{\mu}_\omega\|_\infty = O_{\mathrm{P}}(1)$$

and

$$\max_{t \in \Lambda_\ell} \|\partial^t \widehat{\mu}_\omega - \partial^t \mu_\omega\|_\infty = O_{\mathrm{P}}(n^{-\gamma_\ell}) \text{ for all } \ell \in \{1, \ldots, \max\{\lfloor d/2 \rfloor, 1\}\},$$

with some constants γ_ℓ's satisfying $\gamma_\ell > \max\{1/2 - \ell/d, 0\}$ for $\ell = 1, 2, \ldots, \max\{\lfloor d/2 \rfloor, 1\}$.

Several remarks on the above assumptions align with our conceptual discussion. First, as the quantile transformation preserves all information, Assumption 1 is either equivalent to, or weaker than, the standard assumptions in the matching estimation literature. Second, Assumption 2 accommodates regression model misspecification, and its validity may be verified by leveraging the fundamental projection principles underlying regression techniques (see, also, the discussions in Section 4). Third, Assumption 3 constitutes a mild condition, notably satisfied by distribution families such as the Gaussian (copula) and Cauchy (copula). Lastly, Assumptions 4 through 6 merit more discussion: due to the shift from using X_i's as inputs to $\widehat{U}_i$'s in the regression function, direct verification using standard results from the nonparametric smoothing estimation literature is no longer possible. We return to this technical issue in Section 4 by considering explicitly least squares series estimation (Newey, 1997; Huang, 2003; Cattaneo & Farrell, 2013; Belloni et al., 2015; Cattaneo et al., 2020) to illustrate verification of Assumption 2 and Assumptions 4–6.

We are now ready to present our main theorem for Rosenbaum's rank-based matching estimator.

Theorem 1 (Main Theorem). *(i) (Double robustness of $\widehat{\tau}$) If either Assumptions 1, 2, 3, $M \log n/n \to 0$, and $M \to \infty$ as $n \to \infty$ hold, or Assumptions 1 and 4 hold, then*

$$\widehat{\tau} - \tau \text{ converges in probability to } 0.$$

(ii) (Semiparametric efficiency of $\widehat{\tau}$) Assume Assumptions 1, 3, 5, 6 hold. Define

$$\gamma = \max\left\{\left(1 - \frac{1}{2}\frac{d}{\max\{\lfloor d/2 \rfloor, 1\} + 1}\right), \min_{\ell \in \{1, \ldots, \max\{\lfloor d/2 \rfloor, 1\}\}} \left\{1 - \left(\frac{1}{2} - \gamma_\ell\right)\frac{d}{\ell}\right\}\right\},$$

where the $(\gamma_\ell : \ell = 1, 2, \ldots, \max\{\lfloor d/2 \rfloor, 1\})$ are introduced in Assumption 6. If $M \to \infty$ and $M/n^\gamma \to 0$ as $n \to \infty$, then

$$n^{1/2}(\widehat{\tau} - \tau) \text{ converges in distribution to } N(0, \sigma^2),$$

where

$$\sigma^2 \equiv E\left\{\mu_1(U) - \mu_0(U) + \frac{D(Y - \mu_1(U))}{e(U)} - \frac{(1-D)(Y - \mu_0(U))}{1 - e(U)} - \tau\right\}^2.$$

(iii) (Variance Estimation) If the conditions in part (ii) and Assumption 4 hold, then

$$\widehat{\sigma}^2 - \sigma^2 \text{ converges in probability to } 0$$

where

$$\widehat{\sigma}^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left\{\widehat{\mu}_1(\widehat{U}_i) - \widehat{\mu}_0(\widehat{U}_i) + (2D_i - 1)\Big(1 + \frac{K_M(i)}{M}\Big)\widehat{R}_i - \widehat{\tau}\right\}^2.$$

This theorem establishes three main results. Part (i) shows that the generic bias-corrected Rosenbaum's rank-based matching estimator is doubly robust for a fairly large class of regression estimators based on estimated ranks of the covariates. The result in part (ii) gives general regularity conditions guaranteeing asymptotic normality of the estimator. It follows directly from the second result that the estimator is semiparametrically efficient for estimating the ATE (Hahn, 1998). Finally, part (iii) establishes consistency of the plug-in variance estimator under nearly the same conditions as required for consistency and asymptotic normality.

4. Regression Adjustment using Series Least Squares

The only remaining issue concerning Theorem 1 revolves around the use of pairs $(\widehat{U}_i, Y_i)$ for bias correction, as opposed to (X_i, Y_i), or the idealized oracle (U_i, Y_i) pairs. The dependence among the estimated rank-adjusted $\widehat{U}_1, \ldots, \widehat{U}_n$ poses a challenge, making it hard to apply existing results in nonparametric statistics for the direct verification of Assumption 2 or Assumptions 4–6. This section illustrates how these assumptions can be verified when using series least squares regression estimation, covering both canonical approximating functions (e.g., power series, fourier series, splines, wavelets, and piecewise polynomials) as well as general covariate transformations (e.g., high-dimensional least squares regression with structured regressors).

The main result in this section concerns general estimated transformations of the independent variables, based on the underlying regressors only, and allowing for possible misspecification in both fixed-dimension and increasing-dimension least squares regression settings. Thus, we consider a more general setup where we either observe or have approximate information about n i.i.d. pairs $(Y_1, W_1), \ldots, (Y_n, W_n)$ of (Y, W), and the goal is to estimate the conditional expectation

$$\psi(w) \equiv E(Y \mid W = w),$$

using only the outcome variables $\{Y_i\}_{i=1}^n$ and the *generated covariates* $\{\widehat{W}_i\}_{i=1}^n$ that are "approximately close" to $\{W_i\}_{i=1}^n$, where $\widehat{W}_i$'s are measurable with respect to some sigma field $\mathcal{F}_n$, $n \geq 1$. In practice, as it is the case in the rank-based transformation we consider in this paper, a useful choice is $\mathcal{F}_n = \mathcal{S}(W_1, W_2, \ldots, W_n)$, where $\mathcal{S}(Z)$ denotes the sigma field generated by the random variable Z.

Let $p_K(w) = (p_{1K}(w), \ldots, p_{KK}(w))^\top$ be a K-dimensional vector of basis functions so that their linear combination may approximate $\psi(\cdot)$ well when K is sufficiently large, at least under some specific assumptions. However, a good approximation is not strictly required, as we also consider misspecified regression adjustments. This is important in the context of this paper because Theorem 1 established a double-robust property of the regression adjusted Rosenbaum's rank-based matching estimator, thereby allowing for misspecified or inconsistent estimators $(\widehat{\mu}_0, \widehat{\mu}_1)$ of the regression functions (μ_0, μ_1). Furthermore, we also allow for the possibility of W having a Lebesgue density that may not be bounded away from zero, as it occurs when its support is unbounded. In our application, due to the rank transformation, the support of the rank-based covariates is bounded but their density may not be bounded away from zero in, e.g., the Gaussian copula case.

The following assumption summarizes our setup. Let $\lambda_{\min}(\cdot)$ be the smallest eigenvalue of the input matrix.

Assumption 7. (i) $(Y_1, W_1), \ldots, (Y_n, W_n)$ are i.i.d. draws from $(Y, W) \in \mathcal{Y} \times \mathcal{W} \subseteq \mathbb{R} \times \mathbb{R}^d$.
(ii) $E(Y^2)$ is bounded, and $\lambda_{\min}\big(E(p_K(W)p_K(W)^\top)\big) > 0$ for all K and n.
(iii) There exists a sequence of sigma fields $\{\mathcal{F}_n\}$ such that $(\widehat{W}_1, \widehat{W}_2, \ldots, \widehat{W}_n)$ is measurable with respect to $\mathcal{F}_n$ for all n, $\sup_{i \le n} E\{(Y_i - \psi(W_i))^2 | \mathcal{F}_n\}$ is bounded uniformly in n, and $E\{(Y_i - \psi(W_i))(Y_j - \psi(W_j))|\mathcal{F}_n\} = 0$ for all $i \neq j$ and n.

The series estimator with generated covariates is

$$\widehat{\psi}_K(w) = p_K(w)^\top \widehat{\beta}_K, \qquad \widehat{\beta}_K \in \operatorname*{argmin}_{b \in \mathbb{R}^K} \frac{1}{n} \sum_{i=1}^n (Y_i - p_K(\widehat{W}_i)^\top b)^2,$$

which gives $\widehat{\beta}_K \equiv (\vec{P}_n^\top \vec{P}_n)^- \vec{P}_n^\top \vec{Y}$, where $\vec{P}_n = (p_K(\widehat{W}_1), \ldots, p_K(\widehat{W}_n))^\top \in \mathbb{R}^{n \times k}$, $\vec{Y} \equiv (Y_1, \ldots, Y_n)^\top$, and $\vec{A}^-$ denotes a generalized inverse of the matrix $\vec{A}$. Let $\vec{P} \equiv (p_K(W_1), \ldots, p_K(W_n))^\top$ be what $\vec{P}_n$ shall approximate, and

$$\beta_K \equiv \operatorname*{argmin}_{b \in \mathbb{R}^K} E\{(Y_1 - p_K(W_1)^\top b)^2\} = \operatorname*{argmin}_{b \in \mathbb{R}^K} E\{(\psi(W_1) - p_K(W_1)^\top b)^2\}$$

be what $\widehat{\beta}_K$ shall approximate. It follows that $\psi_K(w) \equiv p_K(w)^\top \beta_K$ is the best L^2 approximation of $\psi(w)$ based on $p_K(w)$, where $\beta_K = \vec{Q}^- E(p_K(W_1)\psi(W_1))$ with $\vec{Q} \equiv E(p_K(W_1)p_K(W_1)^\top) \in \mathbb{R}^{K \times K}$.

Our results rely on the following quantities characterizing different aspects of the series estimator and the approximation errors:

$$\lambda_K \equiv \lambda_{\min}(\vec{Q}), \qquad \zeta_{q,K} \equiv \max_{t \in \Lambda_q} \sup_{w \in \mathcal{W}} \|\partial^t p_K(w)\|,$$
$$\xi_K^2 \equiv E\{(\psi(W_1) - \psi_K(W_1))^2\}, \qquad \vartheta_{q,K} \equiv \max_{t \in \Lambda_q} \|\partial^t \psi - \partial^t \psi_K\|_\infty,$$

and

$$R_n \equiv \|\vec{\Psi} - \vec{\Psi}_n\|^2/n, \qquad B_n \equiv \|(\vec{P} - \vec{P}_n)\vec{Q}^{-1/2}\|_2^2/n,$$

where $\vec{\Psi} \equiv (\psi(W_1), \ldots, \psi(W_n))^\top \in \mathbb{R}^n$, $\vec{\Psi}_n \equiv (\psi(\widehat{W}_1), \ldots, \psi(\widehat{W}_n))^\top \in \mathbb{R}^n$, and $\|\cdot\|_2$ denotes the matrix spectral norm.

Let $\|g\|_{L^2}^2 \equiv \int |g(w)|^2 \mathrm{d}F_W(w)$, and consider first the L^2 rate of approximation of the series-based least squares estimator:

$$\|\widehat{\psi}_K - \psi\|_{L^2}^2 \le 2\|\widehat{\psi}_K - \psi_K\|_{L^2}^2 + 2\|\psi_K - \psi\|_{L^2}^2,$$

where it follows immediately that $\|\psi_K - \psi\|_{L^2}^2 = \xi_K^2 \le \vartheta_{0,K}^2$, implying that the best mean square approximation $\psi_K(w) = p_K(w)^\top \beta_K$ will approximate $\psi(w)$ well if $\vartheta_{0,K} \to 0$, or at least $\xi_K \to 0$, which in turn requires $K \to \infty$ in general. However, in many applications the series estimator may be misspecified or inconsistent in the sense that $\xi_K \not\to 0$. In those cases, it is natural to take $\psi_K(w)$ as the target "parameter". The following theorem establishes two distinct L^2 convergence rates for the series estimator relative to the latter quantity.

Theorem 2 (L_2 Convergence). *Let Assumption 7 hold, $\lambda_K^{-1}\zeta_{0,K}^2 \log(K)/n = o(1)$, and $B_n = o_{\mathrm{P}}(1)$. Then,*

$$\|\widehat{\psi}_K - \psi_K\|_{L^2}^2 = O_{\mathrm{P}}\Big(\frac{K}{n} + \mathsf{A}_n\Big),$$

where the approximation error term can be taken to be either

$$\mathsf{A}_n = \min\left\{B_n + \xi_K^2\,,\; R_n + \vartheta_{0,K}^2\right\},$$

or

$$\mathsf{A}_n = B_n + B_n \min\left\{B_n + \xi_K^2\,,\; R_n + \vartheta_{0,K}^2\right\} + \min\left\{\lambda_K^{-1}\zeta_{0,K}^2\xi_K^2/n\,,\; K\vartheta_{0,K}^2/n\right\}.$$

This theorem provides new results relative to previously known mean square convergence rates for series estimation. More specifically, it allows for generated regressors based on covariates with a possibly vanishing minimum eigenvalue of the expected scaled Gram matrix (λ_K), as it may occur when the Lebesgue density of W is positive but not bounded away from zero on $\mathcal{W}$. Furthermore, the second rate estimate allows for a non-vanishing L^2 approximation error ($\vartheta_{0,K} \geq \xi_K \not\to 0$), thereby offering L^2 consistency results for general least squares approximations.

It is easy to deduce (suboptimal) uniform rates of approximation using Theorem 2 because

$$\begin{aligned}\max_{t\in\Lambda_q}\|\partial^t\widehat{\psi}_K - \partial^t\psi\|_\infty &\leq \max_{t\in\Lambda_q}\|\partial^t p_K^\top(\widehat{\beta}_K - \beta_K)\|_\infty + \max_{t\in\Lambda_q}\|\partial^t\psi_K - \partial^t\psi\|_\infty,\\ &\leq \zeta_{q,K}\lambda_K^{-1/2}\|\widehat{\psi}_K - \psi_K\|_{L^2} + \vartheta_{q,K},\end{aligned}$$

where, as noted before, the first term characterizes the error in approximation when perhaps $\vartheta_{q,K} \not\to 0$, in which case the target "parameter" can be taken to be $\partial^t\psi_K(w) = \partial^t p_K(w)^\top\beta_K$, regardless of whether $K \to \infty$ or not.

Underlying the assumptions imposed in Theorem 2, there are several parameters that need further discussion. From the standard series estimation literature, $\zeta_{q,K} = O(K^{1+q})$ for power series and $\zeta_{q,K} = O(K^{1/2+q})$ for fourier series, splines, compact supported wavelets, and piecewise polynomial regression. Lower bounds for λ_K need to be established on a case-by-case basis when the density of W is not assumed to be bounded away from zero, so we illustrate one such verification further below for the case of rank transformations and a Gaussian copula. As already mentioned, the parameter $\xi_K \leq \vartheta_{0,K}$ captures the degree of approximation (or misspecification) of the series regression estimator, and needs not to vanish, in which case the second rate result in Theorem 2, and the implied uniform convergence rate, must be used. If ψ is s-times differentiable and other regularity conditions hold, then $\xi_K = O(K^{-s/d})$ for all the usual approximation basis functions if appropriately specified. In general, however, the difference between the L^2 and L^∞ approximaton errors, ξ_K and $\vartheta_{0,K}$, depends on the basis functions employed and the data generating features. In particular, for instance, when employing locally supported basis functions, it can be verified that $\xi_K \asymp \vartheta_{0,K}$, in which case $\vartheta_{q,K} = O(K^{-(s-q)/d})$ under regularity conditions. See Newey (1997), Huang (2003), Cattaneo & Farrell (2013), Belloni et al. (2015), Cattaneo et al. (2020), and references therein, for more details.

An important feature of Theorem 2 is that it allows for generated regressors based on the covariates, which introduces two additional quantities characterizing the approximation rate: R_n and B_n. For example, if ψ satisfies the Lipschitz condition $|\psi(a) - \psi(b)| \leq L\|a - b\|$ for some constant L, then

$$R_n = \frac{1}{n}\sum_{i=1}^n\left(\psi(\widehat{W}_i) - \psi(W_i)\right)^2 \leq L^2 \max_{i=1,2,\ldots,n}\|\widehat{W}_i - W_i\|^2 = O_{\mathrm{P}}(r_n),$$

and therefore the convergence rate of R_n is determined by the uniform convergence rate of the transformation of the covariates. Recall that in our application, we consider the empirical rank transformation $(W_i, \widehat{W}_i) = (U_i, \widehat{U}_i)$, and therefore $r_n = 1/n$. A similar calculation can be done to

bound B_n when $p_K(\cdot)$ is smooth, because

$$B_n \leq \lambda_K^{-1} \frac{1}{n} \sum_{i=1}^{n} \|p_K(\widehat{W}_i) - p_K(W_i)\|^2 \leq \lambda_K^{-1} L_K^2 \max_{i=1,2,\ldots,n} \|\widehat{W}_i - W_i\|^2 = O_{\mathrm{P}}(b_n),$$

provided that the basis functions are Lipschitz with constant L_K, and where $L_K^2 = O(\zeta_{1,K}^2)$. Thus, in the particular case of the empirical rank transformation, $b_n = \lambda_K^{-1}\zeta_{1,K}^2/n$.

It remains to illustrate how to lower bound the minimum eigenvalue λ_K. It is well-known that if W admits a Lebesgue density bounded and bounded away from zero over the support $\mathcal{W}$, then λ_K^{-1} is uniformly bounded in K, after possibly rotating the basis functions. The following lemma considers the more interesting case when the density is not bounded away from zero, as it occurs for example when the support of W is unbounded.

Lemma 1 (Lower Bound on λ_K). *Suppose that W admits a Lebesgue density f_W, and $p_{1K}, \ldots, p_{KK}$ are orthonormal with respect to the Lebesgue measure over the support of W. If there exists a universal constant $C > 0$ such that, for all sufficiently small $t > 0$, the Lebesgue measure $\mathsf{Leb}(\{w : 0 < f_W(w) < t\}) \leq Ct^\rho$ for some $\rho > 0$, then $\lambda_K^{-1} = O(\zeta_{0,K}^{2/\rho})$.*

As expected, the additional conditions in the previous lemma restrict the tail of f_W. It remains to illustrate how to verify the result for the case when W_i and $\widehat{W}_i$ are taken to be U_i and $\widehat{U}_i$ as in Section 2. This is done in the following proposition for the case of a Gaussian copula.

Proposition 1 (Sufficient Conditions for Gaussian Copula). *Suppose $U_1, U_2, \ldots, U_n$ follow a Gaussian copula distribution with an invertible parameter correlation matrix Σ. Then, the conditions of Lemma 1 are satisfied with $\rho = \lambda_{\max}^{-1}(\Sigma^{-1} - I_d)/d$, where I_d is the d-dimensional identity matrix and $\lambda_{\max}(\cdot)$ is the largest eigenvalue of the input matrix..*

Using Theorem 2 in general, or Lemma 1 and specific conditions on the underlying copula of the population rank-based transformed covariates, it is possible to verify the conditions of Theorem 1. To be more specific, Theorem 2 gives versatile (albeit suboptimal) uniform consistency rates for series estimators $(\widehat{\mu}_0, \widehat{\mu}_1)$ of either some approximate (misspecified) functions or the population functions (μ_0, μ_1). It is worth noting that our verification aimed for generality in terms of high-level conditions, but for the case of partition-based (locally supported) series estimators it is possible to obtain better (in fact optimal in some cases) uniform convergence rates (Cattaneo & Farrell, 2013; Belloni et al., 2015; Cattaneo et al., 2020). Specifically, Cattaneo et al. (2024) studies the case of rank-based transformations for B-splines when $d = 1$, and establishes optimal uniform convergence rates on compact support. Their results could be extended to obtain sharper uniform convergence rates with generated regressors based on the covariates as required by Theorem 1.

5. Discussion

Rank-based distance is extensively used in the matching literature due to its robustness against various types of contaminations, and because of its computational simplicity. These benefits have been discussed in Rosenbaum (2020a, Chapter 9.3) and Rosenbaum (2020b, Sections 4.5 and 4.6). Furthermore, rank-based matching methods have been applied in numerous studies, including Kang et al. (2013), Keele et al. (2015), Kang et al. (2016), Keele & Morgan (2016), and Yu et al. (2020), to just name a few. Our paper complements that literature by providing a theoretical exploration of rank-based matching, and thereby offering insights into its underlying foundational principles.

More specifically, this paper studied the large sample properties of Rosenbaum's rank-based matching estimator with regression adjustment, and established its consistency, double robustness, asymptotic normality, and semiparametric efficiency. Consistency of a plug-in variance estimator was also established. These results were obtained as a consequence of a more general theorem given in the supplemental appendix, which allows for a class of transformations of the covariates, a leading special case being the empirical rank transformation proposed by Rosenbaum (2005, 2020a). To provide primitive conditions for regression adjustment, novel convergence rates for series estimators with generated regressors and possibly covariate Lebesgue density not bounded away from zero were derived, which may be of independent interest.

ACKNOWLEDGEMENT

We thank Peng Ding, Yingjie Feng, and Boris Shigida for insightful comments. Cattaneo gratefully acknowledge financial support from the National Science Foundation through grants DMS-2210561 and SES-2241575. Han gratefully acknowledge financial support from the National Science Foundation through grants SES-2019363 and DMS-2210019.

REFERENCES

ABADIE, A. & IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74**, 235–267.

ABADIE, A. & IMBENS, G. W. (2011). Bias-corrected matching estimators for average treatment effects. *Journal of Business and Economic Statistics* **29**, 1–11.

BELLONI, A., CHERNOZHUKOV, V., CHETVERIKOV, D. & KATO, K. (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics* **186**, 345–366.

BICKEL, P. J. (2004). Rietz lecture: The frontiers of statistics and computer science. Personal communication.

CALONICO, S., CATTANEO, M. D. & TITIUNIK, R. (2015). Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association* **110**, 1753–1769.

CATTANEO, M. D., CRUMP, R. K., FARRELL, M. H. & FENG, Y. (2024). On binscatter. *American Economic Review* **114**, 1488–1514.

CATTANEO, M. D. & FARRELL, M. H. (2013). Optimal convergence rates, bahadur representation, and asymptotic normality of partitioning estimators. *Journal of Econometrics* **174**, 127–143.

CATTANEO, M. D., FARRELL, M. H. & FENG, Y. (2020). Large sample properties of partitioning-based series estimators. *Annals of Statistics* **48**, 1718–1741.

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–331.

HUANG, J. (2003). Local asymptotics for polynomial spline regression. *Annals of Statistics* **31**, 1600–1635.

JOE, H. (2014). *Dependence Modeling with Copulas*. CRC press.

KANG, H., KREUELS, B., ADJEI, O., KRUMKAMP, R., MAY, J. & SMALL, D. S. (2013). The causal effect of malaria on stunting: a mendelian randomization and matching approach. *International Journal of Epidemiology* **42**, 1390–1398.

KANG, H., KREUELS, B., MAY, J. & SMALL, D. S. (2016). Full matching approach to instrumental variables estimation with application to the effect of malaria on stunting. *Annals of Applied Statistics* **10**, 335–364.

KEELE, L. & MORGAN, J. W. (2016). How strong is strong enough? Strengthening instruments through matching and weak instrument tests. *Annals of Applied Statistics* **10**, 1086–1106.

KEELE, L., TITIUNIK, R. & ZUBIZARRETA, J. R. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical Society Series A: Statistics in Society* **178**, 223–239.

LIN, Z., DING, P. & HAN, F. (2023). Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *Econometrica* **91**, 2187–2217.

LIN, Z. & HAN, F. (2022). On regression-adjusted imputation estimators of the average treatment effect. *arXiv preprint arXiv:2212.05424* .

NEWEY, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics* **79**, 147–168.

ROBINS, J. M., ROTNITZKY, A. & ZHAO, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89**, 846–866.

ROSENBAUM, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**, 515–530.

ROSENBAUM, P. R. (2020a). *Design of Observational Studies*. Springer, 2nd ed.

ROSENBAUM, P. R. (2020b). Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application* **7**, 143–176.

SCHARFSTEIN, D. O., ROTNITZKY, A. & ROBINS, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94**, 1096–1120.

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science* **25**, 1–21.

YU, R., SILBER, J. H. & ROSENBAUM, P. R. (2020). Matching methods for observational studies derived from large administrative databases. *Statistical Science* **35**, 338–355.

[*Received on* X *XX* XXXX. *Editorial decision on* X *XX* XXXX]