The Bias of Matching in the Normal Case[*]

Connor T. Jerzak[†] Nikos Vlassis[‡]

November 21, 2018

Abstract

We derive expressions for the bias of the matching estimator for causal effects when the treated and control units have Gaussian covariates. Our main result shows that the bias grows in magnitude at a polynomial rate with the distance between the treated and control means, which confirms in an analytical way the difficulty of causal estimation under lack of overlap between the two groups. The specific degree of the polynomial, which can be quadratic or cubic, depends on the relationship between the covariance imbalance and the covariate-outcome correlation. We conclude by discussing the implications of these results for applied researchers.

[*] Our thanks to Gary King, Matthew Blackwell, Xiang Zhou, and researchers at Adobe for extremely helpful comments.

[†] PhD Candidate and Carl J. Friedrich Fellow, 1737 Cambridge Street, Harvard University, Cambridge MA, ConnorJerzak.com, cjerzak@g.harvard.edu.

[‡] Senior Data Scientist, Netflix, Los Gatos, CA, https://sites.google.com/site/nikosvlassis/, nikos.vlassis@gmail.com.

1 Introduction

On average, how close are our causal estimates to the truth? This question is of decisive importance to those in the social, natural, or biomedical sciences, who are in the business of understanding the causal linkages which underpine the world. In this paper, we consider this question when restricting our inquiry to matching methods, which are particularly important for making causal inferences because they are model-free and therefore avoid making assumptions about how input covariates are related to the outcome (Ho, Imai, King, and Stuart, 2007). Moreover, important qualitative research protocols (such as the most similar systems design) are explicitly or implicitly motivated with reference to nearest neighbor matching (see King, Keohane, and Verba (1994) or Lijphart (1971), respectively). Nearest neighbor matching is thus a cornerstone of causal analyses in both qualitative and quantitative research.

We also restrict our analysis to the situation when the covariates are Gaussian. This assumption is strong and implies that the variables important for predicting the outcome are well-behaved. Nevertheless, the Normal distribution is arguably the most important one in the statistical sciences, partly because the Central Limit Theorem affirms that sums are asymptotically Gaussian in distribution. To the extent that real-world social or biomedical processes often involve the average contribution of many small effects, Gaussianity is often assumed in statistical or causal analyses.

Our results indicate that, in the Gaussian case, the bias of matching grows at a polynomial rate with the distance between the treated and control group covariate centers. The exact rate depends on the specific relationship between the covariate imbalance and the gradient of the conditional expectation function (the rate can range from nearly quadratic to cubic). We also provide additional results which strengthen previous analyses of the matching bias . To conclude, we illustrate the implications of our results for applied researchers.

2 Notation & Assumptions

Our work builds on the Rubin-Neyman causal model, which was introduced in an early version by Jerzy Neyman in the 1920s and which was more formally developed by Donald Rubin in the 1970s (see Rubin (1974) and Splawa-Neyman, Dabrowska, and Speed (1990)). In this model, each unit has two potential outcomes, only one of which is realized: formally, $Y_i^{\text{Obs}} = W_iY_i(1)+(1-W_i)Y_i(0)$, where W_i denotes the treatment indicator which is 1 if unit i is in the treated group and 0 if unit i is in the control group.

Investigators summarize the effect of the treatment on the outcome by comparing $Y_i(1)$ and $Y_i(0)$. The Average Treatment Effect (ATE) is a popular estimand, and is defined as $\tau_{\text{ATE}} = \mathbb{E}\left[Y_i(1) - Y_i(0)\right]$. The Average Treatment Effect on the Treated (ATET) is also popular in cases when there is a large pool of control units and a smaller supply of treated units. It is defined as $\tau_{\text{ATET}} = \mathbb{E}\left[Y_i(1) - Y_i(0)|W_i = 1\right]$. Because $Y_i(0)$ or $Y_i(1)$ is unobserved for i, each unit's missing value must be imputed. For example, when investigators estimate τ_{ATE} by taking the difference in means between the two groups, they impute the missing values of $Y_i(w)$ with the mean value in treatment class w. That is, if n_0 denotes the number of control units, then for observations having $W_i = 0$, they estimate $\hat{Y}_i(0)$ with the observed value: $\hat{Y}_i(0) = Y_i$. For observations having $W_i = 1$, they estimate $\hat{Y}_i(0)$ using $\hat{Y}_i(0) = \frac{1}{n_0}\sum_{j:W_j=0} Y_j$. $\hat{Y}_i(1)$ is defined analogously and $\hat{\tau}_{\text{ATE, Baseline}} = \frac{1}{n}\sum_{i=1}^{n}[\hat{Y}_i(1) - \hat{Y}_i(0)]$.

Extensions to the Rubin-Neyman causal model have proliferated in recent years. For example, Belloni, Chernozhukov, Fernandez-Val, and Hansen (2017) and Athey, Imbens, and Wager (2016) examine analytically and experimentally the use of machine learning methods for estimating causal relations in the context of many control variables. In addition, Athey and Imbens (2016) use tree-based methods for estimating heterogeneous treatment effects. There has also been concerted effort to improve matching algorithms, which have historically been a method of choice because they are simple to use and model-free (Stuart, 2010). In this paper, we focus on this class of estimators. These estimators impute $Y_i(1)$ for control units (for which $W_i = 0$) and impute $Y_i(0)$ for treated units (for which $W_i = 1$) based on the outcome values of the neighboring m units in Euclidean space. Formally,

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0 \\ \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } W_i = 1 \end{cases}$$

and

$$\hat{Y}_i(1) = \begin{cases} \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if } W_i = 0 \\ Y_i & \text{if } W_i = 1 \end{cases}$$

where $\mathcal{J}_M(i)$ denotes the set of indices for the first M matches for unit i. Like before, $\hat{\tau}_M = \frac{1}{N} \sum_{i=1}^{N} [\hat{Y}_i(1) - \hat{Y}_i(0)]$.

The fact that we use the Euclidean distance metric when calculating the nearest neighbors is restrictive, but in practice, researchers can easily transform the raw data as desired and use the Euclidean distance metric in that new space. Indeed, Euclidean distance matching is typically done in one of three spaces: the space of propensity scores, the space of predictive means, or the raw covariate space (Austin, 2010; Crump, Joseph, Imbens, and Mitnik, 2009; Di Zio and Guarnera, 2009; Ho, Imai, King, and Stuart, 2007; King, Lucas, and Nielsen, 2017; Kum and Masterson, 2008; Little, 1988; Rosenbaum and Rubin, 1985; Rubin, 1986; Wang et al., 2013).

In this paper, we offer new guidance for investigators interested in using matching-based methods by revisiting the analytical bias of such estimators in the Gaussian case. The most immediate forerunner to our paper is Abadie and Imbens (2006), who show that, when the distribution of $X_i|W_i = w$ is defined over a bounded domain, $\mathbb{D}$, the bias of the matching estimator for τ_{ATET} can be written as

$$\begin{aligned}\text{Bias}^{\text{ATET}}_{\text{M, Bounded Domain}} = &-\left(\frac{1}{M}\sum_{m=1}^{M}\Gamma\left(\frac{mk+2}{k}\right)\frac{1}{(m-1)!k}\right)\frac{1}{n_1^{(2r/k)}} \\ &\times \theta^{2/k}\int_{x\in\mathbb{D}}\left(f_0(x)\frac{\pi^{k/2}}{\Gamma(1+k/2)}\right)^{-2/k} \\ &\left\{\frac{1}{f_0(x)}\frac{\partial f_0}{\partial x}(x)\frac{\partial \mu_0}{\partial x}(x)+\frac{1}{2}\text{tr}\left(\frac{\partial \mu_0}{\partial x'\partial x}\right)\right\} \times f_1(x)\,\mathrm{d}x + o(n_1^{-2r/k}),\end{aligned} \tag{1}$$

where M is the number of control matches per treated unit, k is the dimensionality, $f_w(x)$ is the conditional distribution of the covariates given treatment status w, $\mu_0(x) = \mathbb{E}[Y_i(w_i)|X_i = x_i]$, and where, for $r \geq 1, n_1^r/n_0 \to \theta$ for $0 < \theta < \infty$. Their analysis indicates that the bias of nearest neighbor matching is proportional to $(1/n)^{1/k}$. This result suggests that the bias does converge to 0 as n approaches infinity, but does so at a slow rate when the covariate dimension, k, is large.

There is considerable interest in extending the important findings of Abadie and Imbens (2006). For example, those results cannot be applied in situations where the domain of the covariates is unbounded. Thus, the results do not apply when the background covariates are Gaussian or follow a Multivariate t-distribution, which are among the most popular distributions for modeling continuous data. It is also difficult to apply the results to specific bounded continuous distributions. For example, if we assume $f_0(x)$ and $f_1(x)$ are identical multivariate uniform distributions on k-dimensional rectangles, the expression could be evaluated, but there is no unconditional bias in this situation as the treated and control units have identical feature distributions. Conversely, the expression cannot be used if $f_1(x)$ and $f_0(x)$ are uniform but non-overlapping in domain: any non-overlapping regions will cause the bias term to be undefined (since it contains $1/f_0(x)$, which is undefined when $f_0(x) = 0$).

Because Gaussianity is so often assumed in statistical analyses, there is also considerable intrinsic interest in examining the bias of matching estimators when this assumption holds true. The Diaconis-Freedman effect suggest that our results may also be applicable even when a high-dimensional set of covariates are not Gaussian, but have been projected into low dimensional subspace provided that no single covariate carries too much relative weight in the projection (Diaconis and Freedman, 1984; Duembgen and Zerial, 2011; Li, Vlassis, Kawale, and Fu, 2016). Additionally, it would of great value to examine the asymptotic behavior of the matching estimator as it relates to the covariate overlap between treatment and control groups: if we can model how the bias behaves with changes in the overlap between the treated and control classes, we can arrive at better algorithms for extracting causal effects by pruning observations. We hope that analytical results for the Gaussian case can help researchers working in that domain.

With these considerations in mind, we can thus more formally present our assumptions. First, we assume that conditional unconfoundedness holds: W_i is independent from $(Y_i(0), Y_i(1))$ given $X_i = x$. Second, $\{(Y_i, W_i, X_i)\}_{i=1}^{N}$ are independent draws from the distribution of (Y, W, X). Next, $\mathbb{E}[Y_i(w)|X_i = x]$ is finite for all x and has finite first and second derivatives. Fourth, we assume matching is done with replacement, and that nearest neighbor matching is used ($M = 1$). Finally, we assume $X_i|W_i = w \sim \mathcal{N}(\alpha_w, \Sigma_w = \sigma_w^2 I_{k\times k})$. This assumption implies that the covariate distribution can be modeled only with the mean and variance. This is the only major assumption that differs from Abadie and Imbens (2006), who instead assume $X_i|W_i = w$ is governed by a generic distribution over a bounded domain. It is important to note that our results definitely do not imply or assume that there are some units with propensity scores of 0 or 1. In fact, with a linear model for the log-odds, the continuity of the Normal distribution implies that no units will have propensity scores of 0 or 1.[1]

3 Results

3.1 The Gaussian Case

The bias of the matching estimator is difficult to evaluate analytically because it involves a non-smooth action being performed on the data. We use Taylor series expansions, advances in extreme value theory (especially Al-Naffouri (2009)), and geometric reasoning to arrive at THEOREM 1.

THEOREM 1. Let $c_1(k, n_0, \sigma_0^2) = 2\sigma_0^2\Gamma\left(1 + k/2\right)^{2/k} n_0^{-2/k}$. Let $\Sigma_M(t) = \left(\left(\frac{-k}{t}\Sigma_0\right)^{-1} +\right.$

[1] It would be interesting to consider the analytical bias of nearest neighbor matching after pruning observations with propensity scores of "nearly" 0 or 1, but that is left for future study. Moreover, our results suggest that pruning observations with propensity scores of nearly 0 or 1 may not be optimal for reducing the bias of nearest neighbor matching. As will be clear, optimal pruning would minimize $\beta\Delta$, with those terms to be defined later. There does not appear to be an *a priori* reason why pruning observations with propensity scores near 0 or 1 should minimize this quantity.

$\Sigma_1^{-1})^{-1}$, $\alpha_M(t) = \Sigma_M(t)((\frac{-k}{t}\Sigma_0)^{-1}\alpha_0 + \Sigma_1^{-1}\alpha_1)$, and

$$\mathbb{E}[||X_{j_1(i)} - X_i||_2^t] = c_1(k, n_0, \sigma_0^2)^{t/2}\Gamma\left(1 + \frac{t}{k}\right)\frac{\sqrt{|2\pi\Sigma_M(t)|}}{\sqrt{|2\pi\Sigma_1|}}$$
$$\exp\Bigg(- 0.5\bigg[\alpha_0'\left(\frac{-k}{t}\Sigma_0\right)^{-1}\alpha_0 + \alpha_1'\Sigma_1^{-1}\alpha_1 - \alpha_M(t)'\Sigma_M(t)^{-1}\alpha_M(t)\bigg]\Bigg).$$

Then, we can then write an asymptotic expression for the bias of nearest neighbor matching under a Gaussian model as

$$\begin{aligned}\text{Bias}^{\text{ATET}}_{\text{M = 1, Gaussian Model}} &= \mathbb{E}[\hat{\tau}_{\text{M=1}} - \tau_{\text{ATET}}] \\ &= -\Big(\mathbb{E}[||X_{j(i)} - X_i||_2] - 0.5\mathbb{E}[||X_{j_1(i)} - X_i||_2]^{-1}\mathbb{E}[||X_{j_1(i)} - X_i||_2^2]\Big) \\ &\qquad \times \frac{(\alpha_0 - \alpha_1)'}{||\alpha_0 - \alpha_1||_2} \cdot \mathbb{E}[\nabla\mu_0(x)] + \mathcal{O}(\mathbb{E}[||X_{j_1(i)} - X_i||_2^2]). \end{aligned} \quad (2)$$

PROOF OF THEOREM 1. See Appendix. An asymptotic upper bound for the bias of nearest neighbor matching is found in the Appendix as well. Note that the expression only applies when there is some covariate shift (i.e. $\alpha_0 \neq \alpha_1$).

DISCUSSION OF THEOREM 1. There are several interesting implications from Equation 2. First, Abadie and Imbens (2006) found that the bias of matching estimators in bounded domains is $\mathcal{O}((1/n)^{1/k})$, where k is the dimensionality. Our result corroborates this finding: in the Gaussian case, the bias is $\mathcal{O}((1/n_0)^{1/k})$ (see the Appendix for the full derivation). The nearest neighbor matching estimators under a Gaussian model is asymptotically unbiased. Our expression for the bias also contains other terms that help explain the realized behavior of matching estimators.

To illustrate these implications, assume without loss of generality that $\alpha_0 = 0_{k\times 1}$ and $\alpha_1 = \alpha_0 + \Delta_{k\times 1}$. Also, assume that σ_0^2 has been normalized to 1 so that we can parametrize the treated group variance by an inflation factor of ρ and that $\rho = 1$. In that case, the leading term in the bias expression has a dominating polynomial component; the bias can, under a range of conditions specified in the Appendix, be approximated as

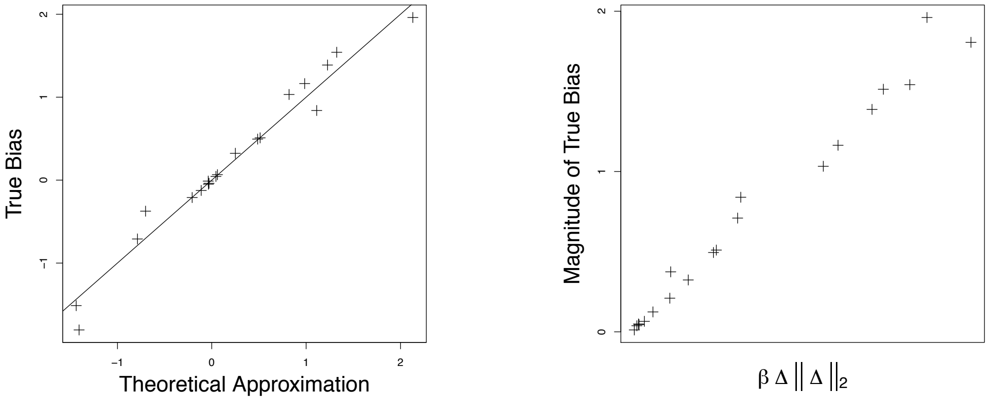
$$\text{Bias}^{\text{ATET}}_{\text{M = 1, Gaussian Model}} \approx -\left(\beta\Delta||\Delta||_2 \cdot d_2(k) + \beta \cdot \frac{\Delta}{||\Delta||_2} \cdot \frac{1}{\sqrt{2}} \right) \times d_1(k, n_0),$$

where $\beta_{1\times k} = \mathbb{E}\left[\frac{\partial\mu_0}{\partial x'}(X_i)\right]$ and $d_1(k, n_0)$ and $d_2(k)$ are the appropriate factors specified in the Appendix.

The key part of this expression is the $\beta\Delta||\Delta||_2$ component. This component signals that the bias of matching in the Gaussian case is approximately proportional in magnitude to a polynomial function of the L_2-norm of the difference in covariate centers. On the one hand, the magnitude of the bias is cubic in $||\Delta||_2$ when $\beta \propto \Delta'$. In that case, $\beta\Delta||\Delta||_2 \propto \Delta'\Delta||\Delta||_2 = ||\Delta||_2^3$. This situation occurs when there is imbalance in each variable that is proportional to the effect of that variable on the outcome. On the other hand, the magnitude of the bias is approximately quadratic in $||\Delta||_2$ when $\beta \propto 1_{1\times k}$. In that case (assuming all entries of Δ are positive), $\beta\Delta||\Delta||_2 \propto ||\Delta||_1 \cdot ||\Delta||_2$. This situation occurs when the imbalance is unrelated to the covariate-outcome correlation. In other situations, the rate is between quadratic and cubic (or could conceivably be lower). In general, for a fixed k, the absolute bias increases at a polynomial rate with the distance between α_0 and α_1. In other words, the matching estimator has quickly deteriorating performance as the treatment group means grow more distant in Euclidean space. In this case, the matching estimator will be systematically biased, with this bias growing more quickly the worse the imbalance is.

We can confirm these theoretical results via Monte Carlo integration. We assume a linear model for the effect of the covariates on the outcome. In addition, $n_0 = 1000$, $n_1 = 100$, $\sigma_0^2 = \sigma_1^2 = 1$, and $k = 10$. In panel (a) of Figure 1, we find an excellent match between the theoretical results of THEOREM 1 and the true bias. The R^2 of the relation is 0.97. The R^2 is not 1 because we use Taylor series approximations. In panel b. of Figure 1, we find support for the idea that $\beta\Delta||\Delta||_2$ is a key determinant of the overall bias (the R^2 is again above 0.96). These results are encouraging and indicate that the theoretical argumentation stands on firm ground.

In sum, our results indicates that the bias of matching estimators is quickly increasing in the lack of overlap between the treatment and control groups. The precise speed of increase is governed by the inner product between the average variable effect and the distance between covariate centers. To our knowledge, no other work provides an analytical result of this type. The result suggests that pruning observations is an important step in matching algorithms, since pruning can help eliminate areas of non-overlap. Although our result is about the Multivariate Gaussian case, we expect the results to generalize to other continuous distributions, although the analysis is expected to be more involved.



(a)

(b)

Figure 1: (a.) There is good match between the theory and the truth. The 45-degree line is also displayed. (b.) $\beta\Delta||\Delta||_2$ term is highly predictive of the overall bias. In both (a.) and (b.), the true bias is obtained via Monte Carlo integration.

3.2 Predictive Mean Matching

Part of the preceding analysis has shown that matching in the raw covariate space has decreasing performance as the dimensionality grows. Predictive mean matching may be an appealing alternative method, since it operates in a space where $k = 1$ (Di Zio and Guarnera, 2009; Kum and Masterson, 2008; Little, 1988; Rubin, 1986). In this section, we consider the analytical bias of this method. Instead of performing nearest neighbor matching in the raw covariate space, we perform matching in the space of predicted values for $Y_i(0)$. Unfortunately, predictive mean matching requires that a model be fit on the observed outcomes under the control regime. However, it can reduce the bias of matching estimators because it gives extra importance to the background covariates which are correlated with the outcome, thereby placing more emphasis on matches that are closer to one another on a subset of substantively relevant covariates. Moreover, even if the covariates are not Gaussian, the projection via a linear model can often be approximated as Gaussian. The approximation is especially good if the number of original covariates is large and a significant number of covariates combine to exert a small individual effect on the outcome.

COROLLARY 1 presents analytical results on predictive mean matching, where the matches have been made by pairing observations in a way that minimizes the distance between $\widehat{\mathbb{E}}[Y_i(0)|X_i = x]$ and $\widehat{\mathbb{E}}[Y_j(0)|X_j = x]$, where the j's are in the alternative testing group relative to i. The imputation step for predictive mean matching is analogous to matching in the raw covariates, and $\hat{\tau}_{\text{PMM}}$ is defined analogously as well. The analytical results assume an unbiased model for the response function under control (i.e. $\mathbb{E}[Y_i(0)|X_i = x]$). We write $\beta_{1\times k} = \mathbb{E}[\nabla\mu_0(x)]$ and assume $\sigma^2_{\hat{Y}(0)|W_i=0} > \sigma^2_{\hat{Y}(0)|W_i=1}$.

Earlier work on predictive mean matching, such as Schenker and Taylor (1996), have primarily relied on simulation studies. Our results indicate that the method can also be analyzed using a more positivistic approach.

COROLLARY 1. If predictive mean matching is done using an asymptotically unbiased model for the response function under control, the asymptotic bias of predictive mean matching can be written as

$$\begin{aligned}\text{Bias}^{\text{ATET}}_{\text{PMM, Gaussian Model}} = -c_1(1, n_0, \sigma^2_{\hat{Y}(0)|W_i=0})^{1/2}\Gamma\left(2\right)\left(\frac{\sigma^2_{\hat{Y}(0)|W_i=0}}{\sigma^2_{\hat{Y}(0)|W_i=0} - \sigma^2_{\hat{Y}(0)|W_i=1}}\right)^{1/2} \\ \times \exp\Bigg(- 0.5\Bigg[\frac{(\beta\alpha_1)^2}{\sigma^2_{\hat{Y}(0)|W_i=1}} - \frac{(\beta\alpha_0)^2}{\sigma^2_{\hat{Y}(0)|W_i=0}} \\ - \frac{\sigma^2_{\hat{Y}(0)|W_i=1}\sigma^2_{\hat{Y}(0)|W_i=0}}{\sigma^2_{\hat{Y}(0)|W_i=0} - \sigma^2_{\hat{Y}(0)|W_i=1}}\left(\frac{\beta\alpha_1}{\sigma^2_{\hat{Y}(0)|W_i=1}} - \frac{\beta\alpha_0}{\sigma^2_{\hat{Y}(0)|W_i=0}}\right)^2\Bigg]\Bigg) \\ \times \frac{\beta(\alpha_0 - \alpha_1)}{|\beta\alpha_0 - \beta\alpha_1|} \times 1 + \mathcal{O}(\mathbb{E}[||X_{j_1(i)} - X_i||^2_2]).\end{aligned} \tag{3}$$

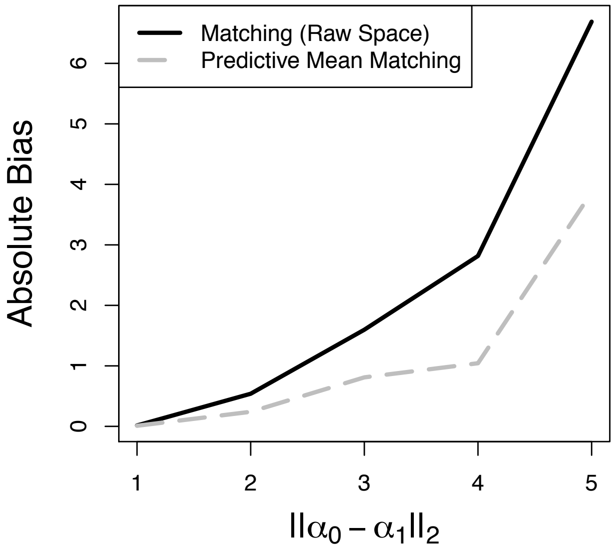
PROOF FOR COROLLARY 1. See Appendix.

DISCUSSION OF COROLLARY 1. Equation 3 suggests that the bias of predictive mean matching will, under a range of conditions, be smaller in magnitude than matching in the raw space. To see this, we can compare some of the constant terms in Equation 3 and Equation 2. One of the key factors in Equation 2 is

$$\sqrt{2\sigma_0^2}\Gamma(1 + k/2)^{1/k}\Gamma(1 + 1/k)n_0^{-1/k} \approx \sqrt{2\sigma_0^2} \times n_0^{-1/k}$$

and the equivalent expression in Equation 3 is

$$\sqrt{2\sigma^2_{\hat{Y}(0)|W_i=0}}\Gamma(1 + 1/2)n_0^{-1}\Gamma(2) \approx 0.89\sqrt{2\sigma^2_{\hat{Y}(0)|W_i=0}} \times n_0^{-1},$$



(a)

Figure 2: Predictive mean matching tends to have better analytical properties than matching in the raw space, provided that an asymptotically unbiased estimator can be found. In addition, note the polynomial growth of the absolute bias as the covariate centers become more distant from one another.

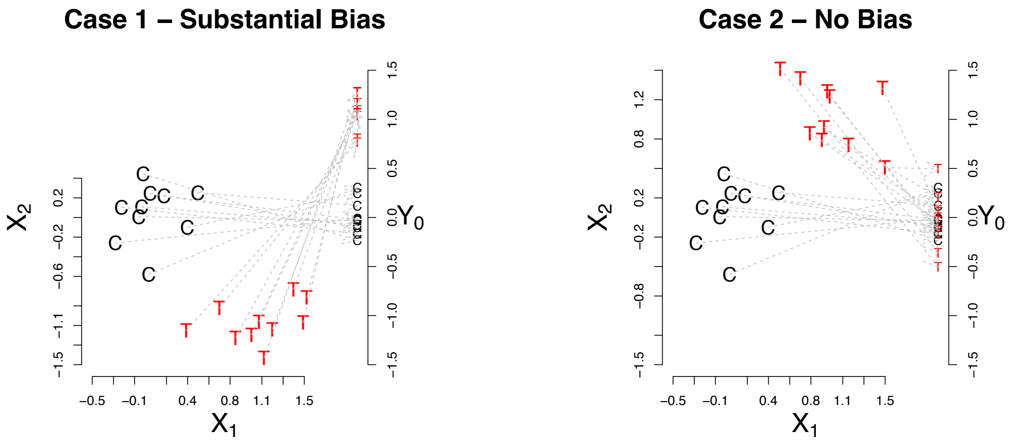
which will in general be considerably smaller if k is large. The exponential component in the first term of the bias is interesting to consider as well. With predictive mean matching, the exponent will contain $\beta\Delta$ instead of $\Delta'\Delta$, which would compare favorably if the covariates are sparse and many entries in β are 0. If many of the covariates have no relationship with the outcome, predictive mean matching decreases the bias, as irrelevant information is no longer considered when comparing candidate pairs. Figure 2 illustrates the improvements of predictive mean matching over matching in the raw space as $||\alpha_0 - \alpha_1||_2$ grows. The Monte Carlo parameters are the same as those from the previous section. Improvements from predictive mean matching get larger as the lack of overlap between covariate centers becomes more severe.

3.3 Limitations

There are several limitations to the expressions derived in this paper. First, they are asymptotic, and may be less accurate when n_0 is small. They are also less accurate when $||\alpha_0 - \alpha_1||_2$ is very large, since in this case the asymptotic approximation of the nearest neighbor distance breaks down and the use of higher order terms is required. Second, we assumed Gaussianity of the background covariates. Although this is an important case, it would be of substantial interest to consider the bias when the covariate distribution has fat-tails, which may be a more realistic assumption for certain processes in the natural or social sciences. A final limitation of our work is that the background covariates are by assumption unrelated to each other. This assumption was necessary to derive the results, but may not be needed in the future in light of coming results in extreme value theory.

4 Implications for Applied Researchers

It is important to consider the implications of the preceding discussion for applied researchers. At the start, it is prudent to emphasize that our results do not deal with the use of matching for as a data preprocessing tool prior to regression, so we do not comment on that topic. The first implication of our discussion is that nearest neighbor matching, on its own, can have unsatisfactory properties. For example, when the covariates which have a positive average relationship with the control outcome, $Y_i(0)$, are systematically higher or systematically lower in one of the treatment groups, the bias is proportional to $||\Delta||_2^3$, which is a cubic function of the distance between the treated and control covariate centers. In other words, as the imbalance in this case increases by a factor of 2, the bias



(a) (b)

Figure 3: (a) The absolute bias is proportional to $||\Delta||_2^3$ when the imbalance is correlated with the average effect of each covariate on the outcome. In other words, $\beta \approx \Delta$ so that $\beta\Delta||\Delta||_2 = \Delta'\Delta||\Delta||_2 = ||\Delta||_2^3$. When this occurs, $Y_i(0)$ differs substantially for treated and control units. In this simulated example, $\beta = (0.5, -0.5)$ and $\Delta = (1, -1)'$ so that $\beta\Delta = 1$. (b) There is no bias when one covariate's contribution (via imbalance) to the bias is exactly offset by another's contribution so that $\beta\Delta \approx 0$. In this fortuitous and presumably rare case, $Y_i(0)$ does not differ substantially for treated and control units. In this simulated example, $\beta = (0.5, -0.5)$ and $\Delta = (1, 1)'$ so that $\beta\Delta = 0$.

increases by a factor of 8. The left panel of Figure 3 illustrates this possibility, where X_1 has a positive average relationship with $Y_i(0)$ and X_2 has a negative average relationship, whereas the control group has a higher mean on X_1 and a lower mean on X_2.

However, our analysis is somewhat more nuanced. For example, we have identified a condition that, if it holds, implies a nearly zero bias of nearest neighbor matching. The right panel of Figure 3 illustrate graphically this possibility: here, X_1 again has a positive average relationship with $Y_i(0)$ and X_2 again has a negative average relationship, but the control group has a higher mean on both X_1 and X_2, leading to a fortuitous cancellation of terms ($\beta\Delta = 0$) and to a bias of nearly zero. The bias is also zero if none of the covariates in fact have an average relationship with $Y_i(0)$ so that $\beta = 0$.

Researchers do not know in advance which of these situations is most relevant for them. More practically, then, one takeaway of the analysis could be that nearest neighbor matching, precisely because it imposes so little structure on the problem, should be seen as a potentially risky inference technique if used alone. There is a chance that, in any given problem, a unit increase in the imbalance leads to a non-linear effect on the bias.

Another implication of the analysis is that, provided a sufficient model for the outcome under control (for $Y_i(0)$) can be found, predictive mean matching is preferable to matching in the raw covariate space. Our analysis has shown that predictive mean matching would be preferable for two reasons. First, predictive mean matching reduces the dimensionality of the problem, which improves greatly the asymptotic rate of convergence of the bias to 0. Second, predictive mean matching allows researchers to obtain balance on the subset of covariate which are predictive of the outcome, while ignoring irrelevant information. We have shown how this fact has a greatly beneficial effect on the bias term.

Finally, this work also can give applied researchers insight into how they might prune observations to improve their inferences. This analysis has identified a key factor which controls the magnitude of the bias. That factor is $\beta\Delta||\Delta||_2$. β denotes the average relationship between the covariates and the outcome. Δ denotes the difference in means between treated and control groups. If researchers, from the literature or from a preliminary analysis, gain information about β, they could then in practice prune observations systematically to reduce $\beta\Delta$ and therefore to minimize the bias. Future theoretical work should also be done to give more explicit guidance about this proposal.

5 Conclusion

In this paper, we have established results about the bias of matching estimators for causal effects in the Gaussian case. In this case, the magnitude of the bias grows polynomially with the distance between the covariate centers of the treatment and control groups. The specific rate of growth is a function of the relationship between the covariate shift and covariate-outcome correlation. The rate can be nearly quadratic or even cubic. Although we also show that the nearest neighbor estimator is asymptotically unbiased, the dimensionality of the feature space slows down the rate of convergence, thus corroborating in the Gaussian case the findings of Abadie and Imbens (2006). Our results motivate the use of techniques that can mitigate the bias, such as predictive mean matching, and we provide novel analytical results for this case as well. Finally, we summarize the implications of these results for applied researchers. △

References

Abadie, Alberto and Guido Imbens (Jan. 2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects". In: *Econometrica*, no. 1, vol. 74, pp. 235–267. DOI: 10.1111/j.1468-0262.2006.00655.x. URL: http://doi.wiley.com/10.1111/j.1468-0262.2006.00655.x.

Athey, Susan and Guido Imbens (July 2016): "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences*, no. 27, vol. 113, pp. 7353–7360. DOI: 10.1073/pnas.1510489113. URL: http://www.ncbi.nlm.nih.gov/pubmed/27382149.

Athey, Susan, Guido Imbens, and Stefan Wager (2016): “Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions”.

Austin, Peter C. (Sept. 2010): “The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies”. In: *Statistics in Medicine*, no. 20, vol. 29, pp. 2137–2148. DOI: 10.1002/sim.3854. URL: http://www.ncbi.nlm.nih.gov/pubmed/20108233.

Belloni, Alexandre, Victor Chernozhukov, Ivan Fernandez-Val, and Christian B. Hansen (2017): “Program Evaluation and Causal Inference with High-Dimensional Data”. In: *Econometrica*, no. 1, vol. 85, pp. 233–298.

Crump, Citation, Hotz Joseph, Guido Imbens, and Oscar A Mitnik (2009): “Dealing with Limited Overlap in Estimation of Average Treatment Effects”. In: *Biometrika*, no. 1, vol. 96, pp. 187–199. DOI: 10.1093/biomet/asn055.

Di Zio, Marco and Ugo Guarnera (June 2009): “Semiparametric predictive mean matching”. In: *AStA Advances in Statistical Analysis*, no. 2, vol. 93, pp. 175–186. DOI: 10.1007/s10182-008-0081-2. URL: http://link.springer.com/10.1007/s10182-008-0081-2.

Diaconis, Persi and David Freedman (Sept. 1984): “Asymptotics of Graphical Projection Pursuit”. In: *The Annals of Statistics*, no. 3, vol. 12, pp. 793–815. DOI: 10.1214/aos/1176346703. URL: http://projecteuclid.org/euclid.aos/1176346703.

Duembgen, Lutz and Perla Zerial (July 2011): “On Low-Dimensional Projections of High-Dimensional Distributions”. URL: http://arxiv.org/abs/1107.0417.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth Stuart (Jan. 2007): “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference”. In: *Political Analysis*, no. 03, vol. 15, pp. 199–236. DOI: 10.1093/pan/mpl013. URL: https://www.cambridge.org/core/product/identifier/S1047198700006483/type/journal%7B%5C_%7Darticle.

King, Gary, Robert O. Keohane, and Sidney Verba (1994): *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press, p. 300.

King, Gary, Christopher Lucas, and Richard A. Nielsen (Apr. 2017): “The Balance-Sample Size Frontier in Matching Methods for Causal Inference”. In: *American Journal of Political Science*, no. 2, vol. 61, pp. 473–489. ISSN: 00925853. DOI: 10.1111/ajps.12272. URL: http://doi.wiley.com/10.1111/ajps.12272.

Kum, Hyunsub and Thomas Masterson (2008): “Statistical Matching Using Propensity Scores: Theory and Application to the Levy Institute Measure of Economic Well-Being”. URL: http://www.ssrn.com/abstract=1136363.

Li, Sheng, Nikos Vlassis, Jaya Kawale, and Yun Fu (2016): “Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns”. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pp. 3768–3774. URL: https://www.ijcai.org/Proceedings/16/Papers/530.pdf.

Lijphart, Arend (Sept. 1971): “Comparative Politics and the Comparative Method”. In: *The American Political Science Review*, no. 3, vol. 65, pp. 682–693.

Little, Roderick J. A. (July 1988): “Missing-Data Adjustments in Large Surveys”. In: *Journal of Business & Economic Statistics*, no. 3, vol. 6, pp. 287–296. ISSN: 0735-0015. DOI: 10.1080/07350015.1988.10509663. URL: http://www.tandfonline.com/doi/abs/10.1080/07350015.1988.10509663.

Al-Naffouri, T.Y. (2009): “Scaling of the minimum of iid random variables”. In: *Signal Processing*, no. 9, vol. 89, pp. 1830–1834. DOI: 10.1016/j.sigpro.2009.02.012. URL: http://www.sciencedirect.com.ezp-prod1.hul.harvard.edu/science/article/pii/S0165168409000668.

Rosenbaum, Paul R. and Donald B. Rubin (Feb. 1985): “Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score”. In: *The American Statistician*, no. 1, vol. 39, pp. 33–38. ISSN: 0003-1305. DOI: 10.1080/00031305.1985.10479383. URL: http://www.tandfonline.com/doi/abs/10.1080/00031305.1985.10479383.

Rubin, Donald B. (1974): “Estimating causal effects of treatments in randomized and non-randomized studies”. In: *Journal of Educational Psychology*, no. 5, vol. 66, pp. 688–701. DOI: 10.1037/h0037350. URL: http://content.apa.org/journals/edu/66/5/688.

— (Jan. 1986): “Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations”. In: *Journal of Business & Economic Statistics*, no. 1, vol. 4, p. 87. ISSN: 07350015. DOI: 10.2307/1391390. URL: http://www.jstor.org/stable/1391390?origin=crossref.

Schenker, Nathaniel and Jeremy M.G. Taylor (Aug. 1996): “Partially parametric techniques for multiple imputation”. In: *Computational Statistics & Data Analysis*, no. 4, vol. 22, pp. 425–446. DOI: 10.1016/0167-9473(95)00057-7. URL: http://linkinghub.elsevier.com/retrieve/pii/0167947395000577.

Splawa-Neyman, Jerzy, D. M. Dabrowska, and T. P. Speed (Nov. 1990): "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9". In: *Statistical Science*, no. 4, vol. 5, pp. 465–472. ISSN: 0883-4237. DOI: 10.1214/ss/1177012031. URL: http://projecteuclid.org/euclid.ss/1177012031.

Stuart, Elizabeth (Feb. 2010): "Matching methods for causal inference: A review and a look forward." In: *Statistical science : a review journal of the Institute of Mathematical Statistics*, no. 1, vol. 25, pp. 1–21. DOI: 10.1214/09-STS313. URL: http://www.ncbi.nlm.nih.gov/pubmed/20871802.

Wang, Yongji et al. (Dec. 2013): "Optimal Caliper Width for Propensity Score Matching of Three Treatment Groups: A Monte Carlo Study". In: *PLoS ONE*, no. 12, vol. 8. Ed. by Robert K. Hills, e81045. DOI: 10.1371/journal.pone.0081045. URL: http://dx.plos.org/10.1371/journal.pone.0081045.