Matching via Dimensionality Reduction for Estimation of Treatment Effects in Digital Marketing Campaigns

Sheng Li
Northeastern University
Boston, MA, USA
shengli@ece.neu.edu

Nikos Vlassis
Adobe Research
San Jose, CA, USA
vlassis@adobe.com

Jaya Kawale
Adobe Research
San Jose, CA, USA
kawale@adobe.com

Yun Fu
Northeastern University
Boston, MA, USA
yunfu@ece.neu.edu

Abstract

A widely used method for estimating counterfactuals and causal treatment effects from observational data is nearest-neighbor matching. This typically involves pairing each treated unit with its nearest-in-covariates control unit, and then estimating an average treatment effect from the set of matched pairs. Although straightforward to implement, this estimator is known to suffer from a bias that increases with the dimensionality of the covariate space, which can be undesirable in applications that involve high-dimensional data. To address this problem, we propose a novel estimator that first projects the data to a number of random linear subspaces, and it then estimates the median treatment effect by nearest-neighbor matching in each subspace. We empirically compute the mean square error of the proposed estimator using semi-synthetic data, and we demonstrate the method on real-world digital marketing campaign data. The results show marked improvement over baseline methods.

1 Introduction

Estimating causal treatment effects from observational (i.e., nonexperimental) data is a key problem in various fields of science [Spirtes *et al.*, 2000; Pearl, 2009b; Morgan and Winship, 2014]. Our domain of interest is *digital marketing*, where a marketer has access to a collection of observational data and seeks answers to business critical questions such as whether a new advertisement attracts more clicks, whether a new campaign drives more conversions, etc. [Chan *et al.*, 2010; Dalessandro *et al.*, 2012; Hill *et al.*, 2015; Barajas *et al.*, 2015; Wang *et al.*, 2015]. Unlike randomized experiments where the covariate distributions of the treatment and control groups are balanced in expectation, in an observational study this balance can be distorted by a confounding bias that is introduced by the systematic assignment of the treatment to the units [Rosenbaum and Rubin, 1983; Pearl, 2010; Hill and Su, 2013]. A simple and popular method to address this bias is to match the distribution of covariates in the two groups in order to create a subsample that is more balanced in the covariates [Rubin, 1973a; Stuart, 2010; Hainmueller, 2012]. The matched counterparts of the treated units in the control group are interpreted as counterfactuals, and the average treatment effect on treated (ATT) is estimated by comparing the outcomes of every matched pair.

One of the widely used matching methods is Nearest Neighbor Matching (NNM) [Rubin, 1973a]. For each treated unit, NNM finds its nearest neighbor in the control group to generate a matched pair, and the ATT is then estimated from the set of matched pairs. Different NNM methods are characterized by the choice of a distance measure for determining such a match. Some of the popularly used distance measures are Exact Matching and its variant Coarsened Exact Matching (CEM) [Iacus *et al.*, 2011], Mahalanobis distance matching [Rubin, 1979], and Propensity Score Matching (PSM) [Rosenbaum and Rubin, 1983]. Although these estimators have been widely applied, they usually exhibit poor performance in high-dimensional data. For example, Exact Matching has to discard a lot of data in order to attain a good matching when a large set of covariates is present and hence it can result in a large bias [Rosenbaum and Rubin, 1985]. Mahalanobis distance matching uses the empirical covariance of the data, but this effectively imposes parametric assumptions on the data and therefore is prone to a bias in high dimensions [Gu and Rosenbaum, 1993]. PSM summarizes each covariate vector by a single number, namely the probability of being treated, but it also suffers from a model misspecification problem in the high-dimensional case [Caliendo and Kopeinig, 2008; King and Nielsen, 2016]. More generally, theoretical results show that the bias of a nearest-neighbor matching estimator increases with the dimensionality of the data [Abadie and Imbens, 2006], which can limit the applicability of matching estimators in real-world domains such as digital marketing.

The above motivates an approach that maintains the practical advantages of NNM while addressing its sensitivity to the high dimensionality of the data. In this work we propose a new estimator that matches the data in a low-dimensional subspace of the original space. Intuitively, a good estimator should preserve as much as possible the neighborhood relationships of the units in the input space, which motivates projections that provide guarantees of that sort. For a total number N of treated and control units, the Johnson-Lindenstrauss (JL) lemma [Johnson and Lindenstrauss, 1984; Ailon and Chazelle, 2006] guarantees that a *random* linear projection to $O(\log N)$ dimensions preserves (withing a con-

trollable factor) all pairwise distances of the points in the high-dimensional input space. Motivated by the JL lemma, we propose a new matching estimator that works as follows. (1) We first project the data to multiple random linear subspaces of dimension $O(\log N)$. (2) Then we estimate the treatment effect by simple nearest neighbor matching in each subspace. (3) Finally we compute the median value of the estimated effects. A theoretical treatment of the statistical properties of the new estimator is out of the scope of this paper, but we provide a careful empirical calculation of the mean square error of the estimator on semi-synthetic data, following a recently proposed experimental design that employs real data to compute statistical properties of ATT estimators [Frölich *et al.*, 2015]. As we demonstrate on semi-synthetic as well as real data from digital marketing campaigns, the proposed estimator is very robust and it consistently improves over baseline methods.

In summary, the main contributions of our work are:

- We propose the estimation of average treatment effects from observational data by means of dimensionality reduction. To the best of our knowledge, this is the first attempt to estimate treatment effects by matching in a low-dimensional subspace of the data.
- We evaluate the proposed estimator on semi-synthetic data as well as real-world data from digital marketing, and show its robustness against baseline methods. To the best of our knowledge, this is the first application of nearest-neighbor matching in the domain of digital marketing. Our approach can also be applied in other domains that involve high-dimensional observational data, such as text analysis [Roberts *et al.*, 2015] and public health [Glass *et al.*, 2013].

2 Background and Related Work

The dominant paradigm for causality in the AI literature is the *structural causal model* of Pearl [Pearl, 2009b], with its $do(\cdot)$ notation for interventions and semantics drawn from the structural equation models and graphical models for probabilistic reasoning [Spirtes *et al.*, 2000]. A contender to the Pearl's paradigm is the *counterfactual* or *potential-outcome* framework of Neyman-Rubin [Neyman, 1923; Rubin, 1974]. The key distinction between Pearl's framework and the counterfactual framework is that the latter allows expressing counterfactual queries as standard probabilistic queries on some probability space defined on both hypothetical and real events, subject to certain consistency constraints between the two types of events [Pearl, 2009a]. The fact that causal queries can be formulated in the familiar probability language has contributed to the popularity of the counterfactual framework for practical applications [Morgan and Winship, 2014], but the framework has also been criticized for its lack of flexibility or intuitiveness when dealing with complex settings [Pearl, 2009a]. The counterfactual framework is sufficient for demonstrating the key ideas of our work, so we have adopted it throughout.

The key assumption of the counterfactual framework is that each individual in the population of interest has a potential outcome under each (binary) treatment state, even though each individual is observed in only one treatment state at any point in time [Morgan and Winship, 2014]. More specifically, for individual (unit) i there are two possible outcomes, $Y_i(1)$ if it undergoes treatment T, and $Y_i(0)$ if it does not. The treatment effect for individual i is defined as

$$\tau_i = Y_i(1) - Y_i(0), \tag{1}$$

and one of the two terms is always missing in the data, since for each individual i we observe either $Y_i(1)$ (if i is treated) or $Y_i(0)$ (if i is in the control group). A naive idea for solving the missing data problem is to directly compare the treated and control groups. This can work if the two groups have balanced distributions, which is approximately the case in randomized trials [King and Nielsen, 2016]. In observational studies, however, the assumption of balanced distributions is unrealistic due to the systematic assignment of treatments.

The idea of matching as a method to estimate causal effects from observational data began as early as the 1940s but gained widespread attention with the work by Rubin [Rubin, 1973a; 1973b; 1974]. Since then it has been widely studied and applied in several fields of science [Stuart, 2010; Morgan and Winship, 2014]. In the methodological literature, matching is typically used as a method to form quasi-experimental contrasts by sampling comparable treatment and control cases from among two larger pools of such cases [Morgan and Winship, 2014]. Matching methods can be broadly classified into nearest neighbor matching, weighting, and subclassification. In this work we focus on Nearest Neighbor Matching (NNM): For each treated unit i, NNM finds the nearest neighbor of i in the control group in terms of a given distance measure in the covariate space, and discards the unmatched control units. The average treatment effect on the treated (ATT) is then estimated by

$$ATT = \frac{1}{N_T} \sum_{i:T_i=1} \big(Y_i(1) - \hat{Y}_i(0)\big), \tag{2}$$

where $N_T = \sum_{i=1}^{N} T_i$ is the number of treated samples, and T_i is a binary treatment indicator. $Y_i(1)$ is the observed outcome of the treated unit i, and $\hat{Y}_i(0)$ is the counterfactual control outcome of unit i, which is estimated from the matched counterpart of unit i in the control group.

Early work on NNM focused on estimating the ATT when dealing with a single covariate [Rubin, 1973a; 1973b; Cochran and Rubin, 1973]. Various distance measures defining the similarity between two units while matching have been proposed in the literature. A straightforward way is to perform exact matching on the covariates, but this clearly becomes impractical when dealing with a large number of covariates or covariates containing continuous values. Mahalanobis distance matching computes the Mahalanobis distance between every pair of controlled and treated units in the covariate space and then matches the units accordingly. This is known to work well when there are relatively few covariates [Rubin, 1979], but it can fail when the number of covariates is large or when the covariates are not normally distributed [Gu and Rosenbaum, 1993].

A key development in this area was the introduction of Propensity Score Matching (PSM) [Rosenbaum and Rubin,

1983]. PSM estimates the probability of receiving treatment (aka propensity score) for each unit, without looking at the outcomes in the sample. The units are matched by simply grouping individuals with similar propensity scores, motivated by the fact that, asymptotically, at each value of the propensity score the distribution of the covariates defining the score is the same in the two groups. PSM has emerged as a very popular method for causal analysis mainly due to its simplicity [Dehejia and Wahba, 2002; Peikes *et al.*, 2008]. However, PSM is reported to be overly sensitive to the choice of model for the propensity score [Caliendo and Kopeinig, 2008], it can increase the imbalance between the two groups [King and Nielsen, 2016], and other issues [Heckman *et al.*, 1998; Bryson *et al.*, 2002; King and Zeng, 2006; Pearl, 2009b].

In a key paper, Abadie and Imbens [Abadie and Imbens, 2006] analyzed the bias of NNM estimators theoretically, and found that the dimensionality of the data can have a negative effect on the bias of the estimator. In particular, they showed that the bias grows, roughly, at a rate $O(N^{-1/d})$, where N is the sample size and d is the number of covariates. This implies that, for finite samples, the bias will increase with the dimensionality of the data. This result has motivated our approach in which matching is performed in a reduced dimension, as we describe next.

3 Randomized Nearest Neighbor Matching

Our approach hinges on the very simple idea that matching can be performed in a reduced subspace of the original covariate space. The principal motivation for such 'matching via dimensionality reduction' is a statistical one, namely to soften the dependence of the estimation bias to the data dimension [Abadie and Imbens, 2006]. As we demonstrate in the experiments section, this idea turns out to be solid.

As we are primarily interested in 'big data' applications such as digital marketing, we are constrained by computational considerations to the use of *linear* dimensionality reduction algorithms. Candidate algorithms here would be Principal Component Analysis (PCA) [Jolliffe, 2002], and Locality Preserving Projections (LPP) [He and Niyogi, 2004]. The LPP algorithm, in particular, aims at preserving the local neighborhood structure of the data in the embedding, and hence it seems a good choice for our problem. However, LPP relies on the choice of a kernel (that would be expensive to optimize over) and moreover it is not clear how to set the dimension of the reduced subspace. We have tried both PCA and LPP in our experiments, and we report the results in the next section.

Since we are aiming at a purely nonparametric approach, a natural choice for dimensionality reduction is via *random* projections. The technical tool we need is the Johnson-Lindenstrauss (JL) lemma [Johnson and Lindenstrauss, 1984] that states that we can (linearly) project high-dimensional data onto a randomly generated subspace of suitable dimension while approximately preserving the original distances between points:

Johnson-Lindenstrauss (JL) lemma. For any $0 < \epsilon < 1/2$ and $x_1, \cdots, x_N \in \mathbb{R}^d$, there exists a mapping $f : \mathbb{R}^d \to \mathbb{R}^k$, with $k = O(\epsilon^{-2} \log N)$, such that

$$\forall i, j \ (1-\epsilon)\|x_i - x_j\|^2 \leq \|f(x_i) - f(x_j)\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2.$$

It is known that the mapping f can be linear, giving rise to particularly simple algorithms for dimension reduction [Hegde *et al.*, 2008; Maillard and Munos, 2012].

Algorithm 1. *Randomized Nearest Neighbor Matching*

Input: Treated group $X_T \in \mathbb{R}^{d \times N_t}$
Control group $X_C \in \mathbb{R}^{d \times N_c}$
Number of random projections m
Response variables Y_T and Y_C
Total sample size N

1: Choose the dimension of subspace $k \approx O(\log N)$
2: ***for*** i from 1 to m
3: Generate a random projection $P_i \in \mathbb{R}^{d \times k}$
4: Project X_T and X_C using P_i
$Z_T^i = P_i^\top X_T, \ Z_C^i = P_i^\top X_C.$
5: Perform NNM between Z_T^i and Z_C^i
6: Estimate the ATT $A(i)$ from (2)
7: ***end for***

Output: Return $median(A)$

Our approach, dubbed *Randomized Nearest Neighbor Matching*, defines an ATT estimator by means of several random projections of the input covariates to subspaces of size $O(\log N)$ in accordance with the JL lemma. In particular, let $X_T \in \mathbb{R}^{d \times N_t}$ and $X_C \in \mathbb{R}^{d \times N_c}$ denote the treatment and control groups, where d is the dimensionality of data, N_t is the size of the treatment group, and N_c is the size of the control group.

The proposed estimator involves the following steps:

1) First, we construct m independent random linear projections, $P_1, \cdots, P_m \in \mathbb{R}^{d \times k}$, where k is chosen according to the JL lemma. Each entry in P_i is drawn from a univariate Gaussian distribution $\mathcal{N}(0, 1)$ (or other distributions [Achlioptas, 2001; Bingham and Mannila, 2001]). Each column of P_i is normalized to unit norm.

2) Second, we project X_T and X_C to low-dimensional subspaces using each of the generated projections. For the i-th subspace, this reads $Z_T^i = P_i^\top X_T$, $Z_C^i = P_i^\top X_C$, where Z_T^i and Z_T^i are the low-dimensional representations of X_T and X_C, respectively. We then perform nearest neighbor matching on Z_T^i and Z_C^i, and estimate a corresponding ATT $A(i)$ using (2). The vector $A \in \mathbb{R}^{m \times 1}$ contains the estimated ATT in all m subspaces.

3) Finally, we choose the median value of the entries of the vector A as our estimate. This enhances the robustness of the estimator and is expected to reduce its total bias (an insight that was corroborated by our experiments).

The above steps are summarized in Algorithm 1. Note that, according to the JL lemma, the embedding dimension k does not depend on the original dimensionality of data. This makes the approach applicable to very high-dimensional data. Moreover, an attractive feature of random projections is that they preserve the privacy of data, which can be critical in applications such as digital marketing where the customer information is sensitive (we do not elaborate on this issue further here; see [Lindell and Omri, 2011; Ahmed *et al.*, 2013]).

4 Experiments

In this section we empirically validate the proposed estimator on synthetic, semi-synthetic, and real-world digital marketing Campaign data.

4.1 Synthetic Dataset

The synthetic dataset is generated by following the procedures introduced in [Sun *et al.*, 2015]. In particular, the sample size N is set to 1000, and the number of features d is set to 200. We define the following basis functions: $f_1(x) = -2\sin(2x)$, $f_2(x) = x^2 - 1/3$, $f_3(x) = x - 0.5$, $f_4(x) = e^{-x} - e^{-1} - 1$, $f_5(x) = (x - 0.5)^2 + 2$, $f_6(x) = \mathcal{I}_{x>0}$, $f_7(x) = e^{-x}$, $f_8(x) = \cos(x)$, $f_9(x) = x^2$, and $f_{10}(x) = x$. The features $x_1, x_2, \cdots, x_d$ are drawn independently from the normal distribution $\mathcal{N}(0, 1)$. The binary treatment variable T is defined as $T|x = 1$ if $\sum_{j=1}^{5} f_j(x_j) > 0$ and $T|x = 0$ otherwise. Finally, the model for generating outcome is $Y|x, T \sim \mathcal{N}(\sum_{j=1}^{5} f_{j+5}(x_j) + T, 1)$. In this synthetic dataset, the first five features are correlated to the treatment and outcome, simulating a confounding effect, while the rest of the features are noisy covariates. The true causal effect (i.e., the ground truth value of ATT) in this dataset is 1.

We compare our matching estimator with the following baseline methods: Raw Space (Euclidean distance), Mahalanobis, PSM, PCA, and LPP. The first two baselines match units in the high-dimensional space by using different distance measures. PSM estimates the propensity score of each unit using logistic regression, and matches units with similar scores. The latter two perform nearest neighbor matching in low-dimensional subspaces computed by PCA and LPP, respectively.

To test the quality of each estimator, we carried out the above data generation process 1000 times and we estimated the ATT for each estimator. In Figure 1 we report the mean square error (MSE) of each estimator. (The standard error is also shown as error bars). For our Randomized NNM, we set the number of random projections m to 50 (we got similar results for a wide range of values for m; results omitted.) From Figure 1 we observe that the proposed estimator achieves lower MSE than all other methods when the dimension is lower than 60. The JL bound for $N = 100$ is about 7, which is consistent with Figure 1; our estimator achieves the lowest MSE when the dimension varies from 5 to 10.

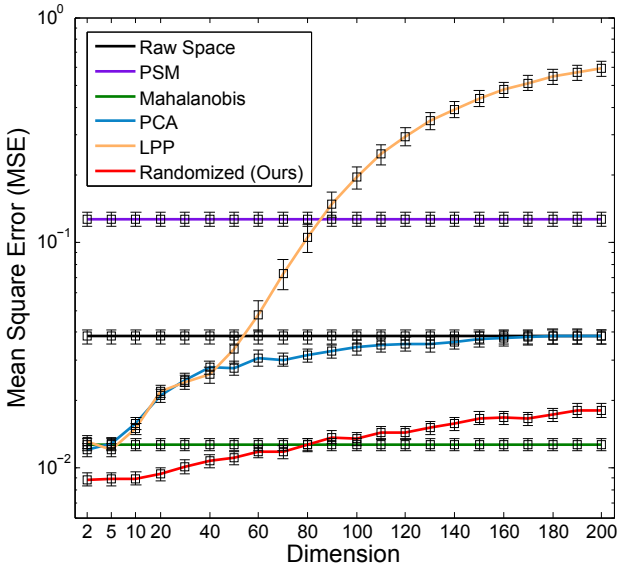
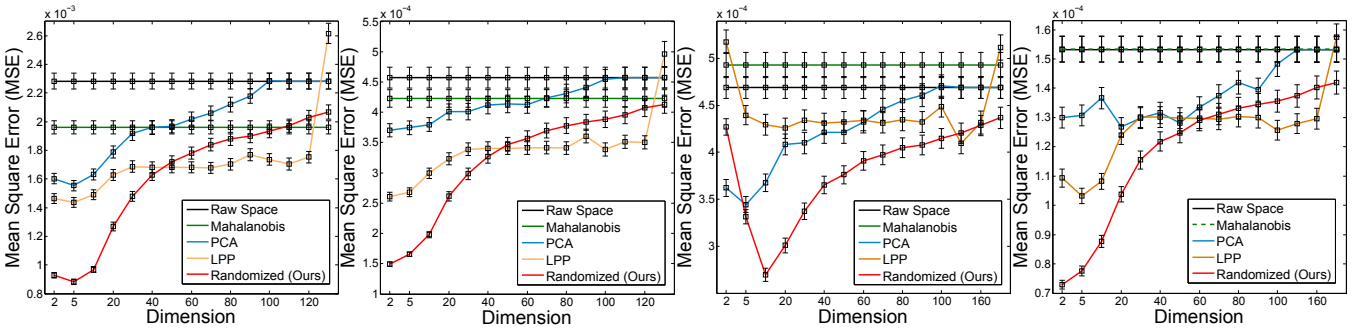


Figure 1: MSE of different estimators on the synthetic dataset. Note that Raw Space and Mahalanobis only involve matching in the original 200 dimensional data space.

4.2 Digital Marketing Campaign Data

We also evaluate the performance of our estimator and baselines on a real-world digital marketing campaign dataset. The dataset contains some covariates, such as customer profile and purchase activities, which are encoded by 209-dimensional binary vectors. It also includes the responses of customers to various promotions offered during the period of about one month. During this period, many campaigns were launched that involved sending promotions to customers via email, aiming at convincing the customers to purchase. In particular, several related promotions were sent to different groups of customers at about the same time. Specifically, two promotional emails were separately sent to two groups of customers, with the only difference between them being the offered discount: The first group involved offers with low discount, while the second group high discount. Due to the non-randomized (and unknown to us) assignment of promotions to customers, it is challenging to answer questions such as which promotion actually performed better than the other. Even worse, we may obtain confusing or misleading results if directly counting the average customer responses in the raw data. To see this, in the above example there were 0.8 million customers who received low discount, and 1.2 million customers who received high discount. The click response rate in the 0.8-million group was 2.24%, while this rate in the other group was only 1.16%. Without a careful causal analysis these rates could give the false impression that customers who received low discount were actually more likely to click the advertisement than those who received high discount, which is counter-intuitive and probably a wrong conclusion.

We apply matching estimators to this campaign data in order to infer the true causal effect of different promotions. Customers who received high discount promotion form the *treatment* group, and those with low discount form the *control* group. The question that we want to answer (and which is very relevant to a marketer) is whether the high discount promotion attracts more responses (e.g., open advertisements, click offers) than the low discount one. We design two empirical studies to compare the performance of our estimator against the baselines. First, we conduct a semi-synthetic experiment by virtue of pseudo-treated samples and we estimate the MSE of each estimator (here the ground truth is known). Second, we perform matching on the campaign data with real responses and we estimate the ATT (here the ground truth is unknown).



(a) N=750, Open Response (b) N=750, Click Response (c) N=3000, Open Response (d) N=3000, Click Response

Figure 2: MSE of semi-synthetic experiment on Campaign Data. Note that the Raw Space and Mahalanobis only perform nearest matching in the original 209 dimensional data space.

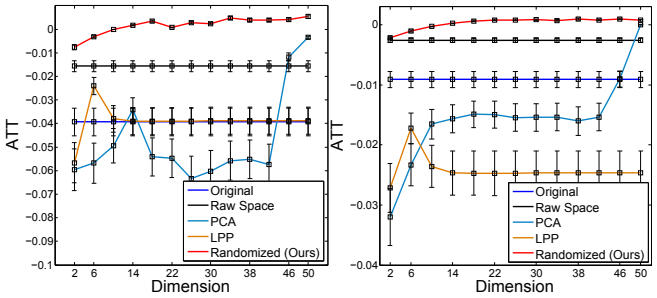
Semi-synthetic Experiments

We follow the experimental design in [Frölich *et al.*, 2015] that goes as follows. First, for each treated sample (i.e., the customer who received high discount promotion), we find its nearest neighbor in the control group (i.e., customers who received low discount promotion) in terms of Euclidean distance of the covariate vectors. The treated samples are then discarded and do not play any further role in the experiment. The matched control samples form the *pseudo-treated* 'population', and the remaining control samples form the control 'population'. Then, we repeatedly draw samples with replacement out of the 'populations', consisting of 50% pseudo-treated and 50% controlled units. Note that the true causal effect in this experiment is zero, as the pseudo-treated units did not (by default) receive any treatment.

In our experiment, the pseudo-treated 'population' contains 250,000 samples and the control 'population' contains 350,000 samples. We randomly choose subsamples from the populations for evaluation. We consider two different sample sizes, $N = 750$ and $N = 3000$, and we draw 4000 times for $N = 750$ and 1000 times for $N = 3000$ (those numbers are suggested by [Frölich *et al.*, 2015] based on theoretical arguments; we refer to their paper for details).

We considered two types of responses, *open* and *click*. The former one indicates whether a customer is attracted by the promotion, and the latter one implies a stronger incentive for purchase. We note that the responses in this campaign data are very sparse. The sparsity of *open* and *click* responses were 98.05% and 86.77%, respectively. In Figure 2 we show the MSE of different matching estimators for the two responses. We observe the following:

- Matching in the original high-dimensional space (such as Raw Space and Mahalanobis) gives poor results in general, which is consistent with the theory [Abadie and Imbens, 2006]. It validates the necessity of reducing the dimensionality of the data prior to matching.
- PCA and LPP can sometimes achieve good performance, but they are quite sensitive to the choice of reduced dimension, which makes them unstable and hence impractical for real-world applications.
- The proposed Randomized NNM estimator achieves the lowest MSE in every case. Interestingly, the largest gap to the other methods seems to appear close to the JL bound. For example, in the case of N=3000 in the Open Response, the JL bound is $\log(3000) \approx 8$, which is roughly the value at which the curve of Randomized NNM is peaked, see Figure 2(c).



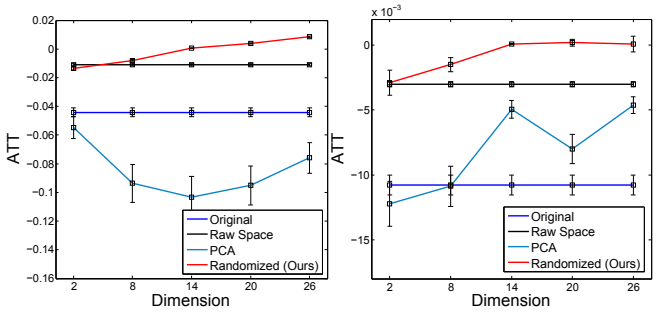
(a) N=20K, Open Response (b) N=20K, Click Response

Figure 3: ATT estimated by different methods on 20K Campaign Data. Note that the Raw Space only performs nearest matching in the original 209 dimensional data space.

Experiments on Data with Actual Treatments

Finally, we quantitatively analyze the performance of different estimators on the above campaign data using the actual treatments. That is, we estimate the causal effects of the campaign by matching the actual treated samples to the control samples. As the ground truth of the causal effect is unknown, we show the estimated ATT for Open and Click responses in Figure 3 (a) and (b), respectively. The sample size is 20,000, which consists of 8,000 treated samples and 12,000 control samples. We also examine the performance of the various estimators by drawing a larger sample set of 200,000 data points, while maintaining the ratio of treated to control samples. Figure 4 shows the results in this case. (LPP is not included because of its large computational cost on large-scale data.)

From Figure 3 and Figure 4 we observe that the original values of ATT (blue curves) are negative, which were calculated by (naively) comparing the response rates in two groups



(a) N=200K, Open Response (b) N=200K, Click Response

Figure 4: ATT estimated by different methods on 200K Campaign Data. Note that the Raw Space only performs nearest matching in the original 209 dimensional data space.

(see discussion in the first paragraph of this section). Such an estimated value of ATT would (wrongly) indicate that the low discount promotion is more effective than the high discount one. All the other baselines estimators also achieve negative ATT values. Only our estimator attains positive ATT, for embedding dimensions that are consistent with the JL lemma. This provides further evidence that the proposed estimator can draw meaningful causal conclusions in practice and it is robust under different settings.

5 Conclusions

We proposed a new matching estimator for estimating causal effects in digital marketing and related applications. The proposed estimator is very simple. It projects the data to several random linear subspaces, and estimates the median treatment effect by nearest-neighbor matching in each subspace. The Johnson-Lindenstrauss lemma guarantees that the neighborhood structure of the data is approximately preserved in each low-dimensional embedding, and hence matching can be done in the reduced spaces without loss. We evaluated the new estimator, as well as several baselines, on synthetic, semi-synthetic, and real-world experiments. The results attest to the quality of the proposed estimator over baseline methods.

Several important problems remain to be addressed in future work, such as dealing with the lack of common support between the treatment and control distributions [Hill and Su, 2013], identifying critical subsets of covariates [Entner *et al.*, 2013; Silva and Evans, 2014], and dealing with highly sparse responses (e.g., purchases in digital marketing campaigns).

Acknowledgments

Part of this work was completed while the first author was an intern at Adobe Research, San Jose CA. The second author would like to thank Ricardo Silva for his feedback and for pointers to the literature. This research is supported in part by the National Science Foundation (NSF) CNS award 1314484, NSF IIS award 1449266, Office of Naval Research (ONR) award N00014-12-1-1028, ONR Young Investigator Award N00014-14-1-0484, Naval Postgraduate School (NPS) award N00244-15-1-0041 and U.S. Army Research Office Young Investigator Award W911NF-14-1-0218, and a Research Award from Adobe.

References

[Abadie and Imbens, 2006] Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

[Achlioptas, 2001] Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 274–281. ACM, 2001.

[Ahmed *et al.*, 2013] Faraz Ahmed, Rong Jin, and Alex X Liu. A random matrix approach to differential privacy and structure preserved social network graph publishing. *arXiv preprint arXiv:1307.0475*, 2013.

[Ailon and Chazelle, 2006] Nir Ailon and Bernard Chazelle. Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform. In *Proceedings of the Thirty-Eighth Annual ACM Symposium on Theory of Computing*, pages 557–563. ACM, 2006.

[Barajas *et al.*, 2015] Joel Barajas, Ram Akella, Aaron Flores, and Marius Holtan. Estimating ad impact on clicker conversions for causal attribution: A potential outcomes approach. In *15th SIAM International Conference on Data Mining*, pages 640–648. SIAM, 2015.

[Bingham and Mannila, 2001] Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 245–250. ACM, 2001.

[Bryson *et al.*, 2002] Alex Bryson, Richard Dorsett, and Susan Purdon. The use of propensity score matching in the evaluation of active labour market policies. Working paper number 4, Department for Work and Pensions, 2002.

[Caliendo and Kopeinig, 2008] Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22(1):31–72, 2008.

[Chan *et al.*, 2010] David Chan, Rong Ge, Ori Gershony, Tim Hesterberg, and Diane Lambert. Evaluating online ad campaigns in a pipeline: causal models at scale. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 7–16. ACM, 2010.

[Cochran and Rubin, 1973] William G Cochran and Donald B Rubin. Controlling bias in observational studies: A review. *Sankhya: The Indian Journal of Statistics, Series A*, pages 417–446, 1973.

[Dalessandro *et al.*, 2012] Brian Dalessandro, Claudia Perlich, Ori Stitelman, and Foster Provost. Causally motivated attribution for online advertising. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, page 7. ACM, 2012.

[Dehejia and Wahba, 2002] Rajeev H Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161, 2002.

[Entner *et al.*, 2013] Doris Entner, Patrik Hoyer, and Peter Spirtes. Data-driven covariate selection for nonparametric estimation of

causal effects. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 256–264, 2013.

[Frölich *et al.*, 2015] Markus Frölich, Martin Huber, and Manuel Wiesenfarth. The finite sample performance of semi-and non-parametric estimators for treatment effects and policy evaluation. *IZA Discussion Paper No. 8756*, 2015.

[Glass *et al.*, 2013] Thomas A Glass, Steven N Goodman, Miguel A Hernán, and Jonathan M Samet. Causal inference in public health. *Annual Review of Public Health*, 34:61–75, 2013.

[Gu and Rosenbaum, 1993] Xing Sam Gu and Paul R Rosenbaum. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420, 1993.

[Hainmueller, 2012] Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

[He and Niyogi, 2004] Xiaofei He and Partha Niyogi. Locality preserving projections. In *Advances in Neural Information Processing Systems*, pages 153–160, 2004.

[Heckman *et al.*, 1998] James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.

[Hegde *et al.*, 2008] Chinmay Hegde, Michael Wakin, and Richard Baraniuk. Random projections for manifold learning. In *Advances in Neural Information Processing Systems*, pages 641–648, 2008.

[Hill and Su, 2013] Jennifer Hill and Yu-Sung Su. Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on childrens cognitive outcomes. *The Annals of Applied Statistics*, 7(3):1386–1420, 2013.

[Hill *et al.*, 2015] Daniel N Hill, Robert Moakler, Alan E Hubbard, Vadim Tsemekhman, Foster Provost, and Kiril Tsemekhman. Measuring causal impact of online actions via natural experiments: application to display advertising. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1839–1847. ACM, 2015.

[Iacus *et al.*, 2011] Stefano M Iacus, Gary King, and Giuseppe Porro. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24, 2011.

[Johnson and Lindenstrauss, 1984] W Johnson and J Lindenstrauss. Extensions of Lipschhitz maps into a Hilbert space. *Contemporary Math*, 26:189–206, 1984.

[Jolliffe, 2002] Ian Jolliffe. *Principal Component Analysis*. John Wiley and Sons, 2002.

[King and Nielsen, 2016] Gary King and Richard Nielsen. Why propensity scores should not be used for matching. Working paper, 2016.

[King and Zeng, 2006] Gary King and Langche Zeng. The dangers of extreme counterfactuals. *Political Analysis*, 14(2):131–159, 2006.

[Lindell and Omri, 2011] Yehuda Lindell and Eran Omri. A practical application of differential privacy to personalized online advertising. *IACR Cryptology ePrint Archive*, 2011:152, 2011.

[Maillard and Munos, 2012] Odalric-Ambrym Maillard and Rémi Munos. Linear regression with random projections. *The Journal of Machine Learning Research*, 13(1):2735–2772, 2012.

[Morgan and Winship, 2014] Stephen L Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge University Press, 2014.

[Neyman, 1923] Jerzy Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–480, 1923.

[Pearl, 2009a] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[Pearl, 2009b] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[Pearl, 2010] Judea Pearl. The foundations of causal inference. *Sociological Methodology*, 40(1):75–149, 2010.

[Peikes *et al.*, 2008] Deborah N Peikes, Lorenzo Moreno, and Sean Michael Orzol. Propensity score matching: A note of caution for evaluators of social programs. *The American Statistician*, 62(3):222–231, 2008.

[Roberts *et al.*, 2015] Margaret E Roberts, Brandon M Stewart, and Richard Nielsen. Matching methods for high-dimensional data with applications to text. Working paper, 2015.

[Rosenbaum and Rubin, 1983] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[Rosenbaum and Rubin, 1985] Paul R Rosenbaum and Donald B Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

[Rubin, 1973a] Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, pages 159–183, 1973.

[Rubin, 1973b] Donald B Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, pages 185–203, 1973.

[Rubin, 1974] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[Rubin, 1979] Donald B Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366):318–328, 1979.

[Silva and Evans, 2014] Ricardo Silva and Robin Evans. Causal inference through a witness protection program. In *Advances in Neural Information Processing Systems*, pages 298–306, 2014.

[Spirtes *et al.*, 2000] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[Stuart, 2010] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1–21, 2010.

[Sun *et al.*, 2015] Wei Sun, Pengyuan Wang, Dawei Yin, Jian Yang, and Yi Chang. Causal inference via sparse additive models with application to online advertising. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 297–303, 2015.

[Wang *et al.*, 2015] Pengyuan Wang, Dawei Yin, Jian Yang, Yi Chang, and Marsha Meytlis. Rethink targeting: detect 'smart cheating' in online advertising through causal inference. In *Proceedings of the 24th International Conference on World Wide Web Companion*, pages 133–134, 2015.