Probabilistic Matching: Incorporating Uncertainty to Correct for Selection Bias

Hui Fen Tan
Cornell University
Ithaca, NY 14850
ht395@cornell.edu

Giles J. Hooker
Cornell University
Ithaca, NY 14850
gjh27@cornell.edu

Martin T. Wells
Cornell University
Ithaca, NY 14850
mtw1@cornell.edu

Abstract

Matching methods such as propensity score matching are commonly used to construct artificial treatment and control groups from observational data, to determine the causal effect of treatment. However, propensity scores, once estimated, are frequently treated as known, and the uncertainty inherent in their estimation is ignored. We introduce probabilistic matching, an improvement on propensity score matching, that incorporates the uncertainty of the estimated propensity score into the subsequent matching process by weighting matches by the estimated probability of matching. Notably, this is equivalent to averaging the estimated treatment effect over the propensity score distribution, given the data. Preliminary results demonstrate that our approach achieves comparable or lower bias and lower variance, when compared to vanilla propensity score matching. While we focus on matching in this paper, the idea of incorporating uncertainty can also be brought into other ways of utilizing estimated propensity scores, such as weighing and substratification.

1 Introduction

Causal inference on observational data is challenging, as observational data is plagued with treatment selection bias [1]. Matching methods, commonly using propensity scores - the probability of treatment selection conditional on observed features [2] - aim to construct artificial treatment and control groups with similar distributions for observed features. If the unconfoundedness assumption holds [2], where treatment selection is independent of outcome given features, any differences between the treatment and control groups must necessarily be only due to the treatment, and hence the desired causal effect of treatment can be obtained.

However, propensity scores, however they are estimated, are frequently treated as known [1], and the uncertainty inherent in the propensity score estimation model is ignored. We motivate our proposed method, probabilistic matching, as a means of accounting for the uncertainty of our propensity score within our estimate of the treatment effect as well as in its uncertainty. Weighting by the estimated probability of matching is equivalent to averaging the estimated treatment effect over the propensity score distribution, given the data. This treats the propensity score as a random effect and we expect our procedure to regularize our estimates.

Previous work to address uncertainty in the estimation of propensity scores has been Bayesian in flavor, including the joint estimation of propensity scores and treatment effects, then sampling from the joint posterior distribution [3, 4], and a fully-Bayesian treatment using Bayesian model averaging [3]. In this paper, we offer a frequentist approach.

Presented at NIPS 2016 Workshop: What If? Inference and Learning of Hypothetical and Counterfactual Interventions in Complex Systems

2 Method

We motivate our method with the illustration in Figure 1 which plots the densities of a randomly selected treated observation and three closest controls in terms of estimated propensity scores. The figure illustrates that the propensity score model is imperfect, but also the difficulty of picking between any of these candidates for matching, especially when the differences in estimated propensity scores appear miniscule.

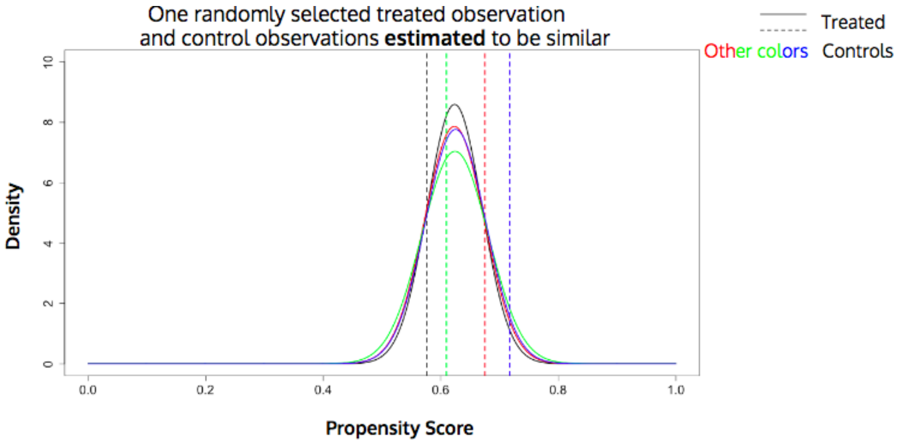

Figure 1: Densities of one randomly selected treated observation and three closest controls in terms of *estimated* propensity scores.. The dotted vertical lines are the *true* propensity scores.

We propose a variation on propensity score matching that incorporates uncertainty into the matching process by comparing the distributions of pairs of candidate matches to estimate the probability of matching.

Estimating the probability of matching. Denote the estimated propensity scores for a treated observation, T, and control candidates for matching, C_j, where $j = 1, \dots, n_c$ where c is the number of control observations in the data set. Considering their asymptotic normality property from the maximum likelihood estimation of the logistic regression coefficients, we can write:

$$T \sim N(\mu_T, \sigma_T^2) \quad C_j \sim N(\mu_{C_j}, \sigma_{C_j}^2) \text{ for } j = 1, \dots, n_c \tag{1}$$

with $Cov(T, C_j) = \sigma_{T,C_j}$. From (1), it follows that:

$$T - C_1 \sim N(\mu_{T-C_1}, \sigma_T^2 + \sigma_{C_1}^2 + 2\sigma_{T,C_1})$$
$$T - C_2 \sim N(\mu_{T-C_2}, \sigma_T^2 + \sigma_{C_2}^2 + 2\sigma_{T,C_2})$$

with

$$\begin{aligned} Cov(T - C_1, T - C_2) &= Var(T) - Cov(T, C_2) - Cov(T, C_1) + Cov(C_1, C_2) \\ &= \sigma_T^2 - \sigma_{T,C_2} - \sigma_{T,C_1} + \sigma_{C_1,C_2}. \end{aligned}$$

We observe that the probability that C_1 instead of C_2 is matched to T is

$$P(T - C_1 < T - C_2) = P((T - C_2) - (T - C_1) > 0). \tag{2}$$

Since we have that

$$(T - C_2) - (T - C_1) \sim N(\mu_{C_1 - C_2}, 4\sigma_T^2 + \sigma_{C_1}^2 + \sigma_{C_2}^2 + 2\sigma_{C_1,C_2}). \tag{3}$$

Equation (2) can be calculated directly or approximated by simulating repeatedly from the distribution in (3) and counting how many times the simulated points exceed zero. Equation (2) is exactly the estimated probability of matching T to C_1 instead of C_2, and we call this probability a weight, $w(C_1, C_2)$, where the ordering of the arguments matters.

The significance of 0.5. When the weight function $w(C_1, C_2) = 0.5$, C_1 and C_2 have equal probability of being selected for matching with T. Hence, we use 0.5 to rank candidates for matching (details below). In the following five variations on our probabilistic matching method, 0.5 plays a role in the first four.

Let C_1 be the control with estimated propensity closest to that of T, and compute $w(C_1, C_j)$ for $j = 1, \ldots, n_c$.

1. k=1: Let $C_2 = \arg\min_{j=1,\ldots,n_c} |w(C_1, C_j) - 0.5|$. The counterfactual outcome for T is calculated as a weighted average of C_1 and C_2's outcomes:
$$\text{Counterfactual } Y_T = w(C_1, C_2)Y_{C_1} + (1 - w(C_1, C_2))Y_{C_2}.$$
2. k=3: Let C_2, C_3, C_4 be the three controls with minimum $|w(C_1, C_j) - 0.5|$. The counterfactual outcome for T is similarly as a weighted average of C_1 to C_4's outcomes:
$$\text{Counterfactual } Y_T = \sum_{j=2}^{4} \{w(C_1, C_j)Y_{C_1} + (1 - w(C_1, C_j))Y_{C_j}\}/3.$$
3. 0.48 ≤weight ≤ 0.52: Let $j = C_2, \ldots, C_{k+1}$ be the set of controls for which $0.48 \le w(C_1, C_j) \le 0.52$.
$$\text{Counterfactual } Y_T = \sum_{j=2}^{k+1} \{w(C_1, C_j)Y_{C_1} + (1 - w(C_1, C_j))Y_{C_j}\}/k. \quad (4)$$
4. 0.45 ≤weight ≤ 0.55: Let $j = C_2, \ldots, C_{k'+1}$ be the set of controls for which $0.45 \le w(C_1, C_j) \le 0.55$. Then we have equation (4) with k' in lieu of k.
5. All weights: All controls are used, making this variation attractive because it does not require parameter tuning for k, w_{lower}, or w_{upper}. The hope is that the effect of bad controls is diminished because their weights going into equation (4) are small.

Here we pick two options for each of k, w_{lower}, or w_{upper} for illustrative purposes, but emphasize the need for tuning.

Estimating ATE. $ATE = E(Y_{\text{treated}} - Y_{\text{control}}) = E(E(Y_{\text{treated}} - Y_{\text{counterfactual}}|X))$.

Propensity Score Estimation. While we estimated propensity scores using logistic regression, any propensity score estimation method with computable uncertainty could be used.

3 Experimental Setup

3.1 Evaluation Criteria

We compare methods by computing absolute bias of estimated ATE compared to ground truth, as well as the standard error of estimated ATE across simulations.

3.2 Competitor Methods

1. **Propensity score [2].** Propensities for treatment estimated using logistic regression, then one-to-one matching with replacement.
2. **Bayesian propensity score [4].** Joint posterior distribution of estimated propensities and outcomes is sampled from. Implementation used: R package IUPS [5].
3. **Mahalanobis distance [6].** Mahalanobis distance on features computed, then one-to-one matching with replacement. Note that treatment is not include in the computation of distance. Implementation used: R package Matching [7].
4. **Covariate-balanced propensity score [8].** The objective of achieving balance in treatment and control samples is built into the propensity score model, with generalized method-of-moments estimation. Implementation used: R package cbps [9].
5. **Boosted-regression estimated propensity score [10].** One of the first papers [1] to use more advanced methods, beyond logistic or probit regression, to estimate propensity scores. Implementation used: R package twang [11].

4 Results

1. **Simulated data** ($n = 100, p = 10$). We follow the simulation setup first developed by Setoguchi et al. [12] and extended by Lee et al. [13] and Austin [14]. In this setup, seven treatment scenarios were introduced, covering a variety of non-additivity and non-linearity in features. We demonstrate our method on two scenarios (italicized below), and leave the remaining for subsequent work:

   A: Additivity and linearity
   B: Mild non-linearity
   C: Moderate non-linearity
   D: Mild non-additivity
   E: Mild non-additivity and non-linearity
   F: Moderate non-additivity
   G: Moderate non-additivity and non-linearity

   For each scenario, we simulated 125 data sets, each of $n = 100$ observations and $p = 10$ features, a mix of continuous and categorical. Of these ten features, three are associated with only treatment selection, and three are associated with only the outcome. The remaining four are associated with both. The inclusion of features associated with not only treatment but also outcome in the treatment selection process is intended to make matching more challenging. The coefficients for the true treatment selection and outcome models are in Appendix A of [6].

Table 1: Methods compared on data A. True ATE equals -0.4. The * gives the value determined by the data.

Method	Estimated ATE			Weight		k
	Mean	SD	Bias	Min	Max	
Mahalanobis	-0.33	0.46	0.07			
Propensity	-0.37	0.61	0.03			
Boosting-estimated propensity	-0.38	1.02	0.02			
Covariate-balanced propensity	-0.38	0.58	0.02			
Bayesian propensity	-0.36	0.61	0.04			
Probabilistic, k=1	-0.35	0.52	0.05	0.5*		1
Probabilistic, k=3	-0.37	0.51	0.03	0.49*	0.51*	3
Probabilistic, $0.48 \leq$ weight ≤ 0.52	-0.39	0.51	0.01	0.48	0.52	8*
Probabilistic, $0.45 \leq$weight ≤ 0.55	-0.33	0.44	0.07	0.45	0.55	21*
Probabilistic, all weights	-0.29	0.46	0.11	0.43*	0.57*	48*

Vanilla propensity score matching performs well in Scenario A - unsurprising since the model is close to the true treatment model, with little nonlinearity or nonadditivity. For both scenarios, certain variations of probabilistic matching achieves bias comparable to vanilla propensity score matching, but with lower standard deviation. In Scenario G, the majority of probabilistic matching variations achieve lower bias. More advanced methods such as covariate-balanced propensities, and boosting-estimated propensities also demonstrate good performance.

Table 2: Methods compared on data G. True ATE equals -0.4. The * gives the value determined by the data.

Method	Estimated ATE			Weight		k
	Mean	SD	Bias	Min	Max	
Mahalanobis distance	-0.27	0.47	0.13			
Propensity	-0.33	0.57	0.07			
Boosting-estimated propensity	-0.54	1.01	0.14			
Covariate-balanced propensity	-0.37	0.57	0.03			
Bayesian propensity	-0.37	0.61	0.03			
Probabilistic, k=1	-0.39	0.54	0.01	0.5*		1
Probabilistic, k=3	-0.36	0.51	0.04	0.49*	0.51*	3
Probabilistic, $0.48 \leq$weight ≤ 0.52	-0.35	0.49	0.05	0.48	0.52	9*
Probabilistic, $0.45 \leq$weight ≤ 0.55	-0.30	0.46	0.10	0.45	0.55	23*
Probabilistic, all weights	-0.26	0.46	0.14	0.44*	0.58*	48*

2. **Semi-simulated data, with simulated ground truth: Schafer Kang health survey data [15]** ($n = 200, p = 14$). To study if dieting makes adolescent girls depressed, Schafer & Kang simulated data based on the marginal distributions of features such as health characteristics and adolescent behavior from the National Longitudinal Study of Adolescent Health [16], a representative sample of American adolescents. To simulate ground truth, a simple, unconfounded treatment selection model, and a confounded model for outcome with interactions were used. Variants of probabilistic matching achieved lowest bias and standard deviation.

Table 3: Methods compared on Schafer Kang data. True ATE equals 0.003. The * gives the value determined by the data, with 0.51 being the maximum.

Method	Estimated ATE			Weight		k
	Mean	SD	Bias	Min	Max	
Mahalanobis distance	-0.03	0.08	0.033			
Propensity	-0.08	0.11	0.083			
Boosting-estimated propensity	-0.17	0.24	0.173			
Covariate-balanced propensity	-0.08	0.10	0.083			
Bayesian propensity	0.02	0.10	0.017			
Probabilistic, k=1	-0.06	0.09	0.063	0.49*		1
Probabilistic, k=3	-0.05	0.07	0.053	0.49*	0.50*	3
Probabilistic, $0.48 \leq$weight ≤ 0.52	-0.03	0.05	0.033	0.48	0.51*	29*
Probabilistic, $0.45 \leq$weight ≤ 0.55	0.03	0.08	0.027	0.45	0.51*	55*
Probabilistic, all weights	-0.01	0.07	0.013	0.44*	0.51*	60*

3. **Real data, with simulated ground truth: Lalonde job training data [6]** ($n = 445, p = 10$). The Lalonde data set is a classic, somewhat controversial data set often used to benchmark methods for causal inference on observational data [6]. The National Support Work program was a 1970s government-run job training program conducted as a randomized experiment, where eligible people were randomly selected to participate or not, with the goal of determining if the program increased participants' earnings. Lalonde made this randomized experiment observational by adding in observations from two observational surveys - PSID [17] and CPS [18] who resembled program participants. To simulate ground truth, a nonlinear treatment selection model, and a simple outcome model were used. With vanilla propensity scores performing badly, almost all other methods performed better. The best results were achieved by modern techniques, i.e. boosted-estimated propensity score.

Table 4: Methods compared on Lalonde data. True ATE equals $1,000. The * gives the value determined by the data.

Method	Estimated ATE			Weight		k
	Mean	SD	Bias	Min	Max	
Mahalanobis distance	401	606	599			
Propensity	235	966	765			
Boosting-estimated propensity	953	2095	47			
Covariate-balanced propensity	213	1036	787			
Bayesian propensity	349	1228	651			
Probabilistic, k=1	459	970	541	0.5*		1
Probabilistic, k=3	379	703	621	0.49*	0.50*	3
Probabilistic, $0.48 \leq$weight ≤ 0.52	360	727	640	0.48	0.52	135*
Probabilistic, $0.45 \leq$weight ≤ 0.55	477	1248	523	0.45	0.55	200*
Probabilistic, all weights	262	974	738	0.45*	0.54*	208*

5 Discussion and Ongoing Work

We estimated propensity scores using logistic regression, however our way of incorporating uncertainty translates to any estimation method with quantifiable uncertainty. Moreover, while we have focused on one-to-one matching with replacement as it is “more versatile" [19], being able to handle the case of more treated observations than controls, and generally has lower bias than one-to-one matching without replacement [1], we note the idea of incorporating the uncertainty in propensity score estimation can be brought into propensity score weighing and sub-stratification as well.

For tractable computation, we compared pairs of matches, but note that triplets or beyond, which give multivariate normal distributions, can also be used. In this case, estimating the weights by sampling from the distribution will be more tractable than calculating probabilities from large multivariate normal systems.

We are working on demonstrating our method on the remaining simulated data scenarios and more real data.

We have proposed a method to account for the uncertainty inherent in estimated propensity scores by weighing possible matches by the estimated probability of matching. By averaging the estimated treatment effect over the propensity score distribution. the propensity score is effectively being treated as a random effect to be marginalized out. Here, we have five variations on our method, but will go on to investigate other variations, vary the tuning parameters k, w_{lower}, or w_{upper}, as well as demonstrate the method on other data sets.

References

[1] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1, 2010.

[2] 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.

[3] David Kaplan and Cassie JS Chen. Bayesian propensity score analysis: Simulation and case study. *Society for Research on Educational Effectiveness*, 2011.

[4] Weihua An. Bayesian propensity score estimators: incorporating uncertainties in propensity scores into causal inference. *Sociological Methodology*, 40(1):151–189, 2010.

[5] Weihua An, Huizi Xu, and Zhida Zheng. Package iups: Incorporating uncertainties in propensity scores. 2013.

[6] Alexis Diamond and Jasjeet S Sekhon. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics*, 95(3):932–945, 2013.

[7] Jasjeet S Sekhon. Multivariate and propensity score matching software with automated balance optimization: the matching package for r. *Journal of Statistical Software*, 42(7), 2011.

[8] Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

[9] Christian Fong, Marc Ratkovic, Chad Hazlett, Xiaolin Yang, and Kosuke Imai. Package cbps. 2016.

[10] Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4):403, 2004.

[11] Greg Ridgeway, Dan McCaffrey, Andrew Morral, Beth Ann Griffin, and Lane Burgette. Package twang: Toolkit for weighting and analysis of nonequivalent groups. 2016.

[12] Soko Setoguchi, Sebastian Schneeweiss, M Alan Brookhart, Robert J Glynn, and E Francis Cook. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6):546–555, 2008.

[13] Brian K Lee, Justin Lessler, and Elizabeth A Stuart. Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346, 2010.

[14] Peter C Austin. Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based g-computation. *Multivariate behavioral research*, 47(1):115–135, 2012.

[15] Joseph L Schafer and Joseph Kang. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods*, 13(4):279, 2008.

[16] J Richard Udry. The national longitudinal study of adolescent health (add health), wave iii, 2001-2002. 2003.

[17] C Brown, VA Freedman, N Sastry, KA McGonagle, FT Pfeffer, RF Schoeni, and F Stafford. Panel study of income dynamics, public use dataset. *Ann Arbor, MI: University of Michigan*, 2014.

[18] Sarah Flood, Miriam King, Steven Ruggles, and J Robert Warren. Integrated public use microdata series, current population survey: Version 4.0.[machine-readable database]. *Minneapolis: University of Minnesota*, 2015.

[19] Jennifer Hill and Jerome P Reiter. Interval estimation for treatment effects using propensity score matching. *Statistics in medicine*, 25(13):2230–2256, 2006.