Statistica Neerlandica (2013) Vol. 67, nr. 2, pp. 211–226
doi:10.1111/stan.12004

An optimization approach for making causal inferences

Wendy K. Tam Cho*†

University of Illinois at Urbana-Champaign, Urbana, IL, USA

Jason J. Sauppe‡

University of Illinois at Urbana-Champaign, Urbana, IL, USA

Alexander G. Nikolaev§

University at Buffalo (SUNY), Buffalo, NY, USA

Sheldon H. Jacobson¶

University of Illinois at Urbana-Champaign, Urbana, IL, USA

Edward C. Sewell††

Southern Illinois University at Edwardsville, Edwardsville, IL, USA

To make causal inferences from observational data, researchers have often turned to matching methods. These methods are variably successful. We address issues with matching methods by redefining the matching problem as a subset selection problem. Given a set of covariates, we seek to find two subsets, a control group and a treatment group, so that we obtain optimal balance, or, in other words, the minimum discrepancy between the distributions of these covariates in the control and treatment groups. Our formulation captures the key elements of the Rubin causal model and translates nicely into a discrete optimization framework.

Keywords and Phrases: causal inference, matching, optimization, subset selection.

1 Experimental versus observational studies

Experimental studies are powerful because the experimental framework allows one to examine causal effects. Applying standard statistical models to non-experimental or observational data, on the other hand, generally allows the researcher to make

*wendycho@illinois.edu

†Professor in the Departments of Political Science and Statistics and Senior Research Scientist at the National Center for Supercomputing Applications.

‡Graduate student in the Department of Computer Science.

§Assistant Professor in the Department of Industrial and Systems Engineering.

¶Professor in the Department of Computer Science.

††Professor in the Department of Mathematics and Statistics.

Published by Wiley Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

associational inferences only. The key difference between experiments and observational studies is that in experiments, when randomization is successful, the treatment effect is isolated from potential confounders. Differences in response can thus be attributed to the treatment.

Experiments can be complex and multifaceted, but let us assume, for simplicity, that a subject is either treated ($T=1$) or not ($T=0$). For subject i, $i=1,\ldots,N$, the two potential outcomes are $Y_i(0)$ and $Y_i(1)$. The causal effect of the treatment, as measured by Y, on a particular subject i, is

$$Y_i(1) - Y_i(0). \tag{1}$$

The fundamental problem of causal inference is that it is impossible to observe the value of both $Y_i(1)$ and $Y_i(0)$ on the same subject because the subject has either been exposed to the treatment or has not. Only one of the terms in Equation (1) is observable (Holland, 1986).

The Rubin causal model reconceptualizes this framework so that either the outcome under treatment or under control, but not both, needs to be observed for each unit (Rubin, 1974; Rubin, 1978). Hence, one statistical solution to the fundamental problem of causal inference is to shift to an examination of an *average* treatment effect (ATE) over all of the subjects,

$$\text{ATE} = E(Y(1) - Y(0)) = E(Y(1)) - E(Y(0)). \tag{2}$$

A remaining issue for observational studies arises from the non-random nature of the subjects in the data set. One observes some set of subjects who have received a treatment, giving us $E(Y(1)\,|\,T=1)$. From this group, the average treatment effect for the treated (ATT) is

$$\text{ATT} = E(Y(1) - Y(0)|T = 1), \tag{3}$$

which quantifies the effect of the treatment on subjects that are treated. In general, $E(Y(1)) \neq E(Y(1)\,|\,T=1)$ and $E(Y(0)) \neq E(Y(0)\,|\,T=1)$. That is, the ATE, $E(Y(1)) - E(Y(0))$, and the ATT, $E(Y(1)\,|\,T=1) - E(Y(0)\,|\,T=1)$, are not generally interchangeable.

The ATE and the ATT would be interchangeable if the independence assumption – exposure to treatment is statistically independent of all other variables, including $Y(1)$ and $Y(0)$ – holds because conditioning on treatment is then irrelevant. This allows us to compute the ATE as $E(Y(1)\,|\,T=1) - E(Y(0)\,|\,T=1)$, but we must still determine how to compute $E(Y(0)\,|\,T=1)$, the untreated outcome for treated individuals. Notice here that if treatment is completely random, then a viable approach is to use the average outcome of similar subjects who were not exposed to treatment. We would then no longer require an observation of $Y_i(1)$ and $Y_i(0)$ from the *same* subject, but are able to use information from *different* subjects. If exposure to treatment satisfies the independence assumption, then those who have been treated give us information about $E(Y(1))$, whereas those who have not been treated give us information on $E(Y(0))$. Hence, the treatment effect can be calculated as

$$\mathrm{ATE} = \mathrm{ATT} = E(Y|T=1) - E(Y|T=0) = \frac{1}{N_t}\sum_{i\in\{T=1\}} Y_i(1) - \frac{1}{N_c}\sum_{i\in\{T=0\}} Y_i(0), \quad (4)$$

where N_t is the number of treated subjects, N_c is the number of control subjects, $\{T=1\}$ denotes the set of treated subjects, and $\{T=0\}$ denotes the set of control subjects.

In observational data, it is unusual for the independence assumption to hold. The treated group almost surely differs systematically from the non-treated group. Hence, if one wishes to make causal inferences from observational data, then the task at hand is to postprocess the observational data so that exposure to treatment satisfies the independence assumption. If this can be satisfactorily accomplished, then the postprocessed data will resemble a randomized experiment, and one can then straightforwardly compute the treatment effect.

2 Matching

'Matching' is a method for postprocessing observational data so that they resemble experimental data by simulating statistical independence of treatment exposure and all other available variables (Rubin, 1974; Rubin, 1977; Rubin, 1978). The problem involves two population groups, treated and control, and a set of pretreatment covariates, **X**. The objective is, given the treatment group, to identify a control group so that the treated and control covariate distributions are statistically indistinguishable, creating the 'appearance of randomization'. If treatment is completely random for similar individuals, then the *unconfoundedness* or the *selection on observables* assumption is satisfied. Formally, if

> Assumption 1: *T is independent of* $Y(0)$ and $Y(1)$, *conditional on* $\mathbf{X}=x$, and
> Assumption 2: $0 < P(T=1\,|\,\mathbf{X}=x) < 1$,

hold, treatment assignment is 'strongly ignorable' (Rosenbaum and Rubin, 1983). The driving goal of matching is to postprocess observational data so that treatment assignment is strongly ignorable.

The first step in matching is to establish a distance metric that quantifies the difference between two subjects on the basis of their covariates. The second step is to match subjects so that this distance metric is minimized across all matches.

Matching methods are variably successful, sometimes failing to replicate the results of corresponding randomized controlled trials (LaLonde, 1986; Dehejia and Wahba, 1999; Smith and Todd, 2001).[1] Diamond and Sekhon (2012) document shortcomings and propose a genetic algorithm to identify a distance metric that results in better covariate balance. They posit that the long-running debate between Dehejia and Wahba (1999, 2002) and Smith and Todd (2001, 2005a, 2005b) is largely a result of researchers using matching methods that have not achieved good, or 'good enough', balance in the covariates. In particular, although the original LaLonde data analysis claimed that the balance was good, subsequent analyses demonstrated that balance could have been better. This debate highlights a key deficiency with matching methods – *there is no baseline to judge the success of the matching procedure in achieving balance*. Matching produces a set of control subjects that are similar to treated subjects, but we are unsure whether we have identified the most similar set of subjects or whether there is a better balanced set that might result in a different estimated treatment effect.

Rosenbaum and Rubin (1985) also illustrate the goal of covariate balance and the uncertainty of having achieved it. They used three different matching methods (nearest available matching on the estimated propensity score, Mahalanobis metric matching including the propensity score, and nearest available Mahalanobis metric matching within calipers defined by the propensity score) to obtain three different matched samples. They stated that '[t]he third matching method—Mahalanobis metric matching within propensity score calipers—appears clearly superior' because it resulted in the best covariate balance (Rosenbaum and Rubin, 1985, p. 38). Rosenbaum, Ross and Silber (2007) explain that 'one can construct several matched samples by different methods and select for use the sample that produces the most satisfactory balance on covariates'. This work highlights that there are multiple methods by which one might obtain a matched sample. The clear message is that any one method may not identify the best balanced matched set; so, different methods should be explored to identify the one that yields the best balance.

Plainly, the various works of Rubin and Rosenbaum make it clear that the critical assessment factors are not in the *individual matches*, but rather on the resulting treatment and control *covariate distributions*. That is, they assess statistical independence of the covariates with treatment exposure by examining covariate balance at the aggregate distribution level, not at the level of individual matches. The individual matches are important insofar as they provide a means for achieving balance in the covariate distributions. Indeed, if the individual matches are not sufficiently similar, then the overall covariate distributions will not be balanced, and the estimate of the treatment effect will not be unbiased (Rosenbaum and Rubin, 1983). We know, however, from logic, that the converse of this statement is not necessarily true. If the covariate distributions are balanced, then the origin of this balance need not be from individual matches. Moreover, individual matching procedures are varyingly successful in this venture to obtain covariate balance. Success depends on a variety of factors such as the closeness of the matches, the specification of the propensity score, the nature of the underlying data, the metric to assess the closeness of a match, and the algorithm for pairing subjects, to name but a few. If individual matching is not necessary and the methods are inconsistently successful, then the door is open for exploring other fruitful methods for obtaining covariate balance.

3 Randomized experiments and twin studies

An important insight from the experimental realm is that there are different ways to conduct experiments to isolate treatment effect. Matching most closely resembles an identical twin framework where the data are comprised of identical twins where one is treated but not the other. In these studies, subjects are not randomly drawn. Covariate balance is attained because the subjects are identical twins. Twins studies comprise an experimental framework but are not representative of all experimental frameworks. The twin framework differs from randomized trials where subjects are randomly selected from a population, and then randomly assigned to treatment or control. A successful randomization process produces treatment and control *groups* with covariate distributions that are statistically indistinguishable. The covariates are balanced; but unlike the twin framework, the subjects in the treated group do not have a matching twin in the control group. Both randomized trials and twin studies are valid experimental frameworks, and when successfully implemented, isolate the treatment effect.

Because the critical assessment of balance is at the covariate distribution level rather than at the level of individual matches, it makes sense to focus on the covariate distributions rather than the individual/twin matches, which are not necessary for obtaining balance. There has been some shift in focus in this direction as Rosenbaum *et al*. (2007, p. 80) make clear that '[b]alance refers to the distribution of the covariate in treated and control groups after matching, rather than to close matches in each and every pair. For instance, there is balance on diabetes if the proportion of diabetics is about the same in treated and control groups after matching, even if diabetics are not always matched to other diabetics'. Ensuring a close match in the covariate distributions is a departure from the traditional manner in which matching has been performed, but comports well with the idea of balance.[2]

Because twin studies comprise only a particular subset of experiments, the results from matching methods (intended to replicate twin studies) are subsumed by results from a method that focuses on covariate distributions. If matching results in balanced covariate distributions, then a method that balances covariate distributions will find the same results that would be obtained by a matching procedure as well as results that are consistent with a randomized experiment but would elude matching procedures. If one is interested in estimating a treatment effect, then a twin study is one way in which one might isolate and estimate a treatment effect, but it is, by no means, the only way to do so. As such, the matching framework is one way in which one might isolate a treatment effect but is also not the only way to do so. In this paper, we move away from a narrow focus on identical twin experiments and present and discuss methods that emulate the larger class of randomized experiments. In particular, we propose a *focus shift from individual matches and twin studies to covariate distributions and randomized experiments writ large*.

4 Balance optimization subset selection

Rather than searching for twins or close individual matches, our method seeks to identify a treatment group and a control group that resemble two groups that have been randomly drawn from a population. In other words, we seek to postprocess the data so that they resemble a randomized control trial rather than an identical twin experiment.

Our study design, Balance Optimization Subset Selection (BOSS), recognizes the matching problem as an optimization problem. One insight is that the goal of causal inference methods is to optimize the level of balance. Matching procedures currently match first then assess the success of the matching later by the level of balance achieved. Without knowing how all matching methods perform, it is difficult to assess if balance is good or 'good enough' because the baseline or optimal level of balance in a particular data set is unknown. In our formulation, our goal is optimal balance, not 'good balance'. The optimal level of balance is the baseline or standard for assessing any particular balance level. BOSS reframes the causal inference problem from a matching problem to a subset selection problem where the goal is to find S^T, a subset of the treatment pool, and S^C, a subset of the control pool, so that a measure of balance, $b(S^T,S^C)$, is maximized. This discrete optimization problem can be addressed using operations research algorithms and heuristics in a flexible formulation where any measure of balance can be incorporated into the objective function. The end goal, balance in the covariate groups, remains the same.

For illustration and proof of concept, we present one simple implementation of this optimization problem that incorporates data bins. In the BOSS with Bins (BOSS-B) framework, we create sets of B uniformly sized data bins for each covariate and assign covariate values to the bin whose value range includes it.

When the number of bins, B, is small, many different covariate values are mapped to the same bin. However, when B is large, the increased granularity results in bins that house a smaller range of covariate values. More formally, for covariate X_p, the covariate values lie in the closed set, $[L_p,U_p]$, where $L_p = \min_{i\in T\cup C} X_{pi}$, and $U_p = \max_{i\in T\cup C} X_{pi}$. We can separate this range into B bins specified with $B+1$ breakpoints given by $L_p = t_0^p < t_1^p < t_2^p < \cdots < t_B^p = U_p$. These bins can be used to approximate the marginal and joint distributions of the covariates. For a set of P covariates, there exists a total of $K = P + \binom{P}{2} + \binom{P}{3} + \cdots + \binom{P}{P}$ marginal and joint distributions because there are P marginal distributions, $\binom{P}{2}$ joint distributions of two covariates, $\binom{P}{3}$ joint distributions of three covariates, and so forth, with $\binom{P}{P} = 1$ joint distribution that includes all P covariates.

The optimization routine seeks control units such that the control and treatment covariate distributions are as similar as possible. If there are two covariates, p_1 and p_2, then there are two marginal distributions and one joint distribution, for a total of $K = 3$ distributions. Using two bins per covariate ($B = 2$), the first marginal distribution is characterized by the set of bins for covariate $p_1\left(\left\{\left[t_0^{p_1}, t_1^{p_1}\right], \left[t_1^{p_1}, t_2^{p_1}\right]\right\}\right)$, whereas the second marginal distribution is characterized by the set of bins for covariate p_2 $\left(\left\{\left[t_0^{p_2,} t_1^{p_2}\right], \left[t_1^{p_2}, t_2^{p_2}\right]\right\}\right)$. The joint distribution is defined by the set of bins, $\left\{\left[t_0^{p_1}, t_1^{p_1}\right] \times \left[t_0^{p_2}, t_1^{p_2}\right], \left[t_1^{p_1}, t_2^{p_1}\right] \times \left[t_0^{p_2}, t_1^{p_2}\right], \left[t_0^{p_1}, t_1^{p_1}\right] \times \left[t_1^{p_2}, t_2^{p_2}\right], \left[t_1^{p_1}, t_2^{p_1}\right] \times \left[t_1^{p_2}, t_2^{p_2}\right]\right\}$. The optimization routine can be formulated to find balance for all, or any subset, of the K distributions. In practice, it is not necessary to optimize over all K distributions because the distributions have overlapping information. In general, any n-way ($n > 1$) distribution subsumes some number of lower-order distributions. The overall joint distribution encapsulates all lower-order marginal and joint distributions.

Suppose the bins of interest are ordered from $b = 1, 2, \ldots, N_B$ (where the specific ordering is inconsequential). Let $\#\{S_b\}$ denote the cardinality of set S with values in bin b. The objective of the BOSS-B optimization problem is to minimize $|\#\left(S_b^C\right) - \#(T_b)|$ over all bins. Any objective function that minimizes these terms may be used to evaluate the distribution fit. More formally, given a treatment group, T, of size N, a set of P pretreatment covariates, $\{X_1, X_2, \ldots, X_P\}$, and a set of N_B bins for the distributions of interest, find a subset $S^C \subset C$ of size N such that

$$\sum_{b=1}^{N_B} \frac{\left[\#\left(S_b^C\right) - \#(T_b)\right]^2}{\max(\#(T_b), 1)} \tag{5}$$

is minimized. This objective function (5) is similar in to the χ^2 goodness-of-fit test statistic.

5 Statistical properties of balance optimization subset selection

We now proceed to explore the statistical properties of the BOSS estimator by examining its performance on two different data sets. The first data set is a simulated data set. The second data set is the LaLonde data set that has been extensively examined in the context of causal inferences from observational data (LaLonde, 1986).

5.1 Simulated data set

For the simulated data set, we randomly generated three $N(0,1)$ pretreatment covariates, $\mathbf{X} = [X_1, X_2, X_3]$ (each of size 100,000) and a positive definite 3×3 variance–covariance matrix, Σ. The covariates in the treatment pool are created by multiplying the covariate matrix and the square root of the variance–covariance matrix ($\mathbf{X}\Sigma^{\frac{1}{2}}$). Covariate i in the control pool is generated with mean zero and variance

$s_{i1}^2 + s_{i2}^2 + s_{i3}^2$, where s_{ij} is the ijth element of Σ. This process ensures the same mean for corresponding covariates in the treatment and control pool, but allows the variances to differ. Next, we generated the response value for both the treatment and control pools through the linear response function,

$$Y = 14 + 7X_1 + 11X_2 - X_3 + \epsilon, \tag{6}$$

where $\epsilon \sim N(0,2)$. Because the same response function is used for both treatment and control, there is no treatment effect in the simulated data.

From our treatment pool, we non-randomly choose a treatment group of size 500 using a thinning algorithm.[3] Our particular algorithm heavily favors units with covariates values at the tails of its distribution. Figure 1 displays the treatment group and control pool covariate distributions.[4] There are three plots. The distribution for covariate X_1 is on the left. The plot for X_2 is in the middle, and the plot for X_3 is on the right. The filled-in bars are for the control group, whereas the unfilled bars are for the treatment group. As we can see, the distribution of covariates in the treatment group are bimodal rather than normally distributed as they are in the control pool. The difference in the distributions mimics a common pattern in observational data where those who choose to be treated are a non-random group with covariate distributions that do not resemble the covariate distribution of non-treated individuals. The entire control pool is used, and the goal is to identify control groups of size 500 with covariates that most closely match the covariate distribution in our treatment group.

We ran experiments for $B = 2$, 4, 8, 16, and 32 uniformly sized bins per covariate. Because these are powers of two, each larger set of bins simply divides the previous bin set in half. That is, for two bins, the thresholds are $t_0 < t_1 < t_2$, where $t_1 = (t_0 + t_2)/2$. For four bins, the thresholds are $t'_0 < t'_1 < \cdots < t'_4$, where $t'_0 = t_0$, $t'_1 = (t_1 + t_0)/2$, $t'_2 = t_1$, $t'_3 = (t_1 + t_2)/2$, and $t'_4 = t_2$. Each unit's covariate values, $\{X_{1i}, X_{2i}, \ldots, X_{ki}\}$, are placed into the bin whose range includes that value, $\{X'_{1i}, X'_{2i}, \ldots, X'_{ki}\}$, where $X'_{ki} = j$ if $t_{j-1}^k \leq X_{ki} \leq t_j^k$. Bins are created for the marginal distributions only.

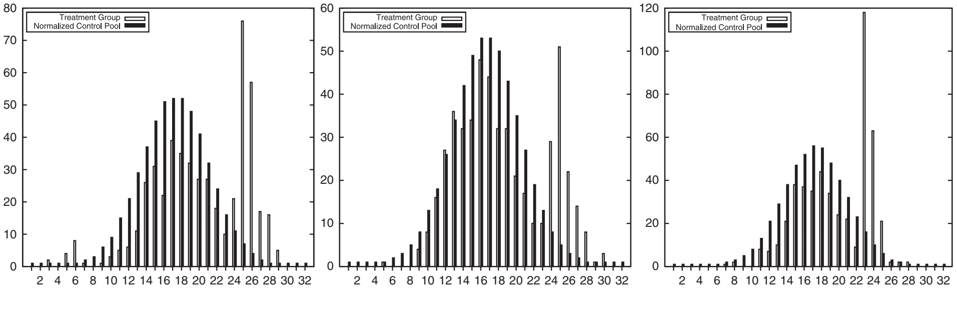

Fig. 1. Covariate distributions of treatment group and control pool (normalized). Left plot shows distribution for covariate $\boldsymbol{X_1}$. Middle plot shows distribution for covariate $\boldsymbol{X_2}$. Right plot shows distribution for covariate $\boldsymbol{X_3}$. The treatment group bars are unfilled, whereas the control pool bars are filled.

A simulated annealing algorithm was used to identify control groups with covariate distributions that were the most similar to the treatment group.[5] Our results are shown in Table 1. The column labeled *Bins* specifies the number of bins used (per covariate). The column *Observations* reports the number of control groups with no imbalance that were identified. The column *Accepts* reports the number of control groups for which we accepted the null hypothesis of no treatment effect. The remaining columns list the means and standard deviations for our estimated treatment effect, the Kolmogorov–Smirnov two-sample test statistic (averaged over the three covariates), and the Anderson–Darling two-sample test statistics (averaged over the covariates).

Several patterns are evident from the results in Table 1. First, as the number of bins for each covariate increases, the estimate of the treatment effect tends toward its true value of zero. This monotonically decreasing pattern is evident and, intuitively, should continue as the number of bins increases, provided that optimal groups can be found with a larger number of bins. Second, as the number of bins increases, the likelihood of accepting the null hypothesis of no treatment effect increases. For 8, 16, and 32 bins, all of the identified control groups lead to the conclusion that there is no treatment effect. Third, as the number of bins increases, the standard deviations for each of our measures of fit tend toward the true underlying standard deviation. Lastly, the Kolmogorov–Smirnov and the Anderson–Darling test statistics indicate an increasingly closer fit for the covariate distributions between the treatment and control groups as the number of bins increases. In short, our estimate of the treatment effect tends toward the true value as the number of bins/granularity increases. This point is underscored by Figure 2 that shows the distribution for the second covariate, X_2 for the treatment group and an optimized control group. In these plots, 4 (leftmost plot), 16 (middle plot), and 32 bins (rightmost plot) were used. In the rightmost plot, the filled and unfilled bars are most similar. In all cases, the control groups had no imbalance (i.e. the objective value was zero), which means that the distributions between the treatment group and control group for the first and third covariates are also similar. As expected, the distribution fits are closer when the number of bins is larger. Notice as well that as the objective function value for our groups′ approaches an optimal level, our estimate of the treatment effect tends toward the true treatment effect. This result (using 32 bins) is shown graphically in Figure 3 as well as in Table 2.[6] Lastly, once a certain objective function value is achieved, our hypothesis test for no treatment effect is accepted for all chosen control groups.

Table 1. Summary of optimal solutions

			Treatment effect		Kolmogorov–Smirnov		Anderson–Darling	
Bins	Observations	Accepts	Mean	SD	Mean	SD	Mean	SD
2	183,103	0	9.712	0.397	0.293	0.006	30.154	1.597
4	11,932	4,680	2.177	0.319	0.133	0.005	8.425	0.740
8	7,079	7,079	0.753	0.199	0.106	0.006	3.523	0.451
16	870	870	0.170	0.126	0.037	0.004	0.268	0.054
32	2	2	0.067	0.121	0.034	0.005	0.156	0.050

Note: Control and treatment group sizes are constrained to be equal. Control groups do not contain any duplicate observations (i.e. individuals are chosen ‘without replacement’). SD, standard deviation.

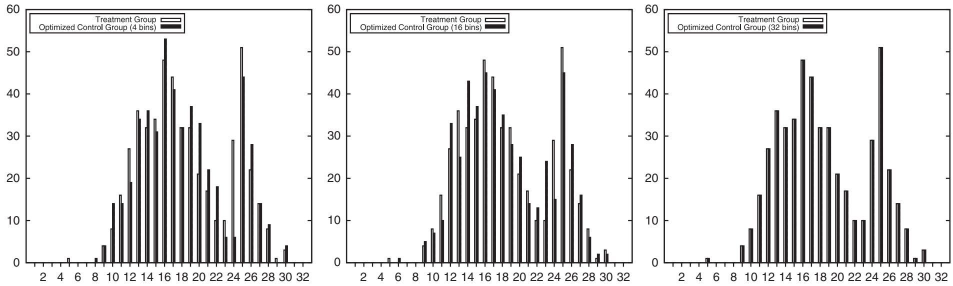

Fig. 2. Distribution of covariate X_2 in treatment and control groups. Left plot shows results from using four bins. Middle plot shows results from using 16 bins. Right plot shows results from using 32 bins. Unfilled bars are used for the treatment group, whereas filled bars are used for the control group.

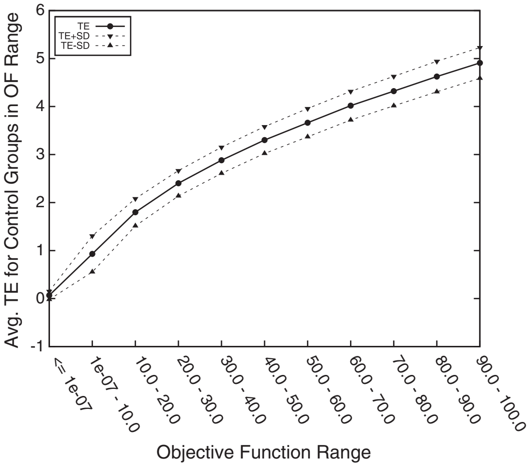

Fig. 3. Average treatment effect by objective function (OF) range (32 bins). Graph shows the estimate of the average treatment effect (TE) and the estimate of its standard deviation (SD).

5.2 *LaLonde data set*

We ran the same simulated annealing heuristic on the well-studied LaLonde data (LaLonde, 1986). These data are from the National Supported Work Demonstration Program, a randomized job training experiment. An experimental benchmark was computed from the experiment, and then the data were augmented with survey data. LaLonde's intention in creating this data set was to examine how well statistical methods would perform in trying to replicate the randomized experiment. In our analysis of these data, we used the Dehejia and Wahba (1999) subsample for the treatment group, which includes pretreatment income in 1974 as a covariate, and the *Current Population Survey* individuals for the control pool. The treatment group contains 185 individuals, and the control pool contains 15,992 individuals. There are eight covariates (some binary, some continuous) in this data set. The experimental benchmark for the treatment effect is $1794.

Table 2. Solutions (using 32 bins) sorted by objective function value

			Treatment effect		Kolmogorov–Smirnov		Anderson–Darling	
OF Range	Observations	Accepts	Mean	SD	Mean	SD	Mean	SD
≤1e-07	2	2	0.067	0.121	0.034	0.005	0.156	0.050
1e-07–10.0	29,006	29,006	0.930	0.376	0.042	0.007	0.352	0.140
10.0–20.0	18,285	16,679	1.797	0.285	0.058	0.005	0.830	0.171
20.0–30.0	13,900	2854	2.402	0.264	0.068	0.004	1.321	0.189
30.0–40.0	11,549	59	2.881	0.273	0.077	0.004	1.822	0.216
40.0–50.0	10,296	1	3.302	0.279	0.085	0.004	2.320	0.236
50.0–60.0	9178	0	3.663	0.295	0.092	0.004	2.809	0.263
60.0–70.0	8176	0	4.018	0.299	0.098	0.004	3.329	0.291
70.0–80.0	8077	0	4.321	0.307	0.104	0.004	3.831	0.319
80.0–90.0	7468	0	4.625	0.319	0.110	0.003	4.357	0.345
90.0–100.0	7030	0	4.909	0.320	0.115	0.003	4.889	0.369

Note: Control and treatment group sizes are constrained to be equal. Control groups do not contain any duplicate observations (i.e. individuals are chosen 'without replacement'). SD, standard deviation; OF, objective function.

A summary of our solution search is shown in Figure 4. We attempted to achieve optimal balance for each of the eight covariates. As we can see from the figure, as balance (as measured by Equation (5)) improves, the estimate of the treatment effect approaches the experimental benchmark. This supports the argument of Diamond and Sekhon (2012) that sufficient covariate balance is necessary to obtain better estimates of the treatment effect. These results mimic those that we found with the simulated data set. One difference between these two data sets is that we were able to obtain better overall balance for the simulated data than for the LaLonde data. This may be an indication of the difficulty of identifying such solutions when the

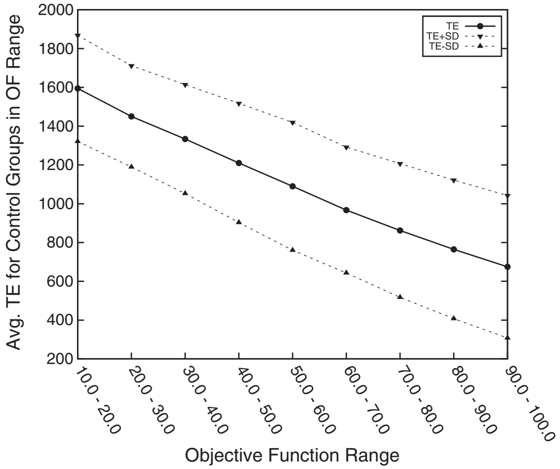

Fig. 4. LaLonde data: average treatment effect by objective function (OF) range (32 bins). Graph shows the estimate of the average treatment effect (TE) and the estimate of its standard deviation (SD).

number of covariates is large or may also be a reflection of the idiosyncrasies of the data set. Not all observational data sets, after all, are well-suited for causal inferences. The LaLonde data are peculiar in that the treatment group is particularly unique and may not have a suitable counterpart among non-treated individuals. All the same, the same pattern was evident in both analyses: as balance improved, our estimate of the treatment effect approached the expected estimate.

For the LaLonde data, our lowest objective function value was 17.17. This particular solution yielded an estimated treatment effect of \$1740.75, just \$53.25 from the experimental benchmark. We do note, however, that for these data, there was a good amount of variance in the treatment effect estimate among the set of solutions with objective values between 10 and 20. As such, the researcher might place more weight on the mean value in this range and the associated standard deviation for the treatment effect estimate among these various solutions. Here, there were 20 solutions in the set with a mean of \$1594.83 and a standard deviation of \$280.88. The experimental benchmark is well within a standard deviation, although this view expresses more uncertainty about the estimated treatment effect than simply using the solution associated with the lowest objective value. Note as well that the grouping of the objective values is largely arbitrary. Any decisions on this realm should be broached with a good understanding of the underlying substantive problem.

5.3 Discussion of results

In any observational data set, there may be both a large number of possible control groups as well as a large number of control groups that have essentially the same level of balance with the treatment group. Plainly, a large number of subsets yields essentially the same level of balance, because swapping out any single unit for another unit changes the balance only marginally. The ability to explore the range of treatment effects that arise from different subsets with essentially identical balance is notable and nicely yields a framework for computing an unbiased treatment effect estimator along with its associated standard error. There is also a nice adherence to statistical theory underlying randomization because repeated, properly randomized trials will produce distinct treatment and control groups, all satisfying random selection, all producing balanced covariates, but producing different estimates of the treatment effect. The ability to find and categorize a large number of essentially equally balanced control groups sets the BOSS framework apart from matching methods that identify a single-matched group. With matching methods, researchers sometimes compute a bootstrapped standard error for the estimated treatment effect. BOSS, on the other hand, allows one to construct a *distribution of control groups* from which we can compute a mean and associated standard error as well as establish a baseline level of balance that can be used to judge the level of balance for any particular control group.

Although our BOSS-B experiments exhibit favorable properties, they also highlight remaining issues. First, although it is desirable to obtain many solutions with comparable balance statistics, how to digest so much data is not always straightforward. One might compute a mean and standard deviation for the set of solutions in the best balanced set, but how to define the different ranges of objective function values where values will be grouped can be arbitrary. At minimum, the principles of design-based research and strong substantive knowledge of the problem should guide decisions on this realm (Angrist and Pischke, 2008). Second, although our balance statistics indicate a close fit, obtaining balance becomes increasingly difficult as the number of covariates increases. It is plain that balancing more covariates is more difficult than balancing fewer covariates. Third, and less obvious, but equally important, including more covariates introduces thorny issues regarding the weight that should be placed on balancing each covariate. The covariates are not likely to be equally well balanced. Instead, balance on one covariate may compete with balance on other covariates. The best 'overall balance' can be achieved by balancing one covariate at the expense of the others or by balancing all covariates at the same level. The best choice is non-obvious and not captured by balance statistics that are averaged over a set of covariates.

The bins approach also highlights an important trade-off. As the number of bins increases, our estimate of the treatment effect tends toward the true treatment effect, but also creates an increasingly difficult optimization problem. The increased complexity points toward a need to improve our optimization tools, whereas the trend in our estimates demonstrates that the BOSS approach presents a viable causal inference framework. To be clear, we hardly advocate the bins approach as the proper implementation. Rather, we began with BOSS-B simply to provide a proof of concept for the novel and promising theory underlying the BOSS framework. Nonetheless, there is plainly much work to be performed before the theory is successfully implemented.

Our message is that the causal inference literature can expand in new, fruitful, and exciting directions by incorporating insights from randomized experiments writ large rather than focusing narrowly on twin experiments and individual matches. Matching, in the best scenario, can closely replicate a single twin study. Our approach searches the entire space of possible control groups and produces a distribution with a large number of control groups that satisfy a balance objective. Fundamentally, we are proposing a paradigm shift from matching that explores sets of individual matches and returns one particular match to a subset selection framework that expands the search universe into the realm of all randomized experiments and returns solutions that easily number in the tens or hundreds of thousands.

6 Research directions and discussion

For future work, rather than using bins, we might optimize directly on a balance measure such as Kullback-Leibler divergence Kolmogorov–Smirnov, Anderson–Darling, a two-sample *t*-statistic for the difference of means, or some simultaneous combination of such distributional goodness-of-fit measures. We may also avoid optimization on all marginal and joint distributions with an approach that incorporates the covariance structure of the covariates into the objective function. These alternative approaches free us from the specific issues associated with the binning model, but raise other problems for optimization. Nikolaev *et al.* (2013) discuss some of these issues.

To be sure, old issues remain. What covariates should be balanced between the two groups? Are all the relevant covariates available? Even a perfect distributional fit between the observed covariates in the control and treatment groups will not yield an unbiased estimate of the treatment effect if unobserved covariates remain unbalanced. These issues, however, will perpetually remain for those wishing to make causal inferences with observational data. No methodology can save us from these data woes. Indeed, there are always a set of issues that arise in any statistical model, and it is always the researcher's charge to understand his or her model, its assumptions, and to interpret his or her statistical output accordingly. That said, we have formulated a new set of models and algorithms that provide a fresh set of practical tools for enhancing our understanding of causal structures by improving the ability to obtain balanced subgroups. Our formulation is flexible and not specific to a particular measure of balance. Any measure of balance can be incorporated, and so the debate surrounding balance measures exists apart from our research. Propensity scores may also be incorporated into our conceptualization as a covariate; so, debates revolving around propensity scores also are not germane to the value of our formulations. The optimization framework provides a novel and neutral method and tool that will help inform, not enter or fuel, these ongoing debates in the causal inference literature.

Our central insight is a discrete optimization framework that yields a more balanced solution to the problem than any existing method. Our approach eliminates the need for a distance measure and does not require a researcher to guess the proper form of a propensity score model. Instead, the quality of treatment effect estimation is now limited just by the complexity of an NP-Hard (non-deterministic polynomial time hard) optimization problem and available computational power. Human bias is replaced with computational constraints. The former is insurmountable. The latter, while certainly not insignificant, becomes less constraining daily.

Acknowledgements

This research has been supported in part by the National Science Foundation (SES-0849223 and SES-0849170). The second author was supported by the Department of Defense (DoD) through the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program (32 CFR 168a). The fourth author was also supported in part by the Air Force Office of Scientific Research (FA9550-10-1-0387). The views expressed in this paper are those of the authors and do not reflect the official policy or position of the National Science Foundation, the United States Air Force, or the United States Government. The computational work was conducted with support from the Simulation and Optimization Laboratory at the University of Illinois.

Thanks to Chris Achen, Jake Bowers, Justin Grimmer, Kosuke Imai, George Judge, Jasjeet Sekhon, Jonathan Wand, and participants at the Methods seminar at Stanford University and Princeton University for helpful comments.

Notes

1. These methods may fail if the selection on observables identifying assumption is not satisfied. Alternatively, linear bias may be worse unless the covariates are distributed ellipsoidally (Rubin, 1976; Rubin, 1976; Rubin and Thomas, 1992). If the covariates are not all ellipsoidally distributed, then we do not have a good understanding of the properties of the matching method. Notably, even if the Equal Percent Bias Reduction (EPBR) property does hold, it may be undesirable if some covariates are more germane to the matching venture than others. Moreover, propensity score matching methods have additional obstacles because they are model dependent. If the wrong propensity score model is used, then propensity score matching may make covariate balance worse. There is much that may go astray, and how or what distance metric to employ is both critical and unclear.

2. The approach in coarsened exact matching is different from propensity score and Mahalanobis metric matching, but the focus on individual matches is the same. In addition, there are other variations such as greedy matching versus optimal matching. These two methods result in different individual matches being made, but both techniques yield individual matches.

3. We employ a variety of functions such as cube and square roots, trigonometric functions, and logarithms to define the likelihood that a unit will be chosen from the treatment pool for inclusion in the treatment group. The specific details and functions are available upon request but are not particularly germane as long as we achieve our purpose of choosing units non-randomly to form the treatment group.

4. In these figures, the covariate values are separated into 32 uniformly sized ranges or bins. The control units are reduced by a factor of 1/200 to account for the difference in size between the treatment group and control pool.

5. Pseudo code for the algorithm and a proof of NP-Hardness (non-deterministic polynomial time hard) for the problem are available in Nikolaev *et al.* (2013).

6. Both the plot and the table omit solutions with objective value greater than 100. The trends in all values continue for these solutions.

References

Angrist, J. D. and J.-S. Pischke (2008), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press, Princeton, N.J.

Dehejia, R. and S. Wahba (1999), Causal effects in non-experimental studies: re-evaluating the evaluation of training programs, *Journal of the American Statistical Association* **94**, 1053–1062.

DEHEJIA, R. H. and S. WAHBA (2002), Propensity score matching methods for nonexperimental causal studies, *The Review of Economics and Statistics* **84**, 151–161.

DIAMOND, A. and J. S. SEKHON (2012). Genetic matching for estimating causal effects: a general multivariate matching method for achieving balance in observational studies, *The Review of Economics and Statistics* Forthcoming.

HOLLAND, P. W. (1986), Statistics and causal inference, *Journal of the American Statistical Association* **81**, 945–960.

LALONDE, R. (1986), Evaluating the econometric evaluations of training programs with experimental data, *American Economic Review* **76**, 604–20.

NIKOLAEV, A. G., S. H. JACOBSON, W. K. T. TAM CHO, J. J. SAUPPE and E. C. SEWELL (2013), Balance optimization subset selection (BOSS): an alternative approach for causal inference with observational data, Technical report University of Buffalo.

ROSENBAUM, P. R. and D. B. RUBIN (1983), The central role of the propensity score in observational studies for causal effects, *Biometrika* **70**, 41–55.

ROSENBAUM, P. R. and D. B. RUBIN (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score, *The American Statistician* **39**, 33–38.

ROSENBAUM, P. R., R. N. ROSS and J. H. SILBER (2007), Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer, *Journal of the American Statistical Association* **102**: 75–83.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

RUBIN, D. B. (1976a), Multivariate matching methods that are equal percent bias reducing, I: some examples, *Biometrics* **32**, 109–120.

RUBIN, D. B. (1976b), Multivariate matching methods that are equal percent bias reducing, II: maximums on bias reduction for fixed sample sizes, *Biometrics* **32**, 121–132.

RUBIN, D. B. (1977), Assignment to a treatment group on the basis of a covariate, *Journal of Educational Statistics* **2**, 1–26.

RUBIN, D. B. (1978), Bayesian inference for causal effects: the role of randomization, *The Annals of Statistics* **6**, 34–58.

RUBIN, D. B. and N. THOMAS (1992), Affinely invariant matching methods with ellipsoidal distributions, *The Annals of Statistics* **20**, 1079–1093.

SMITH, J. A. and P. E. TODD (2001), Reconciling conflicting evidence on the performance of propensity score matching methods. *AEA Papers and Proceedings* **91**, 112–118.

SMITH, J. and P. TODD (2005a), Does matching overcome Lalonde's critique of nonexperimental estimators?, *Journal of Econometrics* **125**, 305–353.

SMITH, J. and P. TODD (2005b), Rejoinder, *Journal of Econometrics* **125**, 365–375.

Received: 30 November 2011. Revised: 18 August 2012.