AN OPTIMIZATION APPROACH TO MATCHING AND CAUSAL INference

Wendy K. Tam Cho*
University of Illinois at Urbana-Champaign

Jason J. Sauppe
University of Illinois at Urbana-Champaign

Alexander G. Nikolaev
University of Buffalo

Sheldon H. Jacobson
University of Illinois at Urbana-Champaign

Edward C. Sewell
Southern Illinois University at Edwardsville

Abstract

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 as well as the matching formulation and translates nicely into a discrete optimization framework.

Keywords: Causal Inference, Matching, Optimization, Subset Selection

*Wendy K. Tam Cho is Professor in the Departments of Political Science and Statistics and Senior Research Scientist at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign. Jason Sauppe is a graduate student in the Department of Computer Science at the University of Illinois at Urbana-Champaign. Alexander Nikolaev is Assistant Professor in the Department of Industrial and Systems Engineering at the University of Buffalo (SUNY). Sheldon Jacobson is Professor in the Department of Computer Science at the University of Illinois at Urbana-Champaign. Edward Sewell is Professor in the Department of Mathematics and Statistics at Southern Illinois University at Edwardsville.
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 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 multi-faceted, 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

\[
Y_i(1) - Y_i(0).
\]  

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 (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, 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,

\[
ATE = E(Y(1) - Y(0)) = E(Y(1)) - E(Y(0)).
\]  

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 \( E(Y(1) \mid T = 1) \). From this group, the average treatment effect for the treated (ATT) is

\[
ATT = E(Y(1) - Y(0) \mid T = 1),
\]
1 EXPERIMENTAL VERSUS OBSERVATIONAL STUDIES

which quantifies the effect of the treatment on subjects that are treated. In general, $E(Y(1)) \neq E(Y(1) \mid T = 1)$ and $E(Y(0)) \neq E(Y(0) \mid T = 1)$. That is, the average treatment effect, $E(Y(1)) - E(Y(0))$, and the average treatment effect for the treated, $E(Y(1) \mid T = 1) - E(Y(0) \mid T = 1)$, 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 $E(Y(1) \mid T = 1) - E(Y(0) \mid T = 1)$, but we must still determine how to compute $E(Y(0) \mid T = 1)$, 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 $Y_i(0)$ on the same subject, but are able to use information on different subjects. If exposure to treatment satisfies the independence assumption, those who have been treated give us information about $E(Y(1))$, while those who have not been treated give us information on $E(Y(0))$. Hence, the treatment effect can be calculated as

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

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, the task at hand is to post-process the observational data so that exposure to treatment satisfies the independence assumption, thus resembling a randomized experiment.
2 Matching

“Matching” is a method for post-processing observational data so that they resemble experimental data by simulating statistical independence of treatment exposure and all other available variables (Rubin, 1974, 1977, 1978). The problem involves two population groups, treated and control, and a set of pre-treatment 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, 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 $X = x$, and

**Assumption 2:** $0 < P(T = 1 \mid X = x) < 1$, 

hold, treatment assignment is “strongly ignorable” (Rosenbaum and Rubin, 1983). A goal of matching is to post-process observational data so that treatment assignment is strongly ignorable.

2.1 The Matching Process and the Issues

The first step in matching is to establish a distance metric that quantifies the difference between two subjects based on 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 (Dehejia and Wahba, 1997; LaLonde, 1986; Smith and Todd, 2001).\textsuperscript{1} Diamond and Sekhon (2006) 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 (2002, 1997)

---

\textsuperscript{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, 1976a,b; 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 EPBR 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, 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.
and Smith and Todd (2005a,b, 2001) is largely a result of researchers using matching models that have not achieved good, or “good enough,” balance in the covariates. 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 results 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, and that one should explore different methods to identify the one that yields the best balance.

It is also clear that the critical assessment factors are not in the individual matches, but rather on the resulting treatment and control covariate distributions. That is, covariate balance and statistical independence of the covariates with treatment exposure is measured not at the level of individual matches, but at the aggregate distribution level. The individual matches are simply a vehicle to obtain balance in the covariate distributions.
There are different ways to conduct experiments to isolate treatment effect. Matching most closely resembles an identical twin framework that identifies 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. Matching methods essentially attempt to find twins for treated subjects. 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. Either randomized trials or twin studies are valid experimental frameworks, and when successfully implemented, isolate the treatment effect.

Since 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, Ross and Silber (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 done, but comports well with the idea of balance. The vehicle is not individual matches, but the goal, to obtain the optimal balance for a set of covariates, remains the same.
4 BOSS: Balance Optimization Subset Selection

The distance metric is necessary if one wants to make individual/twin matches, but superfluous for the goal of balanced covariate distributions, leaving room for the matching framework to be reconceptualized. We propose a focus shift from individual matches and twin studies to covariate distributions and randomized trials. Our vehicle is not individual matching but subset selection. One advantage easily afforded by subset selection is the ability to obtain many different comparably balanced subsets. Plainly, many subsets yield essentially the same level of balance, since 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 since 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.

Our study design, Balance Optimization Subset Selection (BOSS), recognizes the matching problem as an optimization problem. One insight is that the goal of these 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. As we saw in our examples, 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 distributions of the treatment and control, remains the same.

For illustration and proof of concept, we present one simple implementation of this optimization problem that incorporates data bins. In the Balance Optimization Subset Selection with Bins (BOSS-B) framework, we create a set of $B$ uniformly-sized data bins. Each covariate value is assigned to the bin that includes its value range. Clearly, when the number of bins or covariate ranges is small, the optimization problem is simple. As the number of bins increases, the optimization problem becomes more difficult, but the covariate distributions become more similar. More formally, for a set of $P$ covariates, there exists a set of $K = P + \binom{P}{2} + \binom{P}{3} + \ldots + \binom{P}{P}$ marginal and joint distributions because there are $P$ marginal distributions, $\binom{P}{2}$ joint distributions of 2 covariates, $\binom{P}{3}$ joint distributions of 3 covariates, and so forth, with $\binom{P}{P} = 1$ joint distribution that includes all $P$ covariates. For $P$ covariates, $B^P = K$ bins span the entire joint range of covariate values. For covariate $X_p$, the covariate values for the treatment group lie in the closed set, $[L_p, U_p]$, where $L_p = \min_T X_p$, and $U_p = \max_T X_p$.\(^2\) We can separate this range into $B$ bins with breakpoints, $L_p = t_0^p < t_1^p < t_2^p < \ldots < t_{R(p)}^p = U_p$, where $R(p) = B + 1$ is the total number of breakpoints for covariate $p$. The optimization routine seeks control units such that the control and treatment covariate distributions are as similar as possible. If there are 2 covariates, $p_1$ and $p_2$, and 2 bins ($B = 2$), then there are 2 marginal distributions and one joint distribution. The first marginal distribution is characterized by the bins for covariate $p_1$ ($[t_0^{p_1}, t_1^{p_1}]$ and $[t_1^{p_1}, t_2^{p_1}]$), while the second marginal distribution is characterized by the bins for covariate $p_2$ ($[t_0^{p_2}, t_1^{p_2}]$ and $[t_1^{p_2}, t_2^{p_2}]$). The joint distribution is defined by the set of bins, $\{[t_0^{p_1}, t_1^{p_1}] \times [t_0^{p_2}, t_1^{p_2}], [t_1^{p_1}, t_2^{p_1}] \times [t_0^{p_2}, t_1^{p_2}], [t_0^{p_1}, t_1^{p_1}] \times [t_1^{p_2}, t_2^{p_2}], [t_1^{p_1}, t_2^{p_1}] \times [t_1^{p_2}, t_2^{p_2}]\}$. 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 since the distributions

\(^2\)Since we are identifying control groups whose covariate distributions match the treatment group, we are able to consider only the range of covariate values for the treatment group.
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 are ordered from \( b = 1, 2, \ldots, K \) (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 \(|\#(S_b^C) - \#(T_b)|\) 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 \(P\) pre-treatment covariates, \(\{X_1, X_2, \ldots, X_P\}\), and a fixed number of bins, \(b\), find subsets \(S^C \subset C\), each of size \(N\), such that

\[
\sum_b \frac{[\#(S_b^C) - \#(T_b)]^2}{\max(\#(T_b), 1)}
\]

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

5 Simulation Results

We explore the statistical properties of the BOSS estimator through simulations. Here, we randomly generated three \(N(0, 1)\) pre-treatment covariates, \(X = [X_1, X_2, X_3]\) (each of size 100,000) and a positive definite \(3 \times 3\) variance-covariance matrix, \(\Sigma\). The covariates in the treatment pool are created by multiplying the covariate matrix and the square root of the variance-covariance matrix \((X\Sigma^{\frac{1}{2}})\). Covariate \(i\) in the control pool is generated with mean 0 and variance \(s_{i1}^2 + s_{i2}^2 + s_{i3}^2\), where \(s_{ij}\) is the \(ij\)th element of \(\Sigma\). 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,

\[
Y = 14 + 7X_1 + 11X_2 - X_3 + \epsilon,
\]

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

From our control pool, our algorithm chooses sets of control groups of size 500 with covariates that most closely match the covariate distribution in our treatment group. From our treatment pool, we non-randomly choose a treatment group of size 500 using a thinning algorithm. 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. 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.

We ran experiments for $B = 2, 4, 8, 16, 32$ uniformly-sized bins. Since these are powers of two, each larger set of bins simply divide the previous bin set in half. That is, for 2 bins, the thresholds are $t_0 < t_1 < t_2$ where $t_1 = (t_0 + t_2)/2$. For 4 bins, the thresholds are $t'_0 < t'_1 < \ldots < t'_4$ 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 $t_{j-1}^k \leq X_{ki} < t_j^k$.

---

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.
A simulated annealing algorithm was used to identify control groups with covariate distributions that were the most similar to the treatment group. 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 perfectly optimized control groups 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 0. Due to the increased computation induced by a large number of bins, we were not able to find optimal solutions for 32 bins. However, the monotonically decreasingly pattern is evident and, intuitively,
should continue as the number of bins increases. Second, as the number of bins increases, the likelihood of accepting the null hypothesis of no treatment effect increases. For both 8 and 16 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 tends 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 for the treatment group and an optimized control group. In these plots, 4, 16, and 32 bins were used. For 4 and 16 bins, the control groups were perfectly optimized (i.e., the objective value was zero), while for 32 bins, the optimized control groups were nearly perfectly optimized (with an objective value of approximately 0.067). 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.\footnote{The plot includes objective values up to 1000. The Table includes omits solutions with objective values in the 50–1000 range. In the plot, there is a break where the objective function range changes from increments of 10 to increments of 100 between 90.0–100.0 and 100.0–200.0. This break is shown with slanted bars in the plot and on the axis.} Lastly, once a certain objective function value is achieved, our hypothesis tests for no treatment effect are accepted for all chosen control groups.

There is 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. The ability to find and categorize a large number of essentially equally-balanced control 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 \textit{distribution of control groups} from which we can
Figure 3: Average treatment effect by objective function range (32 bins)

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

<table>
<thead>
<tr>
<th>Objective Function</th>
<th>Observations</th>
<th>Accepts</th>
<th>Treatment Effect</th>
<th>Kolmogorov-Smirnov</th>
<th>Anderson-Darling</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mean  SD</td>
<td>Mean  SD</td>
<td>Mean  SD</td>
</tr>
<tr>
<td>$\leq 1e-07$</td>
<td>0</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>1e-07–10.0</td>
<td>29044</td>
<td>29044</td>
<td>0.920 0.384</td>
<td>0.042 0.007</td>
<td>0.352 0.144</td>
</tr>
<tr>
<td>10.0–20.0</td>
<td>18071</td>
<td>16326</td>
<td>1.810 0.282</td>
<td>0.058 0.005</td>
<td>0.830 0.167</td>
</tr>
<tr>
<td>20.0–30.0</td>
<td>14308</td>
<td>2881</td>
<td>2.412 0.269</td>
<td>0.068 0.004</td>
<td>1.324 0.187</td>
</tr>
<tr>
<td>30.0–40.0</td>
<td>11569</td>
<td>56</td>
<td>2.888 0.273</td>
<td>0.077 0.004</td>
<td>1.810 0.211</td>
</tr>
<tr>
<td>40.0–50.0</td>
<td>10020</td>
<td>1</td>
<td>3.302 0.271</td>
<td>0.084 0.004</td>
<td>2.313 0.238</td>
</tr>
<tr>
<td>50.0–60.0</td>
<td>9010</td>
<td>0</td>
<td>3.671 0.290</td>
<td>0.092 0.004</td>
<td>2.828 0.267</td>
</tr>
</tbody>
</table>
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.

While the BOSS-B experiments exhibit favorable properties, they also highlight remaining issues. First, 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. Second, 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 will 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 highlights an important tradeoff. 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, while the trend in our estimates demonstrates that the BOSS 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 done 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 represented by randomized experiments, including but not limited to twin studies, and produces a distribution with a large
number of control groups that satisfy a balance parameter. 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 tens of thousands of solutions.

6 Research Directions and Discussion

Rather than using bins, we might optimize directly on a balance measure such as 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.\footnote{Surely, other optimization issues remain. For instance, and importantly, we also need to ensure that we are traversing the space of possible solutions well. Since the total number of possible solution is prohibitive, we must ensure that the solutions we do cull are representative of the universe of possible solutions.} These alternative approaches free us from the specific complexity that accompanies large numbers of bins, though other sources of complexity still need to be addressed.

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 model, its assumptions, and to interpret his 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 optimal 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 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.
References


