A Parallel Evolutionary Algorithm for Subset Selection in Causal Inference Models

Wendy K. Tam Cho*
University of Illinois at Urbana-Champaign
wendycho@illinois.edu

Yan Y. Liu†
University of Illinois at Urbana-Champaign
yanliu@illinois.edu

ABSTRACT

Science is concerned with identifying causal inferences. To move beyond simple observed relationships and associational inferences, researchers may employ randomized experimental designs to isolate a treatment effect, which then permits causal inferences. When experiments are not practical, a researcher is relegated to analyzing observational data. To make causal inferences from observational data, one must adjust the data so that they resemble data that might have emerged from an experiment. Traditionally, this has occurred through statistical models identified as matching methods. We claim that matching methods are unnecessarily constraining and propose, instead, that the goal is better achieved via a subset selection procedure that is able to identify statistically indistinguishable treatment and control groups. This reformulation to identifying optimal subsets leads to a model that is computationally complex. We develop an evolutionary algorithm that is more efficient and identifies empirically more optimal solutions than any other causal inference method. To gain greater efficiency, we also develop a scalable algorithm for a parallel computing environment by enlisting additional processors to search a greater range of the solution space and to aid other processors at particularly difficult peaks.

CCS Concepts

•**Distributed Systems** → **Distributed Applications;**

*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. 420 David Kinley Hall, 1407 W. Gregory St., Urbana, IL 61801.

†Yan Y. Liu is Senior Research Programmer in the Cyber-Infrastructure and Geospatial Information Laboratory and Department of Geography and Geographic Information Science, and at the CyberGIS Center for Advanced Digital and Spatial Studies and National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign. 1205 West Clark Street, Urbana, IL 61801.

XSEDE '16 Miami, FL, USA

DOI: http://dx.doi.org/10.1145/2949550.2949568

Keywords

Combinatorial Optimization; Parallel Computing, Evolutionary Algorithm, Message Passing

1. INTRODUCTION

Experimental studies have enormous and unique potential. When an experiment is well-designed and flawlessly executed, the experimental framework isolates the treatment effect and allows one to examine causal effects [16, 9, 2]. Indeed, experimental research via randomized control trials has long been an important and key research tool across many fields.

While one is immediately able to see the appeal of experiments and establishing causal effects rather than simple associations, it is also clear that it is sometimes not possible to conduct research via a randomized control trial. Implementing randomized control trials requires significant investment of both time and resources [13]. These resources may be scarce and possibly unavailable. Moreover, even if a researcher does possess the time and money to conduct a randomized experiment, some questions simply do not lend themselves to this framework. Though we may wish, for example, to determine if smoking causes lung cancer, it is unethical to conduct an experiment where one randomly chooses people and then randomly assigns some to a treatment group. This experiment, despite its value to society and science violates moral and ethical standards [8, 4].

When a randomized experiment is not possible, observational data offer a potential path forward. We can, for instance, easily observe a large number of people who have chosen, on their own accord, to smoke. Because observational data are more easily collected than experimental data, it is plain that the ability to use observational data to make causal inferences, usually restricted to experimental research, would be extremely desireable. Indeed, many researchers have embarked on this quest to identify causal inferences from observational data. The topics explored have been wide-ranging and important. Researchers have examined the effectiveness of voter canvassing efforts [12], criminality rates related to gene patterns [26], the effect of generic substitution of presumptively chemically equivalent drugs [25], in utero exposure to phenobarbital on intelligence deficits [19], and the effect of maternal smoking on birth weight [5], to name but a few. Plainly, some of the research areas, e.g. in utero exposure to phenobarbital on intelligence deficits, cannot be studied via randomized experiments.

However, using observational data to make causal inferences is far from trivial or simple. If we wanted to ex-

plore data on observed smokers, we would quickly realize that their attributes differ systematically from non-smokers. These observational data are clearly not akin to experimental data since the smokers have not been randomly assigned to a smoking group. In an experiment, on the other hand, because of the randomization process, the attributes of the treatment/smokers and control/non-smokers do not differ systematically. Indeed, the key difference between experiments and observational studies is that in experiments, when randomization is successful, because the treatment effect is isolated from potential confounders, differences in response can be attributed to the treatment. Applying standard statistical models to non-experimental or observational data generally allows the researcher to make associational inferences only.

The complex problem is whether observational data may be adapted in such a way that they may be examined as experimental data that permit causal inferences to be made. If observational data can be post-processed to successfully mimic experimental data and all of the underlying assumptions are satisfied, then we can theoretically derive causal inferences from observational data [11, 22, 24]. While one may always attempt to post-process observational data, it is never known a priori whether this exercise will be successful. Sometimes even the optimal result is insufficient because critical data are unobserved. Other times, a non-optimal result will ensue because the post-processing method was unable to identify a sufficient and satisfactory solution. Certainly, at minimum, it is useful to have causal inference models that yield the optimal data adjustment so one can then assess whether this solution satisfies the requisite statistical requirements.

In this paper, we present a computational model for post-processing observational data so that they resemble a randomized experimental framework. In Section 2, we explain the statistical framework for these causal inference models. In Section 3, we discuss related work. In Section 4, we discuss our paradigm change that also transforms the problem into one that is amenable to a discrete optimization framework. In Section 5, we present an evolutionary algorithm for the identified optimization problem. In Section 6, we discuss how we parallelize the algorithm to make it more effective and efficient for traversing the solution space. In Section 7, we present some empirical results from the classic LaLonde data set. We provide some concluding discussion in Section 8.

2. MAKING CAUSAL INFERENCES FROM OBSERVATIONAL DATA

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 $Y_i(1)$, where $Y_i(0)$ represents the dependent variable or outcome for subject i under the control condition and $Y_i(1)$ represents the dependent variable or outcome for subject i under the treatment condition. The causal effect of the treatment, as measured by the outcome or dependent variable, 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. Since only one of the terms in (1), either outcome under treatment or outcome under control, is observable for a single unit, the expression cannot be evaluated [11].

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 [22, 24]. That is, one statistical solution to the fundamental problem of causal inference is to shift from an examination of individual treatment effects to an *average* treatment effect (ATE) over *all* of the subjects,

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

This alleviates the fundamental problem of causal inference because we no longer require two observations, $Y_i(0)$ and $Y_i(1)$, from one unit, i. Instead, we need only one of these observations. If the outcome under control for unit i, $Y_i(0)$, is observed, we use that information to inform the average treatment effect under control, $Y(0)$. We forego an ability to measure the treatment effect for a particular individual, $\tau_i = Y_i(1) - Y_i(0)$, in exchange for a measure of the average treatment effect across a range of individuals.

A critical 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 $E(Y(1) \mid T = 1)$ rather than $E(Y(1))$. From this observed group, the average treatment effect *for the treated* (ATT) is

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

which quantifies the effect of the treatment on subjects that are treated. Because $E(Y(1)) \neq E(Y(1) \mid T = 1)$ and $E(Y(0)) \neq E(Y(0) \mid T = 1)$, 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 $Y_i(0)$ from the *same* subject, but are able to use information from *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 $E(Y(0))$ allowing the treatment effect can be calculated as

$$\begin{aligned} ATE = ATT &= E(Y \mid T = 1) - E(Y \mid T = 0) \\ &= \frac{1}{N_t} \sum_{i \in \{T=1\}} Y_i(1) - \frac{1}{N_c} \sum_{i \in \{T=0\}} Y_i(0), \end{aligned} \tag{4}$$

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.

With observational data, it would be very 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, the task at hand is to post-process the observational data so that exposure to treatment satisfies the independence assumption. If this task can be satisfactorily accomplished, the post-processed data will resemble a randomized experiment, and one can then compute the treatment effect in a straightforward manner.

In practice, the problem of adjustment or post-processing of observational data usually involves two population groups, treated (observed subjects who have a particular attribute) and control (observed subjects who do not have a particular attribute), and a set of pre-treatment covariates, $\mathbf{X}$. The objective is, given the observed treatment group, to identify a subset of the control pool so that the covariate distributions of the treated and chosen control group are statistically indistinguishable, creating as-if-randomization. If treatment is random, as measured by randomization tests, we assume that the *unconfoundedness* or the *selection on observables* (SOO) assumption is satisfied. Formally, if

Assumption 1: $P(T, Y(0) \mid \mathbf{X}) = P(T \mid \mathbf{X})\, P(Y(0) \mid \mathbf{X})$, $P(T, Y(1) \mid \mathbf{X}) = P(T \mid \mathbf{X})\, P(Y(1) \mid \mathbf{X})$,

and

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

hold, we claim evidence that the treatment assignment is "strongly ignorable" [21]. Assumption 1 states that the treatment is independent of the outcome, conditional on covariates, $\mathbf{X}$.

The driving goal for post-processing observational data is to create data sets so that treatment assignment is strongly ignorable, i.e. adjust observational data to resemble randomized experimental data. Note that while it is always possible to engage in this endeavor, it is not always possible to be successful in this endeavor. Moreover, the extent to which one might be successful is unknown. It depends both on the available and observable data as well as the statistical and/or computational model's ability to successfully post-process the observed data.

3. RELATED WORK

Traditionally, adjusting observational data so that they resemble experimental data by simulating statistical independence of treatment exposure and all other available variables is done via methods that fall under the term "matching methods" [22, 23, 24]. The term matching is used because the methods attempt to match each treatment unit to a control unit that has attributes that are as similar as possible to that treatment unit. These matching methods identify one solution that may, though is certainly not guaranteed to, ensure a level of covariate balance that would pass a randomization test.

Given considerable discussion in the literature about finding the best covariate balance, scholars have, surprisingly, only recently, proposed optimization methods as a vehicle for achieving covariate balance between the treatment and control groups. Optimization, of course, is a general term and a broad malleable tool. The other optimization proposals [27, 7] retain the same statistical framework as the statistical matching models, which makes our contribution fundamentally different. We both alter the framework and proffer a different modeling paradigm that changes the design and shifts the purpose and objective of the optimization. The paradigm change was the subject of previous work [1], and we have also previously designed a simulated annealing optimization tool for this purpose, but the problem is computationally complex and that simulated annealing optimization tool was sufficiently inefficient to realize the proposed goals of the model [17]. Our innovation here is the formulation of an evolutionary algorithm that is more effective and efficient and better able to achieve the goals of our modeling paradigm.

4. MATCHING VERSUS SUBSET SELECTION METHODS

The first step in a matching method is to calculate a distance or similarity measure for any two observations. The smaller the value of the distance metric between two observations, the more similar the two observations are to one another. The particular metric used by a matching method varies but some metric is essential in these methods.

Once all pairwise distances have been computed, the next step is to match treatment units to control units. This matching may be achieved in a variety of ways, perhaps through greedy matching where each treated unit is sequentially paired with its closest non-matched control unit or a network flow optimization routine that seeks to minimize the total sum of pairwise differences [20]. The set of matched units comprise the control group.

The treated and control group distributions are then compared to one another, and an assessment is made for how closely they resemble one another. If the matching does not result in sufficient covariate balance, the entire procedure is repeated, perhaps with a different distance metric or another pairwise matching procedure with some parameter changed, until, ideally, sufficient balance is obtained so that the data may reasonably be statistically regarded as possibly having emerged from a randomized experiment.

To be sure, there are many valid research designs, including factorial designs (where groups receive a combination of treatment effects), crossover designs (where subjects randomly receive or do not receive treatments over time), and cluster designs (where entire groups are randomly assigned to treatment), as well as the completely randomized experimental design, to name a few [3]. The logic underlying matching methods derives from pair randomized designs. We redesigned the model framework to mimic the completely randomized experimental framework through a subset selection procedure [1]. There is no compromise here since our framework subsumes pair-randomized designs. Moreover, this reconceptualization allows us to optimize *directly* on covariate balance. At the same time, viewing the problem in this way requires an efficient implementation of an optimization algorithm.

Importantly, in our subset selection routine, we output any solution that satisfies a statistical as-if-randomization test. In an optimization framework, this is simple to accommodate. Statistical matching methods, on the other hand, output only a single matched group. This results in a single estimate of the treatment effect, which is a random variable. Though matching methods can be adapted, the prevailing matching paradigm simply does not lend itself well to methods for identifying a large number of as-if-randomized solutions. It is also possible to employ constraint satisfaction al-

gorithms. These avenues should be explored, but they have not been because our subset selection framework is a novel conceptualization for which tool development is nascent.

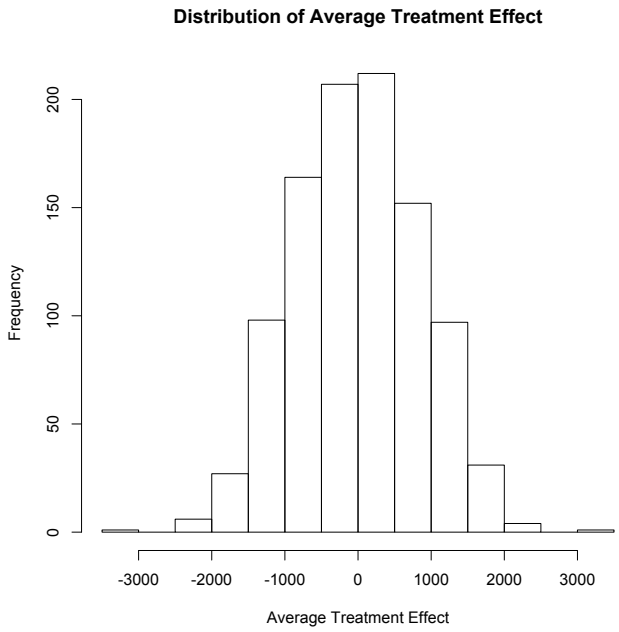
To see why it is interesting to identify thousands of such subsets (rather than simply a single solution which is all other matching techniques identify), consider a randomized experiment. Randomized experiments begin with a randomization process that chooses a set of units from a pool of possible units. These chosen units are then randomly assigned to either the treatment or control condition. If this randomization process were repeated, the next rendition would yield empirically different units and a potentially different estimate of the treatment effect. That variation is embodied within the many potential experiments is plain—it is induced by the randomization process itself.

When one is interpreting results, the magnitude of the variation is dependent on the noisiness of the outcome variable. To see this, we simulate a series of randomized control experiments where we draw 370 individuals at random from the data set. For each chosen individual, we randomly choose whether the individual will be in the treated or the control group. This results in a treatment and control group, each with about 185 subjects. We do not treat these chosen units in any sense—there is no treatment effect because there is no treatment. We repeat this experiment 1000 times, each time drawing a new set of 370 randomly chosen individuals, randomly assigning them to either treatment or control, and then computing the treatment effect. The distribution of the estimated average treatment effects across our 1000 experiments is shown in Figure 1. As we might have expected from knowledge of the Central Limit Theorem, the result is a normal distribution. The mean of this distribution is approximately 6.90, which is very close, and given variability in the simulation, essentially identical to the true treatment effect of zero. Note, however, that the standard deviation of the distribution is quite large (1,037.68), reflecting the noisy nature of outcome variable.

Clearly, while the *mean of the distribution* of estimated treatment effects is quite a good estimate of the true treatment effect, the variance of this estimate is somewhat large. In other words, the result of any particular experiment may not be a good estimate of the treatment effect, despite the randomized design of the experiment. An unbiased estimator need not be a low-variance or efficient estimator—bias is unrelated to variance. The randomized experimental framework provides us with an unbiased estimate, but the noisy outcome variable ensured that we also have a large variance estimate of the average treatment effect. Matching methods supply us with one estimate, somewhere in the distribution, hopefully near the mean. However, there is no way of knowing where the estimate might lie in the distribution.

It is only by conducting many experiments that we are able to contextualize any one experiment. Our computational model integrates this idea, which is plainly the underlying theoretical construct of statistical models. Matching models identify only a single solution and hope that that solution is not unusual and close to the mean of the underlying true distribution. Our series of experiments here also highlights why using a matching method and then an associated bootstrap might be problematic. If our one estimate from matching is unusual and not close to the mean, this problem will be propagated to the subsequent bootstrap, resulting in misleading estimates. We need independent realizations of the underlying phenomena.

Figure 1: Average treatment effect across 1000 randomized control experiments



We develop a computational model to capture the theoretical statistical underpinnings of randomized experiments through an algorithm that optimizes covariate balance via optimal subset selection. The goal is the same as the matching methods, but the procedure is novel. To gain a sense for how difficult this problem is, consider the solution space in question. First, the solution space is enormous. If there are 100 units from which to choose a subset of 10, there are $\binom{100}{10} = 1.73 \times 10^{13}$ possible subsets. Second, at a rough level of granularity, the solution space is rugged. At a finer level of granularity, the solution space is flat. It is evident than if one swaps out a single observation from a subset, the new subset is substantially similar and the covariate balance does not change significantly. For any subset, there is a slew of such minor modifications which behooves the algorithm to embody a method for a diversified search of the solution space that retains a level of independence among identified solutions.

5. EVOLUTIONARY ALGORITHM

We develop an **E**volutionary algorithm in C for the **B**alance **O**ptimization **S**ubset **S**election problem (E-BOSS). Similar to other evolutionary algorithms, E-BOSS has a probabilistic mutation operation, an n-exchange mutator, a crossover operator that enables larger movements to potential solutions, and an evolutionary framework for transforming the initial population to future generations. Together, these operators define how the chromosomes will evolve from generation to generation. Each of these operators has a probabilistic nature, which can be modified and adapted for each

particular application.

In our subset selection problem, we wish to choose n units from the N total control pool units so that the covariate balance between the n treated units and the n chosen control units is maximized. Because the order in which subjects are chosen and placed into the treatment and control groups is inconsequential, we can begin by indexing the N total units arbitrarily and consecutively (from 1 to N). In our algorithm, the chromosome is encoded as a list of n distinct integers (representing the indices) from the set $\{1, \ldots, N\}$. Each allele holds the index number of a chosen control unit.

The fitness of a chromosome represents one measure of balance between the treatment and control groups, which we may define via the objective function as

$$b = \sum_{i=1}^{C} w_i \left(KS_i + \mid t_i \mid + \left| 1 - \frac{\sigma_{ti}^2}{\sigma_{ci}^2} \right| \right) . \qquad (5)$$

where i indexes the variable, C is the number of variables, w is a weight, KS_i is the Kolmogorov-Smirnov statistic, t is the t-statistic for the difference of means, and σ_t^2 and σ_c^2 indicate the variance of the treatment and control groups, respectively. The weights help guide the search routine to areas of the solution space that are less likely to be traversed with uniform weights. Larger weights lead the algorithm to work harder to balance substantively important covariates over less substantively important covariates. Some covariates may also be more difficult to balance, so the weights also aid in distributing the balance more equally among all covariates. Adjusting the weights also helps the algorithm to expand the search to particular parts of the space that may otherwise be more difficult to locate. We seek to maximize the covariate balance, which corresponds to minimizing the objective function value since maximum balance is achieved at the value $b = 0$.

The reproduction and mutation operators are fairly standard. The crossover operator requires some customization from the standard type of genetic crossover. In a standard crossover operation, a random point of division in the chromosome is chosen. If this occurs at position k, then two new strings are created by swapping all alleles between positions $k + 1$ and n inclusively. These two new chromosomes are then part of the next generation. Note, however, that this procedure would not work with our encoding of the problem. Swapping all alleles after a certain position, k, may result in duplicate indices within a single chromosome. Our formulation of the problem does not allow for duplicate alleles within a single chromosome. Each allele needs to be unique. The chromosomes represent subsets that need to be of a fixed length since our chosen control group solution is constrained to be of the same size as the treatment group. This requirement of a fixed-length subset creates a difficulty with a standard crossover because the procedure may not result in n unique alleles. To bypass this difficulty, we pursue a crossover operator via a variation of Random Assortment Recombination (RAR) [18].

In our evolutionary algorithm, we begin with a population of size 100. The algorithm's parameters are fairly standard with a high crossover probability and a low mutation probability. The crossover probability is set to 0.85 while the mutation probability is the reciprocal of the population size or 1/100. Selection of parents for crossover is conducted via roulette wheel selection where a chromosome's share of the roulette wheel is proportional to its fitness. This ensures that the attributes of the more fit solutions are more likely to be propagated to the next generation. To prevent premature convergence, a random restart is invoked when the population becomes too homogeneous.

6. PARALLEL IMPLEMENTATION

Efficiency is needed in the sequential evolutionary algorithm, but some necessary additional searching power is gained through a parallel implementation. Here, our algorithm's basic message passing structure utilizes MPI. A goal of the parallelization is to increase the efficiency of the optimization process by developing diversification and intensification protocols for the search procedure. We probabilistically initiate multiple processors at points of the optimization that appear to be especially difficult. Once the multi-processor search moves beyond a particular optimization threshold, the processors resume with their sequential searching. This mode of operation where multiple processors work collaboratively and then retreat to their own peaks once a threshold has been reached defines the basic structure of the message passing protocol. We also implement some memory into the search via a tabu list and regular message passing to ensure that the various processors are focused on different areas of the solution space, providing a more diversified search and aiding in the identification of independent solutions. A working prototype of this message passing structure has been coded.

To scale the code performance on supercomputers such as Blue Waters, we explore efficient communication structures. The global synchronization often used in MPI for collective operations (e.g., broadcasting) is eliminated by using non-blocking MPI communication functions (i.e., *MPI_Iprobe()* and *MPI_ISend()*). Each processor outputs a list of best solutions found during the evolutionary process. We will use our grant of time on Blue Waters for further development and refinement of how to best use a massively parallel architecture to increase the efficiency of the optimization.

We have already resolved a major communication bottleneck after examining the scalability of our code. By adopting an asynchronous communication mechanism [15], we eliminate the costly global synchronization operation which was originally conducted for incoming message checking after looking at 25 solutions for the crossover and mutation operations in each iteration. With the global synchronization, the message passing time ranged from 42% on 240 cores to 72% on 960 cores, which is not acceptable because of its severe impact on performance. After we implemented the asynchronous communication, the message passing cost declined considerably and is now marginal (0.007% on 240 cores and 0.01% on 960 cores).

Figure 2 shows the results of our weak scaling experiments. The time required to reach different solution quality thresholds (indicated by the separate lines) was recorded using different numbers of cores. The bottom graph shows the results when we used global synchronous communication while the top graph displays the numbers from an asynchronous communication protocol. When the code utilized global synchronization, since more time was needed for synchronization of a larger number of cores, the time spent on the actual evolutionary search by each core was reduced. We can see from the graph that the time required to achieve various solution thresholds increased rather than decreased as

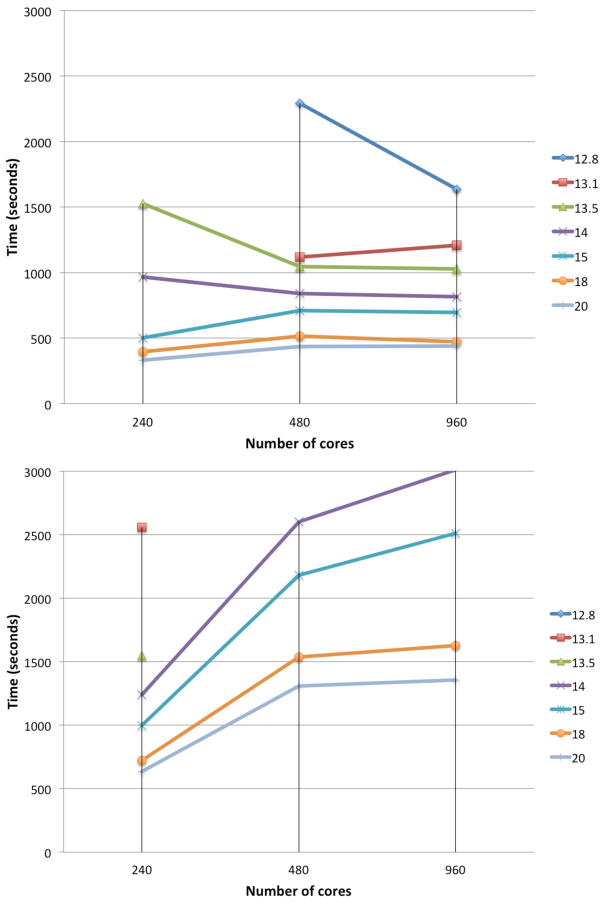


Figure 2: Results from Weak Scaling Experiments. Asynchronous Communication versus Synchronous Communication (260 to 960 processing cores).

we utilized more processors. As a result, using more cores produced worse results. However, with the asynchronous communication mechanism, the numerical performance improves and scales well. The use of additional cores improves the solution quality in a fixed amount of time as the solution quality threshold becomes tighter or smaller.

With Blue Waters resources, we will continue to explore methods for enhancing performance and scalability as well as conduct production runs at larger scales. We will conduct a series of experiments to devise a distributed protocol among participating processors to probabilistically determine the number of responses to a hill climbing request and the total number of responses allowed in the system at any time. This is a trade-off between enlisting more cores to help in searching a difficult solution landscape and allowing a more diversified search to transverse a greater portion of the solution landscape. The working mechanism for this trade-off must be decentralized to avoid, again, global synchronization. As well, this mechanism will be malleable and will switch between different search strategies depending on the stage of the search. At the beginning of the search, we would like to diversify so that we can initially reach as many different areas of the landscape as possible. As the search becomes increasingly difficult, we would then like to intensify the search with the hope of finding a global optima. Understanding the relationship between this working mechanism and the scalability of our algorithm will be important.

Another key component of the algorithm is a mechanism for identifying *independent* solutions. This is non-trivial requirement, especially within standard optimization frameworks, since solution space traversal often moves from one solution to another largely similar solution. Our diversification protocol that is embedded into the parallel structure of the algorithm will be an important component in identifying solutions that have a degree of independence from one another.

7. LALONDE DATA CASE STUDY

We have conducted some empirical tests using the classic LaLonde data set [14]—an observational data set that many matching methods have had notorious difficulty in identifying matches that yield good covariate balance. We use the Deheijia and Wahba [6] 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 in this data set.

Nikolaev et al. [17] implemented a simulated annealing algorithm (BOSS-B) and ran it on modified LaLonde data. Instead of using the actual data values, they put the data into bins to create a smaller problem size and to simplify the optimization task. This was necessary because their algorithm was sufficiently inefficient that it failed to converge and find solutions with the actual covariate values. The best BOSS-B results are summarized in Table 1 (taken from [17]). We can see that while BOSS-B does not find it difficult to identify many (albeit not independent) solutions at poor levels of balance, it has considerable difficulty in identifying solutions that fall within increasingly better objective function ranges. The objective function ranges are shown in the first column of Table 1. In the objective function value range of 40–45, the BOSS-B algorithm was able to identify 72 solutions. It was not able to find any solutions where the objective function value fell below 40. They called for additional work to be done in the computational and optimization realm. Along these lines, our evolutionary algorithm, is able to optimize directly on covariate balance while using the actual (not aggregated and binned) covariate values. In addition, our preliminary tests demonstrate that E-BOSS is able to identify thousands of solutions with objective function values below 20. The best solution identified by BOSS-B has an objective function value of 40.77. The best solution we have thus far identified by E-BOSS on the same data set and problem has an objective function value of 16.48.

We examined the quality of our results via a variety of different graphical measures as well as through a set of statistical tests. Our solutions perform better than BOSS-B under all tests including the omnibus balance test, d^2, that combines the individual difference of means tests [10],

$$d^2(\mathbf{z};\mathbf{x}_1,\ldots,\mathbf{x}_k) = [d(\mathbf{z},\mathbf{x}_1),\ldots, d(\mathbf{z},\mathbf{x}_k)] \left[\text{Cov}\begin{pmatrix} d(\mathbf{Z},\mathbf{x}_1) \\ \vdots \\ d(\mathbf{Z},\mathbf{x}_k) \end{pmatrix}\right]^{+} \begin{bmatrix} d(\mathbf{z},\mathbf{x}_1) \\ \vdots \\ d(\mathbf{z},\mathbf{x}_k) \end{bmatrix}, \quad (6)$$

Table 1: LaLonde data: BOSS-B solutions sorted by objective function value

Objective Function Range	Obs	Treatment Effect			
		Mean	SD	Minimum	Maximum
(40.0 , 45.0]	72	1534.03	259.11	930.84	2201.42
(45.0 , 50.0]	777	1481.05	256.78	695.39	2195.45
(50.0 , 55.0]	2,231	1415.61	269.84	572.07	2359.02
(55.0 , 60.0]	3,041	1318.83	296.90	397.96	2400.89
(60.0 , 65.0]	2,388	1217.10	301.42	207.51	2217.09
(65.0 , 70.0]	714	1140.94	313.55	-60.36	2151.17

Control and treatment group sizes are constrained to be equal.
Control groups do not contain any duplicate observations
Results generated via the BOSS-B simulated annealing algorithm.
(Table from Nikolaev et al. (2013).)

where

$$d(\mathbf{z}, \mathbf{x}) = \frac{\mathbf{z}'\mathbf{x}}{n_t} - \frac{(\mathbf{1} - \mathbf{z})'\mathbf{x}}{n_c}, \tag{7}$$

and $\mathbf{z}$ is a vector of zeros and ones to indicate assignment to the treated or control group. This test, related to Hotelling's (1931) T-test, considers balance on the base set of k covariates $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$, as well as on all linear combinations of the covariates. We also found that the best BOSS-B solution failed statistical significance tests for the d^2 omnibus test of randomization statistic at the 0.20-level. For one of the covariates, the p-value was less than 0.001. As a result, the omnibus test was also statistically significant, with a p-value of 0.013.

Quite clearly, other balance measures could also be assessed. Explicitly multivariate balance measures, for instance, could be utilized. In any particular application, the researcher should apply as many balance tests and measures as he deems necessary to convince himself that the control and treatment groups are similar enough to be regarded as as-if-randomized groups.

Verification is an important and necessary step whether the vehicle for identifying the groups is via E-BOSS or some other algorithm. Once verified, another important observation is that there are many as-if-randomized solutions. In this sense, the search for the optimally balanced solution is not an end to itself. It is a process. In the process, a large number of other independent as-if-randomized solutions are identified, and this entire set of solutions is critical in gaining an understanding of the substantive problem.

8. CONCLUSION

Substantive research that might benefit from the ability to make causal inferences abounds. Causal research questions transcend all scientific fields. BOSS embodies a new paradigm for developing an analytical toolbox that is based in operations research. The causal inference literature has been heavily ensconced in the statistical literature. While the statistical foundations are quite clearly integral and important, the computational modeling approach with firm roots in operations research allows us to create a systematic solution methodology that allows us to bypass some of the assumptions that necessarily underlie the statistical matching models. For instance, guessing the form of a propensity score model or specifying a particular distance function is no longer necessary under the BOSS framework. The associated element of human error in this process is thus also eliminated. This human bias is replaced by the complexity of a non-trivial *NP*-Hard optimization problem.

To be sure, we are replacing one difficult problem with another difficult problem. Substituting available computational power and heuristic algorithmic development are non-trivial replacements. On the other hand, the nature of the problem shifts from human bias in model specification and assumptions underlying statistical models to the need for more computational power and algorithm development. Auspiciously computational power is perpetually rising and the algorithm development is straightforward, albeit challenging. The needs of statistical models, however, are, more insurmountable. All the same, statistical modeling and computational modeling can and ideally do work in tandem with one informing the other and each capitalizing on the strengths of the other.

There are many avenues for the causal inference research agenda ahead. For the particular substantive problem of deriving causal inferences from observational data, identifying an optimal subset is clearly interesting. However, the goal is to identify not only an optimal subset but to identify a host of *independent* subsets that satisfy a randomization test and could thus be regarded as a subset that might have emerged from a randomized experimental design. Together, all of these subsets represent the distribution of the treatment effect. Identifying the distribution of treatment effects would be a significant step forward for causal inference models, and one potentially obtainable via computational modeling. The goal of the optimization process, then, can be seen as two-fold. It is important to be able to search the entire space to identify optimal solutions, but this search must be sufficiently diversified to distinguish independent solutions. This is difficult for optimization routines because the nature of these algorithms tends to identify an area of the solution space where an optimum may lie and then scour that neighbor space for the optimal solution. For causal inference, there are potentially many subsets that would be statistically indistinguishable from the treatment group. An experiment, after all, may be repeated with a different but equally valid randomized group. Indeed, an experiment can be replicated a large number of times, each subsequent iteration offering new information, and each future one as valid as those that preceded it.

The causal inference problem, akin to all interesting problems, embodies its own set of peculiarities. Insights are gained when optimization tools are developed with deep understanding of the statistical theory underlying the process. Domain knowledge and expertise are integral and cannot be replaced with a default algorithm. Incorporating operations

research tools and the insights from computational modeling provide a promising avenue for pursuing significant advances in models for causal inference.

9. ACKNOWLEDGMENTS

This research is part of the Illinois allocation of the Blue Waters sustained-petascale computing project, OCI–0725070 and ACI–1238993, and the state of Illinois, under the project "A Computational Model for Causal Inference via Subset Selection" (project number: *jtt*). Blue Waters is a joint effort of the University of Illinois at Urbana-Champaign and its National Center for Supercomputing Applications. Earlier work on this project was funded by the National Science Foundation.

10. REFERENCES

[1] W. K. T. Cho, J. J. Sauppe, A. G. Nikolaev, S. H. Jacobson, and E. C. Sewell. An optimization approach for making causal inferences. *Statistica Neerlandica*, 67(2):211–226, May 2013.

[2] W. G. Cochran and G. Cox. *Experimental Designs*. Chapman & Hall, London, 1957.

[3] T. D. Cook and D. T. Campbell. *Quasi-Experimentation: Design & Analysis Issues for Field Settings*. Houghton Mifflin, 1979.

[4] T. D. Cook and M. R. Payne. Objecting to the objections to using random assignment in educational research. In *Evidence Matters: Randomized Trials in Education Research*, pages 150–178. Brookings Institution Press, 2002.

[5] P. V. da Veiga and R. P. Wilder. Maternal smoking during pregnancy and birthweight: A propensity score matching approach. *Maternal and Child Health Journal*, 12(2):194–203, March 2008.

[6] R. Dehejia and S. Wahba. Causal effects in non-experimental studies: Re-evaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999.

[7] A. Diamond and J. 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.

[8] E. Erez. Randomized experiments in correctional context: Legal, ethnical, and practical concerns. *Journal of Criminal Justice*, 14(5):389–400, 1986.

[9] R. A. Fisher. *Design of Experiments*. Hafner, New York, 1935.

[10] B. B. Hansen and J. Bowers. Covariate balance in simple, stratified and clustered comparative studies. *Statistical Science*, 23(2):219–236, 2008.

[11] P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[12] K. Imai. Do get-out-the-vote calls reduce turnout? the importance of statistical methods for field experiments. *American Political Science Review*, 99(2):283–300, 2005.

[13] S. Johnston, J. Rootenberg, S. Katrak, W. Smith, and J. Elkins. Effect of a u.s. national institutes of health programme of clinical trials on public health and costs. *Lancet*, 367(9519):1319–1327, April 2006.

[14] R. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76:604–20, September 1986.

[15] Y. Y. Liu and S. Wang. A scalable parallel genetic algorithm for the generalized assignment problem. *Parallel Computing*, 2015.

[16] J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9 (1923). *Statistical Science*, 5(4):465–472, 1923 [1990]. reprint. Transl. by Dabrowska and Speed.

[17] A. G. Nikolaev, S. H. Jacobson, W. K. T. Cho, J. J. Sauppe, and E. C. Sewell. Balance optimization subset selection (boss): An alternative approach for causal inference with observational data. *Operations Research*, 61:398–412, March/April 2013.

[18] N. J. Radcliffe. Genetic set recombination. In L. D. Whitley, editor, *Foundations of Genetic Algorithms 2*. Morgan Kaufmann Publishers, San Mateo, CA, 1993.

[19] L. M. Reinisch, S. A. Sanders, E. L. Mortensen, and D. B. Rubin. In utero exposure to phenobarbital and intelligence deficits in adult men. *The Journal of the American Medical Association*, 274:1518–1525, 1995.

[20] P. R. Rosenbaum. Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032, 1989.

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

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

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

[24] D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *Annals of Statistics*, 6(1):34–58, 1978.

[25] D. B. Rubin. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, 4:1213–1234, 1991.

[26] H. A. Witkin, S. A. Mednick, F. Schulsinger, E. Bakkestrom, K. O. Christiansen, D. R. Goodenough, K. Hirschhorn, C. Lundsteen, D. R. Owen, J. Philip, D. B. Rubin, and M. Stocking. Criminality in xyy and xxy men. *Science*, 193:547–555, 1976.

[27] J. Zubizarreta. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107:1360–1371, 2012.