Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Balance Optimization Subset Selection (BOSS): An Alternative Approach for Causal Inference with Observational Data

Alexander G. Nikolaev, Sheldon H. Jacobson, Wendy K. Tam Cho, Jason J. Sauppe, Edward C. Sewell,

To cite this article:
Alexander G. Nikolaev, Sheldon H. Jacobson, Wendy K. Tam Cho, Jason J. Sauppe, Edward C. Sewell, (2013) Balance Optimization Subset Selection (BOSS): An Alternative Approach for Causal Inference with Observational Data. Operations Research 61(2):398-412. http://dx.doi.org/10.1287/opre.1120.1118

OPERATIONS RESEARCH
Vol. 61, No. 2, March–April 2013, pp. 398–412
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.1120.1118

Balance Optimization Subset Selection (BOSS): An Alternative Approach for Causal Inference with Observational Data

Alexander G. Nikolaev
Department of Industrial and Systems Engineering, University at Buffalo (SUNY), Buffalo, New York 14260, anikolae@buffalo.edu

Sheldon H. Jacobson
Department of Computer Science, University of Illinois at Urbana–Champaign, Urbana, Illinois 61801, shj@illinois.edu

Wendy K. Tam Cho
Departments of Political Science and Statistics and the National Center for Supercomputing Applications, University of Illinois at Urbana–Champaign, Urbana, Illinois 61801, wendycho@illinois.edu

Jason J. Sauppe
Department of Computer Science, University of Illinois at Urbana–Champaign, Urbana, Illinois 61801, sauppe1@illinois.edu

Edward C. Sewell
Department of Mathematics and Statistics, Southern Illinois University Edwardsville, Edwardsville, Illinois 62026, esewell@siue.edu

Scientists in all disciplines attempt to identify and document causal relationships. Those not fortunate enough to be able to design and implement randomized control trials must resort to observational studies. To make causal inferences outside the experimental realm, researchers attempt to control for bias sources by postprocessing observational data. Finding the subset of data most conducive to unbiased or least biased treatment effect estimation is a challenging, complex problem. However, the rise in computational power and algorithmic sophistication leads to an operations research solution that circumvents many of the challenges presented by methods employed over the past 30 years.

Subject classifications: causal inference; balance optimization; subset selection.
Area of review: Optimization.
History: Received September 2011; revisions received February 2012, July 2012; accepted September 2012. Published online in *Articles in Advance* March 19, 2013.

1. Problem Description

Randomized experiments have been used by a diverse swath of researchers to isolate treatment effects and establish causal relationships. Such experiments have informed our understanding of medicine (e.g., the effect of drugs, the causes of cancer, the benefit of vitamins), and have been instrumental in the implementation of public policy (e.g., shedding insight on the effect of racial campaign appeals, testing the effect of get-out-the-vote appeals, determining the impact of new voting technologies). The randomized experimental framework is best suited for exploring causal inferences. In an experiment, a study population is chosen (ideally) at random, or otherwise, by a careful selection of a convenient sample. Another random process determines whether or not each unit will receive a treatment. Because randomization ensures that the treatment and control units are identical in distribution, save that the treatment units have received a treatment, the treatment effect can then be defined as the difference in response (measurable outcome) between the units in the treatment group and those in the control group. In addition to offering tools for measuring estimation accuracy (e.g., calculating p-values, confidence intervals), randomization is powerful because it allows the effect of treatment to be isolated from that of *confounding* factors.

There are numerous situations where conducting a randomized experiment is impractical or not even possible (due to ethical dilemmas). For example, to determine whether smoking causes lung cancer, it would not be possible to randomly select people to smoke. Similarly, although it would be beneficial to understand the perils of radiation exposure, randomly choosing people and exposing them to high levels of radiation is unethical. Although experiments cannot be conducted for these pressing and important research queries, one can often collect observational data. So, although we would not expose people to situations that might put their health in peril, because these situations do occur, we can observe people who *choose* to smoke or find people who have been inadvertently exposed to radiation. This type of data is called *observational data* because it is observed (rather than created via experiments).

Observational data are both more prevalent than experimental data and available for a larger set of important

queries. Indeed, there are already many instances of research attempting to make causal inferences using observational data. In the health field, for example, studies have examined the impact of generic substitution of presumptively chemically equivalent drugs (Rubin 1991), the consequences of in utero exposure to phenobarbital on intelligence deficits (Reinisch et al. 1995), and the effect of maternal smoking on birthweight (da Veiga and Wilder 2008). Public policy applications of causal analysis have included the impact of different voting technologies for counting votes (Herron and Wand 2007), the varying role of information on voters in mature versus new democracies (Sekhon 2004), and the effect of electoral rules on the presence of the elderly in national legislatures (Terrie 2008). At the same time, there is no consensus on how best to proceed if one wishes to make causal inferences with observational data.

The critical difference between experiments and observational studies is that in experiments, because units are randomly assigned to a treatment, the distributions of their covariates (attributes) in the treatment and control groups are identical, isolating the effect of treatment and permitting its determination in expectation. Although various mechanisms have been proposed for random assignment in the statistical literature to handle such issues (Morris 1985), working with observational data sets requires a different set of tools.

It is well recognized that confounding effects in a data set may exist due to both *observed* (those reflected in the data set) and *unobserved* covariates. Dealing with unobservable covariates is a fundamental challenge for causal inference and requires additional information to supplement the available data, whereas the effects of observed covariates can be isolated by data postprocessing, which has received significant interest from practitioners as reported above. A large body of literature has been sparked by the works of Rubin and Rosenbaum, the first to present definitions, assumptions, and discussions to arrive at a technically sound formulation of the causal inference problem with observational data (see individual references in the text below). This paper makes a contribution to this already rich literature, offering an alternative approach to causal analysis.

In order to analyze observational data, where treatment assignment has already been made (a priori nonrandomly), one must postprocess the data with respect to the observed covariates so as to remove confounding effects by creating treatment and control groups with statistically indistinguishable distributions of their covariates. How to best postprocess observational data and assess the success of this venture is an open question.

To transition from a randomized experimental setting to an observational setting, the nuances and similarities of each must be examined. For unit u, let Y_u^1 (Y_u^0) denote a treated (untreated) response; T_u, a treatment indicator (1 means treated, 0 means not treated); and $\mathbf{X}_u = \{X_{1u}, X_{2u}, \ldots, X_{Ku}\}$, a vector of values for K covariates. In both experimental and observational settings, a population of units is under consideration. For a particular unit u, the causal effect of the treatment (relative to the control) is defined as the difference in response that results from receiving and not receiving the treatment, $Y_u^1 - Y_u^0$. The fundamental problem of causal inference is that it is impossible to observe both values Y_u^1 and Y_u^0 on the same unit u (Holland 1986) (e.g., a person either smokes or does not smoke). The outcome of an observation of a unit is termed the observed response, $T_u Y_u^1 + (1 - T_u) Y_u^0$. The Rubin causal model (Rubin 1974, 1978) reconceptualizes this causal inference framework so that the response under either treatment or control, but not both, needs to be observed for each unit. That is, one statistical solution to the fundamental problem of causal inference is to shift to an examination of an average causal effect over all units in the population, $E(Y_u^1 - Y_u^0) = E(Y_u^1) - E(Y_u^0)$, where $E(Y_u^1)$ is computed from the treatment group and $E(Y_u^0)$ is computed from the control group.

An important consideration is how one determines which units will inform the values of Y_u^1 and Y_u^0. In an observational study, one observes some pool of units who have received a treatment, giving $E(Y_u^1 \mid T = 1)$, and some pool of units who have not received a treatment, giving $E(Y_u^0 \mid T = 0)$. In general, $E(Y_u^1) \neq E(Y_u^1 \mid T = 1)$ and $E(Y_u^0) \neq E(Y_u^0 \mid T = 0)$. Moreover, the average treatment effect (ATE), $E(Y_u^1 - Y_u^0)$, is not the same as the average treatment effect *for the treated* (ATT), $E(Y_u^1 \mid T = 1) - E(Y_u^0 \mid T = 1)$. By design, ATE and ATT are interchangeable if the independence assumption holds. That is, if exposure to treatment ($T = 1$) or control ($T = 0$) is statistically independent of response and covariate values, then the units have been properly randomized into treatment and control pools, rendering ATE and ATT to be the same. This situation is not typically the case in observational studies because units are not randomly placed into treatment and control pools. Instead, $\text{ATT} = E(Y_u^1 \mid T = 1) - E(Y_u^0 \mid T = 1) = E(Y_u^1 \mid T = 1) - E(Y_u^0 \mid T = 0) + \mathcal{B}$, where *selection bias* is present, defined as $\mathcal{B} \equiv E(Y_u^0 \mid T = 0) - E(Y_u^0 \mid T = 1)$.

One approach for estimating treatment effects outside the experimental realm relies on multivariate statistical techniques, which fall under the broad rubric of *matching methods* (Rubin 2006). The core of these methods is to employ tools to match units based on their covariate similarity. This results in each *treatment unit* being matched with a *control unit*. If the matching venture is successful, then treatment and control groups are obtained such that the two groups are similar in their covariates, differing only on the treatment indicator value, thereby reducing the bias in the estimation of treatment effects.

Although this set of techniques has been widely used, there remains a lack of consensus on how best to achieve matching or how to assess the success of a matching process. However, a generally accepted principle is that *balance* on the covariates leads to minimal bias in the

estimated treatment effect (Rosenbaum and Rubin 1985). Here, balance has been loosely understood as similarity between *distributions* of covariates in the treatment and control groups. Therefore, whereas most researchers agree that a reasonable goal of matching procedures is to *obtain balance*, there remains disagreement on how to *measure balance*, leading to a difficulty in assessing how a particular matched group compares to other possible matched groups that achieve varying levels of balance. The resulting lack of guidance is a critical omission, because different matched sets can lead to conflicting conclusions.

Interestingly, few of the existing matching methods directly attempt to obtain optimal covariate balance despite claiming that covariate balance is the measure by which to judge the success of the matching procedure. Instead, researchers perform some type of matching (e.g., propensity score matching, Mahalanobis matching), check to see if the groups appear to be roughly similar, and, if unsatisfied, modify parameters of the matching procedure (e.g., distance metric weights or regression model specification) and repeat (see Figure 1). The point at which to end this iterative procedure is at the discretion of the researcher. By design, researchers are unable to objectively assess the quality of their final matched groups because the benchmark, the matched groups with optimal balance, is unknown. Recognizing this issue, recent work of Diamond and Sekhon (2010) attempts to streamline the process of "match—check balance—adjust and repeat as needed" by using a genetic algorithm to adjust the parameters and weights used in the matching algorithm in order to obtain matched samples with the best possible balance measure.

Other researchers have also begun to move towards the idea of direct optimization of balance within a matched samples framework. In particular, Rosenbaum et al. (2007) introduce the notion of *fine balance*, which "refers to exactly balancing a nominal variable, often one with many categories, without trying to match individuals on this variable" (Rosenbaum et al. 2007, p. 75). This relaxation from exact individual matches on a covariate to equal proportions of individuals in the treatment and control groups for each value of the covariate is central to the approach proposed in this paper. Whereas Rosenbaum et al. (2007) consider fine balance for one (nominal) covariate, with matches required on the rest, this paper extends this concept to all covariates.

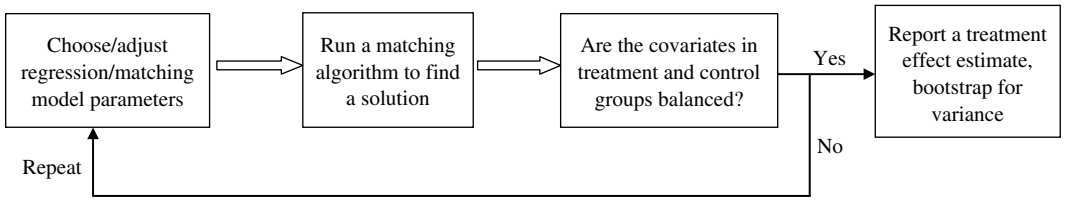
Another recent effort introduced entropy balancing (Hainmueller 2012), which uses a maximum entropy reweighting scheme to adjust weights for each of the control individuals in order to meet user-specified balance constraints placed on the moments of the covariate distributions. For more background on the idea of weighting observations in a data set, see Hellerstein and Imbens (1999).

Matching treatment and control units on an individual level is one method to achieve covariate balance; however it is not a guarantee. We argue that although the focus in the causal inference literature has been on matching, the matching itself of treatment units to control units is not necessary. Notable publications that support the idea of conducting causal analysis on an aggregate, group level include Abadie and Gardeazabal (2003) and Abadie et al. (2010). Matching is not the only way to reduce selection bias, and arguably not even the best way, because one is not interested in *unit* matches per se, but in creating control and treatment *groups* that are *statistically indistinguishable* in the covariates (i.e., featuring covariate balance). Such an observation suggests that a shift in direction is possible in how treatment and control groups can be created.

To realize such a shift, §2 motivates and presents the *Balance Optimization Subset Selection* (BOSS) approach to the problem of causal inference based on observational data. Section 3 reports computational results from one BOSS algorithm for the estimation of treatment effect in a simulated problem. Section 4 offers concluding remarks, discusses the potential of the BOSS approach, raises some theoretical and practical challenges, and outlines several topics for future investigation within the operations research community.

Note that the main contribution of this paper is conceptual and theoretical. The goal of §2 is to present the problem of causal inference in a new light, opening up a field where optimization tools developed within the operations research community can make an impact. By motivating and formalizing an alternative approach to a problem of great importance to multiple domains of modern science, this paper is intended as a seed for more applied, computational-oriented literature. Section 3 is not meant to be comprehensive; instead, it positions itself to illustrate that the proposed theory can shift the problem at hand into the computational realm. It is not intended to deliver comprehensive numerical achievements, but rather supports the call for more intense, goal-driven computational research of BOSS. The electronic companion to this paper is available as supplemental material at http://dx.doi.org/10.1287/opre.1120.1118.

Figure 1. Matching methods logic.



2. BOSS Approach

The presented approach offers an alternative perspective on causal inference using observational data. It exploits the idea that covariate balance leads to minimized bias in the estimated treatment effect by directly optimizing a balance measure without requiring matched samples. As noted in §1, although the success of matching methods is assessed by the degree of balance achieved, very few of the current matching methods directly optimize balance, resorting to different types of optimization problems (e.g., optimal parameter estimation for regression models, optimal assignment for unit matching with calipers). Traditional matching methods simply report balance statistics without a guide to assessing whether the reported balance could be improved upon, is good, or even sufficient. There may be no standard metric to assess the degree of balance achieved; however, a discussion of balance is always presented and perceived as a final verdict, validating a conducted analysis. This simple observation highlights that the problem at hand is a *balance optimization* problem, not a matching problem. Matching is one method to obtain balance, but it unnecessarily restricts the solution space and lacks a measure of balance optimality. Indeed, the end goal is balance, not matching, and hence, optimizing on balance measures is reasonable and preferred.

The BOSS approach to causal inference with observational data reformulates the problem as one of balance optimization (Cho et al. 2011). In so doing, the problem is transformed from matching individual units to a *subset selection* problem, and exploits operations research methodologies (and in particular, discrete optimization) that are ideally suited to model and address the *balance optimization* problem. In essence, BOSS inverts the direction of the solution methodology and redefines the problem structure to directly obtain the goal of covariate balance (see Figure 2). Note that the results of this subset selection approach come at a cost of losing qualitative information of individual matches, which may be useful in some practical situations; however, group-based average quantities can be estimated more precisely.

2.1. The Value of Covariate Balance

To motivate the subset selection problem and explain balance on covariates and why it is required for unbiased estimation of the treatment effect, a formal problem formulation is presented.

Let $\mathcal{S}_N \equiv \{u_i\}_{i=1}^N$ denote a set of N observed units. Define the average treated response $\bar{Y}^1_{\mathcal{S}_N} = (1/N)\sum_{u\in\mathcal{S}_N} Y^1_u$ and the average untreated response $\bar{Y}^0_{\mathcal{S}_N} = (1/N)\sum_{u\in\mathcal{S}_N} Y^0_u$. Given a set of units that have received treatment, *treatment pool* $\mathcal{T}$; a set of units that have not received treatment, *control pool* $\mathcal{C}$; and a set of K covariates, a pair of subsets for comparison is identified: treatment group $\mathcal{S}^{\mathcal{T}}_N \subset \mathcal{T}$ and control group $\mathcal{S}^{\mathcal{C}}_N \subset \mathcal{C}$. To understand the value of covariate balance in causal inference, the following assumption is required (Rosenbaum and Rubin 1983).

Assumption 1 (Strong Ignorability for Groups). *Consider a population of all groups of size N, where $\mathcal{S}_N \equiv \{u_i\}_{i=1}^N$ denotes any such group of N observed units, which are either entirely treated (i.e., $\{T_u = 1\}_{u\in\mathcal{S}_N}$) or untreated (i.e., $\{T_u = 0\}_{u\in\mathcal{S}_N}$). For any set of covariates $\{X_u\}_{u\in\mathcal{S}_N}$, assume*

$$(\bar{Y}^1_{\mathcal{S}_N}, \bar{Y}^0_{\mathcal{S}_N}) \perp\!\!\!\perp \{T_u\}_{u\in\mathcal{S}_N} \mid \{X_u\}_{u\in\mathcal{S}_N}, \tag{1}$$

and

$$0 < P(\{T_u = 1\}_{u\in\mathcal{S}_N} \mid \{X_u\}_{u\in\mathcal{S}_N}) < 1. \tag{2}$$

Expression (1) means that for any group of units, its average responses are independent of treatment, given the units' covariate values. The symbol "$\perp\!\!\!\perp$" signifies conditional independence (Dawid 1979). This implies that the K observed covariates include all the covariates, dependent on the treatment assignment T_u, that have causal effects on the responses Y^1_u and Y^0_u, for every unit u. Additionally, by expression (2), each group with a given set of its units' covariate values is assumed to have a positive probability of appearing in either the treatment pool or control pool. These assumptions are made throughout the statistical literature, albeit for individual units (Rosenbaum and Rubin 1983). Assumption 1 is equivalent to the original assumption of Rosenbaum and Rubin (1983) when $N = 1$. The following proposition captures the objective of any method of postprocessing observational data for causal inference.

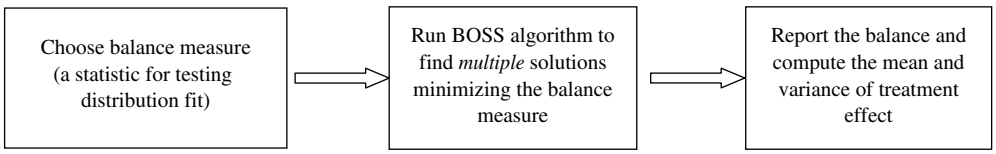
Proposition 1. *Assume that Assumption 1 holds. From the treatment pool, randomly select treatment group $\mathcal{S}^{\mathcal{T}}_N$. Next, randomly select groups of size N from the control pool, until control group $\mathcal{S}^{\mathcal{C}}_N$ is identified such that $\{X_u\}_{u\in\mathcal{S}^{\mathcal{C}}_N} = \{X_u\}_{u\in\mathcal{S}^{\mathcal{T}}_N}$. Then,*

$$E(\bar{Y}^1_{\mathcal{S}^{\mathcal{T}}_N} - \bar{Y}^0_{\mathcal{S}^{\mathcal{C}}_N}) = ATT. \tag{3}$$

Proof. The described mechanism for the selection of $\mathcal{S}^{\mathcal{T}}_N$, and subsequently, $\mathcal{S}^{\mathcal{C}}_N$ ensures that

$$E(\bar{Y}^1_{\mathcal{S}^{\mathcal{T}}_N}) = E_{x\mid\{T_u=1\}_{u\in\mathcal{S}_N}}\big[E(\bar{Y}^1_{\mathcal{S}_N} \mid \{T_u = 1\}_{u\in\mathcal{S}_N} \cap \{X_u\}_{u\in\mathcal{S}_N} = x) \mid \{T_u = 1\}_{u\in\mathcal{S}_N}\big],$$

Figure 2. BOSS logic.



$$E(\bar{Y}^0_{\mathscr{S}^{\mathscr{C}}_N}) = E_{x\,|\,\{T_u=1\}_{u\in\mathscr{S}_N}}\big[E(\bar{Y}^0_{\mathscr{S}_N}\,|\,\{T_u=0\}_{u\in\mathscr{S}_N} \cap \{X_u\}_{u\in\mathscr{S}_N}=x)\,|\,\{T_u=1\}_{u\in\mathscr{S}_N}\big].$$

By definition,

$$\mathrm{ATT} = E(\bar{Y}^1_{\mathscr{S}_N} - \bar{Y}^0_{\mathscr{S}_N}\,|\,\{T_u=1\}_{u\in\mathscr{S}_N}).$$

By conditioning,

$$\mathrm{ATT} = E_{x\,|\,\{T_u=1\}_{u\in\mathscr{S}_N}}\big[E(\bar{Y}^1_{\mathscr{S}_N} - \bar{Y}^0_{\mathscr{S}_N}\,|\,\{T_u=1\}_{u\in\mathscr{S}_N} \cap \{X_u\}_{u\in\mathscr{S}_N}=x)\big],$$

and under Assumption 1,

$$\begin{aligned}\mathrm{ATT} = {} & E_{x\,|\,\{T_u=1\}_{u\in\mathscr{S}_N}}\big[E(\bar{Y}^1_{\mathscr{S}_N}\,|\,\{T_u=1\}_{u\in\mathscr{S}_N} \cap \{X_u\}_{u\in\mathscr{S}_N}=x)\,|\,\{T_u=1\}_{u\in\mathscr{S}_N}\big] \\ & - E_{x\,|\,\{T_u=1\}_{u\in\mathscr{S}_N}}\big[E(\bar{Y}^0_{\mathscr{S}_N}\,|\,\{T_u=0\}_{u\in\mathscr{S}_N} \cap \{X_u\}_{u\in\mathscr{S}_N}=x)\,|\,\{T_u=1\}_{u\in\mathscr{S}_N}\big],\end{aligned}$$

which completes the proof. □

From Proposition 1, the key to causal inference research is the ability to identify control groups with the joint distribution of covariates identical to that of a treatment group. This translates into the property that the probability that (as a group) units in $\mathscr{S}^{\mathscr{C}}_N$ *could be* treated is the same as the probability that units in $\mathscr{S}^{\mathscr{T}}_N$ *are* treated. Note that for individual units (i.e., for $N=1$), this probability is known as the *propensity score*. If the distributions of covariates in groups $\mathscr{S}^{\mathscr{T}}_N$ and $\mathscr{S}^{\mathscr{C}}_N$ are the same, then such groups are said to be *optimally balanced* on the set of the K covariates, rendering $P(\{T_u=1\}_{u\in\mathscr{S}^{\mathscr{C}}_N}) = P(\{T_u=1\}_{u\in\mathscr{S}^{\mathscr{T}}_N})$.

The result of Proposition 1 is for *groups* of units, not individual units. If groups $\mathscr{S}^{\mathscr{T}}_N$ and $\mathscr{S}^{\mathscr{C}}_N$ have one unit each ($N=1$), and these units are perfectly matched ($X_{u\in\mathscr{S}^{\mathscr{C}}_N} = X_{u\in\mathscr{S}^{\mathscr{T}}_N}$), then (3) holds. Similarly, in propensity-score based methods (Rosenbaum and Rubin 1983), regression is used to match units with the same estimated probabilities of being treated, again to have $P(\{T_u=1\}_{u\in\mathscr{S}^{\mathscr{C}}_N}) = P(\{T_u=1\}_{u\in\mathscr{S}^{\mathscr{T}}_N})$ for groups of such units. In all these methods, however, a value assessing covariate balance is judged *after* the data have been postprocessed, with covariate balance not serving as a direct guide for optimal group selection. Although more rigorously designed propensity score models might mitigate this problem to some degree, such potential advances will require deeper statistical design research in the future.

2.2. Modeling and Optimization for Causal Inference

BOSS reframes the causal inference problem as a subset selection problem. The goal is to randomly generate $\mathscr{S}^{\mathscr{T}}$, a subset of $\mathscr{T}$, and find $\mathscr{S}^{\mathscr{C}}$, a subset of $\mathscr{C}$, such that a measure of balance, $M(\mathscr{S}^{\mathscr{T}}, \mathscr{S}^{\mathscr{C}})$, is optimized. This discrete optimization problem can be addressed using operations research algorithms and heuristics. This formulation, moreover, lays the foundation for the development of a new analytical model that exploits the power of ever-increasing computational resources to assess, inform, and improve data analytic techniques.

The BOSS conceptualization is flexible and falls within a general discrete optimization framework. Various measures of balance can be adapted into BOSS. This paper provides a detailed statement of one instance of a balance optimization problem, using a balance measure for a binning model. An intuitive way of comparing distributions is a visual study of histograms based on their *probability mass functions* (pmf) (Imai 2005). Using goodness-of-fit test statistics based on histograms is a more precise and rigorous way of quantifying the difference between covariate distributions for $\mathscr{S}^{\mathscr{T}}$ and $\mathscr{S}^{\mathscr{C}}$.

More formally, for each covariate $k = 1, 2, \ldots, K$, its range $[L_k, U_k]$, with $L_k = \min_{u\in\mathscr{T}\cup\mathscr{C}} X_{ku}$ and $U_k = \max_{u\in\mathscr{T}\cup\mathscr{C}} X_{ku}$, can be broken up by thresholds $L_k = t^k_0 < t^k_1 < t^k_2 < \cdots < t^k_{R(k)} = U_k$. The total number of thresholds $R(k)$ used for covariate $k = 1, 2, \ldots, K$ is typically the number of categories for discrete (categorical) variables and some positive integer for continuous variables. This is similar to the coarsening procedure proposed by Iacus et al. (2012) for coarsened exact matching.

Let covariate cluster D denote a subset of the set of covariates $D \subseteq \{1, 2, \ldots, K\}$. For any covariate cluster $D = \{k_1, k_2, \ldots, k_m\}$ consisting of m covariates, with $1 \leqslant k_1 < k_2 < \cdots < k_m \leqslant K$, define a set of bins $\mathbf{B}^D$ as the set of intervals of the form $[t^{k_1}_{r-1}, t^{k_1}_r] \times [t^{k_2}_{r-1}, t^{k_2}_r] \times \cdots \times [t^{k_m}_{r-1}, t^{k_m}_r]$ that spans the entire joint range of values of the covariates in D. Assuming a given fixed ordering of the elements in $\mathbf{B}^D$, the individual bins are indexed $\{B^D_1, B^D_2, \ldots, B^D_{R_m}\}$, with $R_m \equiv \prod_{j=1}^m R(k_j)$. These bins are used to quantify the difference between the joint distributions of values of covariates in D for groups $\mathscr{S}^{\mathscr{T}}$ and $\mathscr{S}^{\mathscr{C}}$.

Let $\mathscr{N}(\mathscr{S}, B^D_b)$ denote the number of units in group $\mathscr{S}$ with the values of covariates in D contained in bin B^D_b, or the number of units falling into bin b. The objective of the BOSS optimization problem is to minimize the difference between $\mathscr{N}(\mathscr{S}^{\mathscr{C}}, B^D_b)$ and $\mathscr{N}(\mathscr{S}^{\mathscr{T}}, B^D_b)$ over all of the bins for all covariate clusters of interest, where any objective function that simultaneously minimizes these differences can be used to evaluate the distribution fit. The *Balance Optimization Subset Selection with Bins* (*BOSS-B*) problem is now formally stated:

Given: K covariates; a fixed integer N; set $\mathscr{S}^{\mathscr{T}}$, randomly selected from set $\mathscr{T}$ of units represented by vectors $\{X_{1u}, X_{2u}, \ldots, X_{Ku}\}$, $u \in \mathscr{T}$, with $|\mathscr{T}| = N$; set $\mathscr{C}$ of units represented by vectors $\{X_{1u}, X_{2u}, \ldots, X_{Ku}\}$, $u \in \mathscr{C}$, with $|\mathscr{C}| > N$; a set of covariate clusters $\mathbf{D}$; bins $\mathbf{B}^D$ for each $D \in \mathbf{D}$.

Objective: find subset $\mathscr{S}^{\mathscr{C}} \subset \mathscr{C}$ of size N, such that

$$\sum_{D\in\mathbf{D}} \sum_{b=1,2,\ldots,|\mathbf{B}^D|} \frac{(\mathscr{N}(\mathscr{S}^{\mathscr{C}}, B^D_b) - \mathscr{N}(\mathscr{S}^{\mathscr{T}}, B^D_b))^2}{\max(\mathscr{N}(\mathscr{S}^{\mathscr{T}}, B^D_b), 1)} \tag{4}$$

is minimized.

BOSS-B is a balance optimization problem. It exemplifies how the BOSS approach can be used for causal inference, with one measure of balance $M(\mathscr{S}^{\mathscr{T}}, \mathscr{S}^{\mathscr{C}})$ expressed by (4). In BOSS-B, assignments of treatment and control units into groups are determined such that a finite number of preselected marginal and/or joint distributions of covariates are optimally balanced, thereby isolating the effect of treatment from marginal and/or joint effects of these covariates and reducing bias in the estimated expected difference between the treatment and the control responses. The objective function (4) is similar in form to the chi-square test statistic, which provides additional meaning to the formulation. As the distributions get simultaneously balanced, which occurs with an increasing number of bins, the more accurate estimates of the treatment effect can be obtained. However, as more bins are used, resulting in the histogram resolution increasing, optimizing (4) becomes more difficult, because fewer and fewer control groups can be identified as similar to the treatment group. Additionally, the number of required bins for a covariate cluster grows exponentially with the number of covariates in that cluster. Fortunately, this exponential growth is mitigated by the fact that the number of *occupied* bins for any covariate cluster is at most $|\mathscr{T}| + |\mathscr{C}|$.

The decision version of BOSS-B is NP-complete through a polynomial many-one reduction from the "Exact Cover by 3-Sets" problem, which is known to be NP-complete (Garey and Johnson 1979), and hence, the optimization version of BOSS-B is NP-hard (see the online supplement for a formal proof). However, for small-size problem instances, algorithms like simulated annealing are sufficient to deliver good results in reasonable computing time.

Note also that many algorithms solving an instance of BOSS often encounter a large number of optimal or nearly optimal solutions, depending on the binning scheme that is used. As one might intuitively guess, there exist multiple subsets of the treatment and control pools (i.e., solutions to a balance optimization problem) that yield optimal or nearly optimal balance. Swapping out a single unit for another often produces only small changes in the balance function. Often even fairly large differences in subsets result in similar balance values. Accordingly, in addition to finding the optimal balance, it is helpful to also examine the subsets that produce similarly balanced covariates and estimate the *spread* of the distribution of the treatment effect.

2.3. Theoretical Aspects of BOSS-B

This section discusses how solutions to a balance optimization problem can be used to obtain estimates for ATT, and how the estimation bias is reduced as a function of covariate clusters in BOSS-B (more specifically, the number of bins) and the quality of solutions achieved for a given measure of balance. Without loss of generality, assume that $\mathscr{S}^{\mathscr{T}} = \mathscr{T}$. In most real-world observational studies, treated units are rare, and hence, all available such units are included in the treatment group. Therefore, a solution to BOSS-B is a control group that is selected out of a larger control pool of units. Also, for a given instance of BOSS-B, refer to solutions with zero objective function in (4) as *perfectly optimized*. A perfectly balanced solution (i.e., one that has exactly the same joint distribution of covariates in a control group as in the treatment group) is typically perfectly optimized in any measure of balance, though the reverse is not necessarily true. For example, balance on all of the marginal distributions does not generally imply balance on the joint distribution.

Three sources of error are inherent with the application of BOSS-B: error due to noise in the response functions for Y^1 and Y^0; error due to bin size or the number of bins used; error due to nonzero objective function (when a perfectly optimized solution is not found or does not exist).

The first source of error is present in all problems, resulting from the uncertainty inherent in all processes in nature, and hence cannot be eliminated. However, given Assumption 1, the noise in the response has zero mean, and averages to zero for sufficiently large treatment and control groups. The other two sources of error are not so well behaved. However, under certain assumptions, the impact of these errors can be limited. Ideally, one would like to obtain $\mathscr{S}_N^{\mathscr{C}} \subset \mathscr{C}$ that feature perfect balance on the joint distribution of all covariates, $D = \{1, 2, \ldots, K\}$. Note that this condition is equivalent to perfect individual matching, which, if possible, one could find in polynomial time (in the sizes of $\mathscr{T}$ and $\mathscr{C}$, and N) using an assignment algorithm. In practice, however, this is rarely achievable for N large. Therefore, suboptimal solutions may need to be considered, which is why working with observational data is a challenge. Fortunately, perfect balance on the joint distribution of all covariates may not be necessary for accurate inference. This suggests that most real-world causal inference problems can be solved using groups that offer good, albeit not perfect, balance, or using groups that are perfectly balanced on a more limited set of marginal and/or joint distributions of covariates, for making a correct inference. Theorem 1 illustrates the latter point.

THEOREM 1. *Suppose that for any unit u, response $Y_u^{1(0)}$ can be expressed as a sum of functions of individual covariates,*

$$Y_u^{1(0)} = \sum_{k=1,\,2,\,\ldots,\,K} h_k^{1(0)}(X_{ku}) + \epsilon^{1(0)}, \tag{5}$$

where random variable $\epsilon^{1(0)}$ represents noise, with $E(\epsilon^{1(0)}) = 0$. Suppose also that the function $h_k^{1(0)}(X_{ku})$ is locally Lipschitz continuous such that for each $k = 1, 2, \ldots, K$,

$$|h_k^{1(0)}(x_1) - h_k^{1(0)}(x_2)| \leqslant L_k^{1(0)}|x_1 - x_2|, \tag{6}$$

where $L_k^{1(0)}$ is a positive Lipschitz constant for the function $h_k^{1(0)}$, $k = 1, 2, \ldots, K$. Consider an instance of BOSS-B with $\mathscr{S}_N^{\mathscr{T}} = \mathscr{T}$, $N = |\mathscr{T}|$, and $\mathbf{D} = \{\{1\}, \{2\}, \ldots, \{K\}\}$. The bias that arises in the estimation of ATT using an estimator

$\bar{Y}^1_{\mathscr{S}^{\mathscr{T}}_N} - \bar{Y}^0_{\mathscr{S}^{\mathscr{C}}_N}$, *obtained from a perfectly optimized solution* $\mathscr{S}^{\mathscr{C}}_N \subset \mathscr{C}$, *then converges to zero as the number of bins in the sets* $\mathbf{B}^D$, $D \in \mathbf{D}$, *approaches infinity telescopically* (*i.e., the number of bins is increased by uniform sequential subpartitioning*).

Proof. Consider the control group $\mathscr{S}^{\mathscr{C}^{(1)}}_N$, a perfectly optimized solution to an instance of BOSS-B with fixed sets of bins $\mathbf{B}^D$, $D \in \mathbf{D}$. Also, consider control group $\mathscr{S}^{\mathscr{C}^{(2)}}_N$, a perfectly optimized solution to the same instance of the BOSS-B problem, where bin $B^D_r \in \mathbf{B}^D$ for some $D = \{k\} \in \mathbf{D}$, $k \in \{1, 2, \ldots, K\}$, and $r \in \{1, 2, \ldots, R(k)\}$ is partitioned to form bins $B^D_{r_1}$ and $B^D_{r_2}$ such that $B^D_{r_1} \cap B^D_{r_2} = \varnothing$ and $B^D_{r_1} \cup B^D_{r_2} = B^D_r$. Define sets $I_r = \{i\colon i \in \mathscr{S}^{\mathscr{T}}_N,\ X_{ki} \in B^D_r\}$, $J^{(1)}_r = \{j\colon j \in \mathscr{S}^{\mathscr{C}^{(1)}}_N,\ X_{kj} \in B^D_r\}$, and $J^{(2)}_r = \{j\colon j \in \mathscr{S}^{\mathscr{C}^{(2)}}_N,\ X_{kj} \in B^D_r\}$. Let Δ_1, Δ_2, and Δ denote the volumes of bins $B^D_{r_1}$, $B^D_{r_2}$, and B^D_r, respectively. Also, let Z denote the number of control units in $\mathscr{S}^{\mathscr{C}^{(1)}}_N$ falling into bin B^D_r, and let Z_1, Z_2 denote the number of control units in $\mathscr{S}^{\mathscr{C}^{(2)}}_N$ falling into bins $B^D_{r_1}$ and $B^D_{r_2}$, respectively. By design, $\Delta = \Delta_1 + \Delta_2$ and $Z = Z_1 + Z_2$, and $|J^{(2)}_{r_1}| = Z_1$, $|J^{(2)}_{r_2}| = Z_2$ and $|I_r| = |J^{(1)}_r| = Z$.

Proposition 1 describes an approach to select treatment and control groups to ensure that $\bar{Y}^1_{\mathscr{S}^{\mathscr{T}}_N} - \bar{Y}^0_{\mathscr{S}^{\mathscr{C}}_N}$ is an unbiased estimator of ATT. Using this notation, observe that $(1/|I_r|) \cdot \sum_{i \in I_r} Y^1_i$ is an unbiased estimator of $E(\bar{Y}^1_{\mathscr{S}_N} \mid \{T_i = 1\}_{i \in I_r})$, by (5). However, in general, $(1/|J^{(1)}_r|) \sum_{i \in J^{(1)}_r} Y^0_i$ is not an unbiased estimator of $E(\bar{Y}^0_{\mathscr{S}_N} \mid \{T_i = 1\}_{i \in I_r})$, because the exact values in covariate k for the control units falling into a single bin may be different from the values for treatment units in the same bin. As such, an *imbalance* is created within bin B^D_r, because the treatment and control values are not identically distributed within the bin. This imbalance results in a contribution $\mathscr{B}(B^D_r)$ to the bias in the estimation of $E(\bar{Y}^0_N \mid \{T_i = 1\}_{i \in I_r})$ using $\mathscr{S}^{\mathscr{C}^{(1)}}_N$,

$$\mathscr{B}(B^D_r) \equiv \left| \frac{1}{|J^{(1)}_r|} \sum_{j \in J^{(1)}_r} E(Y^0_j) - \frac{1}{|I_r|} \sum_{i \in I_r} E(Y^0_i) \right|.$$

From (5) and (6),

$$\mathscr{B}(B^D_r) = \frac{1}{Z} E\left(\sum_{j \in J^{(1)}_r} h^0_k(X_{kj}) - \sum_{i \in I_r} h^0_k(X_{ki}) \right)$$

$$\leqslant \frac{1}{Z} \sum_{i \in I_r,\, j \in J^{(1)}_r} |h^0_k(X_{kj}) - h^0_k(X_{ki})| \leqslant L^0_k \Delta \equiv U^{(1)},$$

where $U^{(1)}$ is an upper bound on the bias $\mathscr{B}(B^D_r)$. Similarly, by (5), an imbalance within bins $B^D_{r_1}$ and $B^D_{r_2}$ results in contributions $\mathscr{B}(B^D_{r_1})$ and $\mathscr{B}(B^D_{r_2})$, respectively, to the bias in the estimation of $E(\bar{Y}^0_N \mid \{T_i = 1\}_{i \in I_r})$ using $\mathscr{S}^{\mathscr{C}^{(2)}}_N$, with

$$\mathscr{B}(B^D_{r_1}) + \mathscr{B}(B^D_{r_2}) \leqslant \frac{1}{Z} \left(\sum_{i \in I_{r_1},\, j \in J^{(2)}_{r_1}} |h^0_k(X_{kj}) - h^0_k(X_{ki})| + \sum_{i \in I_{r_2},\, j \in J^{(2)}_{r_2}} |h^0_k(X_{kj}) - h^0_k(X_{ki})| \right).$$

Therefore, by (6),

$$\mathscr{B}(B^D_{r_1}) + \mathscr{B}(B^D_{r_2}) \leqslant L^0_k \frac{Z_1\Delta_1 + Z_2\Delta_2}{Z_1 + Z_2} \equiv U^{(2)},$$

which is an upper bound on the bias $\mathscr{B}(B^D_{r_1}) + \mathscr{B}(B^D_{r_2})$. Observe that for $Z_1 > 0$, $Z_2 > 0$, $\Delta_1 > 0$ and $\Delta_2 > 0$, $Z_1\Delta_1 + Z_2\Delta_2 < Z\Delta$, and hence, $U^{(2)} < U^{(1)}$. Moreover, if bin B^D_r is subpartitioned uniformly, which implies $\Delta_1 = \Delta_2$, then $U^{(2)} = U^{(1)}/2$.

Generalizing this argument to a telescopically increasing number of subpartitioned bins, let U denote the bias in the estimation of $E(\bar{Y}^0_{\mathscr{S}_N} \mid \{T_u = 1\}_{u \in \mathscr{S}_N})$ when no optimization is conducted and $\mathscr{S}^{\mathscr{C}}_N \equiv \mathscr{C}$. Observe that because U is finite, then for a perfectly optimized solution $\mathscr{S}^{\mathscr{C}^{|\mathbf{B}|}}_N$ to the instance of BOSS-B with bins $\mathbf{B} = \bigcup_{D \in \mathbf{D}} \mathbf{B}^D$, the total bias can be bounded, and converges to zero as the number of bins, $|\mathbf{B}|$, approaches infinity,

$$\mathscr{B} \equiv \left| \frac{1}{N} \sum_{u \in \mathscr{S}^{\mathscr{C}}_N} Y^0_u - E(\bar{Y}^0_N \mid \{T_u = 1\}_{u \in \mathscr{S}_N}) \right|$$

$$\leqslant \sum_{b \in \mathbf{B}} \mathscr{B}(b) \leqslant \frac{U}{|\mathbf{B}|} \to 0. \quad \square$$

Theorem 1 assumes that the response function (5) is *separable,* meaning that it can be represented as a sum of functions of individual covariates. Although such an assumption may appear restrictive, this class of functions subsumes the class of extensively studied separable models given by

$$Y_u = \beta_0 + \beta_1 \Phi(X_{1u}) + \beta_2 \Phi(X_{2u}) + \cdots + \beta_K \Phi(X_{Ku}) + \epsilon.$$

Furthermore, in the linear modeling literature, if the response function includes a term that is a function of two or more covariates, say $X_{k_1u} * X_{k_2u}$, then the response function can be converted to a linear model by introducing a new covariate that is the product of covariates k_1 and k_2. More generally, if the response function is a function of several covariates, say $\phi(X_{k_1u}, X_{k_2u}, \ldots, X_{k_du})$, with $1 \leqslant k_1 < k_2 < \cdots < k_d \leqslant K$, then the response function can be transformed to satisfy the assumptions of Theorem 1 by introducing a new covariate that is the joint of $X_{k_1u}, X_{k_2u}, \ldots, X_{k_du}$.

Theorem 1 shows that under (5) and (6), as the number of bins in BOSS-B problem grows and perfectly optimized solutions are identified, $\bar{Y}^1_{\mathscr{S}^{\mathscr{C}}_N} - \bar{Y}^0_{\mathscr{S}^{\mathscr{C}}_N}$ monotonically converges to $E(\bar{Y}^1_N - \bar{Y}^0_N \mid \{T_u = 1\}_{u \in \mathscr{S}_N})$, and hence gives the minimally biased estimator of ATT that can be obtained using the available observed data.

3. Computational Analysis

This section illustrates the theory of §2 by presenting a simple numerical example. Note that its contribution to the paper is more illustrative than fundamental. By setting up a computational model for a limited problem and using a

generic optimization algorithm to attack this problem, the reader can visually inspect the dynamics of the proposed balance optimization and the convergence of the proposed estimator to the treatment effect. It also provides grounds to discuss future computational challenges for BOSS.

The simulated experiments presented illustrate that as a balance measure approaches its optimal value, the bias in the estimate of the treatment effect decreases. Additionally, as the number of bins increases, (4) allows for more accurate estimation of the treatment effect.

3.1. Experimental Setup

To illustrate the BOSS-B approach, two data sets were created, designated as *data*3*c*10*k* and *data*10*c*10*k*. Each data set consists of a treatment group of 500 units and a control pool of 10,000 units using 3 and 10 covariates, respectively. The data sets were created by first randomly generating a pool of 5,000 potential treatment individuals and a pool of 10,000 control individuals, with the covariate values for each unit drawn from a normal distribution. Once the units were generated, each unit i was assigned a response value using the expression

$$Y_i^{1(0)} = 10 + 7X_{1i} + 6X_{2i} + 5X_{3i} - 3X_{4i} + 3X_{5i} + 2X_{6i} + X_{7i} - X_{8i} + 0.5X_{9i} + 0.1X_{10i} + \epsilon_i, \quad (7)$$

where $\epsilon_i \sim N(0, 2)$. (The extra covariate terms are omitted for *data*3*c*10*k*.) Under this formulation, there is no treatment effect (i.e., exposure to treatment has no effect on the response): $\mathrm{ATT} = 0$.

Once the individuals were created, a treatment group of 500 units was drawn randomly but nonuniformly from the pool of potential treatment individuals. Individuals with covariate values in the tails of the covariate distribution were drawn with higher probability than those with values in the center of the distributions, ensuring that the resulting treatment and control groups had different covariate distributions. Figure 3 shows the initial distributions in the treatment group and control pool for covariates 1, 2, and 3, respectively, of *data*3*c*10*k*. In these histograms, covariate values are separated into 32 uniformly sized bins. The number of control units in a bin was normalized by a factor of 1/20 to account for the difference in size between the treatment group and control pool. The histograms indicate that the covariate distributions of the treatment group differ from those of the control pool, particularly for the first two covariates.

Optimization was performed using a simulated annealing algorithm (Kirkpatrick et al. 1983). In the experiments, the preselected treatment group was used, and the desired control group size was 500 units. The first step in the algorithm is to bin the data: each unit is converted from a vector of covariate values $\{X_{1i}, X_{2i}, \ldots, X_{Ki}\}$ into a vector of bin numbers $\{X'_{1i}, X'_{2i}, \ldots, X'_{Ki}\}$ where $X'_{ki} = j$ if and only if $t^k_{j-1} \leqslant X_{ki} \leqslant t^k_j$ (i.e., unit i falls into bin j for covariate k). In the experiments, the bin thresholds were uniformly spaced across the covariate distributions, with $R(k)$ set to a given value (an input parameter) for all covariates $k = 1, 2, \ldots, K$. Moreover, a unique covariate cluster was created for each individual covariate. By Theorem 1, these covariate clusters are sufficient for generating an accurate estimate of ATT because of the separability of the response function (7).

After binning the data, the simulated annealing algorithm begins with an initial control group consisting of a random subset of 500 units from the control pool. At each iteration, the algorithm attempts a 1-exchange, replacing one unit in the control group with an unselected unit in the control pool. If the exchange improves (4), then it is accepted unconditionally. Otherwise, it is accepted with some probability according to the input parameters. A random restart is applied when little progress has been made in (4) for some number of iterations or after the algorithm identifies a perfectly optimized control group. The algorithm terminates after performing a preset number of iterations. For more details, see Algorithm 1 in the paper's online supplement.

3.2. Experimental Results

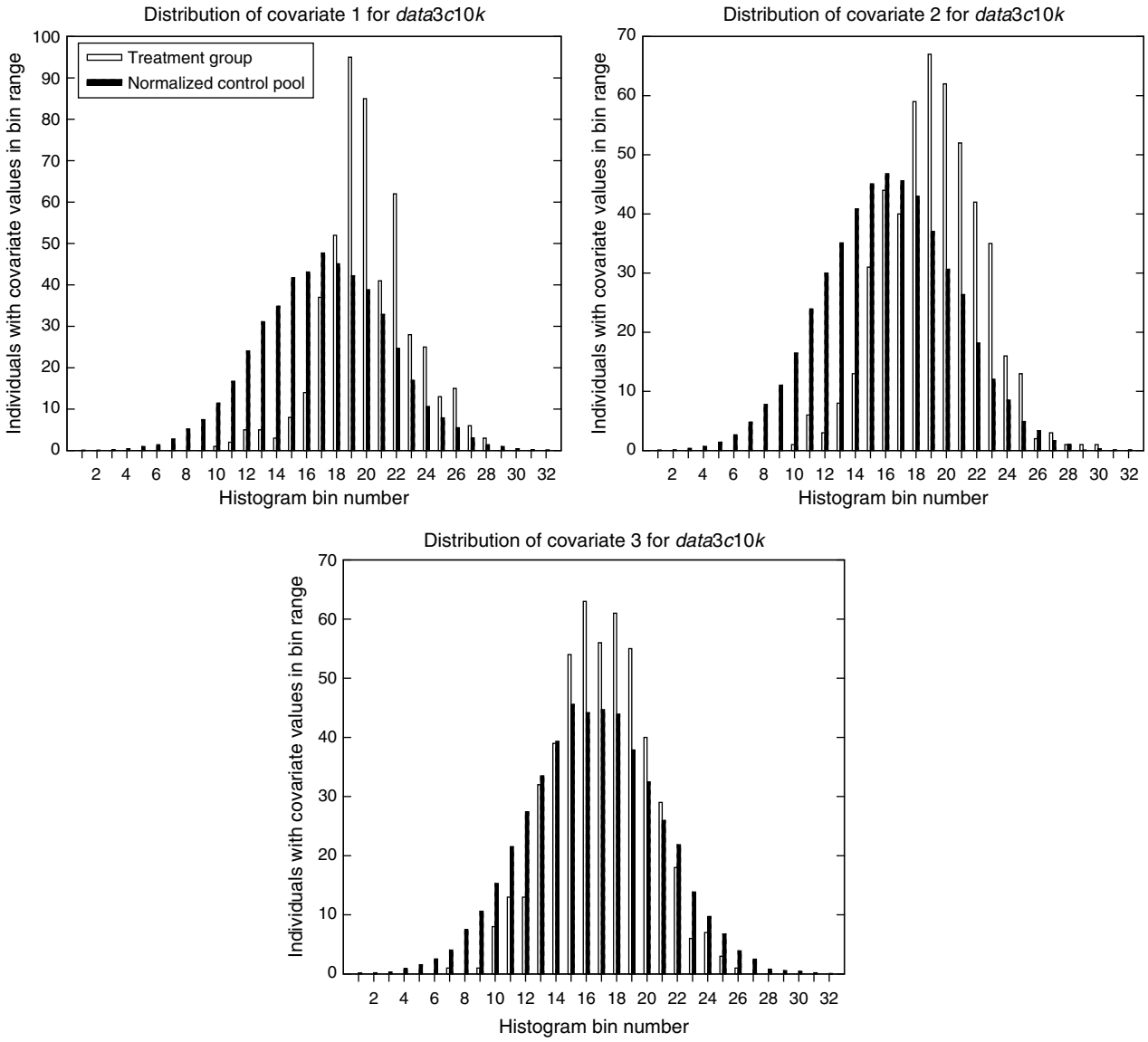
Several experiments were conducted on the two data sets (*data*3*c*10*k* and *data*10*c*10*k*) using uniformly spaced bins with $R(k) = 4, 8, 16$, and 32 for all $k = 1, 2, \ldots, K$. This sequence was chosen because it forms a bin scheme where each successive set of bins simply subdivides the previous set of bins in half, creating a telescopic increase in the number of bins.

For each data set and bin scheme, 25 runs of the simulated annealing algorithm were performed, with a different random seed used for each run. Throughout a run, every 50th identified control group or perfectly optimized control group was processed and stored, along with Kolmogorov-Smirnov (KS) two-sample goodness-of-fit test statistics for the treatment and control covariate distributions. For data sets with multiple covariates, the KS test statistic values were averaged over all the covariates. Upon completion of the experiments, any duplicated control groups were removed. This was implemented by assigning a hash number to each control group based on its units.

Note that because the search process moves by 1-exchange, each successive control group that is reported by the algorithm will have a high degree of overlap with the previously reported control group. To prevent overlap among the perfectly optimized solutions, random restarts were performed after each perfectly optimized solution was identified. This facilitates the generation of perfectly optimized control groups with minimal overlap between them.

Table 1 summarizes the features of optimal solutions obtained in solving the *data*3*c*10*k* instance. In the table, the objective function in (4) is referred to as Difference Squared (*DiffSqr*). Column *Bins* specifies the number of bins used (per covariate), and the column *Observations* reports the number of perfectly optimized solutions that

Figure 3. Initial covariate distributions of treatment group and control pool (normalized) for *data*3*c*10*k*.



were identified. The remaining two columns list the treatment effect and the KS two-sample test statistic (averaged over the covariates), respectively. No results are presented for *data*10*c*10*k* because perfectly optimized solutions were not obtained for this data set when more than four bins per covariate were used.

Table 1. Optimal solutions for *data*3*c*10*k* with respect to DiffSqr objective.

		Treatment effect		Kolmogorov-Smirnov	
Bins	Observations	Mean	SD	Mean	SD
4	25,214	2.2904	0.2684	0.1155	0.0090
8	17,404	1.1434	0.1605	0.0825	0.0072
16	7,689	0.2380	0.1098	0.0369	0.0038
32	833	0.0122	0.0900	0.0274	0.0027
64	0	N/A	N/A	N/A	N/A

Table 1 shows that as the number of bins for each covariate increases, the estimator mean tends toward the true ATT value of zero. The KS test statistic values also indicate an increasingly higher level of balance in the covariate distributions of the treatment and control groups.

Table 2 shows the difference in covariate means for the treatment group and control pool, as well as the difference in covariate means for the treatment group and an optimized control group obtained by solving BOSS-B with $R(k) = 32$ for all $k = 1, 2, \ldots, K$. Observe that the bias due to covariate imbalance in the treatment group and control pool is largely removed by the optimization.

Next, for a given data set and number of bins, all recorded control groups were sorted by their scores in (4).

Table 2. Difference of covariate means for covariates before and after optimization with $R(k) = 32$.

Data set	Covariate	Difference of means: Before optimization	Difference of means: After optimization
*data*3*c*10*k*	1	0.869	0.009
	2	0.862	0.001
	3	0.160	0.007
*data*10*c*10*k*	1	0.539	0.007
	2	0.553	0.014
	3	0.420	0.001
	4	−0.355	0.002
	5	0.446	0.028
	6	0.346	0.007
	7	0.407	0.010
	8	−0.180	0.005
	9	0.208	0.002
	10	0.152	0.009

Then, control groups in a fixed range of scores were aggregated and their estimated treatment effects and other relevant statistic values were averaged. Tables 3 and 4 display these average values obtained with $R(k) = 32$ for all $k = 1, 2, \ldots, K$. Figures 4 and 5 show the trends for the treatment effect and its standard deviation, as the objective function value decreases. In general, as the score for (4) approaches zero, the estimated treatment effect tends toward 0, the true ATT value. Despite the inability to obtain perfectly optimized solutions for *data*10*c*10*k*, accurate ATT estimates are still obtained when the objective function is close to 0.

Table 3. Solutions for *data*3*c*10*k* ranked by DiffSqr objective using 32 bins.

OF range	Observations	Treatment effect: Mean	Treatment effect: SD	Kolmogorov-Smirnov: Mean	Kolmogorov-Smirnov: SD
⩽1e−07	833	0.0122	0.0900	0.0274	0.0027
1e−07–1.0	4,377	0.0679	0.0950	0.0282	0.0028
1.0–2.0	4,675	0.1478	0.1111	0.0294	0.0029
2.0–3.0	3,747	0.2291	0.1173	0.0312	0.0032
3.0–4.0	3,098	0.2948	0.1183	0.0328	0.0034
4.0–5.0	2,751	0.3596	0.1233	0.0344	0.0035
5.0–6.0	2,308	0.4085	0.1304	0.0356	0.0035
6.0–7.0	2,022	0.4666	0.1303	0.0370	0.0036
7.0–8.0	1,873	0.5173	0.1306	0.0381	0.0037
8.0–9.0	1,670	0.5584	0.1315	0.0394	0.0037
9.0–10.0	1,544	0.5881	0.1355	0.0402	0.0038
10.0–20.0	10,937	0.7889	0.1790	0.0449	0.0047
20.0–30.0	8,313	1.1213	0.1828	0.0528	0.0044
30.0–40.0	7,009	1.4045	0.1974	0.0597	0.0046
40.0–50.0	6,148	1.6617	0.1956	0.0659	0.0045
50.0–60.0	5,416	1.8779	0.2050	0.0713	0.0047
60.0–70.0	4,910	2.0778	0.2125	0.0762	0.0048
70.0–80.0	4,437	2.2490	0.2160	0.0808	0.0049
80.0–90.0	3,920	2.4258	0.2159	0.0854	0.0049
90.0–100.0	3,745	2.5803	0.2250	0.0892	0.0052

Table 4. Solutions for *data*10*c*10*k* ranked by DiffSqr objective using 32 bins.

OF range	Observations	Treatment effect: Mean	Treatment effect: SD	Kolmogorov-Smirnov: Mean	Kolmogorov-Smirnov: SD
⩽2.0	0	N/A	N/A	N/A	N/A
2.0–3.0	1	0.2168	0.0000	0.0260	0.0000
3.0–4.0	25	0.2409	0.1056	0.0251	0.0014
4.0–5.0	116	0.2809	0.1113	0.0251	0.0016
5.0–6.0	229	0.3567	0.1065	0.0255	0.0014
6.0–7.0	332	0.4024	0.1198	0.0259	0.0013
7.0–8.0	327	0.4467	0.1189	0.0262	0.0016
8.0–9.0	377	0.4914	0.1200	0.0267	0.0016
9.0–10.0	350	0.5159	0.1225	0.0271	0.0015
10.0–20.0	3,305	0.7416	0.1719	0.0295	0.0021
20.0–30.0	3,105	1.0607	0.1679	0.0328	0.0021
30.0–40.0	2,737	1.3523	0.1748	0.0359	0.0021
40.0–50.0	2,677	1.6002	0.1855	0.0384	0.0022
50.0–60.0	2,608	1.8155	0.1970	0.0409	0.0022
60.0–70.0	2,649	2.0576	0.1899	0.0434	0.0023
70.0–80.0	2,499	2.2616	0.1956	0.0456	0.0024
80.0–90.0	2,527	2.4404	0.2036	0.0477	0.0024
90.0–100.0	2,221	2.6453	0.2113	0.0499	0.0024

Note that in Figures 4 and 5, there is a break where the objective function range changes from increments of 1 to increments of 10 between 9–10 and 10–20. This break is shown with bars in the plot and on the axis. Also, results from control groups with scores for (4) that were greater than 100 are available in the online supplement.

3.3. Comparison with an Alternate Balance Measure

The BOSS framework is not limited to just the BOSS-B formulation presented in §2. Indeed, the goal of the BOSS

Figure 4. *data*3*c*10*k* with 32 bins: Average treatment effect for varying objective function ranges.

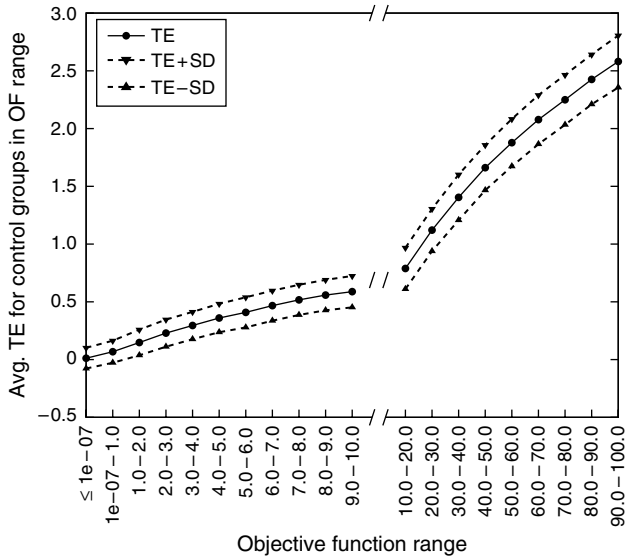
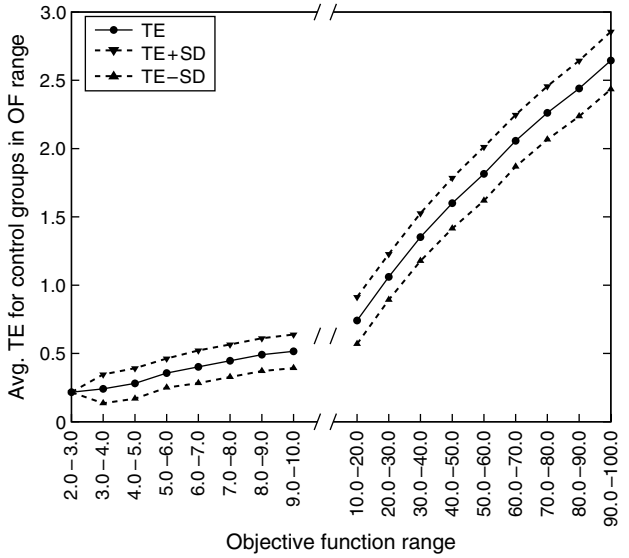


Figure 5. *data*3*c*10*k* with 32 bins: Average treatment effect for varying objective function ranges.



framework is to handle any proposed measure of balance $M(\mathcal{S}^{\mathcal{T}}, \mathcal{S}^{\mathcal{C}})$. For example, one can use a difference of means as an optimization objective. Let $\mu(\mathcal{S}, k) = (1/|\mathcal{S}|)\sum_{s\in\mathcal{S}} X_{ks}$ be the mean value of covariate k across the individuals in $\mathcal{S}$. Then, a BOSS objective is to find a control group $\mathcal{S}^{\mathcal{C}} \subset \mathcal{C}$ with $|\mathcal{S}^{\mathcal{C}}| = |\mathcal{T}|$ that minimizes

$$\sum_{k=1}^{K} \left|\mu(\mathcal{S}^{\mathcal{C}}, k) - \mu(\mathcal{T}, k)\right|. \tag{8}$$

Note that such analysis was done by Rubin (1973) for one covariate, where it was referred to as *mean matching*. With BOSS objective (8), no preprocessing of the data is necessary, because no binning is performed (compared to BOSS-B). Table 5 shows the performance of objective (8), referred to as *DOM* for difference of means, in determining the treatment effect across a wide range of solutions obtained during the simulated annealing algorithm execution. As the score for (8) approaches 0, the estimated treatment effect tends toward the true treatment effect of 0, which is as expected given the linear nature of the response function (7). Results for control groups with scores for (8) greater than 1.00 are available in the online supplement.

Observe that using (8) as a BOSS objective compared to (4) results in more accurate ATT estimation. This observation might lead one to assume that (8) is better than (4) at capturing balance. However, the KS scores are worse with (8), indicating that although the covariate means are close, the covariate *distributions* are not as balanced as those for the solutions obtained with (4). An additional set of experiments was performed to illustrate the importance of balancing the distributions. These experiments used a new data set, *data*3*c*10*kn*, created by taking the same individuals from *data*3*c*10*k* and using the response function

$$Y_i^{1(0)} = 10 + e^{X_{1i}} + X_{2i}^2 + 0.1X_{3i}^3 + \epsilon_i. \tag{9}$$

Table 5. Solutions for *data*10*c*10*k* ranked by *DOM* objective.

		Treatment effect		Kolmogorov-Smirnov	
OF range	Observations	Mean	SD	Mean	SD
⩽0.001	0	N/A	N/A	N/A	N/A
0.001–0.01	12,004	0.0596	0.0857	0.4101	0.0258
0.01–0.02	66,859	0.0789	0.0913	0.4167	0.0276
0.02–0.03	94,364	0.1115	0.0916	0.4201	0.0272
0.03–0.04	94,269	0.1548	0.0920	0.4200	0.0264
0.04–0.05	83,005	0.2015	0.0938	0.4199	0.0265
0.05–0.10	286,406	0.3434	0.1323	0.4236	0.0266
0.10–0.20	374,035	0.7421	0.2066	0.4419	0.0276
0.20–0.30	290,608	1.2774	0.2244	0.4721	0.0291
0.30–0.40	255,131	1.7747	0.2439	0.5027	0.0289
0.40–0.50	238,708	2.2529	0.2560	0.5347	0.0306
0.50–0.60	244,812	2.7030	0.2688	0.5667	0.0301
0.60–0.70	241,576	3.1296	0.2770	0.5999	0.0315
0.70–0.80	226,956	3.5528	0.2829	0.6350	0.0313
0.80–0.90	229,046	3.9600	0.2831	0.6688	0.0312
0.90–1.00	235,354	4.3380	0.2934	0.7032	0.0313

Five runs of the simulated annealing algorithm were performed with *data*3*c*10*kn*, using both (4) with $R(k) = 32$ for all $k = 1, 2, \ldots, K$ and (8). The best solutions obtained from these runs are reported in the first two rows of Table 6. In this case, the best solutions obtained with (4) lead to better estimates of ATT than those obtained with (8). Optimizing (4) results in more accurate estimation because Theorem 1 still holds for (9) due to the separability of the covariate terms. Moreover, the KS scores are better, indicating better balance for the covariate distributions.

The function (8) can be improved by incorporating higher moments of the distributions, such as the variance. Let $s^2(\mathcal{S}, k) = (1/(|\mathcal{S}|-1))\sum_{s\in\mathcal{S}}(X_{ks} - \mu(\mathcal{S}, k))^2$ be the unbiased sample variance of covariate k across the individuals in $\mathcal{S}$. Then two additional BOSS objectives can be defined as

$$\min \sum_{k=1}^{K} \left|\mu(\mathcal{S}^{\mathcal{C}}, k) - \mu(\mathcal{T}, k)\right| + \sum_{k=1}^{K} \left|s^2(\mathcal{S}^{\mathcal{C}}, k) - s^2(\mathcal{T}, k)\right| \tag{10}$$

and

$$\min \sum_{k=1}^{K} \left|\mu(\mathcal{S}^{\mathcal{C}}, k) - \mu(\mathcal{T}, k)\right|^2 + \sum_{k=1}^{K} \left|s^2(\mathcal{S}^{\mathcal{C}}, k) - s^2(\mathcal{T}, k)\right|. \tag{11}$$

These two objectives aim at finding control groups with the first and second moments of the covariate distribution as close as possible to those of the treatment group. Objectives (10) and (11) differ in the weight they place on the difference of means, with (11) squaring this difference for each covariate. For *data*3*c*10*kn*, the results of optimizing these

Table 6. Best solutions for *data*3*c*10*kn* for various objectives.

			Treatment effect		Kolmogorov-Smirnov	
Objective	OF range	Observations	Mean	SD	Mean	SD
DiffSqr(32)	⩽1e–07	156	−0.0170	0.0875	0.0804	0.0078
DOM	⩽0.001	7,086	−1.3889	0.3395	0.2770	0.0226
DOM + *DOV*	⩽0.001	357	0.0392	0.0959	0.1669	0.0179
*DOM*2 + *DOV*	⩽0.001	403	0.0986	0.1057	0.1435	0.0121

two objectives (referred to as *DOM* + *DOV* and *DOM*2 + *DOV*) are much better than those obtained for (8), as shown in Table 6.

In a similar manner, higher moments can be included in the objective being optimized. Including higher moments ensures that the two distributions are closer and closer together, which is exactly what the BOSS-B formulation aims to achieve, albeit in a more direct manner.

3.4. Comparison with Matching Methods

To demonstrate the performance of BOSS with respect to existing matching methods, the *Matching* package (Sekhon 2011) was used. The package allows for matching based on propensity score, matching directly on the values of the covariates, or some combination of the two. For the purposes of testing, a standard logistic regression model was used to estimate the propensity score.

Table 7 compares the best solutions (as defined by the objective function value, with ties broken arbitrarily) obtained by the BOSS procedure for objectives (4) with $R(k) = 32$ for all $k = 1, 2, \ldots, K$, (8), (10), and (11) with the solutions returned by both propensity score matching and matching on the covariates for the *data*3*c*10*kn* data set (with the nonlinear response function (9)). Column *Objective* lists the method used to obtain the solution, column *OF Score* lists the function value of the best solution for the BOSS methods (no objective score is provided by the *Matching* package), column *Treatment Effect* lists the estimate of the treatment effect computed from the best solution, and columns *Kolmogorov-Smirnov Mean* and *Max* list the average and maximum values of the KS test statistic for the covariate distributions in the treatment group and the best control group.

The propensity score model fares the worst in producing accurate estimates of the treatment effect, whereas direct matching and BOSS with objective functions (4), (10), and (11) all produce good results. The reason for the poor performance of the propensity score approach is the use of a linear model for estimating the propensity score, whereas the actual response function is nonlinear. A better model for estimating the propensity score would potentially improve these results. It should also be noted that the propensity score approach produces the worst balance as measured by the KS statistic, whereas BOSS with objective function (8) also produces unsatisfactory levels of balance, with BOSS with objective function (4) and covariate matching performing the best.

A difficulty of matching on the covariates is that close matches become difficult to find as the number of covariates increases. To demonstrate this, the matching procedures were also run on the *data*10*c*10*k* data set. Table 8 shows the best solutions obtained by the BOSS approaches and the matching approaches. Because *data*10*c*10*k* uses a linear response function (7), both propensity score matching and BOSS with (8) perform better than they did in the previous case. This improvement occurs because balancing covariate means for a linear response function produces accurate ATT estimates. Estimating the propensity score with a linear model will accomplish this indirectly, whereas optimizing (8) will accomplish this directly. On the other hand, the effectiveness of covariate matching is greatly reduced due to the difficulty of finding close matches on 10 different covariates. Finally, BOSS with (4) is seen to produce the best covariate balance as measured by the KS test statistic, whereas the matching approaches produce the worst covariate balance.

Table 7. Comparison of single best solutions for BOSS and matching for *data*3*c*10*kn*.

		Treatment	Kolmogorov-Smirnov	
Objective	OF score	effect	Mean	Max
DiffSqr(32)	0.0	−0.1142	0.025	0.026
DOM	1.50e–5	−0.9877	0.093	0.118
DOM + *DOV*	3.77e–4	0.0271	0.062	0.088
*DOM*2 + *DOV*	2.69e–4	0.1154	0.045	0.060
Prop. score	N/A	−1.3434	0.125	0.158
Cov. matching	N/A	0.0943	0.025	0.034

Table 8. Comparison of single best solutions for BOSS and matching for *data*10*c*10*k*.

		Treatment	Kolmogorov-Smirnov	
Objective	OF score	effect	Mean	Max
DiffSqr(32)	2.9502	0.2168	0.026	0.036
DOM	0.0029	0.1294	0.039	0.056
DOM + DOV	0.0157	0.1857	0.037	0.048
*DOM*2 + DOV	0.0158	0.1947	0.045	0.052
Prop. score	N/A	−0.1148	0.066	0.114
Cov. matching	N/A	2.818	0.067	0.088

3.5. Discussion of Results

Inspecting the reported results with the goal of evaluating the potential effectiveness of the BOSS approach, the conducted experiments well illustrate the theory of §2. The simulated annealing algorithm was able to perform well for BOSS-B and several other objectives, which suggests that specialized algorithms could be much more effective and efficient in finding optimal balance. Additionally, the BOSS approach performed favorably when compared with some of the existing matching methods proposed in the literature.

The accurate estimates of ATT produced by BOSS in these experiments suggest that BOSS may be a viable approach to successfully determine whether or not a treatment effect exists in problems that approximate real-world scenarios for which observational data exists. For the BOSS-B formulation in particular, as $R(k)$ increases, (4) provides a better measure of covariate balance, and hence a better estimate of the treatment effect. However, as $R(k)$ increases, it also becomes more difficult to identify control groups that are perfectly optimized with respect to (4). Certainly there are improvements that can be made in terms of the optimization process, but determining the appropriate value for $R(k)$ and even the appropriate bin thresholds will be a major factor as well. For the former, Cochran (1968) states that for one covariate, subclassification with five categories is sufficient to remove about 90% of the existing bias under certain conditions. Rosenbaum and Rubin (1983) present similar results when subclassifying on the propensity score. Determining the appropriate locations for bin thresholds will be dependent upon the nature of the data. See Iacus et al. (2012) for further discussion of these issues.

Another issue is determining which covariate clusters to use. In the experiments presented here, the covariate clusters were chosen based on knowing the separability of the response function. In a real-world problem, the response function will almost certainly be unknown, and therefore some guesswork will be involved in appropriately picking the covariate clusters.

For the general BOSS problem, there remains significant work to be done in determining appropriate balance measures for optimization. In the simulated example problems considered here, the difference of means objective (8) was sufficient for a separable linear response function, but not for a separable nonlinear one. Although incorporating the variance into the objective (10) yielded more accurate results for the nonlinear response function, this may not always be the case. Determining exactly what balance measures should be optimized remains an open problem.

4. Research Directions

BOSS introduces a new paradigm for developing an analytical toolbox based on techniques from operations research to create a solution methodology where human bias, associated, for example, with defining distance measures for matching or guessing the form of a regression model, is eliminated, and the accuracy of treatment effect estimation is limited solely by the complexity of an optimization problem (NP-hard) and available computational power.

To make a connection between the balanced marginal distributions and the balanced joint distributions of covariates, the concept of copulas (Nelsen 1999) may be useful if a copula family can be designed to incorporate continuous and categorical covariate values simultaneously with a sizable number of parameters. In many cases, however, preserving the same covariance structure over the covariate values in the control and treatment groups might suffice. For example, if a treatment group consists only of pairs AA and BB, they would have the same marginal distributions as a control group with pairs AB and BA, because both A and B appear twice; the joint distributions, however, would not align. Examining covariance structures would identify and help alleviate this issue. One approach would be to minimize the covariance matrix difference directly, incorporating it into BOSS as part of the objective function or as a constraint. Note that some widely used matching approaches (e.g., propensity score matching) operate under the Stable Unit Treatment Value Assumption (SUTVA) that is violated when observations on one unit are affected by the particular assignment of treatment to other units. The BOSS approach also relies on this strong assumption, even though it may not hold in real observational studies and randomized experiments.

The issue of space traversal, or how well BOSS explores the space of available control groups, is also a rich area for future exploration. For algorithms that generate a large number of optimal or near-optimal solutions, ensuring that these solutions are sufficiently diverse will allow for better estimates on the distribution of the treatment effect. One way in which this can be accomplished is by iteratively running the BOSS algorithm, finding an optimal control group, removing the members of the control group from the control pool, and then rerunning the BOSS algorithm using the smaller control pool. Alternatively, control individuals can be prevented from being used in a control group after appearing in some number of other identified control groups.

In problems with a large number of covariates and/or covariate clusters to balance, it is unlikely that perfectly optimized control groups exist when using even a moderate number of bins for each covariate. Therefore, further research on binning-based measures of balance is required, and bounds are needed on the quality of a control group when it is not perfectly optimized. In the simulated experiments reported in §3, it was observed that many control groups that were near-optimal led to the correct decision with regards to the effectiveness of treatment, although the exact dynamics of this phenomenon are not completely clear. Alternate ways to assess the quality of a control group in addition to the objectives presented here should also be considered.

Additionally, developing algorithms to optimize directly on covariate balance measures such as the Kolmogorov-Smirnov two-sample test statistic instead of using approximation techniques as binning is a promising direction. In the current implementation, using the KS score instead of objective (4) caused the search process to stall and fail to make significant progress. This suggests that a 1-exchange neighborhood is insufficient when used in conjunction with the KS score.

For BOSS to be useful in practice, computational tools need to be developed that can analyze distribution(s) of the designed estimator(s). Besides point estimation, social scientists often resort to hypothesis testing as well as building confidence intervals, the tasks where estimating standard error becomes important. Although our computational investigations indicate that the distribution of the BOSS estimators presented in this paper appears to be Gaussian, more research is required to establish this result theoretically for the subset-selection based approach.

The challenges presented should be addressed simultaneously by research communities over various domains of science. Statisticians might be interested in developing a copula approach for the balancing of joint distributions, whereas operations researchers and computer scientists might work on more efficient optimization algorithms. Opportunities for interdisciplinary collaboration may prove to be fruitful as this research direction continues to expand and evolve.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/opre.1120.1118.

Acknowledgments

The authors thank Alexander Shapiro, the associate editor, and two anonymous referees for their helpful comments, which greatly improved the presentation of this paper and led to more substantial computational results. This research has been supported in part by the National Science Foundation [SES-0849223 and SES-0849170]. The second author was also supported in part by the Air Force Office of Scientific Research [FA9550-10-1-0387]. The fourth author was supported by the Department of Defense (DoD) through the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program (32 CFR 168a). This material is based upon work supported in part by (while serving at) the National Science Foundation. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views 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.

References

Abadie A, Gardeazabal J (2003) The economic costs of conflict: A case study of the Basque country. *Amer. Econom. Rev.* 93(1):112–132.

Abadie A, Diamond A, Hainmueller J (2010) Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *J. Amer. Statist. Assoc.* 105(490):493–505.

Cho WKT, Sauppe JJ, Nikolaev AG, Jacobson SH, Sewell EC (2011) An optimization approach to matching and causal inference. Technical report, University of Illinois at Urbana–Champaign, Urbana, IL.

Cochran WG (1968) Effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 24(2):295–313.

da Veiga PV, Wilder RP (2008) Maternal smoking during pregnancy and birthweight: A propensity score matching approach. *Maternal and Child Health J.* 12(2):194–203.

Dawid AP (1979) Conditional independence in statistical theory. *J. Roy. Statist. Soc. Ser. B* 41(1):1–31.

Diamond A, Sekhon JS (2010) Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. Technical report, Department of Political Science, University of California, Berkeley, Berkeley, CA. Accessed July 2011, http://sekhon.berkeley.edu/papers/GenMatch.pdf.

Garey MR, Johnson DS (1979) *Computers and Intractability: A Guide to the Theory of NP-Completeness* (Freeman and Company, San Francisco).

Hainmueller J (2012) Entropy balancing: A multivariate reweighting method to produce balanced samples in observational studies. *Political Anal.* 20(1):25–46.

Hellerstein J, Imbens G (1999) Imposing moment restrictions from auxiliary data by weighting. *Rev. Econom. Statist.* 81(1):1–14.

Herron MC, Wand J (2007) Assessing partisan bias in voting technology: The case of the 2004 New Hampshire recount. *Electoral Stud.* 26(2):247–261.

Holland PW (1986) Statistics and causal inference. *J. Amer. Statist. Assoc.* 81(396):945–960.

Iacus SM, King G, Porro G (2012) Causal inference without balance checking: Coarsened exact matching. *Political Anal.* 20(1):1–24.

Imai K (2005) Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments. *Amer. Political Sci. Rev.* 99(2):283–300.

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220(4598):671–680.

Morris C (1985) A finite selection model for experimental design of the health insurance study. *J. Econometrics* 11(1):43–61.

Nelsen RB (1999) *An Introduction to Copulas* (Springer, New York).

Reinisch LM, Sanders SA, Mortensen EL, Rubin DB (1995) In utero exposure to phenobarbital and intelligence deficits in adult men. *J. Amer. Medical Assoc.* 274(19):1518–1525.

Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

Rosenbaum PR, Rubin DB (1985) Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* 39(1):33–38.

Rosenbaum PR, Ross RN, Silber JH (2007) Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Amer. Statist. Assoc.* 102(477):75–83.

Rubin DB (1973) Matching to remove bias in observational studies. *Biometrics* 29(1):159–183.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psych.* 66(5):688–701.

Rubin DB (1978) Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* 6(1):34–58.

Rubin DB (1991) Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics* 47(4):1213–1234.

Rubin DB (2006) *Matched Sampling for Causal Effects* (Cambridge University Press, New York).

Sekhon JS (2004) The varying role of voter information across democratic societies. Working paper, Department of Political Science, University of California, Berkeley, Berkeley, CA. Accessed January 2012, http://sekhon.berkeley.edu/papers/SekhonInformation.pdf.

Sekhon JS (2011) Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *J. Statist. Software* 42(7):1–52. Accessed January 2012, http://www.jstatsoft.org/v42/i07.

Terrie L (2008) Using matching to assess the effect of electoral rules on the presence of the elderly in national legislatures. Poster presented at 2008 Political Methodology Meetings. Society for Political Methodology, American Political Science Association, Washington, DC.

Alexander G. Nikolaev is an assistant professor in the Department of Industrial and Systems Engineering at the University at Buffalo. His research interests include stochastic optimization, statistical inference, and social network modeling.

Sheldon H. Jacobson is a professor and director of the Simulation and Optimization Laboratory in the Department of Computer Science at the University of Illinois. He has a diverse set of basic and applied research interests, including problems related to optimal decision making under uncertainty, discrete optimization, causal inference with observational data, aviation security, public health policy (immunization, transportation and obesity, cell phone ban effectiveness), March Madness bracketology, and forecasting the outcome of the United States presidential election.

Wendy K. Tam Cho is a professor in the Department of Political Science and Department of Statistics, and Senior Research Scientist at the National Center for Supercomputing Applications, all at the University of Illinois at Urbana–Champaign.

Jason J. Sauppe is a Ph.D. candidate in the Department of Computer Science at the University of Illinois. His current research interests include mathematical programming, discrete optimization, and approximation.

Edward C. Sewell is a Distinguished Research Professor of Mathematics and Statistics at Southern Illinois University at Edwardsville. His current research interests are combinatorial optimization and health applications.