Journal of Machine Learning Research 17 (2016) 1-31 Submitted 8/15; Revised 10/16; Published 2/16

Mutual Information Based Matching for Causal Inference with Observational Data

Lei Sun LEISUN@BUFFALO.EDU
Department of Industrial and Systems Engineering
University at Buffalo, Buffalo, NY 14260, USA

Alexander G. Nikolaev ANIKOLAE@BUFFALO.EDU
Department of Industrial and Systems Engineering
University at Buffalo, 312 Bell Hall, Buffalo, NY 14260, USA
Department of Computer Science and Information Systems
University of Jyvaskyla, Jyvaskyla, FIN-40014, Finland

Editor: Peter Spirtes

Abstract

This paper presents an information theory-driven matching methodology for making causal inference from observational data. The paper adopts a "potential outcomes framework" view on evaluating the strength of cause-effect relationships: the population-wide average effects of binary treatments are estimated by comparing two groups of units – the treated and untreated (control). To reduce the bias in such treatment effect estimation, one has to compose a control group in such a way that across the compared groups of units, treatment is independent of the units' covariates. This requirement gives rise to a subset selection / matching problem. This paper presents the models and algorithms that solve the matching problem by minimizing the mutual information (MI) between the covariates and the treatment variable. Such a formulation becomes tractable thanks to the derived optimality conditions that tackle the non-linearity of the sample-based MI function. Computational experiments with mixed integer-programming formulations and four matching algorithms demonstrate the utility of MI based matching for causal inference studies. The algorithmic developments culminate in a matching heuristic that allows for balancing the compared groups in polynomial (close to linear) time, thus allowing for treatment effect estimation with large data sets.

Keywords: Observational Causal Inference, Mutual Information, Matching, Subset Selection, Optimization

1. Introduction

The tools for making inference based on observational data are useful for estimating the effects of binary treatments that are non-randomly assigned to the units of a studied population (Cochran, 1965). Causal investigations are of importance in various domains of science including economics (Abadie and Imbens, 2006), medical research (da Veiga and Wilder, 2008), political science (Ho et al., 2007), sociology (Morgan and Harding, 2006), law (Rubin, 2001), etc. As a conventional recipe, *matching* of treated and untreated units allows one to compare them and distill the effect of the treatment, while blocking the effects of confounding unit covariates.

The most widely adopted conventional matching methods employ various distance metrics (e.g., Mahalanobis distance) and propensity scores (see Section 2 for a detailed review); the success of a matching venture is typically assessed by checking if the compared groups are "well-balanced", i.e., if the distributions of covariates within them are similar. The methods introduced more recently strive to directly optimize balance (Zubizarreta, 2012). In particular, Nikolaev et al. (2013) re-cast matching as a subset selection problem with the objective to optimize a measure of covariate balance across groups (as opposed to individual unit pairs). The approach was coined Balance Optimization Subset Selection, with its applicability illustrated by employing linear programming models (Nikolaev et al., 2013) and simulated annealing heuristics (Tam Cho et al., 2013).

Note, however, that improving balance, expressed via some metric(s) capturing the difference between the distributions of covariates in the compared groups, is just one approach that defines a matching procedure objective. It is as good as any other approach that would achieve the reduction of the dependence between the covariates and the treatment variable in the matched groups. This observation is exploited in the present paper, as it explores a new form of covariate balance and an alternative approach to doing matching.

This paper frames matching as an optimization problem with a mutual information (MI) based objective. The presented methods are non-parametric, and hence, do not suffer from human bias in model selection. The value of information theory in empirical statistics research and computer science has been emphasized over the past decade (Burnham and Anderson, 2002). However, while this thrust has been successful in facilitating hypothesis testing, optimization problems with information measures have proven to be difficult, mainly due to the inherent non-linearity of entropy and MI functions (Shannon, 1948). This paper presents a way to treat such non-linearity in subset selection problems, which arise in applying information theory logic for making causal inference with observational data.

MI has been used to formulate various problems involving feature selection (Estévez et al., 2009), dependency analysis (Kraskov et al., 2004) and chaotic data identification (Fraser and Swinney, 1986). It measures the level of dependence between random variables; e.g., when evaluated for two variables, it takes a high value when one random variable contains much information about the other, signifying high dependence, while zero MI implies that the variables are independent. We show that the difference between the covariate distributions among the treated and untreated units can be directly evaluated MI, exploiting the fact that randomization in treatment assignment implies zero MI between covariates and the treatment variable. To the best of the authors' knowledge, no MI based method has yet been employed for grouping observations (units) to achieve a particular group property – most likely due to the non-linearity in the expression defining MI. This paper tackles this challenge and offers the models and algorithms that make theoretical and practical advances in subset selection, or simply, matching for treatment effect estimation.

While some optimization methods have already been employed for the methodological developments in causal inference with observational data (Hansen, 2004), the use of mathematical programming techniques for statistics-oriented applications is still rare. One such notable contribution is due to Bertsimas and Shioda (2007) who re-framed the classification and regression problems using integer programming. Similar to their efforts, this paper motivates the use of non-linear integer programming techniques in causal inference research.

First, this paper identifies pathways for the effective use of information theoretic measures (namely, MI) in optimization problems. The presented theoretical analysis techniques for treating non-linearity are generic, and hence, can be adopted in other applications, where making assumptions on model/data structures is undesirable. More generally, this paper may open up venues for the application of mathematical programming and optimization techniques in information theory itself.

Second, this paper explains how MI can serve as the basis of a new form of covariate balance. The resulting MI-based matching method for selecting control groups for causal inference is flexible in that it can achieve solutions of pre-specified quality, with pre-set control group size, – moreover, it can optimize the latter. The presented algorithmic developments produce a matching heuristic that runs in polynomial (close to linear) time: it thus allows for causal effect estimation with large data sets that are nowadays becoming available through mining social networks, health records, etc. While this work is not the first effort to employ the information theoretic tools for the needs of causal inference Hainmueller (2012), it appears to be the first where mutual information is used as an optimization objective.

The paper is organized as follows. Section 2 explains the problem of causal inference with observational data, and motivates optimization-driven subset selection approaches to attacking it. Section 3 introduces a class of MI-based matching problems with different objectives. Section 4 derives optimality conditions for matched groups using MI, and presents the mixed integer programming-based and sequential selection-based matching algorithms that work to balance the covariate distributions across the treatment and control groups. Section 5 showcases the practical value of the MI-based matching approach by comparing the designed algorithms' performance against the best previously existing matching methods. Section 6 discusses the MIM limitations and future research directions. Section 7 provides concluding remarks and discusses the promising extensions of this line of work.

2. Causal Inference with Observation Data

Observational studies are often the only source of information about a program, policy, or treatment. For example, people non-randomly choose to participate in economy-boosting programs, political movements, online activities such as post re-tweeting, question answering, service subscription, etc. In estimating any causal effect with such data, the researchers resort to the nonparametric data preprocessing, commonly referred to as matching (Ho et al., 2011).

In a real-world causal inference problem instance, a treatment group (a group of *treated* units) is typically smaller than the size of a pool of available *control* (untreated) units; a control *group* can then be selected by a researcher from this pool. When a matching procedure is performed (Rubin, 2006), a control group is designed to contain the units that are similar in covariate values to those in the treatment group (differing only on the treatment indicators). A rule-of-thumb for evaluating the success of a matching procedure posits that better balance on covariates leads to smaller bias in the treatment effect estimation (Rosenbaum and Rubin, 1985); here, balance is understood as similarity between the empirical covariate distributions in the treatment and control groups. Note that theoretically, if an optimal matching does not exist, no guarantee as to the bias reduction amount can be given.

Among different types of matching recipes, the first proposed and well-used one is the nearest neighbor matching (Rubin, 1973). It prescribes to pair up each observed treatment unit with a control unit so as to minimize a weighted distance between the units' covariate vectors in each such pair. Mahalanobis distance is widely used for this purpose (Rubin, 1980), however, as a measure of divergence, it relies on elliptical distributions of covariates (Sekhon, 2008). Another widely-used recipe prescribes to match units on propensity score (Rosenbaum and Rubin, 1983) defined as the probability of a unit to receive treatment.

The Mahalanobis distance and propensity score based matching methods can be combined in various ways (Rubin, 2001; Diamond and Sekhon, 2013). However, such methods require assumptions on model and/or data structure. As such, true units' propensity score values are generally unknown, and must be estimated via regression on covariates, which makes room for the researcher's bias in data analysis (when one can "tinker with" with an analysis tool to make it output the result that one anticipates, perhaps subconsciously). This weakness has led to controversial exchanges between the authors analyzing the same data and reaching conflicting conclusions (Dehejia and Wahba, 1999, 2002; Smith and Todd, 2005b; Dehejia, 2005; Smith and Todd, 2005a).

Both the Mahalanobis distance and propensity score based matching methods are applied with the objective to minimize the differences between the units in the treatment group and the control group of the same size. In contrast, Iacus et al. (2012) introduce a new class of matching methods, the Monotonic Imbalance Bounding (MIB) matching, which looks to assemble matched control groups consisting of a sufficiently large number of observations with a fixed pre-set level of maximum allowed imbalance. Based on the imbalance level, an algorithm is designed to split the range of each covariate into several coarse categories, so that any exact matching algorithm can be applied to solve this discretized problem.

Methodologies for direct optimization of balance have been proposed by researchers just recently. Rosenbaum et al. (2007) introduce a *fine balance* method, where exact balance is sought on several categorized nominal covariates and approximate matching is conducted on the remaining ones. For the exact matching part, a matrix of Mahalanobis distance values across all pairs of treatment and control units is defined, and then the classic assignment algorithm is used to minimize the total distance. Nikolaev et al. (2013) introduce Balance Optimization Subset Selection (BOSS) approach, optimizing explicit measures of balance and treating several models with exact and heuristic methods. Zubizarreta (2012) builds mixed integer programming models to optimize covariate balance directly by minimizing the total sum of the distances between the treated units and matched control units. The latter two lines of research work to measure the difference between the covariate distributions in the treatment group and control pool by employing chi-square, correlations, quantiles and Kolmogorov-Smirnov statistics, which are fundamentally different from the information theory-driven approach developed in the present paper.

3. Problem Definition

This section begins by presenting several matching problems, using illustrative examples, and explains how mutual information can guide a matching process. Then, relying on the mutual information function, nonlinear integer optimization problems are formally stated.

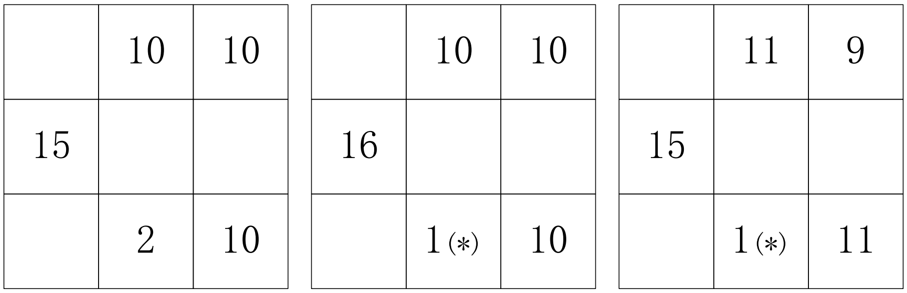
3.1 Motivating the Use of Matching for Causal Inference: Problem Statements

Given a set of observed units that have been treated, termed a treatment pool, and a set of observed untreated units, termed a control pool $\mathcal{C}$, the causal inference problem objective is to evaluate the degree of influence of the treatment on the population units, termed treatment effect. For an observable unit u, let Y_u^1 (Y_u^0) denote a treated (untreated) response and t_u a treatment indicator (1 means treated, 0 means not treated). Per Rubin's model of causal inference, these responses are referred to as *potential outcomes*, reflecting the fact that it is impossible to observe both Y_u^1 and Y_u^0 on the same unit u (Holland, 1986). For this reason, in estimating the population-wide effects of a treatment, researchers have to resort to comparing the averages across the treatment and control groups (Holland, 1986). One commonly targeted quantity of interest in causal inference studies, and the one this paper focuses on, is the average treatment effect for the treated (ATT), $E(Y^1|t=1)-E(Y^0|t=1)$, i.e., the average effect of treatment on the units that actually receive it.

Assume that a treatment group, $\mathcal{T} : |\mathcal{T}| < |\mathcal{C}|$, is given (randomly selected from a treatment pool), so $E(Y^1|t=1)$ can be estimated directly. A decision has to be made about selecting a control subset $\mathcal{S} \subset \mathcal{C}$ so that the units in $\mathcal{T}$ and $\mathcal{S}$ can be compared. If the two groups have the same distribution of covariates, one can use the value $E(Y^0|t=0)$ over $\mathcal{S}$ as an estimate of $E(Y^0|t=1)$ over the entire population (refer Rosenbaum and Rubin (1983) for more statistical fundamental work), and then, obtain an estimate of ATT.

The goal of a matching procedure is to ensure that the covariate distributions in the treatment and control groups are as similar as possible. The key insight this paper exploits is that, if a matching procedure is successful, then it should make it impossible to distinguish the treatment units from the control units based on the covariates, or, in other words, learn the treatment status of an observation based on the information captured by its covariate values. For example, randomization guarantees that the treated and control units are indistinguishable by making the covariate distributions in both groups be identical to that in the whole population; in other words, randomization tends to balance covariates on expectation. The information about the treatment captured in the covariates can be quantified as the MI between the covariates and the treatment variable, and more specifically, expressed using either the joint covariate distribution or the marginal covariate distributions. This paper considers both these formulations, separately.

Let K be the set of covariates. For an observed unit, the $|K|$-dimensional covariate vector is denoted by $\mathbf{X} = \{X_1, X_2, ..., X_{|K|}\}$. Assume that every covariate is or can be made categorical. The discretization of continuous covariates can be accomplished by applying a binning scheme (Iacus et al., 2012; Nikolaev et al., 2013) to divide the range of values for each such covariate into a fixed set of intervals. These categorical or interval bounds partition the covariate hyperspace into subspaces. Define a marginal bin as the largest subspace associated with an interval from a covariate's range, and a joint bin as a covariate subspace that is not further subdivided into any smaller subspaces. Then, by design, a joint bin is an intersection of $|K|$ marginal bins, and every observed unit is contained in one such bin. Let b denote a joint bin, B denote the set of all joint bins, m denote a marginal bin and M denote the set of all marginal bins. With the binning scheme, units with covariate values falling into the same joint bin can no longer be distinguished from each other.



	10	10
15		
	2	10

(a) Treatment group

	10	10
16		
	1 (*)	10

(b) Matching on joint distribution

	11	9
15		
	1 (*)	11

(c) Matching on marginal distribution

Figure 1: Different control groups selected when the perfect matching cannot be achieved due to the lack of control units in bin (*).

Note that the matching problem is trivial if there exists a control group that perfectly matches the treatment group (i.e., the empirical covariate distributions in the groups are identical). Consider the treatment group in Figure 1a, where the two-dimensional grid (built for two covariates) contains in its cells, termed bins, the number of units found in each bin. If a perfect matching of the control units to the treated ones does not exist, then the selection of a good control group becomes challenging. When a joint distribution is used to capture the dependence between the treatment variable and the covariates, the joint bins can be viewed as being independent and all equally important for representing the distribution. A good matching method should select some control units to form a group with a minimum loss in the joint distribution (Figure 1b). On the other hand, since the joint bins are formed as the intersections of $|K|$ multiple marginal bins, the assumption of the independence between the bins may not be well justified. Then, one can take an alternative approach and select the control group that achieves the best matching in all the marginal distributions (Figure 1c), albeit sacrificing some information captured in the copula. In summary, the problems of matching on the joint or marginal distributions each have their pros and cons, which is why the ensuing computational studies use and compare them both for treatment effect estimation (see Section 5).

3.2 Nonlinear Integer Optimization Problems

The objective of our matching problem is to select such a subset $\mathcal{S} \subseteq \mathcal{C}$ that minimizes the MI between the treatment indicator and covariate vector over set $\mathcal{S} \cup \mathcal{T}$. The MI between t and $\mathbf{X}$ (or all the X_k) is denoted by $I(t;\mathbf{X})$ if the computation is based on the full joint distribution of the covariates, and by $\sum_{k\in K} I(t;X_k)$ if the computation is based on the marginal distributions of individual covariates. Since these expressions have similar mathematical forms, only $I(t,X)$ will be used for notation in the following discussion, with X representing either $\mathbf{X}$ or X_k, depending on the context. Note that $I(t,X)$ is an unambiguous notation for MI in a problem with a single covariate. Meanwhile, for a problem with multiple covariates, the units in the joint or marginal bins can be thought of as being projected into a one-dimensional range, and hence, can also be treated as a single-covariate problem, albeit possibly with the additional constraints capturing the copula-based dependencies.

In order to express $I(t;X)$ using the empirical covariate distribution for the units in a given problem, denote the covariate value for any unit contained in bin b by the same variable X_b. Let $p(t)$ be the probability that a unit is treated, and $p(X_b)$ be the probability that its covariate value falls into bin b, with $\sum_{b\in B} p(X_b)=1$. Also, let $p(X_b,t)$ be the probability that the covariate value of a unit with treatment indicator t falls into bin b. Then, the empirical MI between the treatment indicator t and covariate X can be expressed as

$$I(t;X) = \sum_{b\in B}\sum_{t\in\{0,1\}} p(X_b,t)\log\frac{p(X_b,t)}{p(X_b)p(t)}. \tag{1}$$

Let S_b (or T_b, C_b) denote the number of units in group $\mathcal{S}$ (or $\mathcal{T}$, $\mathcal{C}$) with covariate values falling into bin b. From the characteristics of the units in $\mathcal{S}\cup\mathcal{T}$, the probabilities in equation (1) can be estimated. If $t=0$, $p(X_b,t)=\frac{S_b}{|\mathcal{S}|+|\mathcal{T}|}$ and $p(t)=\frac{|\mathcal{S}|}{|\mathcal{S}|+|\mathcal{T}|}$; if $t=1$, $p(X_b,t)=\frac{T_b}{|\mathcal{S}|+|\mathcal{T}|}$ and $p(t)=\frac{|\mathcal{T}|}{|\mathcal{S}|+|\mathcal{T}|}$; also, $p(X_b)=\frac{T_b+S_b}{|\mathcal{S}|+|\mathcal{T}|}$.

In general, an MI estimation bias (which is different from the causal estimation bias discussed above) arises when the MI estimation is done based on a fixed limited number of observations (1) (Panzeri and Treves, 1996; Roulston, 1999). However, this paper analyzes the *empirical* distributions of the variables defined for the units in the control and treatment groups, which are available in their entirety, and hence, by (1), the MI is exactly given,

$$\begin{aligned} I(t;X) = \log\left(|\mathcal{S}|+|\mathcal{T}|\right) + \frac{1}{|\mathcal{S}|+|\mathcal{T}|}\Big(&\sum_{b\in B} S_b[\log S_b - \log(T_b+S_b) - \log|\mathcal{S}|] \\ &+\sum_{b\in B} T_b[\log T_b - \log(T_b+S_b) - \log|\mathcal{T}|]\Big). \end{aligned} \tag{2}$$

Two alternative MI-based objective functions are analyzed in this paper: $I(t;\mathbf{X})$ and $\sum_{k\in K} I(t;X_k)$. Formally, a problem from the class of **Mutual Information based Matching (MIM)** problems is stated:
Given: $|K|$ covariates; treatment group $\mathcal{T}$; control pool $\mathcal{C}$ with $|\mathcal{C}|>|\mathcal{T}|$; for each observed unit $u\in\mathcal{T}\cup\mathcal{C}$, the covariate vectors $\mathbf{X}=\{X_1,X_2,...,X_{|K|}\}$; segmented covariate space with joint bins $b\in B$ and marginal bins $m\in M$; a fixed integer N as the target control group size.
Objective: find a subset $\mathcal{S}\subseteq\mathcal{C}$ such that

- $|\mathcal{S}|=N$ and $I(t;\mathbf{X})$ is minimized (MIM-Joint problem), or
- $|\mathcal{S}|=N$ and $\sum_{k\in K} I(t;X_k)$ is minimized (MIM-Marginal problem).

A matching problem based on either joint or marginal covariate distribution(s) is designed with the decision variables returning the number of control units to be selected from each joint bin. Complete enumeration of feasible solutions in a problem with any of these two objective types would take an exponentially growing number of computing operations in the size of the control pool. Another challenge lies in the nonlinearity of the objective functions, further analysis of which is required in order to arrive at tractable mathematical programming formulations for MIM.

Theorem 1 *The decision version of the MIM-Marginal problem,* $\min_{\mathcal{S}\subset\mathcal{C}} \sum_{k\in K} I(t;X_k)$ *subject to* $|\mathcal{S}| = N$*, is NP-complete.*

Proof See Appendix A. ■

4. Solution Approaches

This section investigates the properties of solutions with the minimum MI, with the goal of developing a method for treating the nonlinearity in the objective function of MIM problems. The derivations presented in this section unfold from the problem of minimizing $I(t;X)$ under the assumption that the contents of the bins capturing the distribution of covariate X are independent. The obtained insights are next extended to the MIM-Joint and MIM-Marginal problems. The mixed integer programming models and matching algorithms are then developed for selecting control subsets for MIM-Joint and MIM-Marginal problems.

4.1 Analyses of Optimality Conditions

Consider the expression of MI in (2); observe that since the treatment group is given, and the target control group size is known, $|\mathcal{S}| = N$, several terms in equation (2) are constant. Also, $\sum_{b\in B} S_b \log|\mathcal{S}| + \sum_{b\in B} T_b \log|\mathcal{T}| = |\mathcal{T}|\log|\mathcal{T}| + |\mathcal{S}|\log|\mathcal{S}|$. Then, the term $\sum_{b\in B} S_b[\log S_b - \log(T_b+S_b)] + \sum_{b\in B} T_b[\log T_b - \log(T_b+S_b)]$ remains the only one to be considered for MI minimization. For the ease of presentation, this term can now be rewritten based not on the bins' aggregate contents but on the individual units' locations in the bins. Because all the observed units, whose covariate values X^u are contained in the same bin, have the same values of T_b and S_b, the minimization of (2) is equivalent to that of

$$R \equiv \prod_{u\in\mathcal{S},X^u\in b} \frac{S_b}{T_b+S_b} \prod_{u\in\mathcal{T},X^u\in b} \frac{T_b}{T_b+S_b}. \tag{3}$$

Consider the MIM problem instance illustrated by Figure 2, where $N-1$ control units have been selected from the control pool into a control group (not necessarily optimally). In order to complete the selection of units into the control group, one last unit has to be selected from any of the bins with $S_b < C_b$. All such bins can be partitioned into three subsets: $B^1 = \{b : S_b < T_b\}$, $B^2 = \{b : S_b \geq T_b, T_b \neq 0\}$, $B^3 = \{b : T_b = 0\}$. Given that the last unit added to the control group is contained in bin b, let I_b denote the resulting MI between t and X, and R_b denote the resulting objective function value in (3).

The following two lemmas provide the guidelines for the optimal selection of the last unit to be included into the control group.

Lemma 2 *Consider an instance where an incomplete control group has* $N-1$ *units in it, and three candidate units (that could complete it) are contained in bins* $b_1 \in B^1$*,* $b_2 \in B^2$ *and* $b_3 \in B^3$*, respectively. Then,* $I_1 < I_2 < I_3$*.*

Proof If the candidate unit from bin $b_1 \in B^1$ is selected, then the value of S_{b_1} increases by 1, while all the other S_b and T_b values stay unchanged. Thus, the objective function value in (3) becomes $R_1 = \hat{R}(\frac{S_{b_1}+1}{T_{b_1}+S_{b_1}+1})^{S_{b_1}+1}(\frac{T_{b_1}}{T_{b_1}+S_{b_1}+1})^{T_{b_1}}(\frac{S_{b_2}}{T_{b_2}+S_{b_2}})^{S_{b_2}}(\frac{T_{b_2}}{T_{b_2}+S_{b_2}})^{T_{b_2}}$,

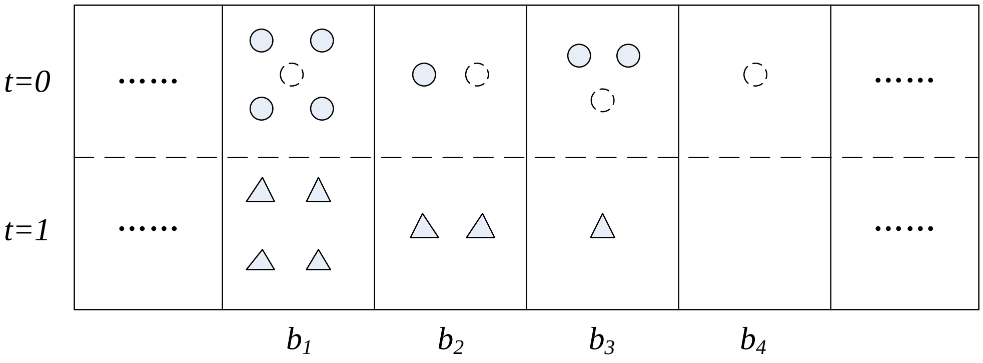


Figure 2: A selection process illustration. Treated units in $\mathcal{T}$, selected control units in $\mathcal{S}$ and unselected control pool units are represented by triangles, full circles and dashed circles, respectively.

where $\hat{R}$ represents the terms unrelated to b_1 or b_2. Similarly, if the candidate unit from bin $b_2 \in B^2$ is selected, then the objective function value in (3) becomes $R_2 = \hat{R}(\frac{S_{b_1}}{T_{b_1}+S_{b_1}})^{S_{b_1}}(\frac{T_{b_1}}{T_{b_1}+S_{b_1}})^{T_{b_1}}(\frac{S_{b_2}+1}{T_{b_2}+S_{b_2}+1})^{S_{b_2}+1}(\frac{T_{b_2}}{T_{b_2}+S_{b_2}+1})^{T_{b_2}}$. By the definitions of B^1 and B^2, observe that $T_{b_1} \geq S_{b_1} + 1 > S_{b_1}$ and $S_{b_2} + 1 > S_{b_2} \geq T_{b_2}$. Therefore, $\frac{R_1}{R_2} < 1$, and hence, $I_1 < I_2$.

If a unit from $b_2 \in B^2$ or $b_3 \in B^3$ is selected to complete the control group, then the corresponding objective function value in (3) is given by $R_2 = \hat{R}(\frac{S_{b_2}+1}{T_{b_2}+S_{b_2}+1})^{S_{b_2}+1}(\frac{T_{b_2}}{T_{b_2}+S_{b_2}+1})^{T_{b_2}}$ or $R_3 = \hat{R}(\frac{S_{b_2}}{T_{b_2}+S_{b_2}})^{S_{b_2}}(\frac{T_{b_2}}{T_{b_2}+S_{b_2}})^{T_{b_2}}$, where $\hat{R}$ represents the terms unrelated to b_2 or b_3, respectively. By the definition of B^2, observe that $S_{b_2} + 1 > S_{b_2} \geq T_{b_2}$, and hence, $0.5 < \frac{R_2}{R_3} < 1$. Consequently, one has $I_2 < I_3$. ■

Lemma 3 *Consider an instance where an incomplete control group has $N - 1$ units in it, and two candidate units (that could complete it) are contained in bins $b_1, b_2 \in B^1$ (or $b_1, b_2 \in B^2$), respectively. Then, $I_1 < I_2$ if and only if $\frac{S_{b_1}-A}{T_{b_1}} < \frac{S_{b_2}-A}{T_{b_2}}$, where $A \approx -0.47$.*

Proof First, consider the case where $b_1 \in B^1$ and $b_2 \in B^1$. Similarly to the proof of Lemma 2, one obtains $R_1 = \hat{R}(\frac{S_{b_1}+1}{T_{b_1}+S_{b_1}+1})^{S_{b_1}+1}(\frac{T_{b_1}}{T_{b_1}+S_{b_1}+1})^{T_{b_1}}(\frac{S_{b_2}}{T_{b_2}+S_{b_2}})^{S_{b_2}}(\frac{T_{b_2}}{T_{b_2}+S_{b_2}})^{T_{b_2}}$, and $R_2 = \hat{R}(\frac{S_{b_1}}{T_{b_1}+S_{b_1}})^{S_{b_1}}(\frac{T_{b_1}}{T_{b_1}+S_{b_1}})^{T_{b_1}}(\frac{S_{b_2}+1}{T_{b_2}+S_{b_2}+1})^{S_{b_2}+1}(\frac{T_{b_2}}{T_{b_2}+S_{b_2}+1})^{T_{b_2}}$, and hence, $\frac{R_1}{R_2} = (1 + \frac{1}{S_{b_1}})^{S_{b_1}} \frac{1}{(1+\frac{1}{T_{b_1}+S_{b_1}})^{T_{b_1}+S_{b_1}}} \frac{S_{b_1}+1}{T_{b_1}+S_{b_1}+1} \frac{1}{\frac{S_{b_2}+1}{T_{b_2}+S_{b_2}+1}} \frac{1}{(1+\frac{1}{S_{b_2}})^{S_{b_2}}} (1 + \frac{1}{T_{b_2}+S_{b_2}})^{T_{b_2}+S_{b_2}}$ $= \{\frac{(1+\frac{1}{S_{b_1}})^{S_{b_1}}}{(1+\frac{1}{T_{b_1}+S_{b_1}})^{T_{b_1}+S_{b_1}}(1+\frac{T_{b_1}}{S_{b_1}+1})}\}/\{\frac{(1+\frac{1}{S_{b_2}})^{S_{b_2}}}{(1+\frac{1}{T_{b_2}+S_{b_2}})^{T_{b_2}+S_{b_2}}(1+\frac{T_{b_2}}{S_{b_2}+1})}\}$. Observe that both the numerator and denominator in this expression have the same form. Therefore, the properties of the ratio $\frac{R_1}{R_2}$ can be analyzed by studying the properties of function $f(x,y) = \frac{(1+\frac{1}{x})^x}{(1+\frac{1}{y+x})^{(y+x)}(1+\frac{y}{x+1})}$; more specifically, if one can show that $f(S_{b_1}, T_{b_1}) < f(S_{b_2}, T_{b_2})$, then one has $\frac{R_1}{R_2} < 1$ and $I_1 < I_2$, and vice versa. Per the properties of $f(x,y)$ (see Appendix B), it is concluded that $f(S_{b_1}, T_{b_1}) < f(S_{b_2}, T_{b_2})$ if and only if $\frac{S_{b_1}-A}{T_{b_1}} < \frac{S_{b_2}-A}{T_{b_2}}$, where $A \approx -0.47$.

The proof for the case with $b_1 \in B^2$ and $b_2 \in B^2$ is constructed in the same manner. ■

Given the arbitrary bins $b_1 \in B^1$, $b_2 \in B^2$ and $b_3 \in B^3$, $I_1 < I_2 < I_3$, by Lemma 2 and by the definition of the subsets B^1, B^2 and B^3, one has $\frac{S_{b_1}-A}{T_{b_1}} < 1$, $\frac{S_{b_2}-A}{T_{b_2}} > 1$, and $\frac{S_{b_3}-A}{T_{b_3}} = +\infty$. Therefore, Lemma 2 can be viewed as a special case of Lemma 3. Having considered the problem of optimally adding a single unit to the existing (incomplete) control group, the obtained results are now generalized to the problem of selecting a whole control group of a given size.

Theorem 4 *(Necessary and Sufficient Condition for Optimality) Consider an instance of minimizing $I(t;X)$. A control group $\mathcal{S}$ of size N is optimal if and only if for any pair of bins b_1 and b_2 with $|C_{b_2} - S_{b_2}| \geq 1$ it holds that*

$$\frac{S_{b_1} - 1 - A}{T_{b_1}} \leq \frac{S_{b_2} - A}{T_{b_2}}, \tag{4}$$

where $A \approx -0.47$.

Proof The proof will proceed by contradiction. For convenience, in the following narrative, any group violating the theorem's condition is termed an *improvable group.* For an improvable group, one can identify at least one pair of bins, b_1 and b_2, such that $|C_{b_2} - S_{b_2}| \geq 1$ and $\frac{S_{b_1}-1-A}{T_{b_1}} > \frac{S_{b_2}-A}{T_{b_2}}$. Any such pair is termed an *improvable pair.*

First, consider the necessary condition for optimality: if a group is optimal, then it is not an improvable group. Suppose that $\mathcal{S}$ is an optimal group with the minimum $I(t;X)$, and $\mathcal{S}$ is also an improvable group. Without loss of generality, assume that b_1 and b_2 are an improvable pair and $\frac{S_{b_1}-1-A}{T_{b_1}} > \frac{S_{b_2}-A}{T_{b_2}}$. Consider an incomplete control group of size $N-1$, obtained by removing a control unit from bin b_1 in $\mathcal{S}$. Because $\frac{(S_{b_1}-1)-A}{T_{b_1}} > \frac{S_{b_2}-A}{T_{b_2}}$ and according to Lemma 3, one can obtain a group with a smaller value of $I(t;X)$ by adding the last unit into bin b_2. Thus, $\mathcal{S}$ is not optimal, and one arrives at a contradiction.

Second, consider the sufficient condition for optimality: if a group is not an improvable group, then it is optimal. Suppose that $\mathcal{S}$ is not an improvable group, $\mathcal{S}^*$ is an optimal group with the minimum $I(t;X)$, and $\mathcal{S}^* \neq \mathcal{S}$. Then, there must exist some bin(s) where $\mathcal{S}^*$ has fewer units than $\mathcal{S}$, and some other bin(s) where $\mathcal{S}^*$ has more units than $\mathcal{S}$. Without loss of generality, assume that b_1 and b_2 are two such bins, respectively. Since $\mathcal{S}^*$ has more units in bin b_2 than $\mathcal{S}$, this implies $|C_{b_2} - S_{b_2}| \geq 1$. Because $\mathcal{S}$ is not an improvable group, one has $\frac{S_{b_1}-1-A}{T_{b_1}} \leq \frac{S_{b_2}-A}{T_{b_2}}$.

If $\frac{S_{b_1}-1-A}{T_{b_1}} < \frac{S_{b_2}-A}{T_{b_2}}$, then according to Lemma 3, if a unit is removed from b_1 and another unit is added into b_2 for $\mathcal{S}$, then one can obtain a group with a greater value of

$I(t;X)$, with $\frac{S_{b_1}-1-\Delta-A}{T_{b_1}} < \frac{S_{b_1}-1-A}{T_{b_1}} < \frac{S_{b_2}-A}{T_{b_2}} < \frac{S_{b_2}+\Delta-A}{T_{b_2}}$ holding for $\forall\Delta > 0$. Note again that b_1 and b_2 were arbitrarily picked. Such unit shuffling (i.e., removal and addition) operations can repeat until $\mathcal{S}$ is modified to become identical to $\mathcal{S}^*$. Since in this process, $I(t;X)$ increases with every shuffle, then $\mathcal{S}^*$ could not be optimal, which is a contradiction.

If $\frac{S_{b_1}-1-A}{T_{b_1}} = \frac{S_{b_2}-A}{T_{b_2}}$, then as a result of removing a unit from b_1 and adding one into b_2, $I(t;X)$ will not change. In such a case, if the updated $\mathcal{S}$ becomes identical to $\mathcal{S}^*$, then this means that $\mathcal{S}$ is an alternative optimal solution with the minimum $I(t;X)$. Otherwise, one can continue shuffling units, with $\frac{S_{b_1}-1-\Delta-A}{T_{b_1}} < \frac{S_{b_1}-1-A}{T_{b_1}} = \frac{S_{b_2}-A}{T_{b_2}} < \frac{S_{b_2}+\Delta-A}{T_{b_2}}$ holding for $\forall\Delta > 0$. Similarly, $I(t;X)$ will continue increasing, leading to $\mathcal{S}^*$ not being optimal, i.e., to a contradiction. ■

Theorem 4 provides the necessary and sufficient optimality conditions for the control groups with the minimum $I(t;X)$. Its value lies in condition (4) being linear in S_b, unlike the minimization problem objective (2). Note, however, that Theorem 4 only works to determine whether a control group is optimal or not; it cannot be used to assess or compare the quality of suboptimal control groups.

In order to effectively apply Theorem 4 in practice, one would like to avoid the exhaustive traversal of bin pairs. Corollaries 5 and 6 allow for tackling this problem and provide a means for efficient optimal control group selection.

Corollary 5 *Consider an instance of minimizing $I(t;X)$ where $N > \sum_{b\in\{b:T_b\geq 1\}} C_b$. Then, a control group $\mathcal{S}$ is optimal if it includes all the control units in all $b \in \{b : T_b \geq 1\}$.*

Proof Follows directly from Lemma 2. ■

Corollary 6 *Consider an instance of minimizing $I(t;X)$, where $N \leq \sum_{b\in\{b:T_b\geq 1\}} C_b$. Then, a control group $\mathcal{S}$ is optimal if and only if for every pair of bins b_1 and b_2 such that*

$$b_1 \in \underset{b\in B}{\operatorname{argmax}}\{\frac{S_b - 1 - A}{T_b}\} \tag{5}$$

and

$$b_2 \in \underset{b\in B}{\operatorname{argmin}}\{\frac{S_b - A}{T_b} : |C_b - S_b| \geq 1\}, \tag{6}$$

one has $\frac{S_{b_1}-1-A}{T_{b_1}} \leq \frac{S_{b_2}-A}{T_{b_2}}$, where $A \approx -0.47$.

Proof Consider any pair of bins, b_3 and b_4, with $|C_{b_4} - S_{b_4}| \geq 1$. In order to determine if $\mathcal{S}$ is optimal, Theorem 4 prescribes to compare the left-hand and right-hand sides of inequality (4) for bins b_3 and b_4. By the statement of this corollary, one has $\frac{S_{b_1}-1-A}{T_{b_1}} \geq \frac{S_{b_3}-1-A}{T_{b_3}}$ and $\frac{S_{b_2}-A}{T_{b_2}} \leq \frac{S_{b_4}-A}{T_{b_4}}$. Then, if inequality (4) holds for bins b_1 and b_2, then it also holds for bins B_3 and B_4, because $\frac{S_{b_3}-1-A}{T_{b_3}} \leq \frac{S_{b_1}-1-A}{T_{b_1}} \leq \frac{S_{b_2}-A}{T_{b_2}} \leq \frac{S_{b_4}-A}{T_{b_4}}$, and vice versa. ■

4.2 Mixed Integer Programming-Based Matching Algorithms

The optimality conditions in Theorem 4 and Corollary 6 allow one to construct an alternative formulation for the problem of minimizing $I(t;X)$ with $N \leq \sum_{b \in \{b:T_b \geq 1\}} C_b$, using the expression $\frac{S_b-1-A}{T_b}$. Note that this ratio is undefined for bins with $T_b = 0$; however, per Lemma 2, an optimal solution can contain control units from such bins only if all the available control units from other bins have been exhausted. In order to reformulate the objective function of minimizing $I(t;X)$, the expression $\frac{S_b-1-A}{T_b}$ should first be revised so that its denominator evaluates to a fixed number, $\alpha \in (0,1)$, small enough to make the selection of control units from bins with $T_b = 0$ very costly. In order to search for a control group satisfying the condition in Corollary 6, the following optimization problem is formulated:

$$\min_{\mathcal{S} \subset \mathcal{C}} \{ \max_{b \in B} \frac{S_b - 1 - A}{\max\{T_b, \alpha\}} \}, \tag{7}$$

where α is a positive parameter small enough to distinguish $T_b = 0$ from other positive values of T_b, e.g., $\alpha = 0.01$.

By solving (7), one can work to construct an optimal control group through minimizing the maximum value of the function in (5). Having found an optimal solution to (7), one can check if (for this solution) the set in (5) is a singleton. If it is, then the condition in Corollary 6 holds. Otherwise, satisfying (7) may not be sufficient for satisfying Corollary 6, since it requires one to check every pair of bins in both the set in (5) and the set in (6). As an example of this situation, suppose that there exist two bins, b_1 and b_3, in the set in (5), and a bin b_2 in the set in (6) such that $\frac{S_{b_1}-1-A}{T_{b_1}} = \frac{S_{b_3}-1-A}{T_{b_3}} > \frac{S_{b_2}-A}{T_{b_2}}$. If a unit is removed from b_1 while another unit is added into b_2, the objective value of (7) does not improve because $\frac{S_{b_3}-1-A}{T_{b_3}}$ does not decrease. Thus, the optimization process based purely on solving (7) would terminate early without guaranteeing an optimal matching.

To handle the situation where the set in (5) is not a singleton for a solution of (7), an algorithm is developed to iteratively solve for the optimal number of units to be selected from each bin. In any iteration, if solving (7) returns multiple bins with values of $\frac{S_b-1-A}{T_b}$ equal to the maximum (over all the bins), then one of these bins is added to a "forbidden bin set", denoted by B^F and initialized at an empty set before the first iteration. Every time B^F is updated, problem (7) is reformulated, with all the bins that are not in B^F, and solved again in the next iteration. After several such iterations, once the set in (5) is found to be a singleton for a solution to (7), one can be sure that an optimal control group has been found. In order to ensure that the unit picking in a given iteration does not mess up the optimality achieved within any bin in the previous iteration(s), a bin with the smallest number of the treatment units in the non-singleton set (5) is always fixed first. In every iteration, (7) is solved as a mixed integer programming (MIP) model,

$$\min \quad q \tag{8}$$

$$s.t. \;\; q \geq \frac{S_b - 1 - A}{\max\{T_b, \alpha\}} \quad \forall b \notin B^F, \tag{9}$$

$$\sum_{b \in B} S_b = N, \tag{10}$$

$$S_b \le C_b \quad \forall b \in B, \tag{11}$$

$$S_b \ge 0 \quad \forall b \in B, \tag{12}$$

$$S_b : integer \quad \forall b \in B, \tag{13}$$

$$B = \{b : C_b + T_b \ge 1\}. \tag{14}$$

The decision variables in this MIP are the numbers of the control units, S_b, to be selected from each bin. Since it is only necessary to consider the bins in $\{b : C_b + T_b \ge 1\}$, then despite the fact that the total number of joint bins grows exponentially with the binning partition granularity, the number of the decision variables is bounded by $|\mathcal{T}| + |\mathcal{C}|$. The contents of the forbidden bin set B^F are updated iteratively in the described algorithm. The minimax optimization problem (7) is formulated with the objective function (8) and the constraint set (9). Constraint (10) ensures that the total number of units in the control group is equal to N. Constraints (11), (12) and (13) restrict the range of S_b to nonnegative integers not exceeding the number of available control units in each respective bin.

Note that for solving any MIM-Joint problem, since the bins' contents can be treated as being independent from each other, the described procedure for minimizing $I(t; X)$ can be exactly followed to minimize $I(t; \mathbf{X})$, with bins b in (8)-(14) being the joint bins.

Algorithm 1 MIP-based matching for MIM-Joint problem

```
1: Initialize the bin set $\{b : C_b + T_b \ge 1\}$ consisting of all the bins occupied by the units
   in $\mathcal{T} \cup \mathcal{C}$; compute $T_b$ and $C_b$; forbidden bin set $B^F = \emptyset$.
2: Update and solve the corresponding instantiation of formulation (8)-(14) to obtain $S_b$
   for every bin $b$.
3: If $\mathrm{argmax}_{b \notin B^F}\{\frac{S_b - 1 - A}{T_b}\}$ is a singleton, go to step 4. Otherwise, add the bin with the
   smallest number of treatment units in $\mathrm{argmax}_{b \notin B^F}\{\frac{S_b - 1 - A}{T_b}\}$ into set $B^F$, record and
   fix the optimal number of control units to be selected in it, and go to step 2.
4: Construct a control group complying with the obtained values of $S_b$ over all the initialized bins $b$. Stop.
```

However, for solving an MIM-Marginal problem, modifications to the above formulation and the algorithm are necessary due to the fact that the marginal bins cannot be assumed independent. The decision variables of an MIP-Marginal model are the numbers of control units to be selected into $\mathcal{S}$ for every joint bin (b still denotes a joint bin), and constraints (10)-(14) remain a part of the optimization problem. Let m denote a marginal bin, M^F denote the forbidden marginal bin set, and T_m, C_m and S_m denote the number of all the treatment units, number of all the control units and number of the selected control units in m, respectively. Equation (9) is then replaced by (15). Also, an additional constraint (16) is added to the formulation to ensure that the number of units in any marginal bin equals the summed total number of units in all the corresponding joint bins.

$$q \ge \frac{S_m - 1 - A}{\max\{T_m, \alpha\}} \quad \forall m \notin M^F, \tag{15}$$

$$S_m = \sum_{b; X_m \in b} S_b \quad \forall m \in \{m : C_m + T_m \ge 1\} \tag{16}$$

Recall that in every iteration of solving MIM-Joint using Algorithm 1, one checks whether the set of bins with the maximum value of $\frac{S_m-1-A}{T_m}$ is a singleton. Because of the dependence between the contents of marginal bins, this condition by itself does not guarantees optimality for MIM-Marginal problem. Specifically, given a feasible solution to an MIM-Marginal problem, if exactly one marginal bin is found to achieve the maximum value of $\frac{S_m-1-A}{T_m}$ and this marginal bin is associated with covariate k, then because the bins in the same covariate are independent and according to Corollary 6, $I(t; X_k)$ is minimized. But the MI in other covariates might still be improved without changing $I(t; X_k)$, e.g., by adding and removing the same number of control units to/from the same marginal bin in covariate k. Thus, while solving for the optimal number of units in each marginal bin, and adding the marginal bins one-by-one to a forbidden bin set, one should not stop until the sets of bins with the maximum values of $\frac{S_m-1-A}{T_m}$ in all the $|K|$ covariates become singletons.

Algorithm 2 MIP-based matching for MIM-Marginal problem

```
1: Initialize the joint bin set $\{b : C_b + T_b \geq 1\}$ and the marginal bin set $\{m : C_m + T_m \geq 1\}$ consisting of all the bins occupied by the units in $\mathcal{T} \cup \mathcal{C}$; compute $T_b$, $C_b$, $T_m$ and $C_m$; forbidden bin set $M^F = \emptyset$.
2: Update and solve the corresponding instantiation of formulation (8), (10)-(16) to obtain $S_b$ for every joint bin $b$.
3: If $\text{argmax}_{m \notin M^F}\{\frac{S_m-1-A}{T_m}\}$ is a singleton for every covariate, go to step 4. Otherwise, add the marginal bin with the smallest number of treatment units in $\text{argmax}_{m \notin M^F}\{\frac{S_m-1-A}{T_m}\}$ into set $M^F$, record and fix the optimal number of control units to be selected in it, and go to step 2.
4: Construct a control group complying with the obtained values of $S_b$ over all the initialized bins $b$. Stop.
```

4.3 Sequential Selection Matching Algorithms

Based on the optimality conditions in Theorem 4, Algorithms 1 and 2 are guaranteed to achieve best matched control groups for MIM-fixed and MIM-marginal problem instances. However, their MIPs may become difficult to solve for problems of large size, which, however, can be avoided by utilizing the result captured in Theorem 7, presented for the problem of minimizing $I(t; X)$.

Theorem 7 *If control group $\mathcal{S}$ has the minimum $I(t; X)$ among all the control groups of size N, then a group with the minimum $I(t; X)$ among all the control groups of size $N+1$ $(N-1)$ can be obtained from $\mathcal{S}$ by adding to it a single unit from bin $b \in \text{argmin}_{b \in B}\{\frac{S_b-A}{T_b} : |C_b - S_b| \geq 1\}$ (removing from it a single unit from bin $b \in \text{argmax}_{b \in B}\{\frac{S_b-1-A}{T_b}\}$).*

Proof Let $\mathcal{S}^+$ denote a control group obtained from $\mathcal{S}$ by adding to it a single unit from bin $b \in \text{argmin}_{b \in B}\{\frac{S_b-A}{T_b} : |C_b - S_b| \geq 1\}$. According to Lemma 3, the MI between T and X over set $\mathcal{T} \cup \mathcal{S}^+$ is minimal among all the groups that can be built on $\mathcal{S}$. The following proof will show $\mathcal{S}^+$ is also globally optimal.

Let S_b^+ denote the number of control units selected into $\mathcal{S}^+$ in bin b. Let b_1 and b_2 be two bins such that $b_1 \in \text{argmax}_{b \in B}\{\frac{S_b^+ -1-A}{T_b}\}$ and $|C_{b_2} - S_{b_2}^+| \geq 1$. If b_1 is the bin where $\mathcal{S}^+$ has one more unit than $\mathcal{S}$, then $S_{b_1}^+ - 1 = S_{b_1}$ and $S_{b_2}^+ = S_{b_2}$, and then $\frac{S_{b_1}^+ -1-A}{T_{b_1}} = \frac{S_{b_1}-A}{T_{b_1}} \leq \frac{S_{b_2}-A}{T_{b_2}} = \frac{S_{b_2}^+ -A}{T_{b_2}}$. Note that by Theorem 4, because $\mathcal{S}$ is optimal, one has $\frac{S_{b_1}-1-A}{T_{b_1}} \leq \frac{S_{b_2}-A}{T_{b_2}}$. If b_2 is the bin where $\mathcal{S}^+$ has one more unit than $\mathcal{S}$, then $S_{b_2}^+ - 1 = S_{b_2}$ and $S_{b_1}^+ = S_{b_1}$, and then $\frac{S_{b_1}^+ -1-A}{T_{b_1}} = \frac{S_{b_1}-1-A}{T_{b_1}} \leq \frac{S_{b_2}-A}{T_{b_2}} < \frac{S_{b_2}^+ -A}{T_{b_2}}$. If $\mathcal{S}$ and $\mathcal{S}^+$ have the same numbers of units in both b_1 and b_2, then $\frac{S_{b_1}^+ -1-A}{T_{b_1}} = \frac{S_{b_1}-1-A}{T_{b_1}} \leq \frac{S_{b_2}-A}{T_{b_2}} = \frac{S_{b_2}^+ -A}{T_{b_2}}$. Therefore, by Corollary 6, $\mathcal{S}^+$ is a group with the minimum $I(t;X)$ among all the control groups of size $N+1$.

Let $\mathcal{S}^-$ denote a control group obtained from $\mathcal{S}$ by removing a single unit from bin $b \in \text{argmax}_{b \in B}\{\frac{S_b-1-A}{T_b}\}$. Let S_b^- denote the number of control units selected into $\mathcal{S}^-$ in bin b. Let b_1 and b_2 be two bins such that $b_1 \in \text{argmax}_{b \in B}\{\frac{S_b^- -1-A}{T_b}\}$ and $|C_{b_2} - S_{b_2}^-| \geq 1$. If b_2 is the bin where $\mathcal{S}$ has one more unit than $\mathcal{S}^-$, then $S_{b_2} - 1 = S_{b_2}^-$ and $S_{b_1}^+ = S_{b_1}$, and then $\frac{S_{b_1}^- -1-A}{T_{b_1}} = \frac{S_{b_1}-1-A}{T_{b_1}} \leq \frac{S_{b_2}-1-A}{T_{b_2}} = \frac{S_{b_2}^- -A}{T_{b_2}}$. Note that due to the optimality of $\mathcal{S}$, $\frac{S_{b_1}-1-A}{T_{b_1}} \leq \frac{S_{b_2}-A}{T_{b_2}}$. If b_1 is the bin where $\mathcal{S}$ has one more unit than $\mathcal{S}^-$, then $S_{b_1} - 1 = S_{b_1}^-$ and $S_{b_2}^+ = S_{b_2}$, and then $\frac{S_{b_1}^- -1-A}{T_{b_1}} < \frac{S_{b_1}-1-A}{T_{b_1}} \leq \frac{S_{b_2}-A}{T_{b_2}} = \frac{S_{b_2}^- -A}{T_{b_2}}$. If $\mathcal{S}$ and $\mathcal{S}^+$ have the same numbers of units in both b_1 and b_2, then $\frac{S_{b_1}^- -1-A}{T_{b_1}} = \frac{S_{b_1}-1-A}{T_{b_1}} \leq \frac{S_{b_2}-A}{T_{b_2}} = \frac{S_{b_2}^- -A}{T_{b_2}}$. Therefore, by Corollary 6, $\mathcal{S}^-$ is a group with the minimum $I(t;X)$ among all the control groups of size $N-1$. ■

Theorem 7 provides a method for finding optimal control groups for MIM problems, without solving any programming models. One can iteratively build control groups of increasing sizes until an optimal solution of the desired size N is obtained. Each control group in this process results in the minimum value of MI among all the groups of the same size. Also, Theorem 7 provides establishes a relationship between the optimal subsets for problems with different target control group sizes. This result will be important for seeking the minimum MI in the problems with an unrestricted (flexible) control group size.

For the MIM-Joint problem, since its bins are treated as independent from each other, Theorem 7 directly applies, and Algorithm 3 guarantees to return an optimal solution. Note that Algorithm 3 is polynomial. In the worst case, it needs to make comparisons of $\frac{S_b-A}{T_b}$ for N multiples of the number of bins occupied by treatment units.

With the MIM-Marginal problem, a challenge arises due to the dependence between the marginal bins' contents. By comparing the terms $\frac{S_m-A}{T_m}$, one can identify the marginal bin to which a control unit should be added, but one still needs to pick some joint bin. Even further, since the marginal bins on the same covariate are independent from each other, the comparison of $\frac{S_m-A}{T_m}$ can reveal the most favorable marginal bin for each covariate, but the bin that lies at the intersection of those $|K|$ marginal bins may not contain any control unit that could be added to the control group. Algorithm 4 offers an organized way

Algorithm 3 Sequential selection matching for MIM-Joint problem

```
1: Initialize the joint bin set $\{b : C_b + T_b \geq 1\}$ consisting of all the bins occupied by the units in $\mathcal{T} \cup \mathcal{C}$; compute $T_b$ and $C_b$; set $S_b = 0$ for all $b$.
2: Select a bin $b \in \operatorname{argmin}_{b \in B}\{\frac{S_b - A}{T_b} : |C_b - S_b| \geq 1\}$, update $S_b$ by adding 1.
3: If $N$ units are selected, go to step 4. Otherwise, go to step 2.
4: Construct a control group complying with the obtained values of $S_b$ over all the initialized bins $b$. Stop.
```

to achieve good (but not necessarily optimal) solutions to MIM-Marginal instances in the following manner. For each joint unit, $|K|$ ratios of the form $\frac{S_m - A}{T_m}$ are evaluated (one per covariate); these $|K|$ ratios are organized in a descending order; then, the joint bin with the lexicographically minimal ratio gets one more unit added to it. Again, the resulting Algorithm 4 is an approximate method, but it is polynomial, and in practice, is found to return solutions of high quality for diverse matching problem instances (see Section 5).

Algorithm 4 Sequential selection matching for MIM-Marginal problem

```
1: Initialize the joint bin set $\{b : C_b + T_b \geq 1\}$ consisting of all the bins occupied by the units in $\mathcal{T} \cup \mathcal{C}$; compute $T_b$ and $C_b$; set $S_b = 0$ for all $b$.
2: Update $\frac{S_m - A}{T_m}$ for each marginal bin, and order all associated $\frac{S_m - A}{T_m}$ by values in descend sequence for each joint bin.
3: Find a bin $b$ in set $\{b : |C_b - S_b| \geq 1\}$ such that its ordered set of $\frac{S_m - A}{T_m}$ is lexicographically minimal; increase the value of the decision variable, $S_b$, corresponding to this bin, by 1.
4: If $N$ units are selected, go to step 5. Otherwise, go to step 2.
5: Construct a control group complying with the obtained values of $S_b$ over all the initialized bins $b$. Stop.
```

The complexity of Algorithm 4 depends on the number of covariates $|K|$, the number of marginal bins $|M|$, treatment group size $|\mathcal{T}|$, control pool size $|\mathcal{C}|$ and target control group size N. In the worst case, every unit (treated or control) occupies one unique joint bin, with each joint bin contributing to $|K|$ marginal bins. The storage of these data requires a space of size $\mathcal{O}(|K|(|\mathcal{T}| + |\mathcal{C}|))$. In the binning step, both the treatment group and control pool are traversed, with each unit being assigned to the appropriate marginal and joint bins: this operation takes $\mathcal{O}(|M|(|\mathcal{T}| + |\mathcal{C}|))$ time. In the matching step, all the occupied joint bins are traversed for the lexicographic comparison, which takes $\mathcal{O}(N(|\mathcal{T}| + |\mathcal{C}|)|K|^2)$ time.

5. Computational Analyses

This section presents the results of the computational experiments with synthetic and real-world (LaLonde, 1986) data sets, evaluating the performance of the MIM method in estimating causal effects, and comparing it to the BOSS method (Nikolaev et al., 2013) and the widely used propensity score based matching method (Rosenbaum and Rubin, 1983).

5.1 Algorithm Performance Assessment

In order to evaluate the performance of the MIM algorithms, a series of tests is first conducted with the data set designed by Sauppe et al. (2014), which was found challenging for the existing matching methods[1]. This synthetic data set with 25 covariates contains 100 treatment units and 10,000 control units. All the covariate values are drawn from normal distributions with mean 0. All the treatment and control units have the same, highly nonlinear response function. Thus, by design, the average treatment effect for the treated (ATT) for the created population is zero.

The experiments were conducted with the number of the considered covariates varied in the range from 1 to 25. In optimizing the covariate balance, Sauppe et al. (2014) uniformly partitioned the range of the observed unit values in each covariate into 20 bins, and used Balance Optimization Subset Selection (BOSS) for control group selection. In order to treat the instances resulting in large MIP formulations, Sauppe et al. (2014) adopted a time limit heuristic. They achieved quite well balanced control groups; however, the limited computational efficiency remains the key challenge for the existing BOSS methods, especially when the data sets to make inference from are very large.

With the same settings as in Sauppe et al. (2014), this section compares the performance of the following matching methods: the Mahalanobis distance-based one, the propensity score-based one, the BOSS methods from Sauppe et al. (2014), and three MIM methods – MIM-Joint, MIM-Marginal MIP, and MIM-Marginal sequential selection. Note that the first two of these methods are widely used and included in several existing matching packages, e.g., MatchIt (Ho et al., 2011) and optmatch (Hansen and Klopfer, 2012). We also used Coarsened Exact Matching (CEM) in MatchIt (Ho et al., 2011) and fullmatch method in optmatch (Hansen and Klopfer, 2012). However, CEM excluded many treatment units from the matched treatment group in the experiments with five or more covariates, and thus, was not found suitable for ATT estimation. Also, under the pre-set control group size, the results of fullmatch were no different from those of the standard Mahalanobis distance or propensity score matching (depending on the selected parameter settings).

The MIP models for MIM were solved using CPLEX. To reduce the runtime in solving some large-size problems, a heuristic was applied that utilized the solution of sequential selection to covert MIP to an integer program. Generally, MIP-based methods (be it for BOSS or MIM) are time-consuming. However, the sequential selection algorithm always runs very quickly: it took about 7 seconds on average to find solutions for the instances with 25 covariates on a desktop with an Intel Xeon E5-2420 1.9GHz CPU and 16G RAM.

Table 1 presents the ATT estimates obtained with the considered methods for the instances with the varied number of covariates; Figure 3 provides a graphical illustration of the results. Recall that by design, the ATT is zero, so the closer an estimate is to zero, the better. Table 2, Figure 4a and Figure 4b report the Kolmogorov-Smirnov (KS) test statistic scores and associated p-values for checking whether the underlying covariate distributions differ in the treated and control groups. In Table 2, column "Avg" reports the average

[1] The data set, named $25c10k$, features a highly nonlinear response function and is available in full in the online supplement of Sauppe et al. (2014). The response function in it is $y = 0.8x_1(1.0-x_1)+0.5x_2(0.7+x_1)+0.27x_3x_2-0.9x_4^2+0.7x_5(0.5+x_5)x_2-0.6x_6x_1+0.4x_7-0.8x_8+0.6x_9(0.9-x_9)+0.2x_1^20(0.3-x_7)+0.5x_{11}^2-1.4x_{12}-0.8x_{13}-0.9x_{14}^2+0.5x_{15}^2(0.1+x_{15})+0.8x_{16}-0.9x_{17}(0.2-x_{13})+1.5x_{18}-1.2x_{19}(1.0+x_{11})+0.7x_{20}^2(0.8-x_{20})-0.5x_{21}-1.3x_{22}(1.0+x_{22})+1.1x_{23}-1.2x_{24}(1.0+x_{23})+0.4x_{25}^2(0.6-x_{25})+N(0,1)$.

Table 1: Estimated treatment effects with different matching methods.

# Covariates	Mahalanobis metric	Propensity score	BOSS	MIM-Joint	MIM-Marginal (MIP)	MIM-Marginal (Sequential)
1	-0.133	-0.117	0.218	0.006	0.006	0.006
5	0.626	-1.223	0.045	6.361	0.005	1.721
10	0.39	5.437	-2.646	16.646	-1.123	0.517
15	6.261	19.126	3.074	30.593	2.590	2.403
20	10.164	-11.849	7.782	35.306	5.927	8.084
25	16.074	-14.753	6.618	50.643	12.170	8.448

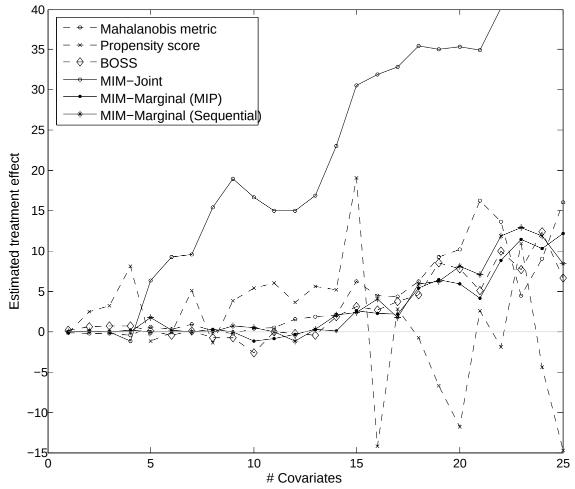


Figure 3: Trends for estimated treatment effects with different matching methods.

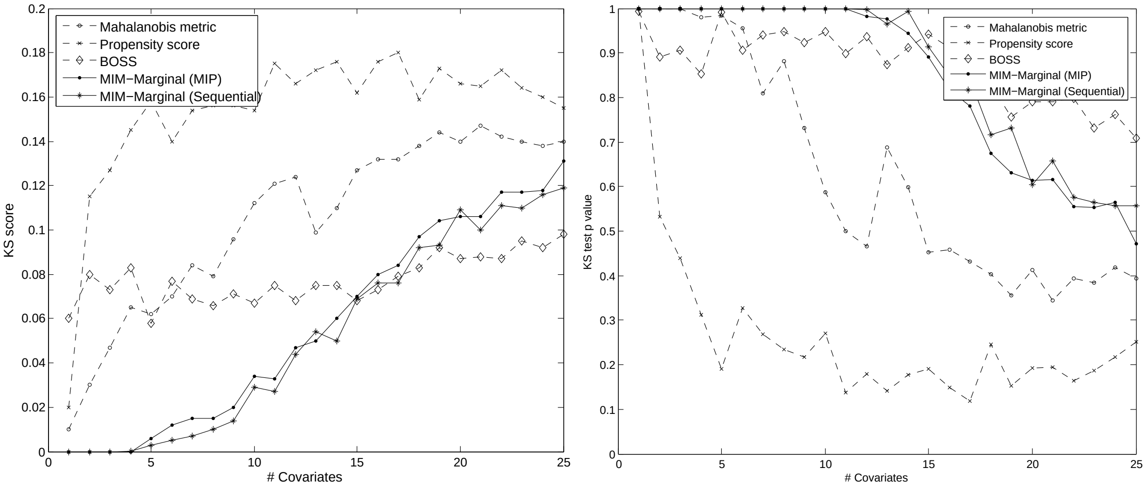
KS distance or p-value over all the covariates, and columns "Max" and "Min" report the maximum KS test statistic and minimum p-values, respectively. The smaller the KS score and the larger the p-value, the better balance is achieved.

In general, the MIM-Joint does not output accurate treatment effect estimates in the multi-covariate cases. If an exact or almost-exact matching solution exists, the MIM-joint performs well, e.g., as in the one-covariate case. However, it is too sensitive to imbalance. As the number of the covariates grows, there remain fewer and fewer bins in which exact matching is possible, making the MIM-Joint formulation not-so-useful for most practical cases. At the same time, the marginal bin-based matching methods succeed in obtaining rather accurate ATT estimates.

Because of the high non-linearity of the response function, propensity score matching does not produce good ATT estimates, and there is no clear trend in its performance as it degrades. Mahalanobis metric matching, BOSS and two MIM-Marginal methods all produce similar estimates, with the MIM-Marginal MIP performing slightly better than the other methods. For the instances with fewer than 15 covariates, the estimates produced by these four methods are close zero (the true ATT value 0), but MIM-Marginal methods achieve much better balance in covariates judging by the KS test scores and p-values. For the

Table 2: Marginal balance quality for matching solutions.

	Mahalanobis metric				Propensity score				BOSS				MIM-Marginal (MIP)				MIM-Marginal (Sequential)			
	KS		pVal		KS		pVal		KS		pVal		KS		pVal		KS		pVal	
# Covariates	Avg	Max	Avg	Min	Avg	Max	Avg	Min	Avg	Max	Avg	Min	Avg	Max	Avg	Min	Avg	Max	Avg	Min
1	0.010	0.010	1.000	1.000	0.020	0.020	1.000	1.000	0.060	0.060	0.994	0.994	0.000	0.000	1.000	1.000	0.000	0.000	1.000	1.000
5	0.062	0.070	0.984	0.967	0.158	0.180	0.190	0.078	0.058	0.070	0.991	0.967	0.006	0.010	1.000	1.000	0.003	0.010	1.000	1.000
10	0.112	0.160	0.588	0.155	0.154	0.200	0.271	0.037	0.067	0.090	0.947	0.813	0.034	0.050	0.999	0.998	0.029	0.040	1.000	1.000
15	0.127	0.180	0.452	0.078	0.162	0.200	0.190	0.037	0.068	0.100	0.942	0.699	0.070	0.120	0.891	0.475	0.069	0.140	0.914	0.281
20	0.140	0.230	0.413	0.010	0.166	0.230	0.193	0.010	0.087	0.150	0.790	0.211	0.106	0.190	0.614	0.080	0.109	0.190	0.605	0.054
25	0.140	0.230	0.393	0.010	0.155	0.230	0.252	0.010	0.098	0.190	0.708	0.054	0.131	0.220	0.471	0.026	0.119	0.240	0.556	0.006



(a) KS scores

(b) KS test p values

Figure 4: Trends for marginal balance quality for matching solutions.

instances with more than 15 covariates, the large number of bins makes for a large variance in the ATT estimates; comparing only the 100 treatment units and the 100 matched control units, one observes significant divergence between the ATT and its estimates obtained with all the methods, even though BOSS achieves the smallest KS scores. Considering both the matching quality and runtime performance, the MIM-Marginal with sequential selection algorithm comes out as the most efficient matching method for practical purposes.

5.2 The Experiences with Using Mutual Information as a Measure of Balance

The next set of experiments reveals that the MI function, employed as the objective in MIM, can be viewed as a surrogate measure of covariate balance. In order to trace the dependence between the MI values, obtained with different control groups, and the corresponding ATT estimates, an additional set of results is reported with the data set of Section 5.1 with 10 covariates. This set is also used to help us assess the impact of the MIM algorithm parameter settings on the matching quality

For a large number of randomly generated control groups, the MI values were recorded together with the resulting treatment effect estimates (see Figure 5). Among these, four groups were found to have the MI less than 0.002 and at least 100 groups fell in each of the other intervals, into which the MI range was divided. Observe that, as the MI grows, the average of the ATT estimates tends to increase, and the standard deviation of the estimate values over the intervals grows as well. This confirms the premise that minimizing MI is a valid approach to guiding the matching process.

Another benefit of using MI lies in the ability to directly compare the matching problem solutions (control groups) of different sizes. Indeed, with the empirical covariate distribution in a given treatment group being fixed, the decision-maker has the freedom of selecting the target control group size. With the same binning scheme, the MI values obtained with

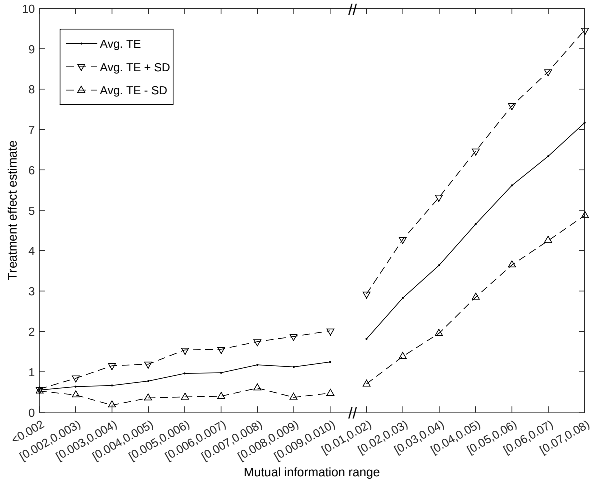


Figure 5: Trends for estimated treatment effects with different mutual information ranges.

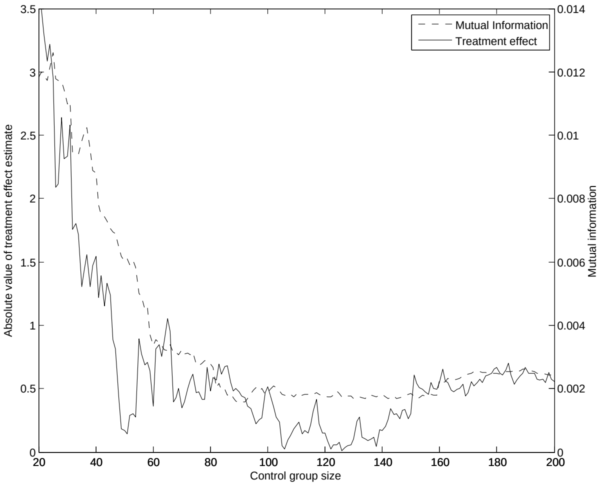


Figure 6: Trends for estimated treatment effects with different control group sizes.

the control groups of different sizes can be compared on the same scale, and hence, the optimization of the control group size becomes possible.

Figure 6 shows the MI values and MIM-based ATT estimates as functions of the control group size in the range from 20 to 200 (of control units). Despite the noise, it is clear that for some control group sizes, the MIM method achieves lower MI values, and simultaneously, higher quality estimates. Most importantly, the group size range, over which the MI is consistently low, coincides with the range, for which the estimates are closest to the truth. As such, in the considered example, the MI gets closer to zero with the lowest noise for the control group of sizes of about 130; a control group of this size returns the ATT estimate value of 0.099.

5.3 Experiences with Large Data Sets

In order to test the performance of the presented MIM methods with large data sets, three suitable real-world data sets were identified. The first one contains weighted census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau (Lichman, 2013): it contains 199,523 records with 41 demographic and employment related variables. The second one was extracted from the 1994 Census database (Lichman, 2013): it contains 32,561 records with 14 variables. The third one was collected in a study focusing on the National Supported Work Demonstration Program (NSW) (LaLonde, 1986), where the randomized job training experiment benchmark was obtained for the treatment effect: it contains 16,177 records with 8 variables. Of the three data sets, only the third one was originally created for a matching purpose. The aim of its creator was to examine how well the statistical methods would perform in trying to replicate the result of a randomized experiment (LaLonde, 1986). To design the test instances with the different number of covariates and varied control pool and treatment group sizes, the first data set was split into a treatment group and a control pool by "US citizenship" and "Business ownership" indicators, respectively; and the second data set was split by "US native" and "Doctorate degree" indicators, respectively. To apply the MIM-Marginal method, each continuous covariate's range was partitioned into 20 bins, while all the categorical covariates kept their original categories. The target control group size was set equal to the treatment group size. Table 3 gives an aggregate view of the data sets' specifics, and MIM results and runtimes.

Overall, the MIM-marginal method achieves very good balance across all the covariates. The average KS scores in all the tests are below 0.01 and the average p-values are all greater than 0.05. Note that since the treated and control data sets in every test are distinct, the results cannot be meaningfully compared across the test instances. For example, the best balance metric values were achieved in the experiment the "Business ownership" data set, even though it had more units and covariates than some other data sets. In the experiment with the "US citizenship" data set, the matching algorithm performed the worst. Indeed, this was a challenging test instance with the target control group size of 13,401, amounting to about 7.2% of the control pool: in such a case, the method is forced to pick non-optimal units to reach the target size, and hence, increases the imbalance. Note, however, that selecting such a large size control group might not be a good idea in practice anyway.

Excellent computational efficiency of the MIM-Marginal method is unparalleled by any other matching method, making it highly practical for data mining; its runtime requirement grows polynomially with the problem size. For example, the "US citizenship" test instance features a very large data set: compared to the training data set, it has 11.6 times more control units, 72.4 times more treated units, significantly larger target control group size, and 5 times more covariates. Yet, the MIM-Marginal runtime with the the "US citizenship" is only 38,040 times larger.

The NSW data set is the most famous one in the matching literature, because an ATT benchmark of 1,794 has been separately obtained for the problem that it addresses (LaLonde, 1986). Dehejia and Wahba (1999) reported the estimate based on propensity score method was 1,691. Tam Cho et al. (2013) used BOSS and obtained the best individual matched solution resulting in the estimate of 1,741, as well as a set of alternative solutions, with mean 1,595 and standard deviation 281.

Table 3: Performance of the MIM-Marginal method on practical data sets.

		Census 94-95	Census 94-95	Extracted 94	Extracted 94	NSW 86
		(US Citizenship)	(Business Ownership)	(US Native)	(Doctorate)	(Training)
Data	# Control Units	186,122	196,825	29,170	32,148	15,992
	# Treated Units	13,401	2,698	3,391	413	185
	# Continuous Covariates	8	8	6	6	4
	# Categorical Covariates	32	32	7	7	4
	# Marginal Bins	645	647	180	206	88
Balance	Avg KS	0.064	0.001	0.020	0.095	0.032
	Max KS	0.412	0.002	0.133	0.200	0.049
	Avg p-value	0.524	1.000	0.769	0.746	0.995
	Min p-value	0.109	1.000	0.518	0.459	0.981
	Avg MI	0.020	0.001	0.004	0.085	0.002
Time	Binning (seconds)	4,410.84	4,319.71	46.23	47.62	1.74
	Matching (seconds)	66,950.58	14,235.49	399.70	49.83	1.76

The runs of MIM-Marginal with the NSW data set produce a solution set with mean 1,851.5 and standard deviation 92.1. The average MI over this set is 0.002383; see the achieved balance metrics in Table 3. Importantly, if one removes the target control group size restriction and allows the MIM-Marginal to optimize over it, then a solution with 169 control units is obtained, with the MI of 0.001824 and ATT estimate set with mean 1,818.2 and standard deviation 91.2.

6. MIM Limitations and Future Research Directions

While the presented computational investigations demonstrate the utility of the MIM methodology, this work has its limitations and desirable directions for further improvement.

First, this paper does not offer an approach to the calculated selection of a binning scheme. The discretization of the covariate space affects the MI-based estimation outputs, however, the present MIM algorithms take the binning scheme as an input and do not work to perturb it to account for the differences in the shapes of the distributions of different covariates or the distances between bins. Intuitively, if the binning is coarse then the MIM cannot be expected to produce high quality solutions. While binning has been a point of research in multiple branches of optimization-based matching literature, the design of binning structures is still an open question. Another point, relevant to the MI based methods specifically, is that mutual information could be employed in its continuous form, in which case the accuracy of matching might be improved without the use of bins.

Second, in its current form as a non-parametric matching methodology, MIM does not differentiate the covariates by relative importance. Moreover, if there is any indispensable information of the form of the response function, covariate relationships, or covariate distributions that is not captured via binning, MIM may underperform. There may even exist circumstances where MIM would be consistently unsuccessful in producing accurate treatment effect estimates: such circumstances, as those exposed by Sauppe et al. (2014) with propensity score-based matching, are yet to be explored with MIM. In any case, prior to using MIM, the researcher must be careful about selecting the covariates to work with. Indeed, data preprocessing has been a topic of research worth much attention. Distance-based matching methods employ weights to emphasize the importance of balancing certain some covariates over others. The developments in propensity score-based methods led to the introduction of the concept of “fine balance”. Expanding the MIM research in a similar direction would add to its value.

Finally, by relaxing the integrality constraints of the MIM problems’ decision variables, one could produce linear (non necessarily integer) solutions allowing for insightful interpretations. The research in this direction might remedy the MIM dependence on binning.

7. Conclusion

The problem of causal inference based on observational data lies in selecting control units from a large unit pool to achieve control groups that are similar in covariate distributions to a given treatment group. To address this problem, this paper presents a set of methods with the objective of minimizing the mutual information between the treatment and covariates over the merged set of selected control and treatment units. Optimal conditions are derived for matching on a single covariate and on the joint distribution of multiple covariates, allowing one to remove non-linear terms from the original mutual information formula and leading to a mixed integer programming formulation of the problem. A sequential selection algorithm is presented that runs in polynomial time and obtains optimal solutions for the problems of matching on a single covariate and matching on a joint distribution of multiple covariates.

Matching problems formulated in this paper for both joint distribution and marginal covariate distributions are analyzed theoretically, and the resulting solution methods tested computationally. The problem of group matching with marginal covariate distributions is proven to be NP-complete, and a fast sub-optimal algorithm is presented. The reported computational study shows that the matching problem formulation with marginal covariate distributions is more valuable than that based on the joint covariate distribution for obtaining accurate causal effect estimates in practice.

Appendix A. Proof of Theorem 1

The decision version of the matching problem on marginal covariate distributions with a fixed target control group size (MIM-Marginal) can be stated as follows. Given a treatment group $\mathcal{T}$, a control pool $\mathcal{C}$ and a set of covariates X_k, $k \in K$. Let m be a marginal bin. (In this proof, it is not necessary to indicate which covariate this marginal bin partitions.) Given parameters γ and N, do there exist subsets $\mathcal{S} \subset \mathcal{C}$ such that $\sum_{k \in K} I(T; X_k) \leq \gamma$ and $|\mathcal{S}| = N$?

First, it has to be proven that MIM-Marginal belongs to the NP class. For any given subset, one can check that the subset contains exactly N units, and then, calculate the mutual information value as in 2 to check if it is smaller or equal to γ. This can be completed in polynomial time, thus MIM-Marginal belongs to NP.

Second, it has to be proven that MIM-Marginal is NP-hard. Let δ_{um} be a binary variable, with $\delta_{um} = 1$ if unit u belongs to m, and 0 otherwise; let η_u be another binary variable, with $\eta_u = 1$ if unit u is selected into $\mathcal{S}$, and 0 otherwise. Let T_m denote the number of units in group $\mathcal{T}$ with the values of covariates falling into bin m. If let $\gamma = 0$ and $N = |\mathcal{T}|$, then problem's objective is to check whether a perfect matching exists, i.e., whether the following constraints can be simultaneously satisfied:

$$\sum_{u=1}^{|\mathcal{C}|} \delta_{um} \eta_u = T_m \quad \forall m, \tag{17}$$

$$\sum_{u=1}^{|\mathcal{C}|} \eta_u = N, \tag{18}$$

$$\eta_u \in \{0, 1\} \quad \forall u,$$

where η_u is the decision variable. Constraint (17) ensures that a perfect matching is achieved in each covariate. Constraint (18) limits the size of a control group that can be selected. Note that, since $N = |\mathcal{T}|$ and each unit belongs to exactly $|K|$ marginal bins, then converting the constraints (17) and (18) into inequalities does not affect the optimal set of the problem, which can now be stated as

$$\sum_{u=1}^{|\mathcal{C}|} \delta_{um} \eta_u \geq T_m \quad \forall m,$$

$$\sum_{u=1}^{|\mathcal{C}|} \eta_u \leq N,$$

$$\eta_u \in \{0, 1\} \quad \forall u.$$

Now, the set cover (SC) problem can be reduced to MIM-Marginal problem. The SC problem is known to be NP-Hard (Garey and Johnson, 1979), and can be stated as follows. Given: an element set J, a collection I of finite subsets of J, and a fixed number n. Question: does I contain a subcollection of sets such that the total number of sets in this subcollection is at most n, and each element of I is included in at least one of the selected sets?

Let δ'_{ij} be a binary variable, with $\delta'_{ij} = 1$ if element j is included in set $I_i \in I$, and 0 otherwise; let η'_i be another binary variable, with $\eta'_i = 1$ if set I_i is selected, and 0 otherwise. The objective of the SC problem is to find η' such that

$$\sum_{i=1}^{|I|} \delta'_{ij}\eta'_i \geq 1 \quad \forall j \in J,$$

$$\sum_{i=1}^{|I|} \eta'_i \leq n,$$

$$\eta'_i \in \{0, 1\} \quad \forall i.$$

Define the following mapping: $T_m = 1$, $N = n$, $u = i$, $m = j$, $\delta_{um} = \delta'_{ij}$ and $\eta_u = \eta'_i$. Thus, the SC problem has a feasible solution if and only if the corresponding MIM-Marginal has a solution. The transformation required to execute the described mapping can be completed in polynomial time in the size of problem inputs. This completes the proof.

Appendix B. Properties of the Ratio Function $f(x, y)$ in Lemma 3

This Appendix analyzes how the values of function $f(x,y) = \frac{(1+\frac{1}{x})^x}{(1+\frac{1}{x+y})^{(x+y)}(1+\frac{y}{x+1})}$ can be compared for arbitrary inputs (x, y), with $x > 0$ and $y > 0$.

Define function $g(x, y) \equiv \log f(x, y) = x \log(1 + \frac{1}{x}) - (x + y) \log(1 + \frac{1}{x+y}) - \log(1 + \frac{y}{1+x})$ $= x(\log(1 + x) - \log(x)) - (x + y)(\log(1 + x + y) - \log(x + y)) - (\log(1 + x + y) - \log(1 + x))$ $= (1 + x) \log(1 + x) - x \log(x) - (1 + x + y) \log(1 + x + y) + (x + y) \log(x + y)$.

Then, $\frac{\partial g}{\partial x} = \log(1+x)-\log(x)+\log(x+y)-\log(1+x+y)$ and $\frac{\partial g}{\partial y} = \log(x+y)-\log(1+x+y)$. Also, $\frac{\partial g}{\partial x} = \frac{1}{f}\frac{\partial f}{\partial x}$ and $\frac{\partial g}{\partial y} = \frac{1}{f}\frac{\partial f}{\partial y}$.

Because $x > 0$ and $y > 0$, one has $1 + x + y > 1 + x > 0$. Also, the logarithm is a monotonically increasing function with the monotonically decreasing slope, and hence, one has $\log(1 + x) - \log(x) > \log(1 + x + y) - \log(x + y) > 0$, which implies that $\frac{\partial g}{\partial x} > 0$ and $\frac{\partial g}{\partial y} < 0$. Moreover, since $f > 0$, then $\frac{\partial f}{\partial x} > 0$ and $\frac{\partial f}{\partial y} < 0$.

To sum up, the function of interest is monotonic along both x and y directions. This prompts one to study its contour lines over the feasible range of inputs (x, y) (Figure 7). Unfortunately, even though the contours look linear, it does not seem possible to produce a closed-form expression for them, of the form $f(x, y) = C$, with constant C.

Since $f(x, y) = C$ is approximately a straight line for every specific C, then the values of $f(x, y)$ under different inputs can be compared by evaluating the slopes of the corresponding contour lines: $\frac{dy}{dx} = -\frac{\frac{\partial f}{\partial x}}{\frac{\partial f}{\partial y}} = -\frac{\frac{\partial g}{\partial x}}{\frac{\partial g}{\partial y}} = \frac{\log(1+x)-\log(x)}{\log(1+x+y)-\log(x+y)} - 1$. Let $h(x, y) \equiv \frac{dy}{dx}$. For arbitrary (x_0, y_0) and (x_1, y_1) such that $f(x_0, y_0) = f(x_1, y_1)$, one has $h(x_0, y_0) = h(x_1, y_1) = \frac{y_1-y_0}{x_1-x_0}$; thus, one can study the linearization $h(x, y)$ of $f(x, y)$. Since these contour lines are straight and do not intersect in the first quadrant, they must have a unique, common intersection point. Thus, one can write $h(x, y) = \frac{y-A_2}{x-A_1}$, where (A_1, A_2) are the coordinates of that unknown intersection point. By numerical approximation, one can derive that $(A_1, A_2) \approx (-0.47, 0)$.

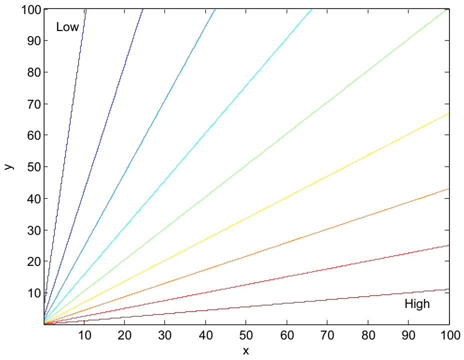


Figure 7: A sketch of the contour lines of $f(x, y)$ on (x, y) plane.

Due to the monotonicity of $f(x, y)$, the greater the slope of a contour line that the point (x, y) lies on, the smaller its corresponding function value. In other words, the smaller $h^{-1}(x, y) = \frac{x-A_1}{y-A_2}$, the smaller $f(x, y)$, and vice versa.

Appendix C. A Complete List of Notations Used

$\mathcal{T}$: Treatment group.
$\mathcal{C}$: Control pool.
$\mathcal{S}$: Control group.
N: A given integer as the target control group size, i.e. $|\mathcal{S}| = N$ for an eligible matched control group.
u: An observable unit, $u \in \mathcal{T} \cup \mathcal{C}$.
t: Treatment indicator (1 means treated; 0 means not treated).
Y_u^1 (or Y_u^0): Treated (or untreated) response of unit u.
b: Joint bin. b_1, b_2 and b_3 are different bins used in proofs.
B: Set of joint bins. B^1, B^2 and B^3 are different bin sets used in proofs. B^F is particular bin set in the MIP-based algorithm.
m: Marginal bin.
M: Set of marginal bins. M^F is particular bin set in the MIP-based algorithm.
k: A covariate.
K: Set of covariates.
X_k: Value of covariate k.
$\mathbf{X}$: Covariate vector $\{X_1, X_2, ..., X_{|K|}\}$.
X: A generalized covariate value to represent $\mathbf{X}$ and X_k for writing convenience.
X^u: Covariate value of unit u.
X_b: Covariate value for any unit contained in bin b.
X_m: Covariate value for any unit contained in marginal bin m.
$p(t)$: Probability that a unit is treated.
$p(X_b)$: Probability that a unit's covariate value falls into bin b.
$p(X_b, t)$: Probability that the covariate value of a unit with treatment indicator t falls into

bin b.
S_b (or T_b, C_b): Number of units in group $\mathcal{S}$ (or $\mathcal{T}$, $\mathcal{C}$) with covariate values falling into bin b. S_{b_1}, S_{b_2}, S_{b_3}, T_{b_1}, T_{b_2}, T_{b_3} and C_{b_2} are the number of units in different groups used in proofs.
S_m (or T_m, C_m): Number of units in group $\mathcal{S}$ (or $\mathcal{T}$, $\mathcal{C}$) with covariate values falling into marginal bin m.
$I(t;X)$ (or $I(t;\mathbf{X})$, $I(t;X_k)$): Mutual information between treatment indicator and covariate value X (or covariate vector $\mathbf{X}$, covariate value X_k).
I_b: Mutual information treatment indicator and covariate value if a unit in bin b is added to the control group, e.g. I_1, I_2 and I_3.
A and α: Constant numbers.
q: Objective value of MIP models.

References

Alberto Abadie and Guido W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Dimitris Bertsimas and Romy Shioda. Classification and regression via integer optimization. *Operations Research*, 55(2):252–271, March 2007. ISSN 0030-364X.

Kenneth P. Burnham and David R. Anderson. *Model selection and multi-model inference: a practical information-theoretic approach*. Springer Verlag, 2002.

W. G. Cochran. The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128(2):234–266, 1965.

Paula da Veiga and Ronald Wilder. Maternal smoking during pregnancy and birthweight: A propensity score matching approach. *Maternal and Child Health Journal*, 12:194–203, 2008.

Rajeev Dehejia. Practical propensity score matching: a reply to smith and todd. *Journal of Econometrics*, 125(1):355–364, 2005.

Rajeev H. Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):pp. 1053–1062, 1999.

Rajeev H. Dehejia and Sadek Wahba. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics*, 84(1):151–161, 2002.

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

Pablo A. Estévez, Michel Tesmer, Claudio A. Perez, and Jacek M. Zurada. Normalized mutual information feature selection. *IEEE Transactions on Neural Networks*, 20(2): 189–201, 2009.

Andrew M. Fraser and Harry L. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33:1134–1140, 1986.

Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1979.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Ben B. Hansen. Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99(467):609–618, 2004.

Ben B. Hansen and Stephanie O. Klopfer. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 2012.

Daniel Ho, Kosuke Imai, Gary King, and Elizabeth Stuart. Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(1):1–28, 2011.

Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236, 2007.

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

Stefano M. Iacus, Gary King, and Giuseppe Porro. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24, 2012.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69:066138, 2004.

Robert J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986.

M. Lichman. UCI machine learning repository, 2013. URL `http://archive.ics.uci.edu/ml`.

Stephen L. Morgan and David J. Harding. Matching estimators of causal effects. *Sociological Methods & Research*, 35(1):3–60, 2006.

Alexander G. Nikolaev, Sheldon H. Jacobson, Wendy K. Tam Cho, Jason J. Sauppe, and Edward C. Sewell. Balance optimization subset selection (BOSS): An alternative approach for causal inference with observational data. *Operations Research*, 61(2):398–412, 2013.

Stefano Panzeri and Alessandro Treves. Analytical estimates of limited sampling biases in different information measures. *Network Computation in Neural Systems*, 7(1):87–107, 1996.

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

Paul R. Rosenbaum and Donald B. Rubin. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38, 1985.

Paul R. Rosenbaum, Richard N. Ross, and Jeffrey H. Silber. Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association*, 102(477):75–83, 2007.

Mark S. Roulston. Estimating the errors on measured entropy and mutual information. *Physica D: Nonlinear Phenomena*, 125:285–294, 1999.

Donald B. Rubin. Matching to remove bias in observational studies. *Biometrics*, 29:159–183, 1973.

Donald B. Rubin. Bias reduction using mahalanobis-metric matching. *Biometrics*, 36: 293–298, 1980.

Donald B. Rubin. Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, 2:169–188, 2001.

Donald B. Rubin. *Matched Sampling for Causal Effects*. Cambridge University Press, New York, 2006.

Jason J. Sauppe, Sheldon H. Jacobson, and Edward C. Sewell. Complexity and approximation results for the balance optimization subset selection model for causal inference in observational studies. *INFORMS Journal on Computing*, 26(3):547–566, 2014.

Jasjeet S. Sekhon. Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*, 42 (i07), 2008.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

Jeffrey Smith and Petra Todd. Rejoinder. *Journal of Econometrics*, 125(1):365–375, 2005a.

Jeffrey Smith and Petra Todd. Does matching overcome lalonde's critique of nonexperimental estimators? *Journal of Econometrics*, 125(1):305–353, 2005b.

Wendy K. Tam Cho, Jason J. Sauppe, Alexander G. Nikolaev, Sheldon H. Jacobson, and Edward C. Sewell. An optimization approach for making causal inferences. *Statistica Neerlandica*, 67(2):211–226, 2013.

Jos R. Zubizarreta. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107 (500):1360–1371, 2012.