Algorithms and Complexities of Matching Variants in Covariate Balancing

Dorit S. Hochbaum, Asaf Levin, Xu Rao

http://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Articles in Advance, pp. 1–15
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Methods

Algorithms and Complexities of Matching Variants in Covariate Balancing

Dorit S. Hochbaum,[a] **Asaf Levin,**[b,*] **Xu Rao**[a]

[a] Department of IEOR, Etcheverry Hall, Berkeley, California; [b] Faculty of Industrial Engineering and Management, The Technion, Haifa, Israel
*Corresponding author

Contact: hochbaum@ieor.berkeley.edu, https://orcid.org/0000-0002-2498-0512 (DSH); levinas@technion.ac.il, https://orcid.org/0000-0001-7935-6218 (AL); xrao@berkeley.edu, https://orcid.org/0000-0003-0260-891X (XR)

Received: December 2, 2020
Revised: May 13, 2021; January 10, 2022
Accepted: January 31, 2022
Published Online in Articles in Advance: April 26, 2022

Area of Review: Optimization

https://doi.org/10.1287/opre.2022.2286

Abstract. Here, we study several variants of matching problems that arise in covariate balancing. Covariate balancing problems can be viewed as variants of matching, or b-matching, with global side constraints. We present here a comprehensive complexity study of the covariate balancing problems providing polynomial time algorithms, or a proof of NP-hardness. The polynomial time algorithms described are mostly combinatorial and rely on network flow techniques. In addition, we present several fixed-parameter tractable results for problems where the number of covariates and the number of levels of each covariate are seen as a parameter.

Funding: This work was supported by National Science Foundation [Grants CMMI-1760102 and NSF 2112533]; Israel Science Foundation (308/18).

Supplemental Material: The e-companion is available at https://doi.org/10.1287/opre.2022.2286.

Keywords: algorithms • complexity • matching • covariate balance • observational studies

1. Introduction

In many modern scenarios, one studies a variant of the matching problem, or the *b*-matching problem, where there is an additional family of side constraints. Such side constraints were studied, for example, in the context of mechanism design where one adds proportional constraints (a.k.a. distributional constraints) for *b*-matching on bipartite graphs. In that case, the set of nodes on one side has some additional property, and the constraints say that the matched nodes to a node on the other side have the same distribution of the additional property (as the entire set of nodes). This is in addition to the upper bound constraint on the total number of nodes matched to that node of the other side; see, for example, Ágoston et al. (2018), Nguyen and Vohra (2019), Bei et al. (2020), and Ashlagi et al. (2020). Other constraints of this type were also studied; see, for example, ratio constraints (Yahiro et al. 2020) and general multidimensional knapsack constraints (Nguyen et al. 2019). Here, we are motivated by applications of design of observational studies that use what is referred to as "matching methods under fine-balanced constraints."

The problem of balancing covariates arises in observational studies in various contexts such as statistics (Rosenbaum 2002, Rubin and Stuart 2006), epidemiology (Brookhart et al. 2006), sociology (Morgan and Harding 2006), economics (Imbens 2004), and political science (Ho et al. 2007). In an observational study, there are two disjoint groups of samples, one of treatment samples and the other of control samples. Each of the samples in the two groups is characterized by several observed covariates, or features.

Even though covariate balancing problems have been extensively studied (Stuart 2010, Rosenbaum 2020), the complexity status of many variants of the problems has not been established in theory. Some of our results demonstrate that two-covariate balancing problems are polynomial time solvable, whereas almost all problems are hard for three or more covariates. These results have practical implications such as justifying the use of implicit enumeration techniques or heuristics for the hard cases. A new approach suggested by our results is to make use of the two-covariate polynomial cases and relax the problem by either selecting two major covariates to represent all covariates or aggregating covariates into two sets. This relaxation would then be solved efficiently and might provide good-quality solutions to the respective hard problems. However, the possible use of such an approach depends on additional aspects that are beyond the scope of our study because they should be based on strong statistical justifications.

Covariate balancing problems arise when estimating causal effects using observational data. It is desirable to replicate a randomized experiment as closely as possible by obtaining treatment and control groups with similar covariate distributions. This goal can often be achieved by choosing well-matched samples of the original treatment and control groups, thereby reducing bias in the estimated treatment effects due to the observed covariates. The matching is to assign each treatment sample to one unique control sample or, in other setups, to assign each treatment sample to a unique set of κ control samples, for κ a prespecified integer, where every control sample is assigned to at most one treatment sample. Detailed reviews of matching-related methods used for covariate balancing problems are given by Stuart (2010) and Rosenbaum (2020).

In this paper, we address various problems of balancing covariates. The covariates here are *nominal* in that they take on discrete values or categories. The set of values of each nominal covariate partitions the treatment and control samples to a number of subsets referred to as *levels*, where the samples at every level share the same covariate value. In an ideal situation known as exact matching, the samples of the treatment and the control in each matched pair or matched set belong to the same levels over all covariates. However, satisfying the requirement that matched samples in each pair or set belong to the same levels over all covariates typically results in a very small selection from the treatment and control group, which is not desirable. To address this, Rosenbaum et al. (2007) introduced a weaker requirement to match all treatment samples to a subset of the control samples, called selection, so that the proportion (or the number, if κ = 1) of control and treatment samples in each level of each covariate is the same. This requirement is known in the literature as *fine balance*.

To formalize the discussion, we introduce essential notation. Let the number of treatment samples be n and the number of control samples be n'. Let the set of all treatment samples be denoted by $\mathcal{T}$, $|\mathcal{T}| = n$. Let P be the number of covariates to be balanced. For $p = 1, \ldots, P$, covariate p partitions both treatment and control groups into k_p levels each. Let the partition of the treatment group under covariate p be $L_{p,1}, L_{p,2}, \ldots, L_{p,k_p}$ of sizes $\ell_{p,1}, \ell_{p,2}, \ldots, \ell_{p,k_p}$. Similarly, let the partition of the control group under covariate p be $L'_{p,1}, L'_{p,2}, \ldots, L'_{p,k_p}$ of sizes $\ell'_{p,1}, \ell'_{p,2}, \ldots, \ell'_{p,k_p}$. Let κ be an integer specifying the ratio of the number of matched control samples to the number of matched treatment samples.

We define the κ-fine-balance constraints for a selection of treatment and a selection of control samples as follows:

Definition 1 (κ-Fine-Balance). For an integer κ, a selection $S \subseteq \mathcal{T}$ of the treatment group and a selection S' of the control group, we say that (S, S')-κ-fine-balance is satisfied if $\kappa \cdot |S \cap L_{p,i}| = |S' \cap L'_{p,i}|$ for $p = 1, \ldots, P$ and $i = 1, \ldots, k_p$.

Obviously for S, S' satisfying (S, S')-κ-fine-balance, the cardinality of S' is κ times as large as the cardinality of S, $|S'| = \kappa\,|S|$.

We are now ready to define the four families of problems investigated here with complexity that varies according to the number of covariates and the value of κ. The *maximum κ-fine-balance selection* (κ-FBS) problem is to select a subset $S \subseteq \mathcal{T}$ and a subset S' of the control group so as to maximize the size of the selection S (equivalent to maximizing the size of S' because $|S'| = \kappa\,|S|$) where the (S, S')-κ-fine-balance constraints are satisfied.

In the second problem family, the fine balance constraints are relaxed and replaced by bounds on the violation of each level size. This problem, previously studied in Zubizarreta et al. (2014), King et al. (2017), and Visconti and Zubizarreta (2018), is to maximize the selection size subject to constraints on the amount of imbalance permissible at each level. We refer to this problem as the Bounded Balance Selection (BBS), or κ-BBS for $\kappa \geq 2$. In a slight generalization of the problem previously studied, we permit different bounds on the excess than the bounds on the deficit at each level. Let $B^{(d)}_{p,i}$ and $B^{(e)}_{p,i}$ be the upper bounds on the deficit and excess, respectively, at level i of covariate p, for $p = 1, \ldots, P$ and $i = 1, \ldots, k_p$. Formally, the relaxed constraints we consider are as follows.

Definition 2 (κ-Bounded-Balance). For an integer κ, a selection $S \subseteq \mathcal{T}$ of the treatment group and a selection S' of the control group, we say that (S, S')-κ-bounded-balance is satisfied if $-B^{(e)}_{p,i} \leq \kappa \cdot |S \cap L_{p,i}| - |S' \cap L'_{p,i}| \leq B^{(d)}_{p,i}$ for $p = 1, \ldots, P$ and $i = 1, \ldots, k_p$.

Using the definition of the κ-bounded-balance constraints, the κ-BBS problem is to find a selection $S \subseteq \mathcal{T}$ of the treatment group and a selection S' of the control group satisfying the κ-bounded-balance constraints so as to maximize the selection size, $|S|$. Note that for this problem, and unlike κ-FBS, the objective is not necessarily equivalent to maximizing $|S'|$. We also consider the variant of this problem, κ-MBBS, where we add the constraint $|S'| = \kappa\,|S|$. We show that our results for κ-FBS also hold for κ-MBBS. One way to formulate this new variant using κ-BBS is to add one auxiliary covariate with only one level with zero upper bounds on its deficit and excess. However, our results depend on the number of covariates, so we will avoid using this reformulation of κ-MBBS. When we compare the output of the κ-FBS problem and the output of the κ-BBS problem, we expect that in the κ-BBS problem when the upper bounds on the deficit and excess are increased the size of the selection of the treatment group is increased. Thus, the κ-BBS problem serves as a tool to examine the solutions obtained as the Pareto-optimal solutions with respect to the objectives of maximizing the selection size and minimizing the upper bounds on the deficit and excess.

Another problem studied here is the *κ-fine-balance matching* (κ-BM) problem, first introduced by Rosenbaum et al. (2007) for one covariate. Here, we are given a distance, or cost, measure between each treatment and each control sample. The κ-BM problem is to minimize the total cost of the assignment of each treatment sample in $\mathcal{T}$ to κ control samples such that the selection of matched control samples S' satisfies $(\mathcal{T}, S')$-κ-fine-balance. Regarding the introduction of this problem by Rosenbaum et al. (2007), it should be mentioned that in that paper the fine balance constraints ensure that the *joint distribution* of the covariates is equated between the matched treatment and the control group. Here, we define the fine balance constraints over the *marginal distributions* of the covariates. The two definitions coincide for the case where there is only one covariate, which is why we say that the κ-BM problem was introduced for one covariate in Rosenbaum et al. (2007). Observe that relaxing the κ-fine-balance constraints when there are several covariates so that they relate to the marginal distributions of the covariates instead of the joint distribution of the covariates has benefits for practitioner. For example, it allows feasible solutions even in cases where there is no selection enforcing the fine balance constraints over the joint distribution.

Another problem family newly introduced here is an optimization where the feasible sets are optimal for another problem. Formally, in the first stage the goal is to find the optimal selections to the κ-FBS problem. In the second stage, among all maximum-sized selections, find the selection that minimizes the total distance of an assignment of each selected treatment sample to exactly κ selected control samples. We refer to this problem as *maximum selection κ-fine-balance matching problem* (κ-MSBM).

Notation-wise, for the case of $\kappa = 1$ we omit the prefix κ, so (S, S')-κ-fine-balance is called (S, S')-fine-balance, κ-FBS problem is called FBS problem, κ-BBS problem is called BBS problem, κ-BM problem is called BM problem, and κ-MSBM problem is called MSBM problem.

A summary of the problems investigated here is given in Table 1.

1.1. Related Literature

The concept of fine balance was first introduced by Rosenbaum et al. (2007), who studied the κ-BM problem for the 1-covariate problem and proposed a network flow algorithm. Sauppe et al. (2014) showed that the BM problem for two or more covariates is NP-hard, and therefore, no polynomial running time algorithm is known for the κ-BM problem. Rosenbaum (2012) considered the problem of finding a subset of the treatment samples of certain cardinality that is matched to a subset of the control samples so as fine balance constraints are satisfied and the total cost is minimized. In Rosenbaum (2012), the two objectives of minimizing the total distance and of maximizing the selection size are seen as a biobjective optimization problem, and a procedure for finding the Pareto-efficient frontier of these objectives is designed.

It is not always feasible to find a selection S' of the control samples that satisfies the $(\mathcal{T}, S')$-κ-fine-balance constraints in the κ-BM problem. To that end, several papers considered the goal of minimizing the violation of this requirement, which we refer to as *imbalance* (Yang et al. 2012, Zubizarreta 2012, Pimentel et al. 2015, Bennett et al. 2020, Hochbaum et al. 2022). The studies in all these papers require the entire treatment group to be selected or matched. Sauppe (2015), Bennett et al. (2020), and Hochbaum et al. (2022) considered finding the selection of the control group that minimizes an imbalance objective, defined as $\sum_{p=1}^{P}\sum_{i=1}^{k_p} ||S' \cap L'_{p,i}| - \kappa \cdot \ell_{p,i}|$. This problem is called *minimum κ-imbalance problem*. The problem is trivial to solve for the 1-covariate problem (see Section 2 for details); the 2-covariate problem was proven to be polynomial time solvable using linear programming in Bennett et al. (2020) and using network flow algorithms in Sauppe (2015) and Hochbaum et al. (2022); for three or more covariates, the problem is NP-hard (Sauppe 2015, Bennett et al. 2020, Hochbaum et al. 2022). Yang et al. (2012) and Pimentel et al. (2015) considered a more complicated problem that minimizes the total assignment cost of the matched sets, each consisting of a single treatment sample and κ control samples, subject to the requirement that the selection of matched control samples is optimal for the minimum κ-imbalance problem. Yang et al. (2012) proposed two network flow algorithms for the case of the 1-covariate problem; Pimentel et al. (2015) proposed a network flow algorithm for the case in which the covariates form a nested sequence. Zubizarreta (2012) considered a different variant that minimizes the total assignment cost of the matched sets with a penalty on the imbalance and presented a mixed integer programming formulation for an arbitrary number of covariates. The bounded balance problem BBS was considered in Zubizarreta et al. (2014), King et al. (2017), and Visconti and Zubizarreta (2018). In King et al. (2017) it was used as a subroutine called multiple times for generating an efficient frontier on the trade-off between the violation of the fine balance and the size of the selection. Additional models for optimization with fine balance constraints were studied in Nikolaev et al. (2013), Tam Cho et al. (2013), Sauppe et al. (2014), Sauppe (2015), Sauppe and Jacobson (2017), Dutta et al. (2017), Kwon (2018), Kwon et al. (2019a, b), Karmakar et al. (2019), Sharma et al. (2020), and Kwon et al. (2020).

1.2. Contributions

We introduce here, for the first time, polynomial time algorithms for several covariate balancing problems. For the FBS and BBS problems on two covariates or less, we provide polynomial time network flow

Table 1. Summary of Problems Studied Here

Problem name	Objective	Constraints
Max fine-balance selection (FBS)	max $\lvert S\rvert$	(S, S')-fine-balance
Max κ-fine-balance Selection (κ-FBS)	max $\lvert S\rvert$	(S, S')-κ-fine-balance
Max bounded-balance selection (BBS)	max $\lvert S\rvert$	(S, S')-bounded-balance
Max κ-bounded-balance selection (κ-BBS)	max $\lvert S\rvert$	(S, S')-κ-bounded-balance
Max modified bounded-balance selection (MBBS)	max $\lvert S\rvert$	(S, S')-bounded-balance and $\lvert S'\rvert = \lvert S\rvert$
Max κ-modified-bounded-balance selection (κ-MBBS)	max $\lvert S\rvert$	(S, S')-κ-bounded-balance and $\lvert S'\rvert = \kappa \cdot \lvert S\rvert$
Fine-balance matching (BM)	min assignment cost	$(\mathcal{T}, S')$-fine-balance
κ-Fine-balance matching (κ-BM)	min assignment cost	$(\mathcal{T}, S')$-κ-fine-balance
Max selection fine-balance matching (MSBM)	min assignment cost	(S, S') optimal for FBS
Max selection κ-fine-balance matching (κ-MSBM)	min assignment cost	(S, S') optimal for κ-FBS

algorithms solving the problems. For problem instances of κ-FBS, κ-BBS, κ-BM, and MSBM, where the number of level intersections is fixed, we show that the problems are polynomial time solvable, using specific mixed integer programming formulations. More specifically, we prove that the κ-FBS, κ-BBS, κ-BM, and MSBM problems are solvable in *fixed-parameter tractable* (FPT) time for the parameter being the total number of covariates levels, yet the κ-MSBM problem is NP-hard for constant $\kappa \geq 3$, even when the numbers of covariates levels are constant.

For the κ-FBS problem, we prove that for three or more covariates, the FBS and κ-FBS problems are NP-hard for any value of κ. For the case of the 2-covariate problem, we present an efficient algorithm for the FBS problem based on an integer programming formulation of the problem in which the constraint matrix has the structure of network flow constraints. For the resulting minimum cost network flow problem, we apply an algorithm with running time $O(n \cdot (\min\{n + n', k_1k_2\} + (k_1 + k_2)\log(k_1 + k_2))$. We also prove that for $\kappa \geq 3$, the 2-covariate κ-FBS problem is NP-hard. For the remaining case in which $\kappa = 2$ and the number of covariates is two, the complexity status of the 2-FBS problem is left open.

The κ-BBS problem generalizes κ-FBS by allowing violations of the fine balance constraints. When the permissible violations are zero, the κ-BBS problem is the κ-FBS problem. Therefore, because it is only harder, for three or more covariates the BBS and κ-BBS problems are NP-hard for any value of κ, and for $\kappa \geq 3$ the 2-covariate κ-BBS problem is NP-hard as well. For the two covariates BBS, we provide a polynomial time algorithm in the form of a minimum cost network flow that runs in time $O(n \cdot (\min\{n + n', k_1k_2\} + (k_1 + k_2)\log(k_1 + k_2))$. Similarly, for the 2-BBS problem on two covariates, the complexity status remains open.

As indicated earlier, the 2-covariate BM problem is NP-hard, and therefore, there is no polynomial time algorithm for the κ-BM problem with two or more covariates unless $P = NP$. The κ-BM problem is shown here to be solvable in polynomial time when the number of level intersections is fixed. However, when there are two covariates and only one of the two covariates has a fixed number of levels (so the number of level intersections in not necessarily fixed), the κ-BM problem is equivalent to the exact matching problem. The exact matching problem is known to have a randomized polynomial time algorithm (Mulmuley et al. 1987), but the existence of a deterministic polynomial time algorithm for the problem is a longstanding open problem. Therefore, the existence of a deterministic polynomial time algorithm for the problem of two covariates κ-BM problem where the first covariate has a fixed number of levels, is an open problem as well.

The κ-MSBM problem is newly introduced here. It relaxes the requirement in the κ-BM problem of selecting all treatment samples and replaces it with a maximum size selection possible while enforcing the κ-fine-balance constraints. This κ-MSBM problem, as shown here, is NP-hard with two or more covariates for any given value of κ. Moreover, it is also proven here to be NP-hard for the 1-covariate problem when $\kappa \geq 3$. We present a polynomial algorithm for the 1-covariate MSBM problem, but the complexity status of the 1-covariate, 2-MSBM problem is left open. We observe here that, for any number of covariates, if the selections of treatment and control samples are fixed, then the optimal assignment among the selected samples, and therefore the optimal solution to the κ-MSBM problem, is attained by solving a minimum cost network flow problem. See Section 2 for details.

A summary of the complexity results for the four problem families, excluding the case of fixed number of level intersections, is given in Table 2.

1.3. Paper Overview

In Section 2, we consider the case of the 1-covariate κ-FBS, κ-BBS, κ-BM, and MSBM problems and provide a compact representation of the sample selections. Then, we present our complexity and algorithmic results of the other cases for the four families of problems separately, that is, the κ-FBS problem in Section 3, the κ-BBS in Section 4, the κ-BM problem in Section 5, and the κ-MSBM

Table 2. Summary of Complexity and Algorithmic Results Derived Here (Here, n is the Size of Treatment Group and n' is the Size of Control Group)

Problem	One covariate	Two covariates	≥ 3 covariates
FBS	$O(n+n')$	$O(n\cdot(\min\{n+n',k_1k_2\}+(k_1+k_2)\log(k_1+k_2)))$	NP-hard
κ-FBS ($\kappa \geq 2$)	$O(n+n')$	NP-hard for $\kappa \geq 3$, open for $\kappa = 2$	NP-hard
BBS or MBBS	$O(n+n')$	$O(n\cdot(\min\{n+n',k_1k_2\}+(k_1+k_2)\log(k_1+k_2)))$	NP-hard
κ-BBS or κ-MBBS ($\kappa \geq 2$)	$O(n+n')$	NP-hard for $\kappa \geq 3$, open for $\kappa = 2$	NP-hard
κ-BM (any κ)	$O((n+n')^3)$ Rosenbaum et al. (2007)	NP-hard Sauppe et al. (2014)	NP-hard
MSBM	$O((n+n')nn')$	NP-hard	NP-hard
κ-MSBM ($\kappa \geq 2$)	NP-hard for $\kappa \geq 3$ open for $\kappa = 2$	NP-hard	NP-hard

problem in Section 6. The fixed-parameter complexity results are provided in Section 7.

2. Preliminaries

Consider first the case of a single covariate, $p = 1$, that partitions the control and treatment groups into, say, k levels each. Let the sizes of levels of the treatment group be $\ell_1, \ldots, \ell_k$ and the sizes of levels of the control group be $\ell'_1, \ldots, \ell'_k$. It is easy to see that there exists a selection S' of control samples that satisfies the $(\mathcal{T}, S')$-κ-fine-balance if and only if $\ell'_i \geq \kappa\ell_i$ for $i = 1, \ldots, k$. If this condition is satisfied, then any subset S^* of the control group with $\kappa \cdot \ell_i$ samples in level i, $i = 1, \ldots, k$, satisfies the $(\mathcal{T}, S^*)$-κ-fine-balance and as such is a feasible selection for the κ-BM problem. With these known numbers of control samples to be selected in each level, the optimal solution to the 1-covariate κ-BM problem is found using a minimum cost network flow formulation, as shown next. Note that a standard linear programming formulation of the minimum cost network flow (MCNF) is given in Appendix A in the e-companion.

The MCNF problem, the solution to which is an optimal solution to κ-BM, is constructed on a bipartite graph, with the treatment samples each represented by a node on one side and the control samples each represented by a node on the other side. The cost on each arc between a treatment sample and a control sample is the "distance" value between the two, and the arc capacity is 1. Each treatment sample has a supply of κ. To account for the requirement that in each level i of control samples there will be $\kappa \cdot \ell_i$ samples matched, we add to the bipartite graph a third layer of k nodes, one for each level. The ith node in the third layer has demand of $\kappa \cdot \ell_i$, and there are arcs to this demand node from all control samples in level i with capacity 1 and cost of 0. In an optimal solution to this MCNF problem, the control sample nodes through which there is a positive flow (of one unit) are the ones selected and matched to the respective treatment sample nodes from which they have a positive flow.

If $\ell'_i < \kappa\ell_i$ for some i, then there is no selection S' of control samples that satisfies the $(\mathcal{T}, S')$-κ-fine-balance. Addressing this context, as mentioned earlier, Yang et al. (2012), Zubizarreta (2012), Pimentel et al. (2015), Bennett et al. (2020), and Hochbaum et al. (2022) considered the problem of minimizing the κ-imbalance, which is the sum of violations for all levels, $\sum_{i=1}^{k} |S' \cap L'_i| - \kappa \cdot \ell_i|$. The solution to this 1-covariate minimum κ-imbalance problem is straightforward. In step 1, select $\min\{\kappa \cdot \ell_i, \ell'_i\}$ control samples in level i; if the number of control samples selected is less than κn in step 1, then we select arbitrary additional control samples such that the selection is of size κn. Another way to address this context is to seek a solution for the (S, S')-κ-fine-balance where, rather than forcing all samples of $\mathcal{T}$ to be included, finding a solution in which the size of the selection S, and equivalently $|S'|$, is maximized. This problem is the κ-FBS problem. The solution to the 1-covariate κ-FBS problem is also straightforward; select $\bar{\ell}_i = \min\{\ell_i, \lfloor \ell'_i/\kappa \rfloor\}$ treatment samples of level i and $\kappa \cdot \bar{\ell}_i$ control samples of level i.

Similarly, the solution to the 1-covariate κ-BBS problem is straightforward. For all levels i, let $\bar{\ell}_i = \min\{\ell'_i, \kappa\ell_i\}$, and if $\bar{\ell}_i < \ell'_i$, then select $\min\left\{\ell_i, \left\lfloor \frac{\bar{\ell}_i + B^{(d)}_{1,i}}{\kappa} \right\rfloor\right\}$ level i treatment samples and $\bar{\ell}_i$ control samples of level i, and otherwise select all level i treatment group and any $\kappa\ell_i$ control samples of level i.

The solution to the 1-covariate κ-MBBS problem is simple as well. First, select $\bar{\ell}_i = \min\{\ell_{1,i}, \lfloor \ell'_{1,i}/\kappa \rfloor\}$ number of level i treatment samples and $\kappa \cdot \bar{\ell}_i$ number of level i control samples. Next, we create a pool of candidate samples. For each level i, if there exists an unselected treatment sample in the level, we add up to $B^{(d)}_{1,i}/\kappa$ of them to the pool; otherwise, we add up to $B^{(e)}_{1,i}$ of remaining control samples in this level to the pool. Let t denote the number of treatment samples in the pool and c denote the number of control samples in the pool. Then, we add to the selected groups $\min\{t, \lfloor c/\kappa \rfloor\}$ number of treatment samples and $\kappa \cdot \min\{t, \lfloor c/\kappa \rfloor\}$ number of control samples chosen arbitrarily from the candidate pool. This is an optimal solution for the 1-covariate κ-MBBS problem.

Again, for the 1-covariate problem of finding an optimal matching, or assignment, among all optimal selections for either the minimum κ-imbalance or the FBS problem, we solve an MCNF problem for the known number of samples to select from each level, similar to the one defined above with the following modifications. For the minimum κ-imbalance, we first need to change the demand of the demand nodes in the above MCNF problem from $\kappa \cdot \ell_i$ to $\min\{\kappa \cdot \ell_i, \ell'_i\}$ for each level i. We also add a dummy demand node in the third layer with demand $\kappa \cdot n - \sum_{i=1}^{k} \min\{\kappa \cdot \ell_i, \ell'_i\}$, which connects with all control nodes each with capacity 1 and cost of 0. For the optimal selections of FBS, in addition to changing the demand from ℓ_i to $\bar{\ell}_i$ for each level i, we also remove the supply on each treatment sample, add for every level i a supply node with supply $\bar{\ell}_i$, and add arcs from this supply node to all treatment samples in level i with capacity 1 and cost of 0. The best assignment found with a selection that is optimal for the FBS problem is an optimal solution for the MSBM problem. However, this method does not apply to the κ-MSBM problem with $\kappa \geq 2$. We further show that even the 1-covariate κ-MSBM problem is NP-hard for $\kappa \geq 3$ (see Section 6).

Hence, all problems discussed here, except for the κ-MSBM problem, are polynomial time solvable for the 1-covariate case. In Section 6, we show that the 1-covariate κ-MSBM does not admit a polynomial time algorithm for $\kappa \geq 3$ unless P = NP.

Consider next the case of multiple covariates. For the κ-FBS problem (and similar arguments hold also for κ-BBS problem), we observe that the selections from the treatment and control groups can be represented compactly in terms of *level-intersections*. For P covariates, the intersection of the level sets $L_{1,i_1} \cap L_{2,i_2} \cap \ldots \cap L_{P,i_P}$, $i_p = 1, \ldots, k_p$, $p = 1, \ldots, P$ form a partition of the treatment group. Similarly, the intersection of the level sets $L'_{1,i_1} \cap L'_{2,i_2} \cap \ldots \cap L'_{P,i_P}$, $i_p = 1, \ldots, k_p$, $p = 1, \ldots, P$ form a partition of the control group. Therefore, instead of specifying which sample belongs to the selection, it is sufficient to determine the number of selected samples in each level intersection for the two groups, because the identity of the specific selected samples has no effect on the fine balance requirement. With this discussion, we have a theorem on the representation of the solution to the κ-FBS problems in terms of the level-intersection sizes.

Theorem 1. *The level-intersection sizes $s_{i_1,i_2,\ldots,i_P}$ and $s'_{i_1,i_2,\ldots,i_P}$ are an optimal solution to the κ-FBS problem if there exists an optimal selection S of treatment samples and S' of control samples such that $s_{i_1,i_2,\ldots,i_P} = |S \cap L_{1,i_1} \cap L_{2,i_2} \cap \ldots \cap L_{P,i_P}|$ and $s'_{i_1,i_2,\ldots,i_P} = |S' \cap L'_{1,i_1} \cap L'_{2,i_2} \cap \ldots \cap L'_{P,i_P}|$ for $p = 1, \ldots, P$, $i_p = 1, \ldots, k_p$.*

We will say that the optimal selection for the covariate problems here is unique if for any optimal selection S and S' the numbers $s_{i_1,i_2,\ldots,i_P} = |S \cap L_{1,i_1} \cap L_{2,i_2} \cap \ldots \cap L_{P,i_P}|$ and $s'_{i_1,i_2,\ldots,i_P} = |S' \cap L'_{1,i_1} \cap L'_{2,i_2} \cap \ldots \cap L'_{P,i_P}|$ are unique. In order to derive an optimal selection given the optimal level-intersection sizes, one selects any $s_{i_1,i_2,\ldots,i_P}$ treatment samples from the intersection $L_{1,i_1} \cap L_{2,i_2} \cap \ldots \cap L_{P,i_P}$ and any $s'_{i_1,i_2,\ldots,i_P}$ control samples from the intersection $L'_{1,i_1} \cap L'_{2,i_2} \cap \ldots \cap L'_{P,i_P}$ for $i_p = 1, \ldots, k_p$, $p = 1, \ldots, P$.

We observe here that, for any number of covariates, if the optimal selection of treatment and control samples in terms of level-intersections is known and unique, then the optimal assignment among the selected samples, and therefore the optimal solution to the κ-MSBM problem, can also be attained by solving an MCNF problem as follows. For each nonempty level intersection of treatment samples there is a source node with supply of $s_{i_1,i_2,\ldots,i_P}$. This source node is connected to all treatment samples in the intersection $L_{1,i_1} \cap L_{2,i_2} \cap \ldots \cap L_{P,i_P}$ with arcs of capacity 1 and cost of 0. For each nonempty level intersection of control samples there is a demand node with demand of $s'_{i_1,i_2,\ldots,i_P}$. This demand node is connected from all control samples in the intersection $L'_{1,i_1} \cap L'_{2,i_2} \cap \ldots \cap L'_{P,i_P}$ with arcs of capacity 1 and cost of 0. The treatment and control sample nodes through which there is a positive flow (of some unit) are the ones selected, and a positive flow between a treatment node and a control node indicates that the two samples are matched. This is a minimum cost network flow problem with a total demand (or supply) bounded by $\min\{n, n'\}$ and $O(nn')$ arcs and $O(n + n')$ nodes. Therefore, the successive shortest paths algorithm, discussed below in Section 3, solves this problem in $O((n + n')nn')$ steps.

3. The Maximum κ-Fine-Balance Selection (κ-FBS) Problem

In this section, we show the complexity and algorithmic results for the κ-FBS problems. We present the results separately first for three or more covariates, then the 2-covariate FBS problem, and finally the 2-covariate κ-FBS problem for $\kappa \geq 3$.

3.1. NP-Hardness for the κ-FBS Problem for Any Constant κ with $P \geq 3$

We show here that even for κ being constant, the κ-FBS problem with three or more covariates is NP-hard. This result excludes (under the assumption that $P \neq NP$) the possibility that there is a value of κ for which the κ-FBS problem (to this value of κ) is polynomial time solvable. The proof via reduction from the three-dimensional matching problem of this technical result is provided in Appendix B in the e-companion.

Theorem 2. *The κ-FBS problem is NP-hard when $p = 3$ for any constant κ.*

Corollary 1. *The κ-FBS problem is NP-hard for any integer $P \geq 3$, for any constant κ.*

For any constant integer κ, any 3-covariate κ-FBS problem instance, and any $p > 3$, we can construct an equivalent P-covariate κ-FBS problem instance as follows; for each sample of the given 3-covariate κ-FBS problem instance, we create a sample for the constructed κ-FBS problem instance such that they have the same level value for covariate $p = 1, 2, 3$. For $p = 4, \dots, P$, set covariate p to have only one level so that all samples in the constructed κ-FBS problem instance have the same value.

Therefore, the NP-hardness of the 3-covariate κ-FBS problem implies that the P-covariate κ-FBS problem is NP-hard for every value of P when $P \geq 3$. □

Because the κ-FBS problem is NP-hard for $P \geq 3$, there is no polynomial time algorithm unless $P = NP$.

In the following subsections, we will discuss the remaining case of the 2-covariate problems.

3.2. The Network Flow Algorithm for FBS with $p = 2$

In this subsection, we present an integer programming formulation with network flow constraints for the 2-covariate FBS problem. We then show how to solve the problem efficiently with a network flow algorithm.

It was noted in Theorem 1 that there is no differentiation between the individual samples selected in each level intersection, only the number of those selected counts. We thus define the decision variables as follows:

- x_{i_1,i_2}: the number of treatment samples selected from the (i_1, i_2) level intersection $L_{1,i_1} \cap L_{2,i_2}$ for $i_1 = 1, \dots, k_1$ and $i_2 = 1, \dots, k_2$;
- u'_{i_1,i_2}: the number of control samples selected from the (i_1, i_2) level intersection $L'_{1,i_1} \cap L'_{2,i_2}$ for $i_1 = 1, \dots, k_1$ and $i_2 = 1, \dots, k_2$.

Let $u_{i_1,i_2} = |L_{1,i_1} \cap L_{2,i_2}|$ and $u'_{i_1,i_2} = |L'_{1,i_1} \cap L'_{2,i_2}|$ for $i_1 = 1, \dots, k_1$, $i_2 = 1, \dots, k_2$. Clearly, x_{i_1,i_2} must be an integer between 0 and u_{i_1,i_2}, and x'_{i_1,i_2} must be an integer between 0 and u'_{i_1,i_2}. With these decision variables, the following is an integer programming formulation for the 2-covariate FBS problem:

(IP – FBS)

$$\max \quad \sum_{i_1=1}^{k_1} \sum_{i_2=1}^{k_2} x_{i_1,i_2} \tag{1a}$$

$$\text{s.t.} \quad \sum_{i_2=1}^{k_2} x_{i_1,i_2} - \sum_{i_2=1}^{k_2} x'_{i_1,i_2} = 0 \qquad i_1 = 1, \dots, k_1 \tag{1b}$$

$$\sum_{i_1=1}^{k_1} x_{i_1,i_2} - \sum_{i_1=1}^{k_1} x'_{i_1,i_2} = 0 \qquad i_2 = 1, \dots, k_2 \tag{1c}$$

$$0 \leq x_{i_1,i_2} \leq u_{i_1,i_2} \qquad i_1 = 1, \dots, k_1, \;\; i_2 = 1, \dots, k_2 \tag{1d}$$

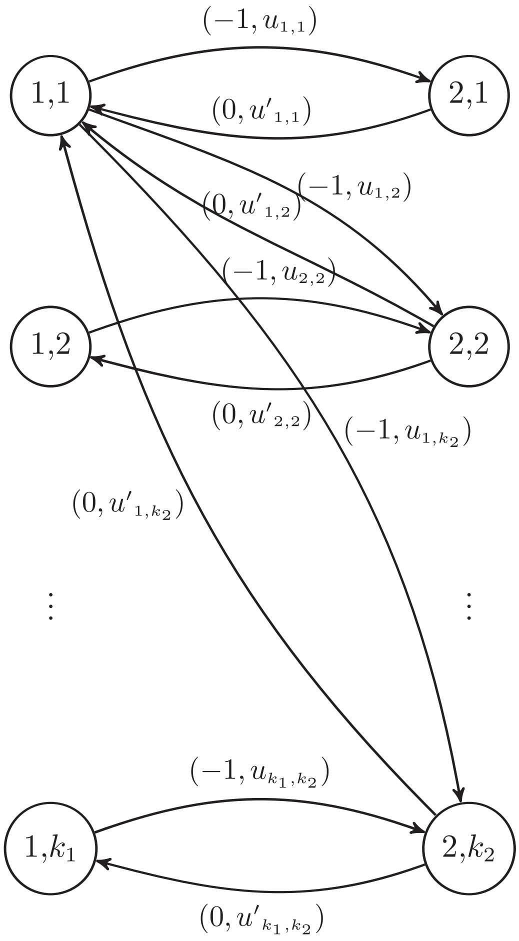
$$0 \leq x'_{i_1,i_2} \leq u'_{i_1,i_2} \qquad i_1 = 1, \dots, k_1, \;\; i_2 = 1, \dots, k_2 \tag{1e}$$

$$x_{i_1,i_2}, x'_{i_1,i_2} \text{ integers} \qquad i_1 = 1, \dots, k_1, \;\; i_2 = 1, \dots, k_2\,. \tag{1f}$$

The objective Equation (1a) is the total number of selected treatment samples. Constraints Equation (1b) are the fine balance requirement under covariate 1, because $\sum_{i_2=1}^{k_2} x_{i_1,i_2}$ equals the number of selected treatment samples in level i_1 under covariate 1 and $\sum_{i_2=1}^{k_2} x'_{i_1,i_2}$ equals the number of selected control samples in the same level. Similarly, constraints Equation (1c) are the fine balance requirement under covariate 2.

Formulation (IP-FBS) is in fact also a network flow formulation. In a minimum cost network flow formulation, each column of the constraint matrix corresponding to a variable that is a flow along an arc has exactly one 1 and one –1. The corresponding MCNF network is shown in Figure 1, where all capacity lower bounds are 0, and each arc has a cost per unit flow and upper bound associated with it. The flow on the arc from node $(1, i_1)$ to node $(2, i_2)$ represents variable x_{i_1,i_2}, which is bounded between 0 and u_{i_1,i_2}, as stated in constraints Equation (1d); arc from node $(2, i_2)$ to node $(1, i_1)$ represents variable x'_{i_1,i_2}, which is bounded between 0 and u'_{i_1,i_2}, as stated in constraints

Figure 1. Min-Cost Network Flow Graph Corresponding to Formulation (IP-FBS)



Note. Arc legend: (cost, upperbound).

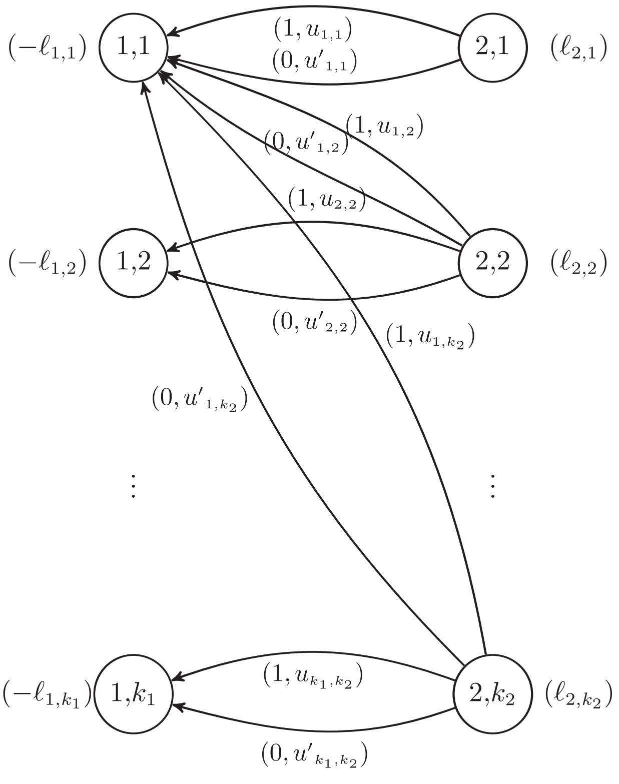
Equation (1e). To get a "minimize" type objective, we take the negative value of $|S| = \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2} x_{i_1,i_2}$ as the objective, so the per unit arc cost should be –1 for arcs from any node in $\{(1,1),(1,2),\ldots,(1,k_1)\}$ to any node in $\{(2,1),(2,2),\ldots,(2,k_2)\}$. All other arcs have cost 0. It is easy to verify that constraints Equation (1b) are corresponding to the flow balance at nodes $(1,i_1)$ for all i_1, and constraints Equation (1c) are corresponding to the flow balance at nodes $(2,i_2)$ for all i_2.

Theorem 3. *The 2-covariate FBS problem is solved as a minimum cost network flow problem in* $O(n\cdot(\min\{n+n',k_1k_2\}+(k_1+k_2)\log(k_1+k_2))$ *time.*

To solve the minimum cost network flow problem of the 2-covariate FBS problem, we choose the algorithm of *successive shortest paths* that is particularly efficient for an MCNF with "small" total arc capacity (see Ahuja et al. (1993), section 9.7). The successive shortest paths algorithm starts with a network graph with no negative cycles, so we first modify the network shown in Figure 1 using a well-known arc reversal transformation in section 2.4 of Ahuja et al. (1993). The resulting network graph is shown in Figure 2.

The successive shortest path algorithm iteratively selects a node s with excess supply (supply not yet sent to some demand node) and a node t with unfulfilled demand and sends flow from s to t along the shortest path in the residual network (Jewell 1958, Iri 1960, Busaker and Gowen 1961). The algorithm terminates when the flow satisfies all the flow balance constraints. Because at each iteration the number of remaining units of supply to be sent is reduced by at least one unit, the number of iterations is bounded by the total amount of supply. For the network in Figure 2, the total supply is n.

Figure 2. Min-Cost Network Flow Graph After Arc Reversal



Note. Arc legend: (cost, upperbound); node legend: (supply).

At each iteration, the shortest path can be solved with Dijkstra's algorithm of complexity $O(|A|+|V|\log|V|)$, where $|V|$ is number nodes and $|A|$ is number of arcs (Tomizawa 1971, Edmonds and Karp 1972). In our formulation, $|V|$ is $O(k_1+k_2)$, which is at most $O(n)$. Because the number of nonempty sets $L_{1,i_1}\cap L_{2,i_2}$ is at most $\min\{n,k_1k_2\}$, the number of unit-cost arcs is $O(\min\{n,k_1k_2\})$. Because the number of nonempty sets $L'_{1,i_1}\cap L'_{2,i_2}$ is at most $\min\{n',k_1k_2\}$, the number of zero-cost arcs is $O(\min\{n',k_1k_2\})$. So the total number of arcs $|A|$ is $O(\min\{n+n',k_1k_2\})$.

Hence, the total running time of applying the successive shortest path algorithm on our formulation is $O(n\cdot(\min\{n+n',k_1k_2\}+(k_1+k_2)\log(k_1+k_2))$. □

In contrast to the 2-covariate FBS problem, which is polynomial time solvable, we show next that the 2-covariate κ-FBS problem is NP-hard when $\kappa \geq 3$.

3.3. NP-Hardness For the 2-Covariate κ-FBS Problem with $\kappa \geq 3$

We prove that the 2-covariate κ-FBS problem is NP-hard for all constant values of κ such that $\kappa \geq 3$. The proof via reduction from the exact 3-cover problem is given in Appendix B in the e-companion.

Therorem 4. *The 2-covariate κ-FBS problem is NP-hard for any constant* $\kappa \geq 3$.

4. The κ-BBS and the κ-MBBS Problems

In this section, we show the complexity and algorithmic results for the κ-BBS problems. We also show that these results hold also for κ-MBBS. Because the κ-FBS problem is a special case of κ-BBS problem with all bounds being 0 (and also of κ-MBBS, because if all bounds are 0 then we also have $|S'| = \kappa\,|S|$), we can infer from the last section that the κ-BBS and κ-MBBS problems are also NP-hard for each of the following cases:

- The number of covariates is at least 3;
- $\kappa \geq 3$.

Next, we present a polynomial time algorithm for the 2-covariate BBS and the 2-covariate MBBS problems.

4.1. The Network Flow Algorithm for BBS and MBBS with $p = 2$

In this subsection, we present an integer programming formulation with network flow constraints for the 2-covariate BBS problem and a similar formulation for the 2-covariate MBBS problem. We then show how to solve the problem efficiently with a network flow algorithm.

We use the decision variables to indicate the number of selected samples from the two groups:

- x_{i_1,i_2}: the number of treatment samples selected from the (i_1, i_2) level intersection $L_{1,i_1} \cap L_{2,i_2}$ for $i_1 = 1,\dots,k_1$ and $i_2 = 1,\dots,k_2$;
- x'_{i_1,i_2}: the number of control samples selected from the (i_1, i_2) level intersection $L'_{1,i_1} \cap L'_{2,i_2}$ for $i_1 = 1,\dots,k_1$ and $i_2 = 1,\dots,k_2$. Additionally, we introduce variables that represent the deficits and excesses:
- $d_{p,i}$: the deficit corresponding to level i under covariate p for $p \in \{1,2\}$ and for $i = 1,\dots,k_p$;
- $e_{p,i}$: the excess corresponding to level i under covariate p for $p \in \{1,2\}$ and for $i = 1,\dots,k_p$.

With these decision variables the following is an integer programming formulation for the 2-covariate BBS problem:

(IP − BBS)

$$\max \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2} x_{i_1,i_2} \tag{2a}$$

$$\text{s.t.} \sum_{i_2=1}^{k_2} x_{i_1,i_2} - \sum_{i_2=1}^{k_2} x'_{i_1,i_2} + e_{1,i_1} - d_{1,i_1} = 0 \qquad i_1 = 1,\dots,k_1 \tag{2b}$$

$$\sum_{i_1=1}^{k_1} x_{i_1,i_2} - \sum_{i_1=1}^{k_1} x'_{i_1,i_2} + e_{2,i_2} - d_{2,i_2} = 0 \qquad i_2 = 1,\dots,k_2 \tag{2c}$$

$$0 \le x_{i_1,i_2} \le u_{i_1,i_2} \qquad i_1 = 1,\dots,k_1, \quad i_2 = 1,\dots,k_2 \tag{2d}$$

$$0 \le x'_{i_1,i_2} \le u'_{i_1,i_2} \qquad i_1 = 1,\dots,k_1, \quad i_2 = 1,\dots,k_2 \tag{2e}$$

$$0 \le d_{p,i} \le B^{(d)}_{p,i} \qquad p \in \{1,2\}, i = 1,\dots,k_p \tag{2f}$$

$$0 \le e_{p,i} \le B^{(e)}_{p,i} \qquad p \in \{1,2\}, i = 1,\dots,k_p \tag{2g}$$

$$x_{i_1,i_2}, x'_{i_1,i_2} \text{integers} \qquad i_1 = 1,\dots,k_1, \quad i_2 = 1,\dots,k_2. \tag{2h}$$

In order to obtain an integer programming formulation for the 2-covariate MBBS problem, we add the constraint

$$\sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2} x_{i_1,i_2} = \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2} x'_{i_1,i_2}.$$

We denote by (IP-MBBS) the resulting integer program.

The two formulations (IP-BBS) and (IP-MBBS) are in fact also network flow formulations. The corresponding MCNF network for (IP-MBBS) is shown in Figure 3. Here, all capacity lower bounds are 0, and each arc has a cost per unit flow and upper bound associated with it. The flow on the arc from node $(1,i_1)$ to node $(2,i_2)$ represents variable x_{i_1,i_2}, which is bounded between 0 and u_{i_1,i_2} as stated in constraints Equation (2d); arc from node $(2,i_2)$ to node $(1,i_1)$ represents variable x'_{i_1,i_2}, which is bounded between 0 and u'_{i_1,i_2} as stated in constraints Equation (2e). To get a "minimize" type objective, we take the negative value of $|S| = \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2} x_{i_1,i_2}$ as the objective, so the per unit arc cost should be −1 for arcs from any node in $\{(1,1),(1,2),\dots,(1,k_1)\}$ to any node in $\{(2,1),(2,2),\dots,(2,k_2)\}$. All other arcs have cost 0. There are two additional nodes 1, 2 (in addition to the nodes $(1,1),(1,2),\dots,(1,k_1),(2,1),(2,2),\dots,(2,k_2)$). The node $p = 1$, 2 is connected to the nodes $(p,1)$, $(p,2),\dots,(p,k_p)$ via edges corresponding to the deficit and excess of the corresponding level of covariate p. It is easy to verify that constraints Equation (2b) correspond to the flow balance at nodes $(1,i_1)$ for all i_1, and constraints Equation (2c) correspond to the flow balance at nodes $(2,i_2)$ for all i_2. Moreover, by summing up constrains Equation (2b), we get that the total deficit corresponding to covariate 1 should equal the total excess corresponding to covariate 1, and thus node 1 in the network should have 0 supply/demand. Similar argument applies to node 2 as well.

The network flow graph for (IP-MBBS) is modified slightly for the 2-covariate BBS problem (IP-BBS). We add two arcs, from node 2 to node 1 and from node 1 to node 2. Both of these arcs have infinite capacity upper bound, zero capacity lower bound, and zero cost. In addition, all supplies/demands of the nodes are set to 0. We these adjustments, the minimum cost flow on the network described in Figure 3 provides the optimal solution to (IP-BBS).

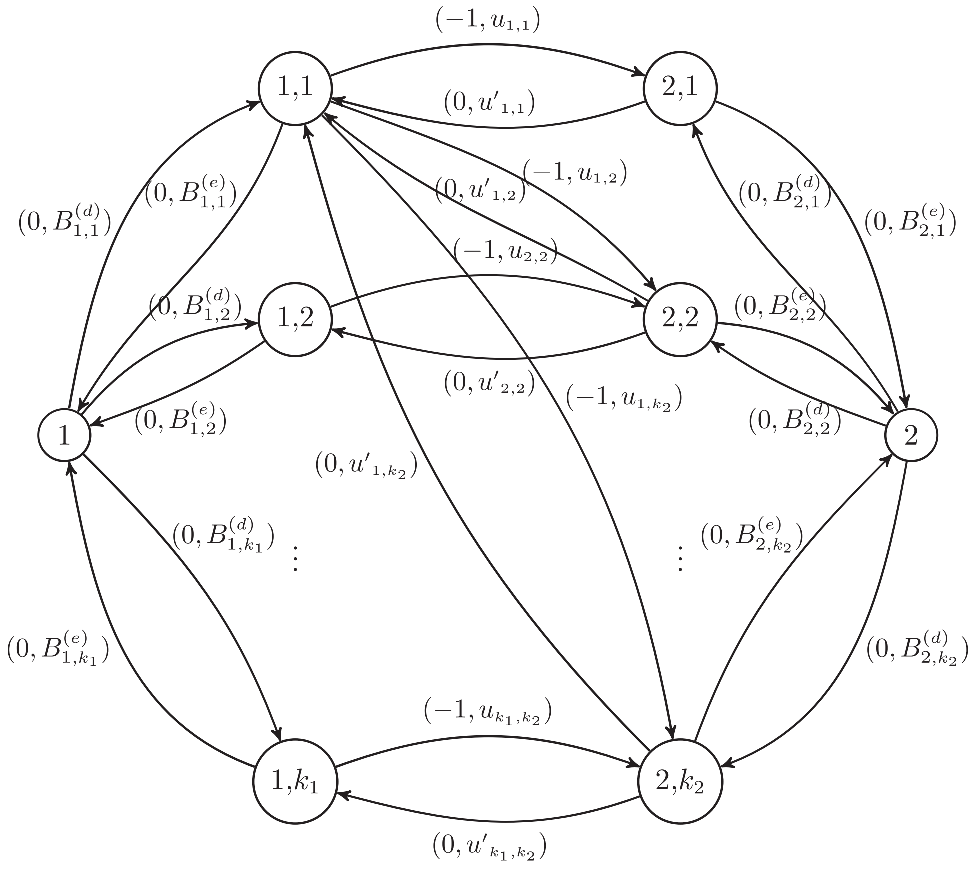
Using the same arguments as in the proof of Theorem 3, we conclude with the following theorem.

Theorem 5. *The 2-covariate BBS problem and the 2-covariate MBBS problem are solved as a minimum cost network flow problem in $O(n \cdot (\min\{n+n', k_1k_2\} + (k_1+k_2)\log(k_1+k_2))$ time.*

5. The 2-Covariate κ-Fine-Balanced-Matching (κ-BM) Problem

The 1-covariate κ-BM problem is solvable in polynomial time (Rosenbaum et al. 2007). However, the BM problem is already NP-hard for two or more covariates (Sauppe et al. 2014). For the 2-covariate BM problem and the 2-covariate κ-BM problem, the complexity status when the numbers of levels of both covariates are constants (or upper bounded by a slowly growing function of n) is discussed in Section 7 together with the other three families of problems.

Figure 3. Min-Cost Network Flow Graph Corresponding to Formulation (IP-MBBS)



Note. Arc legend: (cost, upperbound).

Here, we consider an intermediate case where only one of the covariates has a constant number of levels. We will show here that the 2-covariate BM problem and the 2-covariate κ-BM problem where one of the covariates, say the second covariate, has a constant number of levels can be solved efficiently if and only if the exact matching problem on bipartite graphs can be solved efficiently.

Let LBM be the special case of the 2-covariate BM problem, where the second covariate has a constant number of levels, whereas the first covariate has no restriction on the number of levels. In Section 7, we will establish that if both covariates have constant number of levels, then the 2-covariate BM problem is polynomial time solvable. We show here that the complexity status of the 2-covariate problem in which only one covariate has a constant number of levels is linked to the complexity status of the *exact matching problem* and its weighted version denoted as *weighted exact matching*. In order to present this connection, we assume that the distance matrix is integral, and all distances are given in unary; that is, there is a polynomial π in the variable denoting the input encoding length where $\delta_{ij} \leq \pi$ for all *i*, *j*.

The exact matching in bipartite graph problem is defined as follows.

5.1. Exact Matching

Given an integer number k together with a bipartite graph, $G = (V_1 \cup V_2, E)$ with $|V_1| = |V_2| = q$, such that the edge set E is partitioned into $E_b \cup E_r$, where E_b is the set of blue edges and E_r is the set of red edges. Find a perfect matching that has exactly k blue edges (and all other $q - k$ edges are red).

The complexity status of the exact matching problem is as follows. Whereas Mulmuley et al. (1987) showed that there is a randomized polynomial time algorithm for the problem, the existence of a deterministic polynomial time algorithm is still an important open problem. The correctness of the algorithm of Mulmuley et al. (1987) follows by using their isolating lemma, and the algorithm itself is based on computing the square root of the determinant of a matrix that is randomly obtained from the input graph G with the partition of the edges into blue and red.

The weighted exact matching problem is defined as follows.

5.2. Weighted Exact Matching

Given a target value K and a bipartite graph $G = (V, E)$ together with nonnegative integral distances δ_e for all $e \in E$, there is a polynomial π in the variable that equals $|V| + |E|$ such that $\delta_e \le \pi$ for all $e \in E$. Find a perfect matching of total distance exactly K.

Note that the weighted exact matching problem is a generalization of the exact matching problem, because the later problem can be interpreted as the weighted exact matching problem where the weight of a blue edge is 1 and the weight of a red edge is 0. Thus, a polynomial time algorithm for the weighted exact matching problem gives a polynomial time algorithm for the exact matching problem. On the other hand, it is known that a polynomial time algorithm for the exact matching problem gives a polynomial time algorithm for the weighted exact matching problem (see proposition 1 in Papadimitriou and Yannakakis 1982). If the algorithm for the exact matching is deterministic (randomized), then the algorithm for the weighted exact matching is deterministic (randomized, respectively) as well (Papadimitriou and Yannakakis 1982). Therefore, the complexity of the weighted problem has the same status as the one of the exact matching problems. Namely, the result of Mulmuley et al. (1987) gives a randomized polynomial time algorithm for the weighted exact matching problem, whereas the existence of a deterministic polynomial time algorithm for this problem will result in a deterministic polynomial time algorithm for the exact matching in bipartite graphs problem.

We show in Appendix C in the e-companion the following connections between the exact matching problem (or the weighted exact matching) and problem LBM.

Theorem 6. *If there is a deterministic (or randomized) polynomial time algorithm for LBM, then there is a deterministic (or randomized, respectively) polynomial time algorithm for exact matching in bipartite graphs.*

Theorem 7 . *If there is a deterministic (or randomized) polynomial time algorithm for weighted exact matching in bipartite graphs, then there is a deterministic (or randomized, respectively) polynomial time algorithm for LBM.*

Next, we consider the κ-LBM that is the special case of 2-covariate κ-BM, where the second covariate has a constant number of levels, and once again we assume that the distance matrix is integral and the maximum distance is upper bounded by a polynomial π in the variable that equals the input encoding length. We show in Appendix C that κ-LBM has the same complexity status as LBM (for all $\kappa \ge 2$). That is, we establish the following result.

Theorem 8. *There is a polynomial time algorithm for LBM if and only if there is a polynomial time algorithm for κ-LBM.*

6. The Maximum Selection κ-Fine-Balance Matching (κ-MSBM) Problem

Because any κ-BM problem can be solved as a κ-MSBM problem with the same κ and the same number of covariates, we can infer that the P-covariate κ-MSBM problem is also NP-hard for $P \ge 2$ and any constant κ. And in Section 2, we show that the 1-covariate MSBM problem can be solved as an MCNF problem in polynomial time. We show in Appendix B in the e-companion that the 1-covariate κ-MSBM problem is NP-hard even for any constant $\kappa \ge 3$ by reduction from the Exact-3-cover problem.

Theorem 9. *For any constant value of κ such that $\kappa \ge 3$, the 1-covariate κ-MSBM problem is NP-hard even with only one level.*

For the 1-covariate κ-MSBM problem where $\kappa = 2$, the complexity status remains open.

7. Fixed-Parameter Tractable Algorithms

In this section, we consider the special cases of the κ-FBS, κ-BBS, κ-BM, and MSBM problems, where all covariates have a small number of levels. Let $K = \prod_{i=1}^{P} k_i$ be the number of level intersections. Observe that if the number of covariates is constant and all covariates have a constant number of levels, then K is a constant. In this section, we establish that the problems κ-FBS, κ-BBS, κ-BM, and MSBM can be solved in fixed-parameter tractable (FPT) time with parameter K. In order to state these results, we say that a problem is fixed-parameterized complexity with parameter K and denote it by $FPT(K)$ if it has an algorithm whose time complexity is upper bounded by a function of the form $f(K) \cdot \texttt{poly}$, where $f(K)$ is some computable function of the parameter K and $\texttt{poly}$ is some polynomial in the variable that equals the input binary encoding length. We also say that an algorithm runs in $FPT(K)$ time and mean that its time complexity can be upper bounded by a function of the form $f(K) \cdot \texttt{poly}$, where $f(K)$ is some computable function of the parameter K and $\texttt{poly}$ is some polynomial in the variable that equals the input binary encoding length. Here, we show that these problems, namely κ-FBS, κ-BBS, κ-BM problems for all κ, and MSBM problem, are $FPT(K)$. In particular, our results imply that these problems are polynomial time solvable if K is a constant, but they also provide such polynomial time algorithms for some superconstant values of K like $K = O(\log n / \log\log n)$. As shown in Theorem 9, unless P=NP we cannot obtain similar results for κ-MSBM where $\kappa \ge 3$. The complexity status of the 2-MSBM problem with constant K is open.

Our proof for the $FPT(K)$ results uses the existence of fast algorithms for solving integer programming in fixed dimension and for solving mixed-integer linear programs if the number of integral variables is fixed. In Lenstra Jr. (1983) (also see Kannan 1983 for an improved time complexity of these algorithms), it is shown that the integer linear programming problem with a fixed number of variables is polynomially solvable; Lenstra Jr. (1983) also showed that a mixed-integer linear program with a fixed number of integer variables can be solved in polynomial time. In fact, these algorithms run in *FPT* time, with parameter being the number of integral variables. Therefore, to prove our results, we show either an integer programming (IP) formulation with the number of decision variables $O(K)$ or a mixed-integer linear program (MILP) with $O(K)$ integer variables such that solving this MILP to optimality ensures that the resulting solution is integral and solves the corresponding problem.

7.1. The κ-FBS Problem

First consider the κ-FBS problem. For this problem, we use an integer program with dimension $O(K)$ that is based on (IP-FBS). Let $u_{i_1,i_2,\ldots,i_P} = |L_{1,i_1} \cap L_{2,i_2} \cap \ldots \cap L_{P,i_P}|$ and $u'_{i_1,i_2,\ldots,i_P} = |L'_{1,i_1} \cap L'_{2,i_2} \cap \ldots \cap L'_{P,i_P}|$ for $i_p = 1, \ldots, k_p$, $p = 1, \ldots, P$. The decision variables are as follows:

$x_{i_1,i_2,\ldots,i_P}$: the number of treatment samples selected from the $(i_1, i_2, \ldots, i_P)$ level intersection $L_{1,i_1} \cap L_{2,i_2} \cap \ldots \cap L_{P,i_P}$ for $i_p = 1, \ldots, k_p$, $p = 1, \ldots, P$;

$x'_{i_1,i_2,\ldots,i_P}$: the number of control samples selected from the $(i_1, i_2, \ldots, i_P)$ level intersection $L'_{1,i_1} \cap L'_{2,i_2} \cap \ldots \cap L'_{P,i_P}$ for $i_p = 1, \ldots, k_p$, $p = 1, \ldots, P$.

The integer programming formulation is as follows:

$$\max \quad \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\cdots\sum_{i_P=1}^{k_P} x_{i_1,i_2,\ldots,i_P} \tag{3a}$$

$$\text{s.t.} \quad \kappa \cdot \sum_{i_1=1}^{k_1}\cdots\sum_{i_{p-1}=1}^{k_{p-1}}\sum_{i_{p+1}=1}^{k_{p+1}}\cdots\sum_{i_P=1}^{k_P} x_{i_1,i_2,\ldots,i_P}$$

$$= \sum_{i_1=1}^{k_1}\cdots\sum_{i_{p-1}=1}^{k_{p-1}}\sum_{i_{p+1}=1}^{k_{p+1}}\cdots\sum_{i_P=1}^{k_P} x'{}_{i_1,i_2,\ldots,i_P} \qquad p = 1, \ldots, P \;\; i_p = 1, \ldots, k_p \tag{3b}$$

$$0 \le x_{i_1,i_2,\ldots,i_P} \le u_{i_1,i_2,\ldots,i_P} \quad p = 1, \ldots, P, \quad i_p = 1, \ldots, k_p \tag{3c}$$

$$0 \le x'_{i_1,i_2,\ldots,i_P} \le u'_{i_1,i_2,\ldots,i_P} \quad p = 1, \ldots, P, \quad i_p = 1, \ldots, k_p \tag{3d}$$

$$x_{i_1,i_2,\ldots,i_P}, x'_{i_1,i_2,\ldots,i_P} \text{ integers} \quad p = 1, \ldots, P, \quad i_p = 1, \ldots, k_p. \tag{3e}$$

Note that this integer programming formulation has $2K$ decision variables and $O(K)$ constraints, and thus the algorithm that constructs it and solves it to optimality runs in $FPT(K)$ time. The optimal solution for this integer program encodes the optimal solution for κ-FBS similarly to the proof of Theorem 1.

7.2. The κ-BBS Problem and the κ-MBBS Problem

We introduce an integer programming formulation for the problem based on (IP-BBS) with $O(K)$ decision variables and $O(K)$ constraints. We use the decision variables to indicate the number of selected samples from the two groups:

$x_{i_1,i_2,\ldots,i_P}$: the number of treatment samples selected from the $(i_1, i_2, \ldots, i_P)$ level intersection $L_{1,i_1} \cap L_{2,i_2} \cap \cdots \cap L_{P,i_P}$ for $i_p = 1, \ldots, k_p$ and for $p = 1, \ldots, P$;

$x'_{i_1,i_2,\ldots,i_P}$: the number of control samples selected from the $(i_1, i_2, \ldots, i_P)$ level intersection $L_{1,i_1} \cap L_{2,i_2} \cap \cdots \cap L_{P,i_P}$ for $i_p = 1, \ldots, k_p$ and for $p = 1, \ldots, P$.

Additionally, we use variables that represent the deficits and excesses:

$d_{p,i}$: the deficit corresponding to level i under covariate p for $p \in \{1, 2, \ldots, P\}$ and for $i = 1, \ldots, k_p$;

$e_{p,i}$: the excess corresponding to level i under covariate p for $p \in \{1, 2, \ldots, P\}$ and for $i = 1, \ldots, k_p$.

With these decision variables, the following is an integer programming formulation for the κ-BBS problem:

$$\max \quad \sum_{i_1=1}^{k_1}\sum_{i_2=1}^{k_2}\cdots\sum_{i_P=1}^{k_P} x_{i_1,i_2,\ldots,i_P}$$

$$\text{s.t.} \quad \kappa \cdot \sum_{i_1=1}^{k_1}\cdots\sum_{i_{p-1}=1}^{k_{p-1}}\sum_{i_{p+1}=1}^{k_{p+1}}\cdots\sum_{i_P=1}^{k_P} x_{i_1,i_2,\ldots,i_P} -$$

$$-\sum_{i_1=1}^{k_1}\cdots\sum_{i_{p-1}=1}^{k_{p-1}}\sum_{i_{p+1}=1}^{k_{p+1}}\cdots\sum_{i_P=1}^{k_P} x'_{i_1,i_2,\ldots,i_P} +$$

$$+ e_{p,i_p} - d_{p,i_p} = 0 \qquad p = 1, \ldots, P \;\; i_p = 1, \ldots, k_p$$

$$0 \le x_{i_1,i_2,\ldots,i_P} \le u_{i_1,i_2,\ldots,i_P} \qquad p = 1, \ldots, P \;\; i_p = 1, \ldots, k_p$$

$$0 \le x'_{i_1,i_2,\ldots,i_P} \le u'_{i_1,i_2,\ldots,i_P} \qquad p = 1, \ldots, P \;\; i_p = 1, \ldots, k_p$$

$$0 \le d_{p,i} \le B^{(d)}_{p,i} \qquad p \in \{1, 2, \ldots, P\}, \quad i = 1, \ldots, k_p$$

$$0 \le e_{p,i} \le B^{(e)}_{p,i} \qquad p \in \{1, 2, \ldots, P\}, \quad i = 1, \ldots, k_p$$

$$x_{i_1,i_2,\ldots,i_P}, x'_{i_1,i_2,\ldots,i_P} \text{integers} \qquad p = 1, \ldots, P \quad i_p = 1, \ldots, k_p$$

$$d_{p,i}, e_{p,i} \text{integers} \qquad p \in \{1, 2, \ldots, P\}, \quad i = 1, \ldots, k_p$$

Because this integer programming formulation has $O(K)$ decision variables and $O(K)$ constraints, it can be solved to optimality in $FPT(K)$ time. For the κ-MBBS problem, we use the fact that it is a special case of κ-BBS problem with one additional covariate, but this transformation does not change the value of K, so the same result holds for κ-MBBS problem as well.

7.3. The κ-BM Problem

Next, consider the κ-BM problem. In Section 2, we describe an MCNF formulation when the level intersection sizes $s'_{i_1,i_2,\ldots,i_P}$ for $p=1,\ldots,P$ and $i_p=1,\ldots,k_p$ are given. Observe that if we treat the sizes $s'_{i_1,i_2,\ldots,i_P}$ for all p and i_p as decision variables, then by enforcing the integrality of these K variables and adding the constraints saying that

$$\sum_{i_1=1}^{k_1}\cdots\sum_{i_{p-1}=1}^{k_{p-1}}\sum_{i_{p+1}=1}^{k_{p+1}}\cdots\sum_{i_P=1}^{k_P} s'_{i_1,i_2,\ldots,i_P}=\kappa\cdot\ell_{p,i_p},$$

$$i_p=1,\ldots,k_p \quad p=1,\ldots,P,$$

forcing the κ-fine balance constraints to the MCNF formulation, we get a MILP formulation of κ-BM with K integral variables. In fact, if we restrict ourselves to common integral values of these K variables, then the other decision variables are integral, as we argue next. By considering the values of these K integral variables as constants, the resulting linear programming formulation is in fact an MCNF LP formulation whose supply/demand vector depends on the values of these K integral variables. Thus, the optimal solution for the MILP is without loss of generality integral, and even if it does not satisfy this integral requirement, it can be transformed to another optimal solution that is integral in polynomial time.

Because the number of variables of the resulting mixed-integer program is at most $n\cdot n'+K$, the number of integer variables is K, and the number of constraints is $O(n\cdot n')$, we conclude that the algorithm that formulates this MILP and solves it to optimality, guaranteeing that the optimal solution is integral, runs in $FPT(K)$ time.

7.4. The MSBM and κ-MSBM Problems

We know from Theorem 9 that the 1-covariate κ-MSBM problem for $\kappa\geq 3$ is NP-hard already if the unique covariate has only one level.

We consider next the MSBM problem. In Section 2, we describe an MCNF formulation if all of the level intersection sizes $s_{i_1,i_2,\ldots,i_P}$ and $s'_{i_1,i_2,\ldots,i_P}$ are given. Observe that if we treat $s_{i_1,i_2,\ldots,i_P}$ and $s'_{i_1,i_2,\ldots,i_P}$ as decision variables, then by enforcing the integrality of these $O(K)$ variables and adding the constraints saying that

$$\sum_{i_1=1}^{k_1}\cdots\sum_{i_{p-1}=1}^{k_{p-1}}\sum_{i_{p+1}=1}^{k_{p+1}}\cdots\sum_{i_P=1}^{k_P} s_{i_1,i_2,\ldots,i_P}$$
$$=\sum_{i_1=1}^{k_1}\cdots\sum_{i_{p-1}=1}^{k_{p-1}}\sum_{i_{p+1}=1}^{k_{p+1}}\cdots\sum_{i_P=1}^{k_P} s'_{i_1,i_2,\ldots,i_P},$$

$$i_p=1,\ldots,k_p\ p=1,\ldots,P,$$

That is, the fine balance constraints, in addition to the constraint saying that the sum over all $s_{i_1,i_2,\ldots,i_P}$ equals the objective function value of FBS, to the MCNF formulation, we get a MILP formulation of MSBM with $2K$ integral variables. In fact, if we restrict ourselves to common integral values of these $2K$ variables, then the other decision variables are without loss of generality integral as well, as we argue next. By considering the values of these $2K$ integral variables as constants, the resulting linear programming formulation is in fact an MCNF LP formulation whose supply/demand vector depends on the values of these $2K$ integral variables. Thus, the optimal solution for the MILP is without loss of generality integral, and even if it does not satisfy this integral requirement, it can be transformed to another optimal solution that is integral in polynomial time.

Because the number of variables of the resulting mixed-integer program is at most $n\cdot n'+2K$, the number of integer variables is $2K$, and the number of constraints is $O(n\cdot n')$, we conclude that the algorithm that formulates this MILP and solves it to optimality, guaranteeing that the optimal solution is integral, runs in $FPT(K)$ time. The existence of such an algorithm implies that MSBM is solvable in polynomial time for a fixed number of level intersections.

Hence, we proved that MSBM is fixed-parameter tractable; the fixed-parameter tractability of κ-MSBM for $\kappa\geq 3$ is NP-hard, and the fixed parameter tractability of 2-MSBM is open.

8. Conclusions

This paper presents a comprehensive complexity study of several problems related to covariate balancing. For the problems of fine balance selection and bounded balance selection, for two covariates, these problems of maximizing the size of the treatment selection subject to the fine balance constraint and the bounded balance constraints, respectively, are both polynomial time solvable with network flow. The respective two-covariate problems, with κ factor selection of the control samples, are hard for $\kappa\geq 3$ and open for $\kappa=2$. These problems and the other problems studied here are NP-hard for three or more covariates. We further show that the problems with fixed numbers of covariates and levels are fixed-parameter tractable, including for general κ, except for κ-MSBM, which is fixed-parameter tractable only for $\kappa=1$ and NP-hard for $\kappa\geq 3$. The practical implications of these complexity results are that for a small number of level intersections and mostly for 2-covariates, the problems can be efficiently solved. These facts can be used in relaxations that aggregate the level intersections to a small number or aggregate covariates to two representative covariates. However, the possible use of such relaxations depends on additional aspects that are beyond the scope of our study because they should be based on strong statistical justification. Such justification need not exist in all cases, but we believe that our work will initiate studies concerning the question of under what cases such procedures have sound statistical justifications.

References

Ágoston KC, Biró P, Szántó R (2018) Stable project allocation under distributional constraints. *Oper. Res. Perspect.* 5:59–68.

Ahuja RK, Orlin JB, Magnanti TL (1993) *Network Flows: Theory, Algorithms, and Applications.* (Prentice-Hall, Hoboken, NJ).

Ashlagi I, Saberi A, Shameli A (2020) Assignment mechanisms under distributional constraints. *Oper. Res.* 68(2):467–479.

Bei X, Liu S, Poon CK, Wang H (2020) Candidate selections with proportional fairness constraints. *Proc. 19th Internat. Conf. Autonomous Agents and MultiAgent Systems*, 150–158.

Bennett M, Vielma JP, Zubizarreta JR (2020) Building representative matched samples with multi-valued treatments in large observational studies. *J. Comput. Graph. Stat.* 29(4):744–757.

Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T (2006) Variable selection for propensity score models. *Am. J. Epidemiol.* 163(12):1149–1156.

Busaker R, Gowen PJ (1961) A procedure for determining minimal-cost network flow patterns. Technical report, ORO Technical Report 15, Operational Research Office, John Hopkins University.

Dutta S, Jacobson SH, Sauppe JJ (2017) Identifying NCAA Tournament upsets using balance optimization subset selection. *J. Quant. Anal. Sports.* 13(2):79–93.

Edmonds J, Karp RM (1972) Theoretical improvements in algorithmic efficiency for network flow problems. *J. ACM* 19(2):248–264.

Ho DE, Imai K, King G, Stuart EA (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit. Anal.* 15(3):199–236.

Hochbaum DS, Rao X, Sauppe J (2022) Network flow methods for the minimum covariate imbalance problem. *Eur. J. Oper. Res.* 300(3): 827–836.

Imbens GW (2004) Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* 86(1):4–29.

Iri M (1960) A new method of solving transportation-network problems. *J. Oper. Res. Soc. Japan.* 3(1):2–87.

Jewell WS (1958) Optimal flow through networks. *Oper. Res.* 6:633–633.

Kannan R (1983) Improved algorithms for integer programming and related lattice problems. Johnson DS, Fagin R, Fredman ML, Harel D, Karp RM, Lynch NA, Papadimitriou CH, Rivest RL, Ruzzo WL, Seiferas JI, eds., *Proc. 15th Annual ACM Sympos. on Theory of Comput., 25-27 April, 1983, Boston, Massachusetts, USA*, 193–206 (ACM), http://dx.doi.org/10.1145/800061.808749.

Karmakar B, Small D, Rosenbaum P (2019) Using approximation algorithms to build evidence factors and related designs for observational studies. *J. Comput. Graph. Statist.* 28(3):698–709.

King G, Lucas C, Nielsen RA (2017) The balance-sample size frontier in matching methods for causal inference. *Am. J. Pol. Sci.* 61:473–489.

Kwon HY (2018) New developments in causal inference using balance optimization subset selection. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Kwon HY, Sauppe JJ, Jacobson SH (2019a) Bias in balance optimization subset selection: Exploration through examples. *J. Oper. Res. Soc.* 70(1):67–80.

Kwon HY, Sauppe JJ, Jacobson SH (2019b) Treatment effect decomposition and bootstrap hypothesis testing in observational studies. *Annals of Data Science* 6(3):491–511.

Kwon HY, Sauppe JJ, Jacobson SH (2020) Duality in balance optimization subset selection. *Ann. Oper. Res.* 289:277–289.

Lenstra Jr. HW (1983) Integer programming with a fixed number of variables. *Math. Oper. Res.* 8(4):538–548.

Morgan SL, Harding DJ (2006) Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociol. Methods Res.* 35(1):3–60.

Mulmuley K, Vazirani U, Vazirani V (1987) Matching is as easy as matrix inversion. *Combinatorica.* 7(1):105–113.

Nguyen T, Vohra R (2019) Stable matching with proportionality constraints. *Oper. Res.* 67(6):1503–1519.

Nguyen T, Nguyen H, Teytelboym A (2019) Stability in matching markets with complex constraints. *Management Sci.* 67(12):7291–7950.

Nikolaev AG, Jacobson SH, Cho WKT, Sauppe JJ, Sewell EC (2013) Balance optimization subset selection (boss): An alternative approach for causal inference with observational data. *Oper. Res.* 61(2):398–412.

Papadimitriou CH, Yannakakis M (1982) The complexity of restricted spanning tree problems. *J. ACM.* 29(2):285–309.

Pimentel SD, Kelz RR, Silber JH, Rosenbaum PR (2015) Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *J. Am. Stat. Assoc.* 110 (510):515–527.

Rosenbaum PR (2002) *Overt Bias in Observational Studies.* (Springer, New York), 71–104.

Rosenbaum PR (2012) Optimal matching of an optimally chosen subset in observational studies. *J. Comput. Graph. Statist.* 21: 57–71.

Rosenbaum PR (2020) Modern algorithms for matching in observational studies. *Annu. Rev. Stat. Appl.* 7:143–176.

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

Rubin DB, Stuart EA (2006) Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *Ann. Statist.* 34(4):1814–1826.

Sauppe JJ (2015) Balance optimization subset selection: A framework for causal inference with observational data. Ph.D. thesis, University of Illinois at Urbana-Champaign.

Sauppe JJ, Jacobson SH (2017) The role of covariate balance in observational studies. *Naval Res. Logist.* 64(4):323–344.

Sauppe JJ, Jacobson SH, Sewell EC (2014) Complexity and approximation results for the balance optimization subset selection model for causal inference in observational studies. *INFORMS J. Comput.* 26(3):547–566.

Sharma D, Willy C, Bischoff J (2020) Optimal subset selection for causal inference using machine learning ensembles and particle swarm optimization. *Complex Intel. Sys.* 7:41–59.

Stuart EA (2010) Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics* 25(1):1. http://dx.doi.org/10.1214/09-STS313.

Tam Cho WK, Sauppe JJ, Nikolaev AG, Jacobson SH, Sewell EC (2013) An optimization approach for making causal inferences. *Stat. Neerl.* 67(2):211–226.

Tomizawa N (1971) On some techniques useful for solution of transportation network problems. *Networks* 1(2):173–194.

Visconti G, Zubizarreta JR (2018) Handling limited overlap in observational studies with cardinality matching. *Observational Studies.* 4:217–249.

Yahiro K, Zhang Y, Barrot N, Yokoo M (2020) Strategyproof and fair matching mechanism for ratio constraints. *Auton. Agent. Multi Agent Syst.* 34(1):1–29.

Yang D, Small DS, Silber JH, Rosenbaum PR (2012) Optimal matching with minimal deviation from fine balance in a study of obesity and surgical outcomes. *Biometrics* 68(2): 628–636.

Zubizarreta JR (2012) Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Am. Stat. Assoc.* 107(500):1360–1371.

Zubizarreta JR, Paredes RD, Rosenbaum PR (2014) Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Ann. Appl. Stat.* 8:204–231.

Dorit S. Hochbaum is a distinguished professor of Industrial Engineering and Operations Research (IEOR) at UC Berkeley. Her research interests are in the areas of design and analysis of computer algorithms, approximation algorithms, and discrete and continuous optimization. Her recent work focuses on efficient techniques related to network flows with novel applications in ranking, pattern recognition, data mining, and image segmentation problems. In particular her, research advances machine learning techniques with efficient combinatorial algorithms. Professor Hochbaum is the author of more than 180 papers that appeared in the *Operations Research, Management Science* and *Theoretical Computer Science* literature. She was awarded an honorary doctorate of sciences via the University of Copenhagen recognizing Hochbaum's groundbreaking achievements and leadership in optimization in general and in the field of approximation algorithms for intractable problems in particular. Professor Hochbaum was the winner of the 2011 INFORMS Computing Society prize for best paper dealing with the Operations Research/Computer Science interface. She is an INFORMS fellow and a fellow of the Society of Industrial and Applied Mathematics (SIAM).

Asaf Levin is an associate professor of operations research at the faculty of industrial engineering and management at the Technion – Israel Institute of Technology. His research interests are in the areas of design and analysis of algorithms for combinatorial optimization problems. Professor Levin is an author of more than 120 papers that appeared in the *Operations Research*, *Discrete Mathematics*, and *Theoretical Computer Science* literature.

Xu Rao received her doctoral degree in Industrial Engineering and Operations Research at the University of California, Berkeley. Her PhD thesis is on integer programming formulations and efficient algorithms. Now she is a data scientist in the Operations Data Science team at Google.