Submitted to the Annals of Applied Statistics

EFFICIENT MULTILEVEL MATCHING USING MAXIMUM FLOWS

BY JULIAN BERNADO [1,a], KATHERINE BRUMBERG[1] AND BEN HANSEN[1]

[1]*Department of Statistics, University of Michigan* , [a]*bernado@umich.edu*

Multilevel data structures, where treatment is assigned at the group level but outcomes are measured at the unit level, pose significant challenges in causal analyses. Traditional matching methods that rely solely on group-level characteristics often fail to ensure unit-level comparability, leading to biased estimates. Existing approaches, such as those by Keele and Zubizarreta and Pimentel et al., address this issue by optimizing unit-level matches for each group pair but face substantial computational burdens when dealing with large datasets.

In this paper, we introduce a novel method for multilevel matching that preserves the integrity of unit-level comparability while significantly reducing computational demands. Our approach utilizes a prognostic model fitted on historical data to predict unit-level outcomes based on covariates, incorporating both group-level and mean-centered unit-level covariates. We then compute a simplified yet informative measure of unit-level comparability for each pair of groups using calipers based on the pairwise contrast standard errors. This allows us to estimate balance and effective sample size without performing exhaustive optimal matching for every group pair.

1. Introduction. Multilevel data structures are common in causal analyses: students within schools, patients within hospitals, households within villages. Often treatment is assigned at the group-level rather than the unit-level in structures of this type for both practical and theoretical reasons. Practically, varying some education intervention between students in the same school is often infeasible, and from a theoretical standpoint, if units are in close enough proximity to interact then we lose an often-needed assumption of no-interference. Even if our outcome is unit-level, such as student test scores or patient diagnostics, we still need to contend with the hierarchical structure of the data. In matching-based analyses in this setting, we often wish to match groups that are similar to one another with hopes that the downstream matched sets of units are similar as well. However, two groups demonstrating closeness on group-level characteristics does not necessarily imply that there will be a suitable number of comparable unit within these two groups.

In this paper, I introduce a method for multi-level matching that avoids the naivete of matching only based on group-level characteristics while also avoiding the computational burden of calculating all optimal unit-level matches. I do so by considering a simple yet highly informative measure of unit-level comparability for each pairing of groups. Modern multi-level matching approaches successfully match groups that will have promising unit-level matches by ”looking ahead” to the best possible matching in each pairing of schools. However, the work in [4] demonstrates the method’s success for a number of schools on the order of one county, raising questions for scalability when analysis is done on all schools in a given state. Our method maintains the ”look ahead” ethos of modern strategies, but rather than finding an optimal match in each pair we take a ”coarser” look at potential unit-matches to save on computation time. I demonstrate the effectiveness and speed of this method in a case-study using real data from an ongoing collaboration with the Texas Education Agency. All code is available at the linked GitHub repository, however data is private and cannot be shared.

Keywords and phrases: First keyword, second keyword.

2. Background and problem setting. Consider a setting with S groups $s = 1, ..., S$ each containing N_s units $u = 1, ..., N_s$ for an overall sample size of $N = \sum_{s=1}^{S} N_s$. Each student has p observed covariates $X_{si} \in \mathbb{R}^p$ and treatment indicator $T_{si} \in \{0, 1\}$. We limit our focus to designs where treatment is assigned at the group-level: for all $1 \leq i, j \leq N_s$ we have $T_{si} = T_{sj}$ and hence denote treatment T_s. However, we assume we are in the observational setting where the distribution of T_s is unknown. We denote the set of treated and control groups as S_T and S_C respectively with sizes N_T and N_C.

We use the potential outcome framework; each unit has potential outcomes $\{Y_{si}(0), Y_{si}(1)\}$ corresponding to response under control and treatment respectively. We only observe the realized outcome $Y_{si} = Y_{si}(T_s)$. We consider observational settings analogous to paired clustered randomized control trials where $|S_T|$ pairs of groups are formed and treatment is randomized within each pair. While treatment is assigned at the group-level, we are interested in estimating the unit-level difference in potential outcomes:

$$\tau = \mathbf{E}[Y_{si}|T_s = 1] - \mathbf{E}[Y_{si}|T_s = 0].$$

To estimate τ, we mirror the idealized experiment by forming first-stage matched pairs $M_1, ..., M_{N_T}$ each containing one treatment and one control group. Then, within each group-level matching, we perform an second-stage optimal full matching of treated to control units. We then use [some procedure] to estimate unit-level treatment effect τ.

The novelty of our method comes from the procedure used to perform the first-stage matching of groups to one another. A naive approach to matching groups before units might compare group-aggregated covariates $\overline{X}_{s\cdot} \in \mathbb{R}^p$ between treated and control groups and match treated group s to control group c if $\overline{X}_{s\cdot} \approx \overline{X}_{c\cdot}$. However, Zubizarreta shows this can lead to poor matches [cite]. In particular, these matches may be poor because they do not demand closeness at our desired resolution: unit-level comparisons. Recent work in multi-level matching addresses this concern by looking ahead to unit-level matches before deciding first-stage group-level matches. Pimentel et. al introduce an algorithm for producing group-level matches that works as follows:

1. Enumerate all pairs $P = \{(t, c) : t \in S_T, c \in S_C\}$ of treated and control groups.
2. For each pair $(t, c) \in P$, perform an matching of students in t to students in c.
3. Calculate a distance $d(t, c)$ between groups t and c for all pairs $p \in P$ based on the results of the unit-level matching.
4. Match each group in S_T according to the distance matrix D with $D_{tc} = d(t, c)$.
5. Perform unit-level matches within each pair of matched groups, then estimate τ using these matches.

We accept that comparing each treated to control school is a necessary part of creating well-formed matches, but focus on an improvement to the most costly part of their procedure: step 2.

3. Methodology. In this section, we discuss the

3.1. *Practical almost-exact matching.* In the idealized setting,

3.2. *Novel matching distances.*

- Motivate the distances by talking about balance and ess
- Introduce the definitions of e_1, e_2, and e_3 and define $d_e(s, c)$ as a function of them.
- Mention valuable properties: the ordering, the big O save, the lack of complicated tuning parameters, the fact that it's actually adaptive so as to reduce unnecessary calculation.

4. Comparative analysis using TEA data.

4.1. *TEA data.*

- motivate problem setting
- describe the multi-year multi-grade nature
- provide some basic data description

4.2. *Experimental setup.*

- Write up how the prognostic scores and calipers are actually attained in our data
- Write up computational experiment procedure
- Write up placebo test procedure (incl. throughline matching description)
- Describe what Katherine's doing

4.3. *Results.*

- computational comparison results. Plots showing comparative performance of the two: Scatter plot with school size on x-axis, time to complete on y-axis and Pie chart showing how many of our distances stopped after e_1, after e_2 and e_3.
- placebo test results. Plot: 12 confidence intervals stacked on one another. One for each grade and subject 3 through 5 and one for each method, colored differently. They all should be centered around zero hopefully. A table showing the bias and variance of each estimate next to one another to see when each method wins. Each method here is: random matching, naive matching, matchAhead matching, then matchMulti matching.
- katherine's work results (tbd)

5. Discussion.

- tbd, but obviously some summary, some highlighting of what's better about matchAhead distances

6. Questions.

- Broadest question is whether this outline makes sense for the whole thing
- Too late to add to the method? (for the bias score, it pretty much adds nothing to the comparison for us to look at the average difference between those "matched" by the maxflow calc).
- Is this what the authors list will look like or something else?