Beyond Covariate Overlap: An Analysis of Genetic Matching and Entropy Balancing

Vinícius Miranda*
College of Computer Sciences, Minerva Schools at KGI

January 4, 2018

Abstract

Covariate balance, assuming a robust design, is the holy grail of causal inference in observational studies. Nothing can save poor designs, but often even strong ones need to be supplemented by adequate statistical methods. I investigate the synergies between matching and reweighting methods exemplified by genetic matching and entropy balancing. Now that full overlap can be practically achieved, several theoretical questions gaps must be filled before researchers can fully trust their balanced estimates.

Keywords: Causal inference, observational studies, genetic matching, entropy balancing, covariate balance.

*I am grateful and indebted to Alexis Diamond that, through his distinguished professorship and kind support, inspired me to follow this line of research. A replication code is available at `https://github.com/viniciusmss/CS112/blob/master/Final%20Project`

1 Introduction

Assignment mechanisms keep researchers awake at night, wondering whether they should be making the conclusions they wish to publish. They wonder if the mechanism is indeed ignorable, if they genuinely selected on all relevant observables, and if error terms are actually normally distributed. When an experimental design is not available, researchers need to confront these questions in claiming plausible treatment effects by trying to assuming inferential conditions that they could not otherwise guarantee. A prominent task in this direction is to produce overlapping covariate distributions of treatment and control groups, resulting in their comparability.

Matching methods have become famous for aiming to accomplish such overlap. Units in the treatment group are matched to units in the control group by some measure of similarity or proximity. Although these methods usually work by dropping control units and blocking or pairing treatment to control units, entropy balancing sets different weights to specific observations to model the covariate distribution found in the treatment group. This procedure builds off of a common strategy in survey studies where unit weights can be reset based on known population characteristics.

I am not aware of any research that investigated, either in theory or practice, the potential synergies between matching and reweighting methods such as entropy balancing. This investigation is the focus of this paper. I select genetic matching (Diamond and Sekhon, 2013) and entropy balancing (Hainmueller, 2012) as the matching and reweighting approaches I analyze. With due caution, the insights might be generalizable to other methods.

This paper is organized as follows. Section 2 provides technical details to each method. In Section 3, I entertain theoretical directions of how these approaches might work together or be adapted. Section 4 presents an empirical application of this work. Next, I discuss these results and the last section concludes.

2 Methods

2.1 The Rubin Causal Inference Model

The model developed by Rubin (1974) frames causal inference around potential outcomes — what would happen to a unit under different treatment conditions. The fundamental problem of causal inference (Holland, 1986) is that we can only ever observe one potential outcome. For a unit in the control group, for example, we see the outcome under control and can only infer what would have happened under treatment. However, in practical applications, one is not usually concerned with unit-level treatment effects. Instead, researchers focus on uncovering the average treatment effect (ATE), defined as the difference between mean outcomes in the treatment and control group. In randomized experimental designs, the control group outcomes serve as an unbiased counterfactual to the outcomes for the treatment group given that they are comparable in expectation for both measured and unmeasured characteristics.

In observational studies, the difference-in-means estimator necessitates two special assumptions to become proper for causal inference. Because the researcher has no control over the assignment mechanism, one can neither know nor control for unobserved covariates. Therefore, we must assume that the program selected units for treatment (or they selected themselves) based on characteristics that were measured, instead of unobserved attributes. The second assumption of consequence is overlap: the covariate distribution of treatment and control groups should be the same, thus ensuring that they are comparable. Overlap almost never occurs in practice given that, in observational studies, units self-select to receive treatment. However, if selection on observables holds, the measured information can be used to produce this overlap by matching comparable units, dropping units that cannot be matched or setting weights across the sample. Therefore, the underlying theoretical goal of matching and reweighting procedures in to secure that the covariate distribution overlaps and thus that we are comparing apples to apples in producing causal estimates.

2.2 Genetic Matching

Genetic matching finds optimal matches by changing the relative weights among covariates in the sample dataset. The approach thus takes Mahalanobis Distance (MD) matching (Rosenbaum and Rubin, 1985) to a higher level of abstraction. In MD matching, the distance metric used to compare units takes into account the covariance among the distribution of different covariates in the sample. Genetic matching generalizes this distance metric by implementing a search algorithm that investigates the optimal set of relative weights W that optimizes balance in the matched dataset. The insight is that one covariate may import more or less to the task of producing the overlap between the treatment and control groups. The most common application of genetic matching requires that the algorithm finds nearest-neighbor matches to every unit in the treatment group while allowing for replacement in the control group. This strategy produces the highest degree of overlap (Diamond and Sekhon, 2013) and lowers conditional bias (Abadie and Imbens, 2006). Notice that the estimate, in this case, changes from ATE to ATET as, usually, many units in the control group are dropped.

2.3 Entropy Balancing

Entropy balancing, in turn, is a scheme that assigns weights to every unit in the control group such as to satisfy a set of balance constraints. Naturally, the most critical balance constraint is securing that the means of the covariate distributions between treatment and control groups converge. However, entropy balancing can further guarantee that the overlap assumption holds by balancing for higher sample moments, such as the variance but also skewness and kurtosis. One subsidiary constraint of consequence that entropy balance optimizes for is the variability of the vector of weights, given that they should be greater than zero. The algorithm imposes a penalty on the variability of weights and enables the researcher to trim the weight vector. The reason for this design choice is to avoid the information loss and analytical efficiency (Hainmueller, 2012). This approach is thus more versatile and is usually able to produce perfect overlap in covariate distribution when measured by standard difference-in-means t-tests.

3 Theoretical Questions

I analyze whether there are ways to combine GM and EB and whether the constraints of the former algorithm are optimal. I conceptualize the optimization problem both methodologies are tackling and critically discuss how their restrictions bear of the solutions they find.

First, matching methods are a particular case of reweighting approaches in a technical sense, differing in their set of constraints. Most matching methods attribute the weight of 0 to all control units that do not pass a threshold of comparability to units in the treatment group. The consequence of this constraint is a change in the estimate from ATE to ATET (average treatment effect on the treated), which is not consequential to most practical applications. An alternative and often accompanying specification is *replacement*, which specifies that good control units can be "reselected" if they are the most appropriate to a match (according to the chosen balance metric), rather than selecting a less comparable unit. The consequence of this assumption is that we are not particularly concerned with the overrepresentation of control units in the matched data set. Besides guaranteeing greater overlap, Abadie and Imbens (2006) have found that replacement decreases conditional bias.

The constraints in entropy balancing go almost in the opposite direction. The algorithm is concerned with guaranteeing the smallest variability in the set of weights, penalizing both the under and overrepresentation of control units (i.e., weights much smaller or much larger than the mean). Notice that this constraint is a methodological choice not strictly necessary for overlap. However, because of the higher methodological flexibility in assigning unit weights, entropy balancing is widely successful in securing covariate balance. Is entropy balancing, then, the best way to achieve overlap?

I claim that there should be ways to optimize entropy balancing in and beyond the attainment of covariate balance. As an example, imagine a job training program for which we have a large observational dataset that poorly represents the pre-intervention characteristics of the treatment group. For example, the mean age of controls is higher due to the presence of elders, a demographic non-existent in the treatment group. The constraint that penalizes variability in the vector of weights is counter-intuitive. For the same level of balance, a solution that underweights or dismisses these units altogether is seemingly preferable to a solution that maintains them.

There are two explanations for this intuition. First, when the treatment is restricted to a section of the population as defined by the treatment group covariate distributions, there is no reason why the ATE for a large sample would be preferable to the ATET for the for the subpopulation which the program targets. Second and more technically, more general results should weaken the assumption of selection of observables because the inference for the treatment effect includes a subpopulation for which the program was not designed. In the example of the supported work program, not only could the treatment effect vary across age populations, but we should not be sure that the covariates that explain treatment assignment for one subpopulation are the same as those of the other.

For these reasons, I claim the design choice of penalizing the exclusion or small weighting of control units is shaky enough to merit further analysis. Given that genetic matching is a prominent method that accomplishes high levels of balance by dropping several control units, I outline several possible synergies between matching and entropy balance below, some of which I test below in Section 4.

3.1 Genetic Matching to Entropy Balancing

The first option is to have genetic matching precede entropy balancing, which would receive the matched instead of the full dataset as an input. In this approach, the permission of zero weights is preserved given that the balancing algorithm never gets the dropped control units. One shortcoming is that the pairing information is lost once the EB algorithm processes the matched dataset. Namely, GM paired treatment unit A with control B because they are the most comparable, but EB ignores this information and can set different weights for these units. The extent to which weight variability in penalized is not due to the matching but instead the whole weight distribution. This procedure is, however, a partial solution that does not require the modification of the existing algorithms.

3.2 Entropy Balancing to Genetic Matching

The reverse procedure is also possible where the weights of entropy balancing might be used to preprocess the dataset or inform genetic matching. There are two immediately available implementations. First, the GM algorithm in R accepts individual observation weights as inputs. However, there is no public information of how these data affect the algorithm.[1] Therefore, this option looks unpromising and would necessitate an investigation into the source code of GM. The second

[1] I searched Diamond and Sekhon (2013) and the paper that describes the Matching CRAN package, besides having investigated the R help files.

path is to trim the vector of weights outputted by entropy balancing to drop observations whose weights fall far below the mean. GM would receive the trimmed data set. The intuition is that underweighted units are likely to be non-similar to those in the treatment group. The mathematical implications of this procedure could be analyzed based on the theorems underpinning entropy balancing, but I restrict myself to an analysis of its impact in the simulations below.

3.3 Match with Higher Tolerance

Given that one of the advantages of entropy balancing is the retention of information, it is possible that by passing the most balanced dataset produced by genetic match, we have already dropped many control units. It could be the case that in doing so, we prevent entropy balancing from being most useful. We could relax the constraint of nearest-neighbor matching by applying exact matching with a high distance tolerance. GM can accept a parameter of tolerance below which it considers the distance between units as zero and performs "exact" matching. The choice of tolerance is arbitrary and thus represents the quintessential trade-off between the two algorithms — how many controls to preserve and how many to drop. It is crucial that the estimate is the same to allow for comparability. Thus, the distance tolerance should not lead to the loss of any treatment units. This reasoning provides a provisionally workable cutoff distance, namely the smallest distance that does not reduce the treatment sample.

3.4 Alternative Loss Function

A promising direction is the adaptation of the constraints of entropy balancing to account for the matching information. This restriction could theoretically substitute the one that requires that the weights be both larger than 0 and the most alike. The algorithm would penalize the weight variation among previously matched units instead of in the whole vector of weights. The solution necessitates the modification of the algorithm to account for the matching information while also redefining the mathematical theorems underpinning entropy balancing. It is undetermined whether the field of solutions to the redesigned optimization problem remains globally convex so that the implementation quickly converges. If that is not the case, research would have to investigate the answer to the redesigned problem.

3.5 Optimize the Division of Labor

Finally, there is an inherent push-and-pull between the matching and entropy balancing approaches that can only be somewhat arbitrarily approached. Once multiple solutions attain full covariate overlap, there is no definitive way of choosing among them. The optimal choice, concerning causality, lies somewhere between dropping too many or too few units. Two topics motivate this problem. The first is the retention of valuable information existing is the dataset leading to the question of how to define "valuable." Next is the assumption of selection on observables and how it is weakened by the existence of incomparable units in the dataset — not regarding covariate overlap, but in unobserved or unmeasurable characteristics. The optimal division of labor or solution to the push-and-pull problem is arguably the most distant and theoretically consequential research question for causal inference.

4 Empirical Application

Here, I use data from the National Supported Work (NSW) program combined with observational data, the problem both Diamond and Sekhon (2013) and Hainmueller (2012) used to showcase the power of their algorithms. The NSW was a randomize job training program that became famous for the work of LaLonde (1986). He showed that econometric and statistical methods using observational data as controls could not recover the experimental estimate, a shocking result at the time. LaLonde used survey data that provided information for the same covariates measured for program participants. Besides, the survey controls were very different from the sample of program participants. I use the DW sample (Dehejia and Wahba, 1999) with Current Population Survey 1 controls to replicate the results of Diamond and Sekhon (2013) and Hainmueller (2012). The same samples are used to test some of the variations suggested above.

4.1 Genetic Matching

I replicate Diamond and Jeekhon (2013) based on their original specifications. I matched on all observed covariates, interaction terms, and quadratic terms including an estimated propensity score. Notice that genetic matching is stochastic, and there it is not expected that I replicate perfectly the original results of $1734. The algorithm produced a causal estimate of $1843 with a

95% confidence interval of [-84, 3777], close to the experimental benchmark of $1794. The smallest p-value found was of 0.21.

4.2 Entropy Balancing

I replicate Hainmueller (2012) also based on his original specifications, which includes two additional dummy covariates and excludes interactions terms that are nonsensical, as they would preclude a solution from being found.[2] The implementation of entropy balancing is deterministic, and therefore I obtain the same results as the original. The treatment effect estimate is of $1571 with a 95% confidence interval of [97, 3044]. The trimming function contained in the replication files produced an error when trying to trim the weights of the result vector. Hainmueller (2012) claimed that these results were "slightly more efficient" than the ones uncovered by the best run of genetic matching reported by Diamond and Sekhon (2013) due, apparently, to the width of the confidence intervals. I will contest this claim in the discussion.

4.3 Genetic Matching + Entropy Balancing

I apply entropy balancing upon the optimized dataset outputted by genetic matching. To be faithful to the specifications of the authors above, I organize the genetically matched dataset in the same way Hainmueller (2012) built the original DW sample with CPS-1 controls. The treatment estimate obtained through this process is of $1924 with a 95% confidence interval of [668, 3180]. Notice that although this procedure secures full overlap, the treatment effect point estimate is farther away from the experimental benchmark when compared to the previous GM estimate of $1843. We should note that the confidence interval did become more precise, an effect I account for in the discussion.

4.4 Entropy Balancing + Genetic Matching

Here, I test two ways of having entropy balancing precede genetic matching. First, I pass the weights outputted by entropy balancing into the genetic matching algorithm. As expected, the

[2] Examples include the quadratic term of the covariate black or the interaction term between black and hispanic.

procedure does not work, and genetic matching seemingly does not use this information as would be expected. For the same specifications used by Diamond and Sekhon (2013), the algorithm is incapable of elevating several p-values above 0. How genetic matching treats these weights should be the subject of further research. Next, the implementation of GM upon a trimmed dataset for deviant weights leads to surprising results. The CPS-1 sample has 15992 controls. In the vector of weights Hainmueller (2012) uses to derive his estimate, 79.7% of controls received weights of less than 0.001, whereas 0.2% of controls are weighted more than 1. The total weight held by the 16 heaviest controls (57.76) is only smaller to that held by 15868 lightest ones (57.86). I discuss these results below. For now, I select the top two percentiles of the weight distribution and pass 320 of the most represented controls to genetic matching. Notice that, after most of the applications of genetic matching above, the algorithm would keep about 285 controls. Therefore, I am here implicitly testing whether that two algorithms seemingly converge on their choice of "good" controls. I find, however, that genetic matching now performs worse than before, with a minimum p-value of 0.08 but with six p-values below 0.10 and 25 p-values below 0.20, leading to an estimate of 1744 with a 95% confidence interval of [-187, 3675].

5 Discussion

First, I will address how entropy balancing produces more efficient or tighter intervals in all applications above. This efficiency is an unmerited advantage of EB. Hainmueller (2012) calculated the confidence intervals for the causal estimate through the standard error outputted by a generalized linear model of outcomes regressed upon treatment while accounting for unit weights. Conversely, Diamond and Sekhon (2013) used the Abadie-Imbens standard error which accounts for the uncertainty of the matching procedure. If they had not taken this uncertainty into account, the GM estimate reported in Section 4.1 would have a 95% confidence interval of [461, 3225], 29% tighter and more precise than the EB confidence interval. Therefore, the claim that EB is more precise than GM is unwarranted and future implementations of EB need to account for the uncertainty of the reweighting procedure.

Also, the application of entropy balancing upon a genetically matched dataset led the estimate to become further away from the experimental benchmark. This result could be due to chance but also could attest to an increase of bias due to the reweighting process. Monte Carlo simulations and additional empirical applications are needed to gather further evidence to this phenomenon. More consequentially, we should not greater overlap, at least as attained by entropy balancing, does not necessarily lead to better estimates.

Another surprising finding is the variability in the distribution of weights in the solution reported by Hainmueller (2012). Sixteen controls accounted for the same total weight as 15868 other controls. This discrepancy is arguably the most problematic result given that it challenges a cornerstone methodological advantages proclaimed by the author. Indeed, EB is not retaining valuable information contained in the dataset but indeed ignoring it to an extent possibly more significant than GM and other matching methods. This contradiction would be excusable if we came to conclude that the overrepresented units in the vector of weights are particularly good "matches." However, once we apply genetic matching to the observations in the top two percentiles in the distribution of entropy-balanced weights, we find that genetic matching achieves a poorer balance than it would have otherwise attained. The question of how the two methods come to prefer different controls and how the selected or heavily weighted units affect the final estimate is a cornerstone suggestion for further research.

Finally, I do not test several of the theoretical suggestion made in Section 3. It is my hope that they present exciting direction for others researchers to tackle.

6 Conclusion

Covariate balancing represents a major development in the statistical methodology that aims to facilitate causal inference in observational settings. More research needs to be done, however, to secure the reliability of the method's estimates and how it could, or should, be used in conjunction with other approaches.

References

Abadie, A., & Imbens, G. (2006). Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica, 74*, 235-267.

Dehejia, R., & Wahba, S. (1999). Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs. *Journal of the American Statistical Association, 94*(448), 1053-1062.

Diamond, A., & Sekhon, J. S. (2013). 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.

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

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association, 81*(396), 945-960.

LaLonde, R. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *American Economic Review, 76*, 604-620.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *The American Statistician, 39*(1), 33-38.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology, 66*, 688-701.