9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

Matching Using Feature Importance: An Auditable Approach to Causal Inference

Keywords: Causal Inference, Matching, LASSO, Nearest Neighbors, Distance Metric

Abstract

Matching methods are a popular approach to causal inference on observational data due to their conceptual simplicity. These methods aim to emulate randomized controlled trials by pairing similar treated and control units, allowing for treatment effect estimation [1]. One significant benefit of using matching methods is their *auditability*. Auditability allows domain experts to validate the estimation procedure, argue about the violation of key assumptions, and determine whether the analysis is trustworthy. [2] and [3] showed that the audit of matched groups using external unstructured data is crucial in high-stakes healthcare and social science scenarios. Since causal analyses often depend on untestable assumptions, it is critical to determine whether all important confounders are accounted for, if data are processed correctly, and whether the matched units are cohesive enough to be comparable. In high-stakes scenarios, audibility enables domain experts to make data-driven and *trustworthy* decisions.

We would ideally match units that are identical to one another [4]. However, such matches are almost impossible in high-dimensional settings with continuous covariates [5]. Several matching methods have been developed to overcome this. The most popular are propensity score matching [4] and prognostic score matching [6]. These approaches can be sensitive to modeling choices that affect the accuracy of the treatment effect estimates and their matched groups are generally not auditable because units within a matched group can be far from each other in covariate space [5]. To date, the only matching techniques that optimize accuracy while maintaining auditability are those stemming from the almost-exact matching framework, namely optimal matching (optMatch) [3], genetic matching (GenMatch) [7], FLAME/DAME [8], MALTS [5] and AHB [9]. FLAME/DAME can scale to large datasets but handles only categorical variables. GenMatch, MALTS, and AHB can handle continuous variables but do not scale well, thereby limiting their usefulness.

We introduce a new *almost-exact matching* approach that provides *accurate* causal estimates, ensures *auditability* so we can evaluate and troubleshoot, and is *scalable* to large observational datasets. Our method uses variable importance from prognostic score models to learn a distance metric. The method has three steps. First, we use machine learning to estimate outcomes and use the measured variable importance to construct a distance metric. Second, we use the learned distance metric to match treatment and control units into matched groups. Third, we use the matched groups to estimate conditional average treatment effects (CATEs). A special case of our method is called *LCM – LASSO Coefficient Matching*, which uses LASSO coefficients to identify important variables. LCM benefits from both the efficiency of parametric modeling and the power of nonlinear modeling by leveraging a parametric method to learn which features to match on and then using a nonparametric approach for estimation.

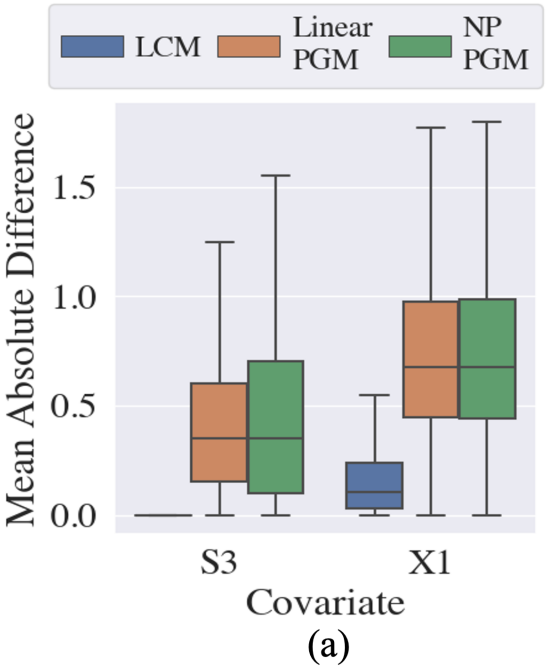
We derive performance guarantees for settings where LASSO outcome modeling consistently identifies all confounders (importantly without requiring the linear model to be correctly specified). We perform extensive empirical studies to highlight the *auditability*, *robustness*, and *scalability* of LCM. For *auditability*, we show on a semi-synthetic dataset based on the

National Study of Learning Mindsets that LCM matched groups are more similar on important covariates than prognostic score matched groups (Figure 1) [10]. We use synthetic data to highlight LCM's *robustness* to model-mispecification (Figure 2) and *scalability* in performance and runtime (Figures 3 and 4). We further examine two adaptations to our method. First, we consider a metalearner LCM which learns a separate distance metric for each treatment assignment (Figure 5). Second, we propose augmenting prognostic score matching with LCM to guarantee tight matches and accurate CATE estimates. In this second approach, we first create large matched groups using prognostic scores and then create smaller matched groups from within each large matched group using the distance metric learned via LCM (Figures 6 and 7). Finally, while we focus on LCM, we suggest Gini feature importance measures as a potential substitute to LASSO coefficients (Figure 8) and note that any variable importance metric can be used.

Our proposed methodology's flexibility makes it adaptable to a variety of practical problems. We are currently working with doctors to apply our method on electronic health records to understand which treatment approaches are best for mothers with preeclampsia. In future work, we will study the use of different variable importance metrics and implement our approach on other large, real-world data such as genome studies and living standards measurement studies.

References

1. Stuart, E. A. Matching methods for causal inference: A review and a look forward. *Statistical Science* **25,** 1 (2010).
2. Parikh, H. *et al. Effects of Epileptiform Activity on Discharge Outcome in Critically Ill Patients: A Retrospective Cross-Sectional Study* 2022.
3. Yu, R., Small, D. S., Harding, D., Aveldanes, J. & Rosenbaum, P. R. Optimal Matching for Observational Studies That Integrate Quantitative and Qualitative Research. *Statistics and Public Policy* **8,** 42–52 (2021).
4. Rosenbaum, P. R. & Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55 (1983).
5. Parikh, H., Rudin, C. & Volfovsky, A. MALTS: Matching After Learning to Stretch. *Journal of Machine Learning Research* **23,** 1–42 (2022).
6. Hansen, B. B. The prognostic analogue of the propensity score. *Biometrika* **95,** 481–488 (2008).
7. Diamond, A. & Sekhon, J. S. Genetic Matching for Estimating Causal Effects: A General Multivariate Matching Method for Achieving Balance in Observational Studies. *The Review of Economics and Statistics* **95,** 932–945 (2013).
8. Wang, T., Roy, S., Rudin, C. & Volfovsky, A. FLAME: A Fast Large-scale Almost Matching Exactly Approach to Causal Inference. *Journal of Machine Learning Research* **22** (2017).
9. Morucci, M., Orlandi, V., Roy, S., Rudin, C. & Volfovsky, A. *Adaptive hyper-box matching for interpretable individualized treatment effect estimation* in *Conference on Uncertainty in Artificial Intelligence* (2020), 1089–1098.
10. Carvalho, C., Feller, A., Murray, J., Woody, S. & Yeager, D. *Assessing Treatment Effect Variation in Observational Studies: Results from a Data Challenge* 2019.



	T	S3	X1
Query	1	5	-.05
LCM	0	5	-.05
	0	5	-.05
	0	5	-.05
	0	5	-.05
Linear PGM	0	5	-.05
	0	5	-.05
	0	**6**	**1.26**
	0	**6**	**1.26**
NP PGM	0	5	-.05
	0	5	-.05
	0	**6**	**.98**
	0	**6**	**.98**

(b)

Figure 1: Closeness in important covariates for matched groups produced by LCM, linear prognostic score matching (Linear PGM), and non-parametric prognostic score matching (NP PGM). *(a)* shows the mean absolute difference between a query unit and its matched group's covariate values. Smaller values imply better and tighter matches. *(b)* shows, for a random sample, the four nearest neighbors of opposite treatment under LCM, linear PGM, and NP PGM. In *(b)*, the text in red indicates values that are far from the query's value. Covariate `S3` indicates the self-reported prior achievements of students and is important for selection into treatment, and covariate `X1` indicates school-level average mindset score of the students and is an effect modifier.

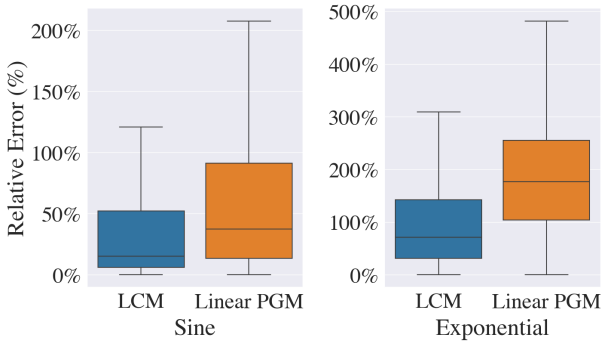


Figure 2: Absolute CATE estimation error relative to the true ATE for LCM and linear prognostic score matching (PGM). Datasets are nonlinear and synthetically generated where the outcome is a **Sine** or **Exponential** function of the covariates. This figure shows that LCM is more robust to nonlinear outcome functions than linear prognostic score matching.

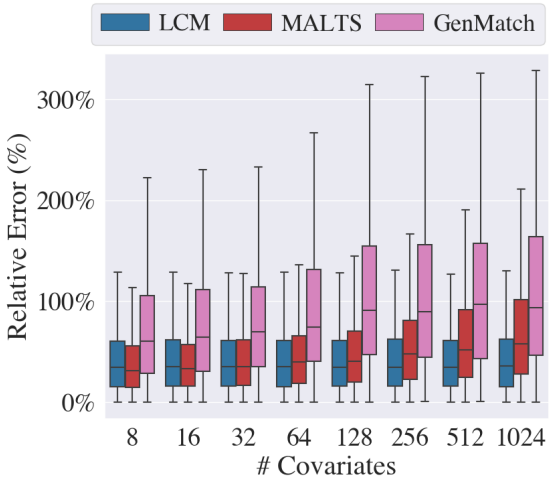


Figure 3: Absolute CATE estimation error relative to the true ATE for LCM, MALTS, and GenMatch as the number of irrelevant covariates increases. We keep the number of samples and the number of covariates relevant to the outcome constant at 1024 and 8 respectively. LCM excels at producing accurate CATE estimates as the number of irrelevant covariates grows.

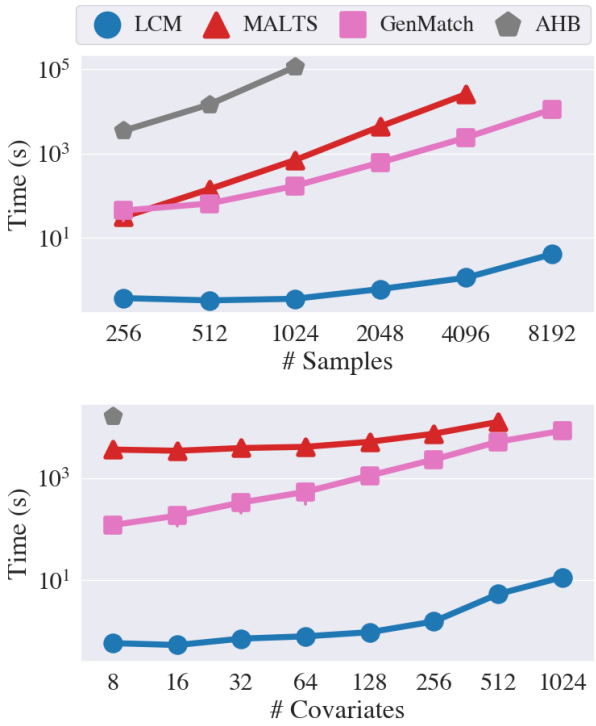


Figure 4: Runtime scalability in the number of samples (n) and the number of covariates (p) for LCM, MALTS, GenMatch, and AHB. To measure scaling runtime in n, we keep the number of covariates constant at 64 and increase the number of samples from 256 to 8192. To measure scaling in p, we set the number of samples to be 2048 and vary the number of covariates from 8 to 1024. This figure highlights the multiple-order-of-magnitude runtime disparity between LCM and other almost-exact matching methods.

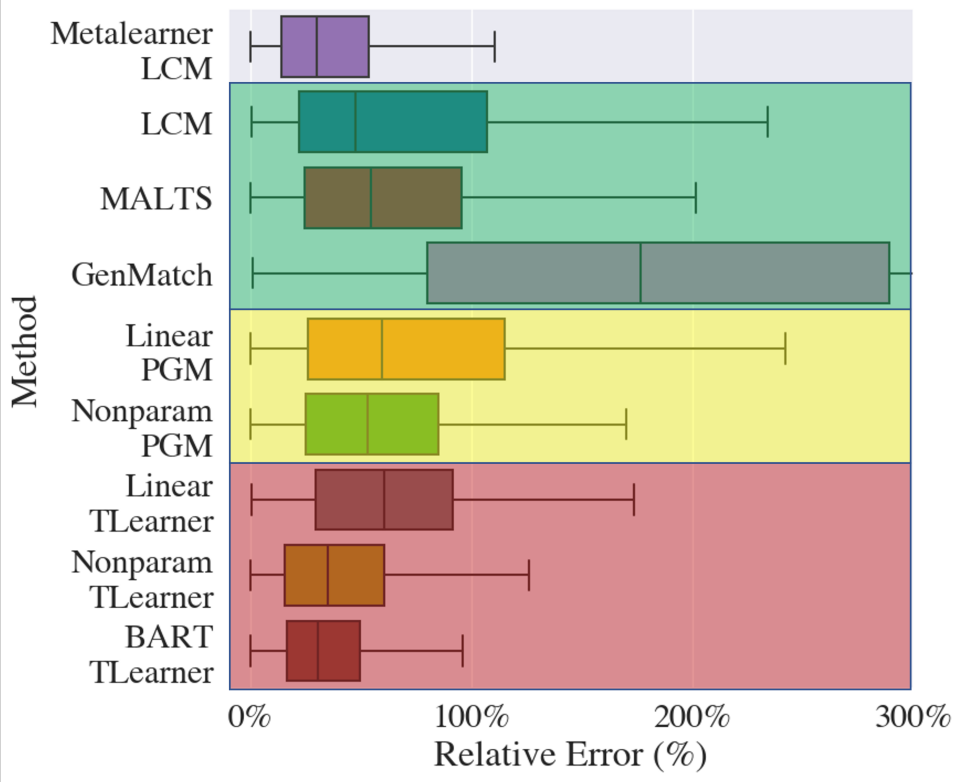


Figure 5: Absolute CATE estimation error relative to the true ATE for various methods. In this dataset, different covariates are important to the outcome depending on if a unit is in the control or the treatment group. The transparent boxes separate the methods into different categories. **Green**: Almost-exact matching methods. **Yellow**: Other matching methods. **Red**: TLearner methods. Metalearner LCM improves upon LCM, which already outperforms other matching methods, and is comparable to T-Learners.

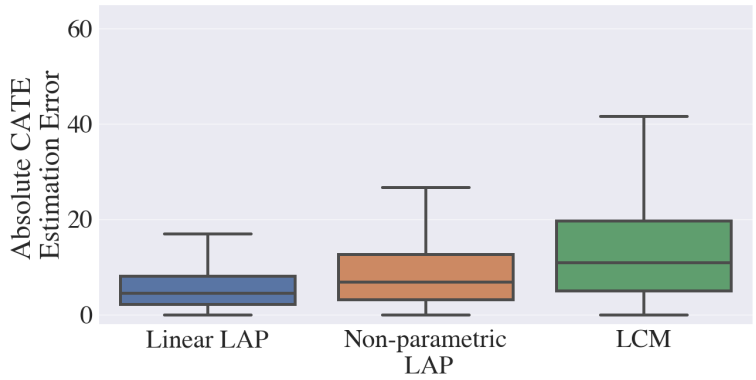


Figure 6: Absolute CATE estimation error for linear LCM augmented prognostic score matching (Linear LAP), non-parametric LCM augmented prognostic score matching (Non-parametric LAP) and LCM. The data is synthetically generated where the expected outcome is a linear model of the covariates with a quadratic treatment effect.

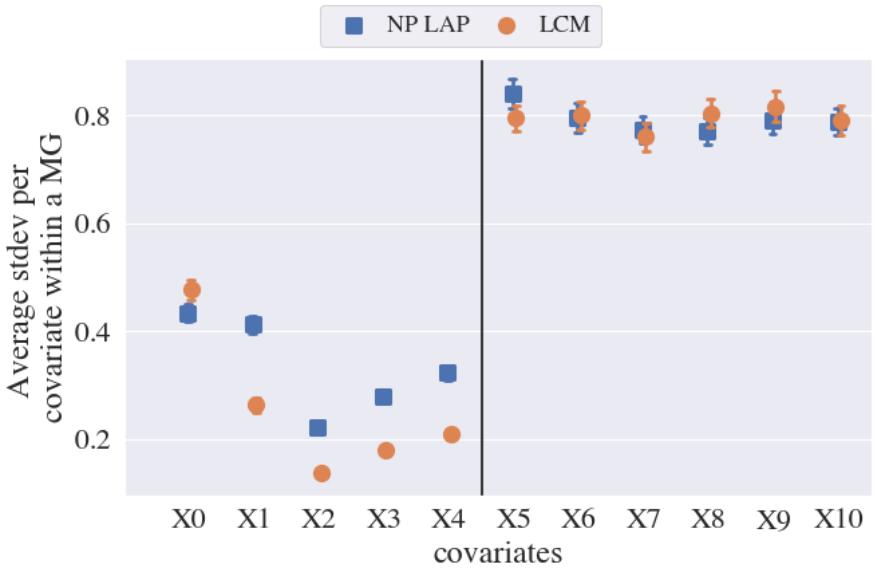


Figure 7: Average standard deviation for each covariate inside the matched groups for non-parametric LAP and LCM. The smaller the standard deviation, the tighter the match on those covariates. The dataset has 20 covariates where only the first 5 are important. We show the first 11 covariates (5 important and 6 unimportant) for ease of presentation. This figure shows that the matched groups created using non-parametric LAP are almost equally as tight as LCM's matched groups on the 5 important covariates (X0-X4) and do not prioritize matching on irrelevant covariates (X5-X10).

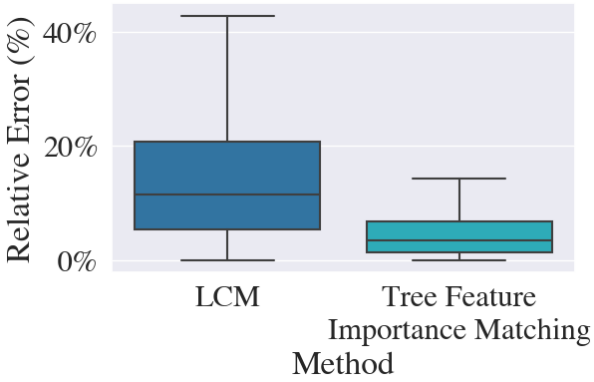


Figure 8: Absolute CATE estimation error relative to the true ATE for LCM vs. feature importance matching with Gini feature importance from a classification and regression decision tree (CART). Here, the outcome function is such that a linear model can not recover the correct covariates, so an alternative feature importance method works better.