Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

Discussion of "Penalized Spline of Propensity Methods for Treatment Comparison"

Georgia Papadogeorgou & Fan Li

To cite this article: Georgia Papadogeorgou & Fan Li (2019) Discussion of "Penalized Spline of Propensity Methods for Treatment Comparison", Journal of the American Statistical Association, 114:525, 32-35, DOI: 10.1080/01621459.2018.1543120

To link to this article: https://doi.org/10.1080/01621459.2018.1543120

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 525, 33–35: Applications and Case Studies
https://doi.org/10.1080/01621459.2018.1543120

Discussion of "Penalized Spline of Propensity Methods for Treatment Comparison"

Georgia Papadogeorgou and Fan Li

Department of Statistical Science, Duke University, Durham, NC

We congratulate Zhou, Elliott, and Little (ZEL hereon) for providing an easy-to-implement and widely applicable method for causal inference (PENCOMP). The method is grounded on imputing the missing potential outcomes, and unites the propensity score and outcome model approaches for confounding adjustment. We regard the main advantage of this method as its flexibility, namely, physically imputing the missing potential outcomes allows one to estimate any causal estimand of interest. This is in contrast to the popular weighting methods, which usually only target at mean estimation. As shown in ZEL, the flexibility is particularly desirable in complex settings such as longitudinal treatments, making it attractive as an off-the-shelf approach in applied work.

Below we discuss three conceptual and operational issues that are fundamental to causal inference in longitudinal settings: (i) target population and overlap, (ii) double robustness, and (iii) the form of sequential ignorability. While ZEL has supplied some insights on these sometimes thorny issues, we feel the audience might benefit from a deeper and more focused discussion.

1. Causal Estimand and the Target Population

The starting point of most causal studies is to define a causal estimand—a contrast of the potential outcomes of the same set of units under different treatments. Causal estimands are tied to the specification of the *target population* over which the contrast of potential outcomes is considered. The most common causal estimand is the average treatment effect (ATE), which is also the target in ZEL. We find that the automatic focus on ATE often leads researchers to overlook a more fundamental question: What target population is the causal estimand defined on, and is that population *scientifically interpretable*? For example, by focusing on the ATE, researchers make the implicit assumption that the study sample is representative of a wider population of intrinsic interest. However, many observational studies rely on "convenience samples" that result from ad hoc and ambiguous inclusion and exclusion criteria (e.g., access to treatment). These samples usually do not represent a scientifically interpretable target population. Moreover, differences in inclusion/exclusion criteria across datasets could lead to large differences in estimated causal effects, since estimated quantities correspond to contrasts of potential outcomes over different target populations.

Below, we discuss how weight trimming in weighting approaches, and dropping observations in matching methods and PENCOMP, implicitly change the target population. As a related point, we discuss that imprecise weighting estimators are not solely a statistical inefficiency issue, but they should instead be viewed as an indication of a *poorly specified target population* and a corresponding malformed causal estimand. We argue that limited overlap should be addressed by defining a meaningful target population over which efficient inference can be made.

The most popular method for drawing causal effects in longitudinal treatments is the inverse probability of treatment weighting (IPTW), originating from the Horvitz–Thompson estimator in sample survey. In causal inference the main idea of weighting is to reweigh the treatment and control groups to create a pseudo-population, resembling the target population, where the two groups are balanced in observed covariates. IPTW and its augmented version (AIPTW) are often criticized for leading to estimates with high variance in the presence of extreme propensity scores. A common remedy of this is trimming of the extreme scores (e.g., Crump et al. 2009). The other popular general approach to causal inference is matching on the propensity score. In one of its forms, matching leads to matched *pairs* of observations with similar covariates and opposite observed treatment levels. Observations with extreme propensity score estimates are usually dropped from the analysis since no sufficiently similar units of the opposite treatment level can be found.

The above two approaches of dealing with extreme propensity scores in weighting and matching are of similar flavor and are rooted in the issue of *limited overlap*, also known as violation to the *positivity* assumption. Both lead to estimation of treatment effects over populations that are different from the target one, and can be hard to interpret. For example, dropping treated units in propensity score matching leads to the estimation of the ATE over only a subset of the treated units, the ones that were matched. Indeed, Rubin (2001) claims that in some cases it is not possible to draw causal conclusions from a dataset without relying heavily on outcome modeling assumptions, and such efforts should be avoided in search of a more comprehensive and representative dataset.

CONTACT Fan Li *fl35@duke.edu* Department of Statistical Science, Duke University, Durham, NC 27708.

Longitudinal treatments magnify the issue of limited overlap. Similarly to matching methods, PENCOMP drops observations with extreme propensity scores. Specifically, ZEL suggested excluding observations based on a sequential overlapping estimated propensity score condition. Specifically, if $Z_{it}, \overline{Z}_{it}$ denote the observed treatment and the observed treatment history of unit i at time point t, and $z_t, \overline{z}_t$ denote potential values of that treatment, let $p_i(\overline{z}_t)$ denote the estimated propensity score for unit i to have received $\overline{z}_t$. Then, the sets A_t, B_t discussed in ZEL correspond to $S_t(\overline{z}_t), S_t(\overline{0}_t)$ where

$$S_t(\overline{z}_t) = \left\{ i : p_i(\overline{z}_t) \in \left(\min_{j \in \mathcal{U}_t} p_j(\overline{z}_t), \max_{j \in \mathcal{U}_t} p_j(\overline{z}_t) \right) \right\},$$
$$\mathcal{U}_t = \left\{ j : \overline{Z}_{jt} = \overline{z}_t \right\}.$$

Inference for a T-dimensional treatment proceeds over the units in $S_1(\overline{z}_1) \cap S_1(\overline{0}_1) \cap \cdots \cap S_T(\overline{z}_T) \cap S_T(\overline{0}_T)$. Therefore, the observations maintained are those with longitudinal propensity score estimates within the range of the propensity score estimates for units which actually received the treatment path of interest. Such trimming of observations renders it hard (if not impossible) to *characterize the population* to which the results generalize to. The issue of limited overlap grows for a growing number of treatment time points T. The authors avoid to directly address that by considering only year-long intervals.

Even though dropping observations with extreme propensity scores in PENCOMP leads to a change in the target population (and therefore the causal estimand), ZEL motivated their approach as a way to address the imprecision of weighting estimators. However, we believe that extreme propensity scores, large weights, and imprecise weighting estimators should be viewed as an indication of limited overlap, where at least one of the treatment levels is unlikely to be assigned to units with a specific covariate profile. Therefore, imprecision of estimators should not be viewed as lack of efficiency, rather than as a representation of *limited information about the treatment effect* for a subset of the target population.

The issue of lack of overlap is particularly hard to tackle in longitudinal treatments. Positivity of all possible treatment vectors, rather than a simple binary treatment, for all units in the sample is likely to be violated, leading to estimands representing the whole population with limited interpretability. Estimands like the ATE of head-to-head comparisons of treatment vectors could lead to vast generalizability doubts when part of the population is unlikely to receive one or both of the compared treatment vectors. Li, Morgan, and Zaslavsky (2018) provide a step—the overlap weights—toward defining estimands over populations resembling an interpretable part of the sample, namely, the subpopulation with the most overlap in observed covariates. The overlap weights are proportional to the probability of being assigned to the opposite group (i.e., $e(x)$ for control units and $1 - e(x)$ for treated units), and thus automatically avoid the high variance problem of IPTW. Indeed, the overlap weights minimize the asymptotic variance of the nonparametric estimator of the weighted ATE among the general class of balancing weights (which includes the IPTW and trimming as special cases). More importantly, as the overlap weights move the focus to the target population of the most overlap, arguably they lead to more generalizable causal estimates that are replicable across datasets. In summary, we feel that exploring alternative weighting schemes (and thus changing the target estimands) is a promising direction in causal inference with longitudinal treatments.

2. Double-Robustness

In observational studies, the true propensity score and outcome models are unknown and need to be estimated. Double-robustness is an attractive property of an estimator because it arguably gives the analysts two chances to "get it right." The AIPTW estimator is doubly robust in the sense that it is consistent if either the outcome model or the propensity score model is correctly specified, but not necessarily both. Double-robustness in PENCOMP is more complicated. Specifically, PENCOMP postulates the outcome model as the sum of two components: (a1) a nonparametric function (e.g., penalized splines) of the propensity score, and (a2) a parametric function of the propensity score and covariates. PENCOMP is doubly robust in the sense that it is consistent if the outcome model is correctly specified, or both the propensity score model and the outcome–propensity relationship is correct. Therefore, correct specification of the propensity score model alone does not guarantee the estimator to be consistent. In this sense, the double-robustness property of PENCOMP is weaker than that of AIPTW, which may render it less attractive in some settings.

Further, modeling the outcome–propensity relationship has important interpretability caveats. Specifically, it is not clear how one can conceive an outcome-generating mechanism that uses the propensity score as a predictor, instead of the covariates directly. Using a non or semiparametric model as splines does provide flexibility, but it does not overcome the challenge of specifying more complex relationships such as covariate–propensity interactions. The reliance of PENCOMP on the outcome–propensity score relationship is evident in the simulation study in ZEL. PENCOMP outperforms AIPTW in situations of "high confounding," that is, large covariate-treatment associations. In such cases, the presence of strong predictors of the treatment intensifies the issue of limited overlap, leading to high variance of weighting methods. Instead, PENCOMP includes the propensity score as a predictor in the outcome model, and implicitly extrapolates based on the outcome–propensity score relationship.

Therefore, PENCOMP depends on the correct specification of at least part of the outcome model. This raises concerns related to the approach's performance in the most prevalent and worst-case scenario that both the outcome and the propensity score model are wrong. Our own experience has suggested that misspecification of the outcome model leads to larger bias in empirical studies (Li, Zaslavsky, and Landrum 2013) than methods based on the propensity score. The reliance of PENCOMP on the outcome–propensity relationship indicates that misspecification of this component (in addition to the rest) may lead to worse results than weighting approaches.

The above discussions lead us to ask: why not directly postulate a flexible nonparametric outcome model such as Gaussian process or tree-based models? Nonparametric estimation leads

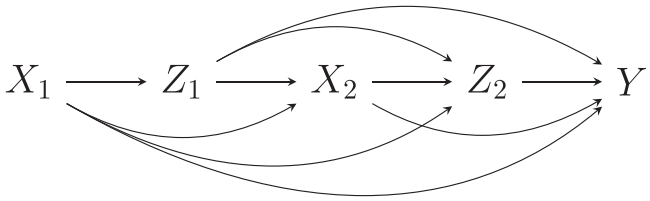

Figure 1. Directed acyclic graph of two point treatment.

to slower rates of convergence if only one model is correctly specified, but it is more likely to lead to consistent estimators. Studies have also shown that adding the estimated propensity score as an additional predictor in the outcome model improves efficiency. Therefore, it might be more attractive to specify the outcome model in PENCOMP as a nonparametric model of covariates and the estimated propensity score. Compared to the current "two-stop" approach of PENCOMP where one has to specify the outcome model as the sum of one parametric and one nonparametric component, we feel that this "one-stop" procedure is more straightforward conceptually and easier to implement practically. Robustness of such estimators is directly supplied by the nonparametric outcome model, while only efficiency is supplied by correctly specified propensity scores.

3. The Form of Sequential Ignorability

The literature on causal inference with longitudinal treatments has nearly exclusively assumed sequential ignorability. To facilitate discussion, Figure 1 shows the casual directed acyclic graph (DAG) of a two time-point treatment. For this DAG, and following ZEL's notation, sequential ignorability states:

Assumption 1 (*Sequential ignorability*).

(a) $\{Y^{z_1 z_2}, X_2^{z_1},$ for all $z_1, z_2\} \perp Z_1 \mid X_1$
(b) $\{Y^{z_1, z_2},$ for all $z_1, z_2\} \perp Z_2 \mid Z_1, X_1, X_2.$

That is, the treatment assignment at time t is ignorable conditional on the *observed* history, which includes the baseline covariates, past treatments, and the observed intermediate outcomes until time $t-1$. Following Assumption 1, analysts usually include the observed history in the propensity score model, and check balance of the distributions of those variables between treated and control units. For example, ZEL checked the standardized difference of means of these variables between treatment groups at each time point; in the IPTW-based marginal structural models, analysts would check the weighted standardized difference of means of these variables.

We stress that sequential ignorability is not the only possible assignment mechanism in longitudinal treatments. An alternative assumption particularly relevant to PENCOMP is the *latent* sequential ignorability assumption proposed in Mattei, Ricciardi, and Mealli (2017), where the conditioning set includes the *potential outcomes* of the intermediate variables.

Assumption 2 (*Latent sequential ignorability*).

(a) $\{Y^{z_1 z_2}, X_2^{z_1},$ for all $z_1, z_2\} \perp Z_1 \mid X_1$
(b) $\{Y^{z_1, z_2},$ for all $z_1, z_2\} \perp Z_2 \mid Z_1, X_1, X_2, \{X_2^0, X_2^1\}.$

Assumption 2 is weaker than Assumption 1 in that the conditioning set in 2 is larger. The joint potential outcomes of the intermediate variable $\{X_2^0, X_2^1\}$ in Assumption 2(b) are essentially principal strata (Frangakis and Rubin 2002). Mattei, Ricciardi, and Mealli (2017) discuss that the latent sequential ignorability assumption allows the treatment assignment to depend on unmeasured factors reflected in the principal strata, which is arguably more plausible in many applications.

PENCOMP explicitly imputes the missing potential outcome (and therefore the principal strata) at all time points. For this reason, we feel one potential advantage of PENCOMP is to assume the weaker latent sequential ignorability than the conventional sequential ignorability. In fact, assuming latent sequential ignorability, analysts could include the imputed intermediate values in both the outcome model and the propensity score model at future time points. Moreover, covariate balance of the units with different observed treatment vectors should be checked based on baseline covariates, the observed intermediate outcome, *and* the potential values of the intermediate variables.

In situations of a treatment applied at many time points, the conditioning sets under both assumptions include the whole treatment and covariate history. However, the conditioning set under latent sequential ignorability further includes the potential values of *all* past intermediate outcomes. Assuming sequential ignorability over latent sequential ignorability could represent an assumption aiming at dimension reduction, since specification of the propensity score and balance check would be performed over a smaller set of variables. If unmeasured confounding is believed to be present, an alternative approach to dimension reduction grounded on latent sequential ignorability could make a Markovian assumption, where the conditional independence at time point t requires only baseline covariates and (observed or potential) intermediate outcomes *at* time $t-1$.

References

Crump, R. K., Joseph Hotz, V., Imbens, G. W., and Mitnik, O. A. (2009), "Dealing With Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187–199. [33]

Frangakis, C. E., and Rubin, D. B. (2002), "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29. [35]

Li, F., Morgan, K. L., and Zaslavsky, A. M. (2018), "Balancing Covariates via Propensity Score Weighting," *Journal of the American Statistical Association*, 113, 390–400. [34]

Li, F., Zaslavsky, A. M., and Landrum, M. B. (2013), "Propensity Score Weighting With Multilevel Data," *Statistics in Medicine*, 32, 3373–3387. [34]

Mattei, A., Ricciardi, F., and Mealli, F. (2017), "Bayesian Inference for Sequential Treatments Under Latent Sequential Ignorability," arXiv no. 1608.07180, pp. 1–28. [35]

Rubin, D. B. (2001), "Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation," *Health Services & Outcomes Research Methodology*, 2, 169–188. [33]