Abstract. A population member is a row of a dataset, or a sample with $n = 1$. A member's propensity score can include additional conditioning on member specification covariates which can collapse the propensity score to 0 if a specification covariate value is incompatible relative to the cause. The causal structure of spacetime, implicit in causal inferences from nonrandomized datasets, suggests impossibility of causal presence if a member's spacetime specification covariate value is outside the future light cone of the earliest relevant spacetime point of the cause. The propensity score with additional conditioning on specification covariates retains all theorems, properties, and extensions of the propensity score with no additional conditioning on specification covariates. Consideration of the propensity score with specification covariates is useful in complicated causal settings where it is difficult to understand which members actually have 0 probability of causal presence. Specification covariates can guide problems of specification by helping determine which members in causal populations should be considered misspecified with an actual propensity score of 0.

Key words and phrases: propensity score, future light cone, problems of specification.

1. INTRODUCTION

Fisher defines problems of specification as "those in which it is required to specify the mathematical form of the distribution of the hypothetical population from which a sample is to be regarded as drawn" [3]. The Glivenko-Cantelli theorem relates the empirical distribution and the hypothetical population implying that the "mathematical form of the distribution of the hypothetical population" is impacted by member-level considerations. A population member is a row of a dataset, or a sample with $n = 1$. If some members comprising the sample dataset of the hypothetical population are upon further consideration removed from the dataset then these considerations are related to the mathematical form studied by the problem of specification. Considerations which can lower relevant sample size can be considerations of population specification.

The propensity score is often used to estimate causal effects from nonrandomized datasets. It is known that members with 0 propensity score should not be included in the causal population [9, Section 12.2.4], but the causal structure of spacetime implies explicit conditioning on member covariates indicating 0 propensity score. The propensity score can include additional conditioning on a set of member covariates called specification covariates which collapse the member's propensity score to 0 if any specification covariate has an incompatible value relative to the cause, for instance if the relevant spacetime of the member is outside the future light cone of the cause. The notion of "relative to the cause" implies that specification covariates are motivated by physical considerations of spacetime causal structure. Explicit conditioning on specification covariates in the propensity score definition can help better specify and analyze members of causal populations.

Section 1 continues with definitions of the propensity score without and with conditioning on specification covariates. Section 2 discusses theoretical aspects of the definition with specification covariates. Section 3 presents the physical motivation based in the causal structure of spacetime. Section 4 suggests some applications of the definition with specification covariates. Section 5 concludes with a general discussion of the role of conditioning in the development of statistical methodology.

The propensity score is the member's probability of causal presence given the set of member covariates: $e(X_i) = P(Z_i = 1|X_i)$ [16]. There are many types of propensity scores [8, 7, 11, 13, 4, 5, 19] but there is consensus that $e(X_i)$ remains the central definition. The theorems, properties, and extensions known of $e(X_i)$ can be retained in a definition which also clarifies the role of the propensity score in causal population specification:

$$e(X_i, S_i) = P(Z_i = 1|X_i, S_i) = \begin{cases} 0, & \prod_{s\in S_i} I(s) = 0 \\ e(X_i), & \prod_{s\in S_i} I(s) = 1, \end{cases}$$

where $I(s)$ for an $s \in S_i$ is an indicator function of specification covariate s equipped with a set of values of s which indicate incompatability relative to the cause and collapse the propensity score to 0. If assumed that the same specification covariates are relevant for each member and that S_i has m elements then there are $n \times m$ indicator functions each defining which specification covariate values are incompatible relative to the cause.

2. DISCUSSION

$e(X_i, S_i)$ collapses a member's probability of causal presence to 0 if any $s \in S_i$ has incompatible value relative to the cause. If the observed causal population satisfies $\sum_{i=1}^{n} \prod_{s \in S} I(s) = n$ then the population is considered to have no misspecification contribution from specification covariates since consideration of all relevant specification covariates does not indicate any member as contributing to misspecification. $\sum_{i=1}^{n} \prod_{s \in S} I(s) \neq n$ could imply estimator bias since some members' outcomes previously used in causal estimation should actually be considered unrelated to the cause. Misspecification from specification covariates also tends to increase causal estimator variance by lowering n. If the population is correctly specified after consideration of specification covariates and the indicator functions then $e(X_i, S_i) = e(X_i)$ since $\sum_{i=1}^{n} \prod_{s \in S} I(s) = n$.

The set X_i contains all member covariates related to confounding and useful for propensity score estimation since the set S_i contains only specification-relevant member covariates. Specification covariates cannot inform issues of confounding since specification covariates indicate unrelatedness between cause and member outcome while confounders, without relevant indicator functions, cannot indicate unrelatedness. Conditioning on S_i is a tool for problems of specification and provides no information about situations where assignment probability is greater than 0; therefore S_i cannot inform propensity score estimation.

The main properties of $e(X_i)$ – such as the being the coarsest balancing score, or unconfoundedness given $e(X_i)$ [16], or the requirement of strict overlap of $e(X_i)$ for efficient semiparametric estimation [6, 10, 1] – are retained by the additional conditioning in the propensity score definition. The main properties of $e(X_i)$ require the assumption of probabilistic assignment – $0 < P(Z_i = 1|X_i, Y_i(0), Y_i(1)) < 1$ where $Y_i(0), Y_i(1)$ are the member's potential outcomes under causal presence and absence [9, Section 12.2] – and S_i seems to provide no additional information about situations where assignment probability is greater than 0. Since S_i is uninformative about situations with nonzero assignment probability the main properties of the extension of $e(X_i)$ to causal settings beyond causal presence and absence [8, 7] are also retained by conditioning on S_i in the propensity score definition. When conditioning on specification covariates S_i it may help to write the propensity score overlap assumption as $e(X_i, S_i) < 1$ to emphasize problems of specification.

3. PHYSICAL MOTIVATION

Considerations from special relativity, especially future light cones of the causal structure of spacetime, motivate additional conditioning with specification covariates in the propensity score definition. The metric of Minkowski spacetime can be written $ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2$ where c is the speed of light, dt is the change in time coordinates between events, and dx, dy, dz are changes in spatial coordinates between events. The Lorentz transformation preserves the invariance of ds^2 between events for observers in different inertial reference frames. Events with $ds^2 < 0$ are called spacelike separated since they cannot have causal connection, events with $ds^2 > 0$ are called timelike separated since they can have causal connection, and events with $ds^2 = 0$ are called lightlike separated since they define light cone boundaries with paths travelled by light. If $ds^2 = 0$ then $dx^2 + dy^2 + dz^2 = c^2dt^2$ is the equation of a sphere representing the paths travelled by light for an elapsed time between events. As the time between events increases the sphere with boundary of light grows but retains its center. Spacelike separation, or the impossibility of causal presence, is classified by the sphere's surface. If the sphere of light at each elapsed time between events is seen in two-dimensions for visual simplicity then the future and past paths of light from an event appear as light cones in opposite directions.

The main physical concept required to motivate specification covariates are future light cones of a cause. In general relativity light cones are curved. The curvature of light cones from gravitation is not a key aspect in motivating specification covariates though curved future light cones may be useful for some types of problems of causal specification. The flat Minkowski spacetime of special relativity can be considered sufficient to motivate specification covariates since Minkowski spacetime is the simplest spacetime in which light cones are causally relevant. Minkowski spacetime is closely related to the mathematical statistics of physical quanta [12, Chapters 3-4] since quantum field theories are Lorentz invariant [2].

The causal structure of spacetime is implicit in all causal inference from nonrandomized datasets since the assignment mechanism is not randomized [18]. Phrases such as natural experiments or observational studies emphasize that datasets in nonrandomized causal inference require spacetime. Any estimation of causal effects from nonrandomized data can assume that the cause can be associated to some earliest spacetime point; before this earliest spacetime point there is no cause which could impact the member's outcome. This earliest spacetime point can be related to an inertial reference frame from which to describe the member's relevant spacetime point. If each cause studied with nonrandomized data has some earliest spacetime point then there is a future light cone describable for any duration from the earliest spacetime of the cause. The existence of a relevant future light cone for each member of the causal dataset implies the member's relevant spacetime point as a value of an element

of a specification covariate set S_i. If $I(s \in S_i) = 0$ then the member is incompatible relative to the cause since the future light cone classifies the member as misspecified. For instance, if the members of a supposed nonrandomized causal population are each located on separate distant galaxies then a cause occurring on earth could not impact any member's outcome for a long duration until the future light cone of the cause entered the galaxy and classified the member's relevant spacetime as timelike separated. The supposed causal population would be fully misspecified with $n = 0$ until $\exists i$ such that $I(s \in S_i) = 1$ where $I(s)$ determines whether the spacetime metric is causally valid.

In most uses of the propensity score it is not essential to make explicit the specification covariate related to future light cones since the relevant spatial distances between cause and member are small. In these cases it is often easily agreeable that $I(s \in S_i) = 1 \forall i$. But causal inference from a nonrandomized dataset requires that the causal structure of spacetime applies to the members of the causal population. Therefore any use of $e(X_i)$ is physically also using $e(X_i, S_i)$ with S_i containing an indicator function of a specification covariate determining whether the relevant spacetime metric is spacelike separated. Since nonrandomized causal inference requires the causal structure of spacetime it seems the propensity score contains the possibility of additional conditioning useful for causal population specification.

4. APPLICATIONS

When the cause of an observational study is complicated it can be difficult to understand which members should not be included in the population specification; $e(X_i, S_i)$ is useful for these types of problems of specification. Each additional specification covariate creates n indicator functions whose product must be 1 for the population to remain correctly specified. Defining each indicator function requires careful consideration of which specification covariate values collapse the propensity score. Previous analysis of nonrandomized causal data might require revision if an argument introducing a relevant specification covariate lowers n by revealing some members with actual propensity score of 0 who were previously included in causal estimation. $e(X_i, S_i)$ could help find large numbers of members who should not be included in the causal population. Each indicator function of a specification covariate expresses a relationship between the member and causal presence. Some expressions of these relationships can provide new perspectives on previously complicated causal specification problems. The possibility of new perspectives about members' causal presence from use of specification covariates implies situations where a large proportion of members in a supposed causal population actually have an incompatible specification covariate value relative to the cause. The addition of a specification covariate or modification of its indicator functions, if capturing a previously important overlooked aspect of members' relations to the cause, could result in a large decrease of n during causal effect re-estimation. In observational studies with interference classifying a single member as misspecified could significantly impact network specification covariate values resulting in cascades of members classified as misspecified.

Understanding which members should be considered compatible relative to the cause can clarify causal structure. The causal structure expressed by specification covariates is a different type than structural aspects of causal identification diagrams [14, 15] since specification covariates consider causal structure through an emphasis on member-level indicator functions, and specification covariates are focused on problems of specification while causal identification diagrams are focused on identification for problems of causal estimation. Problems of specification are related to problems of identification since modifying the causal population can invalidate identification of a previously identifiable causal estimand. There may be similarities between the view of causal population specification as stemming from member-level causal structure and Fisher's suggestion that "any set of independent measurements ... from an infinite population ... are a random sample from the totality of numbers produced by the same matrix of causal conditions" [3, Section 2]. The structures expressed by conditioning on specification covariates are causal since they specify causal populations, and they are member-level causal structures since they classify members as correctly or incorrectly specified.

When using $e(X_i, S_i)$ to discuss causal population specification in complicated applied problems the most useful estimator class to guide discussion may be inverse propensity weighting estimators. $e(X_i, S_i)$ can be helpful when the causal setting of the problem is so complicated that it is nontrivial to classify which members have propensity score 0. Inverse propensity weighting estimators express most clearly that if a single member is included in estimation with an actual propensity score of 0 then the estimation is invalid since the member's outcome cannot be causally weighted and is undefined from division by $e(X_i, S_i) = 0$. The severity of the possibility of division by 0 from misspecification may imply that causal estimators requiring development with $e(X_i, S_i)$ must show theoretical relation to inverse propensity weighting estimators.

Varying specification covariate values and varying collapsing-values of indicator functions can lead to different sets of members of a supposed causal population that actually satisfy $e(X_i, S_i) = 0$. Variation in member specification sets leads to variation in causal estimates since

the outcome distribution and n can differ between member specifications sets. Studying choices of specification covariates and indicator functions which lead to stable causal estimates [20, 17] may help in problems of specification. Modeling uncertainty about specification covariates and uncertainty about the collapsing-action of the indicator functions may help investigate causal stability.

5. CONCLUSION

Applied statisticians using the propensity score for analysis of nonrandomized data implicitly condition on the specification covariate of the member's relevant spacetime; spacelike separation from the cause implies impossibility of causal connection and violates the assumption of probabilistic assignment. In most problems the indicator functions classifying spacelike separation will trivially be 1, but observational studies require the causal structure of spacetime so any consideration of the propensity score conditioning on member covariates X_i also conditions on at least the member spacetime specification covariate S_i with future light cone indicator function.

Explicit conditioning is recommended in development of statistical methodology [13, Section 2.2]. Beyond X_i and S_i there may be statistical arguments supporting the inclusion of further additional explicit conditioning in the propensity score definition. Statistical methodology seems to always allow for additional explicit conditioning which retains previous theory. Physical theories are open to development because of the complexity of physical phenomena; perhaps statistical methodologies are open to development because of the generality of the notion of conditioning coupled with the complexity of phenomena. Perhaps in observational studies additional explicit conditioning can be based in physical argument.

REFERENCES

[1] D'AMOUR, A., DING, P., FELLER, A., LEI, L., and SKHON, J. (2021). Overlap in observational studies with high-dimensional covariates. *J. Econom.* **221**, 644-654.

[2] DIRAC, P. A. M. (1928). The quantum theory of the electron. *Proc. Math. Phys. Eng. Sci.* **117**, 610-624.

[3] FISHER, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc.* **222** 309-368.

[4] FORASTIERE, L, AIROLDI, E. M., and MEALLI, F. (2021). Identification and Estimation of Treatment and Interference Effects in Observational Studies on Networks. *J. Am. Stat. Assoc.* **119**, 901-918.

[5] FORASTIERE, L., DEL PRETE, D., and SCIABOLAZZA, V. L. (2024). Causal Inference on Networks under Continuous Treatment Interference. *Soc. Netw.* **76**, 88-111.

[6] HIRANO, K, IMBENS, G. W., and RIDDER, G. (2003). Efficient Estimation of Average Treatment Effects using the Estimated Propensity Score. *Econometrica* **71**, 1161-1189.

[7] HIRANO, K. and IMBENS, G. W. (2004). The propensity score with continuous treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. Wiley, 73-84.

[8] IMBENS, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87**, 706-710.

[9] IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge.

[10] KHAN, S, and TAMER, E. (2010). Irregular Identification, Support Conditions, and Inverse Weight Estimation. *Econometrica* **78**, 2021-2042.

[11] LEUNG, M. P. and LOUPOS, P. (2023). Graph Neural Networks for Causal Inference Under Network Confounding.

[12] VON NEUMANN, J. (1955). *Mathematical Foundations of Quantum Mechanics.* Princeton University Press, Princeton.

[13] OGBURN, E., SOFRYGIN, O. , DÍAZ, I. and VAN DER LAAN, M. J. (2022). Causal Inference for Social Network Data. *J. Am. Stat. Assoc.* **119**, 597-611

[14] PEARL, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press, Cambridge.

[15] PEARL, J. 2012. The Causal Foundations of Structural Equation Modeling. Technical report, California Univ Los Angeles Dept of Computer Science.

[16] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41-55.

[17] ROTHENHÄUSLER, D., MEINSHAUSEN, N., BÜHLMANN, P., PETERS, J. (2020). Anchor Regression: Heterogeneous Data Meet Causality. *J. R. Stat. Soc. Series B Stat. Methodol.* **83**, 215-246.

[18] RUBIN, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *Ann. Stat.* **6**, 34-58.

[19] TCHETGEN, E. J. T., and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Stat. Methods Med. Res.* **21**, 55-75.

[20] YU, B. (2013). Stability. *Bernoulli* **19**, 1484-1500.