Observational Studies 9.1 (2023): 119-124 Published Jan 2023

Bridging the design and modeling of causal inference: A Bayesian nonparametric perspective

Xinyi Xu **xinyi@stat.osu.edu**
Steven N. MacEachern **snm@stat.osu.edu**
Department of Statistics, College of Arts and Sciences
Bo Lu **lu.232@osu.edu**
Division of Biostatistics, College of Public Health
The Ohio State University
Columbus, OH 43210, USA

Abstract

In their seminal paper first published 40 years ago, Rosenbaum and Rubin crafted the concept of the propensity score to tackle the challenging problem of causal inference in observational studies. The propensity score is set up mostly as a design tool to recreate a randomization like scenario, through matching or subclassification. Bayesian development over the past two decades has adopted a modeling framework to infer the causal effect. In this commentary, we highlight the connection between the design- and model-based perspectives to analysis. We briefly review a Bayesian nonparametric framework that utilizes Gaussian Process models on propensity scores to mimic matched designs. We also discuss the role of variation as well as bias in estimators arising from observational data.

Keywords: Propensity Score, Prognostic Score, Gaussian Process, Heterogeneous Treatment Effects

1. Introduction

Four decades ago, prevailing methods to address causality centered on designed experiments and randomized trials. In both settings, inference about causality was driven by the presence of randomization to control for the effect of confounders on treatment assignment. Rosenbaum and Rubin elevated observational studies to the front stage by crafting a simple but beautiful concept – the propensity score – and by laying out the assumptions on treatment assignment ignorability needed to make causal inferences (Rosenbaum and Rubin, 1983).

As Rubin (1974) and Holland (1986) pointed out, the fundamental problem in causal inference is the missingness of potential outcomes. In the context of treatment and control groups, we observe only one of the two potential outcomes. Causal effect estimation based on conventional regression models and observational data makes implicit, untestable assumptions about the relationship between the treatment and the unobserved potential outcomes. A particular point of concern is apparent bias introduced through the relationship between the observed covariates and treatment assignment. If there is apparent bias in the observed covariate-treatment relationship, it is natural to believe that there is bias in the unobserved covariate-treatment relationship, that there are important unmeasured

confounders, and that naive estimators of the causal effect of treatment will have substantial bias.

The propensity score attempts to fix the broken link of biased treatment assignment, which is due to the lack of randomization. With this focus, most conventional uses of the propensity score focus on design-based inference, including matching, subclassification, and weighting methods. Technically, the two key features of the propensity score are covariate balancing and dimension reduction. As laid out in Rosenbaum and Rubin's paper, the propensity score is a balancing score, which makes the entire distribution of all measured covariates comparable between treated and control groups, as would be achieved with randomization. To move further and obtain a causal interpretation, Rosenbaum and Rubin require the treatment assignment ignorability assumption. Loosely speaking, this follows when the conditional distribution of unmeasured covariates given measured covariates is the same in treatment and control populations. More importantly, the same ignorability also holds if conditioning only on the propensity score (a balancing score), which turns the high-dimensional adjustment problem into a scalar one, which opens the door for more refined modelling options.

2. From Design to Modeling

When working with the propensity score (denoted as $e(X)$ hereafter), one could consider both design-based inference (e.g., matching methods) and model-based inference techniques. At first glance, the two types of inference appear to be very different, but there is an interesting relationship between them. Under appropriate conditions, matching-based inference arises as a limiting case of model-based inference. Consideration of the conditions that lead to this agreement suggests when design-based inference will suffice and when inference may be improved by utilizing a model-based framework. Briefly, the model-based framework, especially a Bayesian nonparametric framework, allows us to separate various parts of the inference problem to obtain more accurate estimates.

When participants are naturally matched on the propensity score, a model may also be constructed directly for the differences. The decision to model only the difference but not the members of the pair yielding the difference has a long history in Bayesian statistics where reducing the scope of modelling is tied to robustness of the model (e.g., Box and Tiao (1962)). The possible nonlinearity of mean difference in $e(X)$ motivates the use of a flexible functional form for the regression on $e(X)$. The propensity $e(X)$ effectively plays the role of a single index, resulting a single index regression.

Exact matches on the propensity score are rare, especially if the underlying covariates are continuous. In this case, one cannot pass to the differences within pairs, and so there is a need to model both treatment and control response surfaces. Details of these models will vary. One natural strategy is to identify one group (typically control) for which it is relatively easy to build a model, to model this response surface and to simultaneously model the difference in response surfaces between two treatment groups. The dimension reduction from X to $e(X)$ greatly simplifies the effort. A second strategy may target the average of the means for treatment and control and the difference between the means. A key part of both strategies is the decision of how smooth the response surfaces should be.

3. Propensity Score-based Gaussian Process Model

This section provides a brief overview of our propensity score-based Gaussian Process model, which provides a flexible semiparametric model while allowing for heterogeneous treatment effects. More details can be found in Xin et al. (2022).

Suppose an observational dataset contains (Y_i, X_i, T_i), where Y_i is a response variable, X_i is a vector of observed covariates and T_i is a binary treatment indicator variable for the ith individual, $i = 1, \cdots, n$. Let $T_i = 1$ denote the treated group and $T_i = 0$ denote the control group. Our development uses the following nonparametric conditional mean regression to flexibly model the outcomes:

$$Y_i = g(X_i, T_i) + \varepsilon_i,$$

where for ease of illustration, the random error ε_i is assumed to be normally distributed with mean zero and unknown variance σ^2. Under the ignorability assumption, the potential outcomes depend on the covariates only through the propensity score through an argument similar to that in Wu et al. (2021). Therefore, we study the simplied form of the regresion based entirely on the propensity score

$$g(X_i, T_i) = f(e(X_i)) + \Delta(e(X_i)) \times T_i, \tag{1}$$

where the function f captures the response surface for the control group and the function Δ captures the treatment effect, that is, the difference between the treatment and control surfaces. Note that the treatment effect may be heterogeneous with respect to $e(X)$.

We assume that these functions are continuous and quantify our uncertainty about them with Gaussian process (GP) priors. That is, letting $e = (e(X_1), \cdots, e(X_n))$,

$$f(e) \sim \mathcal{GP}(\mu_f(e), K(e, e')) \qquad \text{and} \qquad \Delta(e) \sim \mathcal{GP}(\mu_\Delta(e), K(e, e')) \tag{2}$$

where $\mu_f(\cdot)$ are $\mu_\Delta(\cdot)$ are the mean functions and $K(\cdot, \cdot)$ is the covariance function. Let the mean functions take a flexible polynomial form, namely, $\mu_f(e) = H(e)^T\alpha$ and $\mu_\Delta(e) = H(e)^T\beta$, where $H(e)^T = (1, e, e^2, e^3)^T$. Also, let the covariance function $K(e, e')$ be of squared-exponential form, namely,

$$K(e, e') = \gamma^2 \exp(-\frac{1}{2l^2}|e - e'|^2),$$

where $|e - e'|^2$ is the squared Euclidean distance between e and e'. The parameter γ determines the magnitude of the departure between the nonparametric functions and their mean functions, and the parameter l governs the local dependence, with smaller l corresponding to more wiggliness in $f(e)$. Furthermore, we place weakly informative normal priors on the hyperparameters α and β, that is, $\alpha \sim N(0, B_\alpha)$ and $\beta \sim N(0, B_\beta)$. The parameters γ, l, σ^2 and the functions f and Δ can be updated through an MCMC algorithm. Since the priors are conjugate, the computation can be done with a Gibbs sampler and is very efficient.

The *Bayesian estimator* of the average treatment effect under the average squared error loss is

$$\hat{\Delta}^{GP} = \frac{1}{n}\sum_{i=1}^{n} \tilde{\mu}_\Delta(e(X_i)),$$

where $\tilde{\mu}_\Delta(e(X_i))$ is the posterior mean of the treatment effect function Δ at $e(X_i)$. To understand the performance of our semiparametric GP approach, we compared it with the conventional propensity score matching method. When the treated and control subjects are 1:1 matched with respect to X (or $e(X)$), the *exact matching estimator* is:

$$\hat{\Delta}^m = \frac{1}{n}\sum_{i=1}^{n}[Y_{i,t} - Y_{i,c}],$$

which is unbiased for the population average treatment effects provided the matched sample is simple random sample from the desired population.

In Xin et al. (2022), we show that if the prior mean function for Δ is a constant with a diffuse normal prior, i.e.,

$$\mu_\Delta(e) \equiv \beta \sim N(0, \tau^2),$$

then

$$\hat{\Delta}^{GP} \to \hat{\Delta}^m, \qquad \text{as } \tau^2 \to \infty \text{ and } l \to 0.$$

This result provides an important connection between the Bayesian modeling approach and the propensity score matching approach. The limiting $\tau^2 \to \infty$ corresponds to a "flat" mean prior and $l \to 0$ corresponds to "locally independence" of the treatment effect at different propensity scores. Thus, the corresponding Bayesian estimator behaves as the the matching estimator. However, in practice, there is little or no reason to believe that this limiting prior is sensible. One implication of the prior is that the analyst has *no* information about the magnitude of the treatment effect. A second implication is that the treatment effect at a given propensity tells us nothing about the treatment effect at nearby propensities. This latter implication destroys the motivation for the use of approximate matches on the propensity score. Rather than working with a uniform prior exhibiting local independence, we prefer to to incorporate prior knowledge for both mean function and covariance function. When we do so, the Bayesian estimator differs from the matching estimator, sometimes substantially.

4. Further Refinement with Prognostic Scores

Rosenbaum and Rubin focused on estimation of the population average treatment effect, the difference in treatment and control units, averaged across a particular distribution of the observed covariates. Exact matching on any balancing score yields unbiased estimates for the average treatment effect. A more complete look at the estimator may also consider its variance.

The most refined balancing score available is X itself, which is usually of high dimension for real problems. Consider any low dimensional balancing score $b(X)$. Following the usual decomposition of the variance, we can represent the conditional variance of the treatment effect Δ by

$$\begin{aligned} V[\Delta|b(X)] &= E[V(\Delta|b(X), X)|b(X)] + V[E(\Delta|b(X), X)|b(X)] \\ &= E[V(\Delta|X)|b(X)] + V[E(\Delta|X)|b(X)]. \end{aligned} \tag{3}$$

The second line of (3) follows from X being a refinement of $b(X)$. Maximal variance reduction due to pairing occurs when $E(\Delta|X)$ shows no variation, conditional on $b(X)$.

The propensity score is the coarsest balancing score and is effective at removing bias from the estimator. However, when the causal effect is heterogeneous, matching on a refinement of the propensity score may reduce the variance of the estimator. A natural approach is to add the prognostic score, which is a balancing score and contains outcome information (Hansen, 2008). Leacy and Stuart (2014) incorporated a prognostic score into a matching procedure in addition to propensity score, and showed that this would improve inference over matching solely on the propensity score. Let the propensity score of the ith individual be $e_i = e(X_i)$, the prognostic score $\psi_i = \psi(X_i)$ and $s_i = (e_i, \psi_i)$. Adding the prognostic score to the propensity score in our nonparametric regression model yields

$$Y_i = f(s_i) + \Delta(s_i) \times T_i + \epsilon_i, \qquad i = 1, \ldots, n \tag{4}$$

The prognostic score can be one-dimensional (e.g., just for subjects under control) or two-dimensional (e.g., subjects under treatment and subjects under control). Even with two prognostic scores, the balancing score is a set of three scores (including the propensity score), which is still of low dimension compared to the (typically) high dimensional covariate space. This would substantially simplify the nonparametric modeling. For real data analysis, practitioners still need to make many important decisions regarding the modeling. For example, how smooth the response surface of the difference in means is. Similarly, the choice of whether to include strong prior information on the magnitude of the treatment effect and the variation in treatment effect across covariate values is up to the analyst. Finally, the propensity score and prognostic score may be highly correlated, in which case some orthogonization may be desirable before including both scores in the model.

Acknowledgments

We acknowledge support for this project on grant DMS-2015552 from the National Science Foundation.

References

G.E.P. Box and G.C. Tiao. A further look at robustness via Bayes's theorem. *Biometrika*, 49:419–432, 1962.

B.B. Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95:481–488, 2008.

P. W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81:945–970, 1986.

F. Leacy and E. A. Stuart. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: A simulation study. *Statistics in Medicine*, 33:3488–3508, 2014.

P.R. Rosenbaum and D.B. Rubin. The central role of propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

P. Wu, X. Xu, X. Tong, Q. Jiang, and B. Lu. A propensity score-based spline approach for average causal effects. *Journal of Statistical Planning and Inference*, 212:153–168, 2021.

Y. Xin, B. Lu, S.N. MacEachern, L. Wang, X. Xu, and R. Zhang. Heterogeneous Causal Effects Estimation via Semiparametric Bayesian Models. Technical report, The Ohio State University, Department of Statistics, 2022.