Published in final edited form as:

Stat Methods Med Res. 2015 December ; 24(6): 1003–1008. doi:10.1177/0962280212437451.

Consistent causal effect estimation under dual misspecification and implications for confounder selection procedures

Susan Gruber¹ and Mark J van der Laan²

¹Department of Epidemiology, Harvard School of Public Health, 677 Huntington Avenue, Kresge 820, Boston, MA, USA

²Division of Biostatistics, School of Public Health, University of California at Berkeley, 101 Haviland Hall, Berkeley, CA, USA

Abstract

In a previously published article in this journal, Vansteeland et al. [Stat Methods Med Res. Epub ahead of print 12 November 2010. DOI: 10.1177/0962280210387717] address confounder selection in the context of causal effect estimation in observational studies. They discuss several selection strategies and propose a procedure whose performance is guided by the quality of the exposure effect estimator. The authors note that when a particular linearity condition is met, consistent estimation of the target parameter can be achieved even under dual misspecification of models for the association of confounders with exposure and outcome and demonstrate the performance of their procedure relative to other estimators when this condition holds. Our earlier published work on collaborative targeted minimum loss based learning provides a general theoretical framework for effective confounder selection that explains the findings of Vansteelandt et al. and underscores the appropriateness of their suggestions that a confounder selection procedure should be concerned with directly targeting the quality of the estimate and that desirable estimators produce valid confidence intervals and are robust to dual misspecification.

Keywords

collaborative double robustness; TMLE; collaborative targeted maximum likelihood estimation; propensity score; confounder selection; causal effect; causal inference; dual misspecification

1 Introduction

In a statistical analysis of observational data, a number of events, including differential selection into exposure groups, informative treatment switches, and drop-out over time, can bias causal effect estimates if not appropriately handled. Moreover, unless one is willing to rely on unstable modeling assumptions, there must be experimentation within strata defined by combinations of covariates causally related to both treatment and outcome (confounders) in order to adjust a causal effect estimate in a manner that reduces bias. A finite sample from an observational study may contain borderline-sufficient information for

identifying the desired causal effect. An exposure effect estimate from such a dataset will tend to be highly variable and often remains biased. Confounder selection is thus an especially important issue in causal inference when there is sparsity in the data,¹⁻⁶ and estimator performance depends on employing a principled strategy. A theme running through our previous work on targeted minimum loss based learning and targeted maximum likelihood estimation (TMLE) is that estimation procedures should be tailored to provide high-quality answers to questions of scientific interest. From a statistical perspective, this means making a bias variance trade-off that is targeted to yield maximally efficient, unbiased estimation of a parameter of a statistical distribution that provides an answer to the scientific research question.⁷⁻¹¹

2 Collaborative double robust estimation

Double robust (DR) estimators solve an estimating equation defined by a gradient of the pathwise derivative of the target parameter viewed as mapping from the statistical model to the parameter space. In particular, if the estimating equation corresponds with the so-called canonical gradient (also called the efficient influence curve), then these DR estimators are also tailored to be asymptotically efficient. These estimators have been shown to be consistent for coarsened at random data structures when either the full data distribution (Q_0) or censoring mechanism (g_0) is consistently estimated.¹²⁻¹⁴ In a simple binary point treatment (exposure) setting where the data consists of n independent and identically distributed copies of data structure $O = (W, A, Y)$ drawn from joint probability distribution $P_0 = (Q_0, g_0)$, g_0 corresponds to the conditional distribution of treatment indicator A , given baseline covariate vector W , (i.e. the conditional propensity score distribution), and Q_0 factorizes into the conditional distribution of outcome Y , given A and W and the distribution of W , ($Q_0 = (Q_{0Y}, Q_{0W})$). This observed data can be viewed as a missing data structure $Y = (W, A, Y = Y(A))$ on the full data $X = (W, Y(0), Y(1))$, and one might assume the randomization assumption, $A \perp X \mid W$, so that target parameters of Q_0 can be interpreted as causal effects.

Consider the additive treatment effect (ATE) target parameter, defined non-parametrically as $E_0(Y(1) - Y(0))$. This causal quantity is identified by the statistical mapping $\psi(P_0) = E_0(E_0(Y \mid A = 1, W) - E_0(Y \mid A = 0, W))$ defined on a non-parametric statistical model which maps the probability distribution to a real number. An asymptotically linear estimator has an influence curve that describes the behavior of the estimator under perturbances in the empirical distribution of the data. Among all the influence curves generated by the class of regular asymptotically linear estimators, the one with the minimum variance is known as the efficient influence curve $D^*(P)$. The efficient influence curve can be calculated for any given target parameter mapping: IR and statistical model (i.e. class of probability distributions), $\mathcal{M}$, at any $P \in \mathcal{M}$. An estimator is efficient at P if and only if it is asymptotically linear with an influence curve equal to $D^*(P)$. Continuing our example, the efficient influence curve for the ATE parameter is given by

$$D^*(P) = \frac{2A - 1}{g(A|W)} \left(Y - \bar{Q}(A, W) \right) + \bar{Q}(1, W) - \bar{Q}(0, W) - \Psi(Q)$$

where $\bar{Q}(A, W) = E_P(Y|A, W)$ and $g(1|W) = P(A=1|W)$.

All DR estimators based on D^* solve $P_n D_n^* \equiv 1/n \sum_{i=1}^n D_n^*(O_i)$ in some way. For example, an estimating equation approach defines ψ_n the solution of $P_n D^*(Q_n, g_n, \psi) = 0$ in ψ for given estimators Q_n, g_n , (where the subscript n indicates an estimate of the truth). A targeted minimum loss based estimator $\Psi(Q_n^*)$ (TMLE) involves constructing an estimator Q_n^* of Q_0 that also satisfies the equation $P_n D^*(Q, g_n, \psi(Q)) = 0$ in Q . TMLEs are substitution estimators $\Psi(Q_n^*)$ obtained by plugging in a targeted estimate Q_n^* of Q_0 in the parameter mapping. By construction of the TMLE Q_n^* , the linear span of the score equations solved by TMLE includes the efficient influence curve estimating equation, which explains why the double robustness result also applies to TMLEs.

In this setting, a DR estimator is consistent if either the outcome regression

$\bar{Q}_0(A, W) = E_0(Y|A, W)$ or the treatment assignment mechanism $g_0(1|W) = P(A=1|W)$ is consistently estimated. Beyond this, we have previously shown that for estimators satisfying $P_n D^*(\psi_n, Q_n, g_n) = 0$, given a limit Q of Q_n , there exist a specified set of possible limits g of g_n for which this estimator ψ_n remains consistent for ψ_0 .^{8,10} Let $\mathcal{G}(Q, P_0)$ be the set of all conditional distributions satisfying this condition: that is, for each data distribution P_0 , and Q , we define $\mathcal{G}(Q, P_0) = \{g: P_0 D^*(Q, g, \psi_0) = 0\}$ as the candidate censoring/treatment mechanisms that would result in an unbiased estimating function for the target ψ_0 . At a minimum, this set of conditional distributions, $\mathcal{G}(Q, P_0)$, contains g_0 . It also contains any additional conditional distribution that is sufficient for removing residual bias in the estimate. For example (Theorem 2, van der Laan and Gruber⁸), if residual bias

$(\bar{Q}_0(A, W) - \bar{Q}(A, W)) = f_0(A, W(Q))$ only depends on W through $W(Q)$ and $g_s(Q)$ is a conditional distribution of A , given $W(Q)$ (or more), then $P_0 D^*(Q, g_s(Q)) = 0$ and thus $g_s(Q) \in \mathcal{G}(Q, P_0)$.

In addition, for the ATE parameter $E(Y(1) - Y(0))$, a conditional distribution of A , given $S(W)$ with

$$H_g(\bar{Q} - \bar{Q}_0 \equiv (\bar{Q} - \bar{Q}_0)(1, W)/g(1|W) + (\bar{Q} - \bar{Q}_0)(0, W)/g(0|W))$$

being a function of $S(W)$, is also an element of $\mathcal{G}(Q, P_0)$. In fact, our general result presented in a paper⁸ in 2010 and described below, applied to this example shows that we just need

that g solves the single score equation $P_0 H_g(\bar{Q} - \bar{Q}_0)(W)(A - g(1|W)) = 0$, A would which would be solved by a logistic regression with offset $\text{logit}(g)$ and clever covariate

$$H_g(\bar{Q} - \bar{Q}_0).$$

A DR relying on (Q_n, g_n) is asymptotically unbiased when g_n will converge to an element in $\mathcal{G}(Q, P_0)$, with Q being the limit of Q_n , but the finite sample efficiency of the estimator of ψ_0 varies with the choice of estimator g_n . This fundamental collaborative double robustness

of the efficient influence curve has important implications for nuisance parameter estimation procedures, which should be tailored for effective estimation of the parameter of scientific interest.

In previous papers inspired by this collaborative double robustness of the efficient influence curve, we presented an estimator within the targeted minimum loss based estimation (TMLE) framework that we refer to as a collaborative targeted maximum likelihood estimator (C-TMLE).^{8,10} We use the term *collaborative* to draw attention to the fact that the fits for the outcome regression and the propensity score work in tandem to achieve a full bias reduction for the target parameter. Specifically, candidate updates of a fit of the propensity score (e.g. corresponding with adding a variable to the model for the propensity score) are evaluated by the goodness of fit of the corresponding targeted maximum likelihood update of the current estimator of Q_0 . In this manner, g_n is indeed constructed in response to residual bias $Q_n - Q_0$, so that g_n is aimed to converge to an element in $\mathcal{G}(Q, P_0)$.

In our previous work, we gave a general characterization of $\mathcal{G}(Q, P_0)$ as follows.⁸ In the case where the efficient influence curve $D^*(P)$ can be represented as $D^*(\psi, Q, g)$, the efficient influence curve estimating equation for ψ at a (Q, g) is given by $P_0 D^*(\psi_0, Q, g) = 0$. Classical double robustness theory tells us that this equation is solved at the true ψ_0 when $Q = Q_0$, at some $g \neq g_0$ or if $g = g_0$ at any Q . For consistency of an estimator ψ_n solving $0 = P_n D^*(\psi_n, Q_n, g_n)$, we require that the limits (Q, g) of (Q_n, g_n) satisfy $P_0 D^*(\psi_0, Q, g) = 0$. Equivalently, we can write

$$P_0 [D^*(\psi_0, Q, g) - D^*(\psi_0, Q_0, g)] = 0 \quad (1)$$

Recall that the efficient influence curve can be decomposed as $D^*(\psi, Q, g) = D_{IPTW}(\psi, g) - D_{CAR}(Q, g)$ in terms of an inverse probability of treatment weighted (IPTW) estimating function and a score D_{CAR} of g in the model that only assumes coarsening at random¹² (see Theorem 1.3 in van der Laan and Robins¹⁴). Here, D_{CAR} is a function of O with conditional mean zero, given full data X , and $Q \rightarrow D_{CAR}(Q, g)$ is linear in Q . For many statistical models and target parameters, this representation of the efficient influence curve exists. Substituting this representation into equation (1) yields,

$$P_0 [D_{IPTW}(\psi_0, g) - D_{CAR}(Q, g) - (D_{IPTW}(\psi_0, g) - D_{CAR}(Q_0, g))] = 0$$

or equivalently, $P_0[D_{CAR}(Q_0 - Q, g)] = 0$. Thus, $P_0 D_{CAR}(Q_0 - Q, g) = 0$ implies $P_0 D^*(\psi_0, Q, g) = 0$, and thereby under regularity conditions, any estimator, ψ_n , that solves $P_n D^*(\psi_n, Q_n, g_n) = 0$ for (Q_n, g_n) converging to (Q, g) satisfying $P_0 D_{CAR}(Q - Q_0, g) = 0$ will be consistent for ψ_0 (see Theorem 1 in van der Laan and Gruber⁸). In particular, we can define $\mathcal{G}(Q, P_0) = \{g: P_0 D_{CAR}(Q - Q_0, g) = 0\}$, which in practice is saying that the estimator g_n needs to approximately solve the score equation $P_n D_{CAR}(Q_n - Q_0, g_n)$ so that in the limit $P_0 D_{CAR}(Q - Q_0, g) = 0$. Note that in our additive effect example, we have that

$$D_{CAR}(Q - Q_{0,g}) = H_g \left(\bar{Q} - \bar{Q}_0 \right) (A - g(1|W)).$$

3 Causal effect estimation

If we now turn our attention to the estimation of causal effects, knowledge of the collaborative double robustness property helps us to understand why even for non-DR estimators (e.g. estimators that solve the efficient influence curve equation $P_n D^*(Q, g_n, \psi_n) = 0$ at an intentionally misspecified Q , such as $Q = 0$), the likelihood for g is *not* the most relevant guide for selecting confounders into propensity score models. Predictors of treatment are not necessarily strong predictors of the outcome, and because the goal is to achieve an optimal bias/variance trade-off for the target parameter, the mean squared error for the target parameter should factor into confounder selection for any estimation procedure. Other researchers have reached a similar conclusion and suggest propensity score estimators are best evaluated with respect to their effect on estimation of the causal effect of interest, not by metrics such as likelihoods or classification rates.¹⁵⁻¹⁷ Vansteelandt et al.¹⁸ propose a stabilized propensity score estimator and report a limited set of conditions under which consistent estimation of a marginal treatment effect is possible even when Q and g are both misspecified. We recognize this as a specific instance of collaborative double robustness.

Section 3.2 of Vansteelandt et al.¹⁸ focuses on a space of semi-parametric models of the form $Y = \beta A + \theta(W) + \varepsilon$, $E(\varepsilon | A, W) = 0$, and a target parameter β . If the conditional variance of Y , given A, W , only depends on W , the efficient influence curve for this parameter is upto a standardizing constant given by

$$D^*(\theta, g, \beta) = (A - g(W))(Y - \beta A - \theta(W)),$$

where $g(W) = E(A|W)$ (for formal derivations and theorems, see, e.g. Yu and van der Laan¹⁹ and Robins and Rotnitzky²⁰). Given θ , we define a set $\mathcal{G}(\theta, P_0)$ of conditional distributions, g , that satisfy $P_0 D^*(\theta, g, \beta_0) = 0$, where β_0 is the true parameter value. Recall the decomposition $D^*(\theta, g, \beta) = D_{IPTW}^*(g, \beta) + D_{CAR}^*(\theta, g)$, where $D_{CAR}^*(\theta, g)$ is an element of the tangent space of A , given W . The equation for D_{CAR} in this semi-parametric model is given by $D_{CAR}(\theta, g) = (A - g(1 | W)) \theta(W)$ (page 24 of the original paper⁷ on TMLE). Thus, $\mathcal{G}(\theta, P_0)$ contains all conditional distributions g such that $P_0(A - E(A|W))(\theta - \theta_0)(W) = 0$. If, for example, g is fitted with logistic regression with covariate $\theta - \theta_0$, (with θ_0 being the truth), then this remains an unbiased estimating function in. In the special case described in Vansteelandt et al.¹⁸ that P_0 is restricted such that $\theta(W) = \gamma(W)$, with $\gamma(W)$ linear, and g is fitted with logistic linear regression using W , the estimated g_n is asymptotically a member of $\mathcal{G}(\theta, P_0)$ for all θ linear in W , including $\theta = 0$. This special case of collaborative double robustness corresponds exactly with the insight provided in Section 3.2 of Vansteelandt et al.¹⁸

A general C-TMLE introduced in our earlier papers has been implemented and applied to point treatment and longitudinal data.^{21,22} The development of a targeted forward selection algorithm to select covariates to include in the propensity score model is guided by the theory outlined above and fully presented in our above referenced articles on C-TMLE, which is DR, and inference can be based on bootstrap variance estimates as well as the

variance of the efficient influence curve. Results when C-TMLE is applied to data generated as described in Section 3.3 of Vansteelandt et al.¹⁸ were presented at the WNAR 2011 Spring Meeting²³ and are described in a forthcoming paper.

Acknowledgments

Funding

This study was supported by the National Institutes of Health (grant no. 5R01AI74345-5) and the National Institutes of Health/National Heart, Lung, and Blood Institute (grant no. R01HL080644).

References

1. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Math Model.* 1986; 7:1393–1512.
2. Robins JM. addendum to: A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect [*Math Model* 1986; 7(9-12): 1393-1512; MR 87m:92078]. *Comput Math Appl.* 1987; 14(9-12):923–945.
3. Robins JM. Robust estimation in sequentially ignorable missing data and causal inference models. *Proc Am Stat Assoc Sec Bayesian Stat Sci.* 2000:6–10.
4. Tan Z. Bounded, efficient, and doubly robust estimation with inverse weighting. *Biometrika.* 2008; 94:1–22.
5. Tan Z. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika.* 2010; 97(3):661–682.
6. Petersen ML, Porter KE, Gruber S, et al. Diagnosing and responding to violations in the positivity assumption. *Stat Meth Med Res.* Published Online 28 October 2010. DOI: 10.1177/0962280210386207.
7. van der Laan MJ, Rubin D. Targeted maximum likelihood learning. *Int J Biostat.* 2006; 2(1) article 11.
8. van der Laan MJ, Gruber S. Collaborative double robust penalized targeted maximum likelihood estimation. *Int J Biostat.* 2010; 6(1) article 17.
9. Gruber S, van der Laan MJ. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *Int J Biostat.* 2010; 6(1) article 17.
10. Gruber S, van der Laan MJ. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int J Biostat.* 2010; 6(1) article 18.
11. van der Laan, MJ.; Rose, S. Targeted learning: prediction and causal inference for observational and experimental data. Springer; New York: 2011.
12. Robins, JM.; Rotnitzky, A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewell, N.; Dietz, K.; Farewell, V., editors. *AIDS epidemiology: methodological issues.* Birkhäuser; Boston: 1992. p. 297-331.
13. Robins JM, Wang N. Inference for imputation estimators. *Biometrika.* 2000; 87:113–124.
14. van der Laan, MJ.; Robins, JM. Unified methods for censored longitudinal data and causality. Springer; New York: 2003.
15. Lee BK, Lessler J, Stuart EA. Improved propensity score weighting using machine learning. *Stat Med.* 2009; 29:337–346. [PubMed: 19960510]
16. Schneeweiss S, Rassen JA, Glynn RJ, et al. Highdimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology.* 2009; 20:512–522. [PubMed: 19487948]
17. Westreich D, Cole SR, Funk MJ, et al. The role of the c-statistic in variable selection for propensity scores. *Pharmacoepidemiol Drug Saf.* 2011; 20:317–320. [PubMed: 21351315]
18. Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. *Stat Methods Med Res.* 2012; 21(1):7–30. [PubMed: 21075803]

19. Yu, A.; van der Laan, MJ. U.C. Berkeley Division of Biostatistics Working Paper Series. Sep. 2003 Measuring treatment effects using semiparametric models. Working Paper 136
20. Robins JM, Rotnitzky A, Bickel PJ, Kwon J. Comment on Inference for semiparametric models: some questions and an answer. *Stat Sinica*. 2001; 11:920–935.
21. Porter KE, Gruber S, van der Laan MJ, et al. The relative performance of targeted maximum likelihood estimators. *Int J Biostat*. 2011; 7(1) article 31.
22. Stitelman OM, van der Laan MJ. Collaborative targeted maximum likelihood for time to event data. *Int J Biostat*. 2010; 6(1) article 21.
23. Gruber, S.; van der Laan, MJ. Collaborative targeted maximum likelihood estimation; WNAR 2011, Spring Meeting; Anchorage, AL. 6-14 May, 2011;