Machine Learning for Effect Estimation: The Curse of Dimensionality

Ashley I Naimi

Oct 2022

Contents

1 Introduction

Machine learning methods consist of a wide range of analytic techniques that do not require hard to verify modeling assumptions. Because of this, they are often assumed to be less biased than their standard parametric counterparts. This perceived property has motivated many to either recommended or use machine learning methods to estimate causal quantities via g computation or IP weighting (Lee et al., 2010, Westreich et al. (2010),Snowden et al. (2011),Oulhote et al. (2019)). However, it is generally not recognized that machine learning methods are subject to problems that arise from the curse of dimensionality, a term first coined by Bellman Bellman (1957) to refer to a set of problems encountered when estimating models with many variables (Wasserman, 2006). Such problems can include high bias, high mean squared error, and low confidence interval coverage.

Doubly robust estimators are so named because these methods allow two chances for adjustment (Robins and Rotnitzky, 1995, Robins and Rotnitzky (2001),Bang and Robins (2005)). In the case of confounding adjustment, these chances arise because the analyst must fit two models: a model for the outcome conditional on the exposure and all confounders (outcome model); and a model for the exposure conditional all confounders (the propensity score model). These are then combined to estimate the effect of interest (Rotnitzky and Vansteelandt, 2014).

The benefits of doubly robust methods have been explained by pointing out that if a confounding variable is left out of either the exposure or the outcome model (but not both), unbiased estimates can still be obtained (Jonsson-Funk et al., 2011). While true, analysts would not typically leave confounding variables out of either the exposure or outcome model. Such justifications ignore a critically important benefit conferred by doubly robust estimators: under relatively mild conditions, they remain unbiased, with asymptotically nominal confidence interval coverage, even when machine learning methods are used to fit the exposure and outcome models (van der Laan and Rubin, 2006, Kennedy and Balakrishnan (2017)). In effect, doubly robust methods can mitigate or resolve problems caused by the curse of dimensionality.

2 Estimating Causal Effects

We consider a simple setting with a single binary exposure (X), a set of continuous confounders ($\mathbf{C} = \{C_1, C_2, C_3, C_4\}$) measured at baseline, and a single continuous outcome (Y) measured at the end of follow-up. In an observational cohort study to estimate the effect of X on Y, $\mathbf{C}$ might be assumed a minimally sufficient adjustment set (Greenland et al., 1999), and the outcome and exposure would be assumed generated according to some unknown models, for example:

$$E(Y \mid X, \mathbf{C}) = g(X, \mathbf{C}), \qquad \text{(Model 1)}$$

$$P(X = 1 \mid \mathbf{C}) = f(\mathbf{C}). \qquad \text{(Model 2)}$$

In the above equations, we use $g(\bullet)$ and $f(\bullet)$ to emphasize that the expected outcome conditional on X and C, and the probability of the exposure given C need not be considered standard linear or logistic regression functions. Rather, $g(\bullet)$ and $f(\bullet)$ represent arbitrary functions relating the exposure and confounders to the outcome, and the confounders to the exposure. Importantly, **in an observational cohort study assuming a correct confounder adjustment set,** these arbitrary functions usually represent the extent of what is known about the exposure and outcome models (Robins, 2001) That is, while these models may typically be assumed to be in the family of generalized linear models (Nelder and Wedderburn, 1972), we note below why this may not often be ideal.

Say we are interested in the average treatment effect:

$$\psi = E(Y^{x=1} - Y^{x=0})$$

where Y^x is the outcome that would be observed if X were set to x. This estimand is (point) identified under positivity, consistency, no interference, and exchangeability (Robins and Hernán, 2009, Naimi et al. (2017)). If these assumptions hold, ψ can be estimated using a number of approaches. In the equations that follow, we let i index sample observations which range from 1 to N, $\hat{g}_i(X = x, \mathbf{C})$ and $\hat{f}_i(\mathbf{C})$ are individual sample predictions for $E(Y \mid X = x, \mathbf{C})$ and $P(X = 1 \mid \mathbf{C})$, respectively.

With predictions from Model 1, ψ can be estimated via g computation (Naimi et al., 2017), as we did in the previous section. Mathematically, we

can write g computation for a time fixed exposure as:

$$\hat{\psi}_{gComp} = \frac{1}{N}\sum_{i=1}^{N}\{\hat{g}_i(X=1,\mathbf{C}) - \hat{g}_i(X=0,\mathbf{C})\}. \quad (1)$$

With predictions from Model 2, ψ can be estimated via inverse probability weighting (Hernán and Robins, 2006) as:

$$\hat{\psi}_{ipw} = \frac{1}{N}\sum_{i=1}^{N}\left\{\left[\frac{X_i Y_i}{\hat{f}_i(\mathbf{C})}\right] - \left[\frac{(1-X_i)Y_i}{1-\hat{f}_i(\mathbf{C})}\right]\right\}. \quad (2)$$

Both approaches 1 and 2 are "singly robust' ' in that they typically rely entirely on the correct specification of the appropriate single regression model. If these models are misspecified, the estimators will not generally converge to the true value (they will be "biased").

Technical Note:
Often when we use the word "bias", particularly in epidemiology, we actually mean "inconsistent" in the statistical sense. Technically, an estimator $\hat{\theta}$ is consistent if, for some arbitrarily small $\epsilon > 0$:

$$\lim_{n\to\infty} P(|\hat{\theta} - \theta| > \epsilon) = 0.$$

When we have unadjusted confounding, selection, information bias, the estimator will not converge to the truth no matter how large a sample we have.
In contrast, we say that an estimator is biased (in finite samples) if:

$$E(\hat{\theta} - \theta) \neq 0.$$

That is, we can have zero confounding (i.e., a consistent estimator), but still have a biased estimator because of how it performs at using the data to estimate the effect at a given sample size. One example of this is the partial likelihood estimator used to quantify parameters of a Cox regression model (see Johnson et al. (1982)). Usually, this statistical bias will disappear as the sample size increases.

2.1 Parametric Estimation

For continuous Y and binary X, it is customary to specify models Model 1 and Model 2 parametrically using linear and logistic regression, respectively. Doing so effectively states that we know enough about the form of $g(X,\mathbf{C})$ and $f(\mathbf{C})$

to define them as:

$$g(X, \mathbf{C}) = E(Y \mid X, \mathbf{C}) = \beta_0 + \beta_1 X + \beta_2 C_1 + \beta_3 C_2 + \beta_4 C_3 + \beta_5 C_4, \tag{3}$$

$$Y \mid X, \mathbf{C} \sim \mathcal{N}\Big(E(Y \mid X, \mathbf{C}), \sigma^2\Big)$$

$$f(\mathbf{C}) = P(X = 1 \mid \mathbf{C}) = \text{expit}(\alpha_0 + \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \alpha_4 C_4), \tag{4}$$

$$\text{expit}(\bullet) = 1/(1 + \exp[-\bullet])$$

Imposing these forms on $g(X, \mathbf{C})$ and $f(\mathbf{C})$ permits use of maximum likelihood for estimation and inference (Cole et al., 2013).

Equation 3 imposes several parametric constraints on the form of $g(X, \mathbf{C})$: (i) Y follows a conditional normal distribution with constant variance not depending on X or $\mathbf{C}$; and (ii) the conditional mean of Y is related to the covariates X and $\mathbf{C}$ additively, as defined in equation 3. If these constraints on $g(X, \mathbf{C})$ are true, and other identification and regularity conditions hold (Longford, 2008)$^{(ch2)}$ the maximum likelihood estimates of $\boldsymbol{\beta}$ are asymptotically efficient (Rencher, 2000)$^{(p144)}$ Relatedly, under the model constraints and identification and regularity conditions, as the sample size increases, the estimates of $g(X, \mathbf{C})$ and/or $f(\mathbf{C})$ will converge to the true values at an optimal (i.e., $\sqrt{N}$) rate, and their distribution will be such that confidence intervals can be easily derived.

If constraint (i) is violated, the maximum likelihood estimator is no longer the most efficient, but can still be used to estimate ψ consistently. If constraint (ii) is violated, then the maximum likelihood estimator is no longer consistent. Depending on the severity to which constraint (ii) is violated, the bias may be substantial. Unfortunately, in an observational study the true form of equation 3 is almost never known. This means that such maximum likelihood estimates are almost always biased, with the degree of bias depending on the (unknown) extent to which the model is mis-specified (Box, 1976).

One way to avoid relying on correct outcome model specification is to use a parametric approach for Model 2, and estimate ψ via $\hat{\psi}_{ipw}$. Specifically, with IP-weighting, one need not model the interactions between the exposure and any covariates (Hernán et al., 2001). Such an estimator is not as efficient as $\hat{\psi}_{gComp}$, and can be subject to important finite-sample biases when weights are very large, or when there are no observations to weight in certain exposure-

confounder strata. But as the sample size increases, the inverse probability weighted estimator converges at the same standard $\sqrt{N}$ rate as the g computation estimator (Westreich et al., 2012). Unfortunately, as with the outcome model, the true form of Model 2 will almost never be known in an observational study. Mis-specification of equation 4 will also lead to biased estimation of ψ, again with the degree of bias depending on the unknown extent of model mis-specification.

2.2 Causes of Misspecification

It's important to understand what mis-specification bias is and where it comes from. A misspecified model form can occur if the analyst fails to correctly account for the manner in which exposure and confounders relate to the outcome. For a generalized linear model, this would occur if chosen link function is not compatible with how the data were actually generated (Weisberg and Welsh, 1994), if the analyst fails to account for curvilinear relations between the covariates and the outcome, or fails to include important exposure-confounder or confounder-confounder interactions. Unfortunately, in an observational study the true nature of these relations is typically not known, which is one reason underlying the increasing popularity of machine learning methods. However, misspecification resulting an incomplete confounder adjustment set, or incorrectly adjusting for a mediator, cannot be fixed with doubly robust machine learning methods (Keil et al., 2018).

Again, recall that for this simple model would could have up to:

$$2^5 - 5 - 1 = 26$$

k-way interactions (including 2, 3, 4, and 5 way). This means that for this simple modely, from only the perspective of variable interactions, there are 26 possible chances for us to induce potential misspecification.

The point of this is not to emphasize interactions per se, but rather to point out that, in any given analysis, there will always be choices that can potentially lead to mistakes. These choices, particularly about the form of parametric models, pervade empirical regression analyses, and represent a potential weak link in the process of scientific investigation. This is one reason why machine learning methods have become so popular.

3 Machine Learning for Causal Effect Estimation: The Curse of Dimensionality

Nonparametric methods are an alternative to parametric models. For example, nonparametric maximum likelihood estimation (NPMLE) for Model 2 or Model 1 would entail fitting equations 3 or 4, but with a parameter for each unique combination of values defined by the cross-classification of all covariates (i.e., saturating the model).

Consider another simple simulated setting with four continuous confounders, one binary exposure, and a sample size of 250:

```
set.seed(123)
n = 250

c1 <- factor(round(rnorm(n), 2))
c2 <- factor(round(rnorm(n), 2))
c3 <- factor(round(rnorm(n), 2))
c4 <- factor(round(rnorm(n), 2))
x <- factor(rbinom(n, 1, 0.5))

dat_ <- data.frame(x, c1, c2, c3, c4)

head(dat_, 10)
```

```
##    x    c1    c2    c3    c4
## 1  0 -0.56 -0.38  -0.6  1.54
## 2  0 -0.23 -0.56 -0.99 -0.11
## 3  0  1.56 -0.34  1.03  0.51
## 4  1  0.07  0.09  0.75  0.21
## 5  0  0.13   1.6 -1.51 -0.19
## 6  1  1.72 -0.09  -0.1 -0.12
## 7  0  0.46  1.08  -0.9  1.01
## 8  0 -1.27  0.63 -2.07  -0.2
## 9  0 -0.69 -0.11  0.15 -2.04
## 10 0 -0.45 -1.53 -0.08  -0.2
```

```
mod_mat <- model.matrix(~., data = dat_)

dim(mod_mat)
```

```
## [1] 250 728
```

```
mod_mat_int <- model.matrix(~.^2, data = dat_)

dim(mod_mat_int)
```

```
## [1]    250 199067
```

Even if we (1) round all continuous variables to two decimal places, and (2) ignore any potential interactions, and include a parameter for every level of all variables in the model, we end up with a total of 728 parameters in the model. Adding two way interactions alone increases this number to a whopping 199067 parameters for a sample size of N = 250! Consequently, in any realistic setting, the NPMLE will be undefined, particularly in a finite sample with a continuous confounder, since there will be no covariate patterns containing both treated and untreated subjects. In these settings, while the NPMLE makes no assumptions about model form, it will not be possible to use it for quantifying the average treatment effect.

Alternatively, one can use "machine learning' ' methods like kernel regression, splines, random forests, boosting, etc.. These approaches exploit smoothness across covariate patterns to estimate the regression function, without imposing arbitrary parametric forms such as what is articulated in Models 3 or 4. However, for any nonparametric approach there will always be an explicit bias-variance trade-off that arises in the choice of tuning parameters; less smoothing yields smaller bias but larger variance, while more smoothing yields smaller variance but larger bias (parametric models can be viewed as an extreme form of smoothing).

This tradeoff has important consequences. In particular, there is no generally optimal solution for estimating regression functions nonparametrically at the standard $\sqrt{N}$ rates attained by correctly specified parametric estimators (van der Vaart, 2000). These slow rates generally require sample sizes that

are exponentially larger than those required for (fast converging) parametric methods to maintain the same degree of accuracy.

Convergence rates for nonparametric estimators become slower with more flexibility and more covariates. For example, a standard rate for estimating smooth regression functions is $N^{-\beta/(2\beta+d)}$, where β represents the number of derivatives of the true regression function, and d represents the dimension of, or number of covariates in, the true regression function. This issue is known as the curse of dimensionality (Györfi et al., 2002, Robins and Ritov (1997),Wasserman (2006)). Sometimes this is viewed as a disadvantage of nonparametric methods; however, it is just the cost of making weaker assumptions: if a parametric model is misspecified, it will converge very quickly to the wrong answer.

In addition to slower convergence rates, confidence intervals are harder to obtain. Specifically, even in the rare case where one can derive asymptotic distributions for nonparametric estimators, it is typically not possible to construct confidence intervals (even via the bootstrap, as it requires certain convergence rate conditions to hold) without impractically undersmoothing the regression function (i.e., overfitting the data) (Hahn, 1998).

These complications (slow rates and lack of valid confidence intervals) are generally inherited by the singly robust estimators 2 and 1 (apart from a few special cases which require simple estimators, such as kernel methods with strong smoothness assumptions and careful tuning parameter choices that are suboptimal for estimating f or g). For general nonparametric estimators $\hat{f}$ and $\hat{g}$, the estimators 2 and 1 will converge at slow rates, and honest confidence intervals (defined as confidence intervals that are at least nominal over a large nonparametric class of regression functions) (Li, 1989) will not be computable.

We recently conducted a simulation study (Naimi et al., 2022) that demonstrates some of the consequences of these issues. Figure 1 shows the absolute bias of g computation and inverse probability weighting, compared to two double robust estimators, when machine learning methods are used.

As can be seen in the Figure, when machine learning methods are used, g computation or IP-weighting perform poorly relative to the double-robust approaches (Indeed, in the simulation scenario presented, using a machine learning approach with g computation or IP weighting yielded a bias higher than when we used a misspecified parametric model). Similarly, when ma-

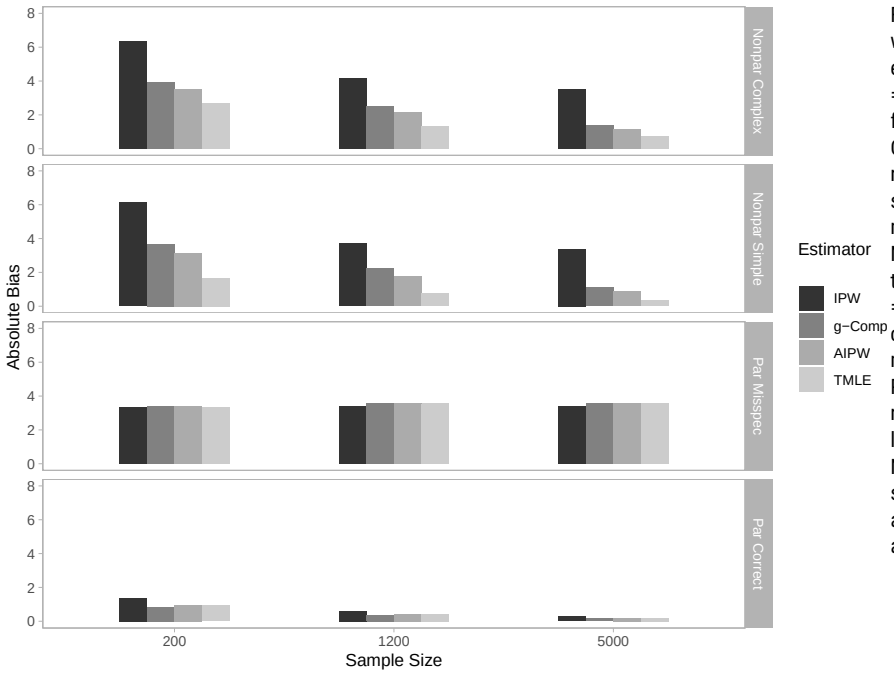

Figure 1: Absolute bias of inverse probability weighted, g-computation, and doubly robust estimators for sample sizes of N = 200, N = 1200, and N = 5000. Bar color intensity, from black to light gray, represent IPW, g Computation, AIPW, and TMLE estimators, respectively. Plot labels refer to the following scenarios: Nonpar Complex = nonparametric method fit to the transformed confounders; Nonpar Simple = nonparametric method fit to the untransformed confounders; Par Misspec = parametric method fit to transformed confounders; Par Correct = parametric method fit to untransformed confounders. Parametric regression included logistic regression for the exposure model, and linear regression for the outcome model. Nonparametric method consisted of a stacked generalization with random forests and extreme gradient boosting algorithms, and no sample splitting.

chine learning methods were used with g computation or IP-weighting, 95% confidence interval coverage was as low as 0%, and typically ranged between 20-30% (Naimi et al., 2022).

4 Takeaway

The important takeaway here is that, even though machine learning methods do not require strict parametric modeling assumptions in the way that standard regression (e.g., generalized linear models) do, they do not necessarily deliver "better" results than standard regression modeling approaches.

Now, it's important to recognize that we are ignoring a lot of important concepts in the statistical theory of estimation here. What it means for one estimation approach to be "better" than another, or to say that an estimation approach does not "work" or that another does, must be understood in a specific mathematical context. That is, the notions of "better" or "work" are often formalized mathematically. We have not covered these formalities, though we have alluded to them in various ways (e.g., bias, confidence interval coverage, mean squared error).

5 Bonus Material

To see the difficulty, consider our estimated average treatment effect using marginal standardization in the NHEFS data. However, instead of using the GLM function in the code we used for marginal standardization with a random forest algorithm using the `ranger` function.[1]

[1] We will discuss random forests and `ranger` in a subsequent section.

```
library(ranger)
#' Marginal Standardization with Random Forest
model0 <- ranger(modelForm, num.trees = 500,
    mtry = 3, min.node.size = 50, data = subset(nhefs,
        qsmk == 0))
model1 <- ranger(modelForm, num.trees = 500,
    mtry = 3, min.node.size = 50, data = subset(nhefs,
        qsmk == 1))
mu1 <- predict(model1, data = nhefs, type = "response")$pred
mu0 <- predict(model0, data = nhefs, type = "response")$pred

marg_stand_RDrf <- mean(mu1) - mean(mu0)

bootfunc <- function(data, index) {
    boot_dat <- data[index, ]
    model0 <- ranger(modelForm, num.trees = 500,
        mtry = 5, data = subset(boot_dat,
            qsmk == 0))
    model1 <- ranger(modelForm, num.trees = 500,
        mtry = 5, data = subset(boot_dat,
            qsmk == 1))
    mu1 <- predict(model1, data = boot_dat,
        type = "response")$pred
    mu0 <- predict(model0, data = boot_dat,
        type = "response")$pred

    marg_stand_RD_ <- mean(mu1) - mean(mu0)
    return(marg_stand_RD_)
```

```
}

#' Run the boot function. Set a seed to obtain reproducibility
set.seed(123)
boot_res <- boot(nhefs, bootfunc, R = 2000)
boot_RDrf <- boot.ci(boot_res)

marg_stand_RDrf
```

```
## [1] 0.1351086
```

```
boot_RDrf$bca
```

```
##      conf
## [1,] 0.95 46.62 1947.39 0.08393356 0.1849261
```

When we use parametric generalized linear models to estimate this effect, we get a risk difference of 0.14, with 95% (bca) confidence intervals of 0.08, 0.2. When we use random forest to estimate this effect, we get exactly the same risk difference of 0.14, with 95% (bca) confidence intervals of 0.08, 0.18.

Unfortunately, even though we get the same point estimates when we switch from the `glm` function to the `ranger` function, this does not actually suggest that the random forest approach works as well as the parametric approach. This can be a subtle point, and it makes it difficult to understand why we simply shouldn't use machine learning with g computation or IP weighting.

References

Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962—973, 2005.

R Bellman. *Dynamic Programming*. Rand Corporation research study. Princeton University Press, 1957. ISBN 9780691079516. URL https://books.google.it/books?id=wdtoPwAACAAJ.

G. E. P. Box. Science and Statistics. *JASA*, 71(356):791–99, 1976.

Stephen R Cole, Haitao Chu, and Sander Greenland. Maximum likelihood, profile likelihood, and penalized likelihood: A primer. *Am J Epidemiol*, 179(2): 252–260, 2013.

Sander Greenland, Judea Pearl, and JM Robins. Causal diagrams for epidemiological research. *Epidemiol*, 10(1):37–48, 1999.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, NY, 2002.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*, 60(7):578–586, 2006.

Miguel A Hernán, Babette Brumback, and James M Robins. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J Am Stat Assoc*, 96(454):440–448, 2001.

M E Johnson, H D Tolley, M C Bryson, and A S Goldman. Covariate analysis of survival data: a small-sample study of cox's model. *Biometrics*, 38(3): 685–698, Sep 1982.

Michele Jonsson-Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M. Alan Brookhart, and Marie Davidian. Doubly robust estimation of causal effects. *Am J Epidemiol*, 173(7):761–767, 2011.

Alexander P Keil, Stephen J Mooney, Michele Jonsson Funk, Stephen R Cole, Jessie K Edwards, and Daniel Westreich. Resolving an apparent paradox in doubly robust estimators. *Am J Epidemiol*, 187(4):891–892, Apr 2018.

Edward H Kennedy and Sivaraman Balakrishnan. Discussion of "Data-driven confounder selection via Markov and Bayesian networks" by Jenny Häggström. *Biometrics*, In Press, 2017.

Brian K. Lee, Justin Lessler, and Elizabeth A. Stuart. Improving propensity score weighting using machine learning. *Stat Med*, 29(3):337–346, 2010.

Ker-Chau Li. Honest confidence regions for nonparametric regression. *The Annals of Statistics*, 17(3):1001–1008, 1989.

NT Longford. *Studying Human Populations: An Advanced Course in Statistics*. Springer, New York, 2008.

AI Naimi, A Mishler, and Edward H. Kennedy. Challenges in obtaining valid causal effect estimates with machine learning algorithms. *Am J Epidemiol*, kwab201, 2022.

Ashley I Naimi, Stephen R Cole, and Edward H Kennedy. An Introduction to G Methods. *Int J Epidemiol*, 46(2):756–62, 2017.

J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *JRSS-A*, 135 (3):370–384, 1972.

Youssef Oulhote, Brent Coull, Marie-Abele Bind, Frodi Debes, Flemming Nielsen, Ibon Tamayo, Pal Weihe, and Philippe Grandjean. Joint and independent neurotoxic effects of early life exposures to a chemical mixture: A multi-pollutant approach combining ensemble learning and g-computation. *Environmental Epidemiology*, 3(5):e063, 2019.

Alvin C. Rencher. *Linear Models in Statistics*. Wiley, New York, 2000.

J M Robins and Y Ritov. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Stat Med*, 16(1-3):285–319, Jan 1997.

James M Robins and Miguel Á Hernán. Estimation of the causal effects of time-varying exposures. In G Fitzmaurice, M. Davidian, G Verbeke, and G Molenberghs, editors, *Advances in Longitudinal Data Analysis*, pages 553–599. Chapman & Hall, Boca Raton, FL, 2009.

JM Robins. Data, design, and background knowledge in etiologic inference. *Epidemiol*, 12(3):313–320, 2001.

JM Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *JASA*, 90(429):122–9, 1995.

JM Robins and Andrea Rotnitzky. Comment: Inference for semiparametric models: Some questions and an answer. *Statistica Sinica*, 11(4):920–936, 2001.

Andrea Rotnitzky and Stijn Vansteelandt. Double-robust methods. In Geert Molenberghs, Garrett Fitzmaurice, Michael G. Kenward, Anastasios Tsiatis, and Geert Verbeke, editors, *Handbook of Missing Data Methodology*, chapter 9, pages 185–209. CRC Press, 2014.

Jonathan M. Snowden, Sherri Rose, and Kathleen M. Mortimer. Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. *Am J Epidemiol*, 173(7):731–738, 2011.

Mark J. van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *Int J Biostat*, 2(1):Article 11, 2006.

A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, 2000.

Larry Wasserman. *All of nonparametric statistics*. Springer, New York; London, 2006.

S. Weisberg and A. H. Welsh. Adapting for the missing link. *The Annals of Statistics*, 22(4):1674—1700, 1994.

Daniel Westreich, Justin Lessler, and Michele Jonsson Funk. Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol*, 63(8):826 – 833, 2010.

Daniel Westreich, Stephen R. Cole, Enrique F. Schisterman, and Robert W. Platt. A simulation study of finite-sample properties of marginal structural Cox proportional hazards models. *Stat Med*, 31(19):2098–2109, 2012.