Semiparametric Causal Sufficient Dimension Reduction of Multidimensional Treatments

Razieh Nabi[1] **Todd McNutt**[2] **Ilya Shpitser**[3]

[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, USA
[2]School of Medicine, Johns Hopkins University, Baltimore, Maryland, USA
[3]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA

Abstract

Cause-effect relationships are typically evaluated by comparing outcome responses to binary treatment values, representing two arms of a hypothetical randomized controlled trial. However, in certain applications, treatments of interest are continuous and multidimensional. For example, understanding the causal relationship between severity of radiation therapy, summarized by a multidimensional vector of radiation exposure values and post-treatment side effects is a problem of clinical interest in radiation oncology. An appropriate strategy for making interpretable causal conclusions is to reduce the dimension of treatment. If individual elements of a multidimensional treatment vector weakly affect the outcome, but the overall relationship between treatment and outcome is strong, careless approaches to dimension reduction may not preserve this relationship. Further, methods developed for regression problems do not directly transfer to causal inference due to confounding complications. In this paper, we use semiparametric inference theory for structural models to give a general approach to causal sufficient dimension reduction of a multidimensional treatment such that the cause-effect relationship between treatment and outcome is preserved. We illustrate the utility of our proposals through simulations and a real data application in radiation oncology.

1 INTRODUCTION

In causal inference, the exposure of interest is commonly assumed to be either binary (e.g., comparing treatment vs placebo) or continuous (e.g., effect of treatment dosages on viral load.) In the latter cases, in addition to contrasts of responses to two specific doses, we may be interested in the entire dose-response relationship, and choose to model it via a simple functional, for example a logarithmic or sigmoidal function. In other applications, we might be interested in assessing causal relationships between outcomes and treatments with values that lie in a multidimensional space. For instance, in natural language processing interest lies in causal analyses that involve high dimensional text data [Gentzkow et al., 2019, Feder et al., 2021]. Another example is the neuroimaging data used to relate neuronal network activity to cognitive processing and behavior [Ramsey et al., 2010, Mather et al., 2013].

As our motivational example, we focus on an application in radiation oncology. In neck and head cancers, minor variations in dose and direction of radiation may result in similar tumor reduction but vastly improve secondary outcomes, such as weight loss, or dysfunction induced by radiation therapy, such as dysphasia or xerostomia [Robertson et al., 2015]. Thus, understanding the causal relationship between a multidimensional radiation exposure and downstream side effects in cancer patients undergoing radiation therapy is of clinical interest. Unlike standard treatments, radiation therapy is complex and is represented by three dimensional voxel maps of radiation doses in different parts of the body. Since this representation is very high dimensional, the exact dose localization information in the voxel map is sometimes represented by cumulative dose-volume histograms and summarized by a multidimensional vector of exposure dosages. Even such summaries complicates establishing clinically relevant causal relationships.

Since we are interested in dimension reduction for the sake of explicating a particular relationship between treatments and outcomes, approaches that do not take outcomes into account in the right way run the risk of distorting the estimate of this relationship, or even falsely concluding the relationship is absent. Therefore, seemingly natural approaches to dimension reduction, such as principal component analysis (PCA), are not appropriate in our setting. On the other hand, there is a line of research in statistics on sufficient dimension reduction (SDR) [Li, 1991] with the objective of reducing

Accepted for the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022).

the dimension of covariates by preserving the associational relations between covariates and outcome. However, due to spurious associations introduced by confounding which is ubiquitous in observational data sources, naive use of SDR approaches to discern causal relationships between treatments and outcomes leads to bias.

We are interested in applying SDR core ideas to reduce dimension of a treatment in a way that preserves a *causal* rather than *associational* relationship with the outcome. In addition, we are interested in doing so under the weakest possible assumptions, which entails generalizing the semiparametric approaches in the SDR literature [Ma and Zhu, 2012]. In this paper, we provide a framework for structural (causal) models based on semiparametric inference theory developed for marginal structural models [Robins, 1999] to give what we believe is the first approach to causal SDR of a multidimensional treatment.

2 PRELIMINARIES

Sufficient dimension reduction. Given an outcome variable Y and a p-dimensional covariate vector X, the goal of SDR is to find a known function $g_X(.;\beta)$ parameterized by β with a much smaller range than domain such that Y depends on X only through $g_X(X;\beta)$. Often this function is assumed to be linear, in which case the goal is to find $\beta \in \mathbb{R}^{p\times d}$, where $d < p$, such that Y depends on X only through $X^T\beta$, i.e., $\mathbb{E}[Y|X] = \mathbb{E}[Y|X^T\beta]$. Often, proposed solutions to SDR rely on strong parametric assumptions that are unlikely to hold in practical applications, such as the linearity condition where $\mathbb{E}[X|X^T\beta]$ is assumed to be a linear function of X, or the assumption that $\text{cov}(X \mid X^T\beta)$ is constant rather than a function of X, [Li, 1991, Cook and Weisberg, 1991, Hardle and Stoker, 1989, Ichimura, 1993, Cook and Li, 2002].

Ma and Zhu [2012] introduced a new approach to SDR by recasting the problem in terms of estimation in a semiparametric model. Crucially, this approach relies on far weaker assumptions than is typical in SDR, and is thus much more generally applicable. To obtain the relevant semiparametric model, we rewrite the above condition as $Y = \ell(X^T\beta) + \epsilon$, where $\ell(X^T\beta) := \mathbb{E}[Y|X^T\beta]$ is an unspecified smooth function and $\mathbb{E}[\epsilon|X] = 0$, while the distribution $p(\epsilon|X)$ remains otherwise unrestricted. Ma and Zhu [2012] derived the class of all influence functions for β, a.k.a. the orthogonal nuisance tangent space denoted by $\Lambda_\eta^\perp$, as: $\Lambda_\eta^\perp = \{(Y - \mathbb{E}[Y|X^T\beta]) \times (\alpha(X) - \mathbb{E}[\alpha(X)|X^T\beta])\}$, where $\alpha(X)$ is any function of X.

The main objective of semiparametric theory of influence functions is to derive an estimator that has the characteristics of parametric models (such as a $\sqrt{n}$-consistency and asymptotic normality) without the restrictive parametric assumptions. Influence functions are arguably the most important approach to estimation in causal inference, especially in observational studies where we do not have prior knowledge on what the true likelihood is; see Appendix A for a brief overview of influence functions, and [Van der Vaart, 2000, Bang and Robins, 2005, Tsiatis, 2007] for more details.

A well-known property of semiparametric models is that all elements of $\Lambda_\eta^\perp$ are mean 0 under the true distribution. Hence, a general class of estimating equations can be obtained using the sample version of

$$\begin{aligned}&\mathbb{E}[U(\beta)] \qquad (1)\\ &= \mathbb{E}\Big[\big(Y - \mathbb{E}[Y \mid X^T\beta]\big) \times \big(\alpha(X) - \mathbb{E}[\alpha(X) \mid X^T\beta]\big)\Big] = 0,\end{aligned}$$

where $U(\beta)$ is an arbitrary element in $\Lambda_\eta^\perp$. The estimator obtained from (1) is doubly robust under any choice of models for $\mathbb{E}[Y|X^T\beta]$ and $\mathbb{E}[\alpha(X)|X^T\beta]$, meaning that the estimator remains consistent if either of these two models is correctly specified [Ma and Zhu, 2012].

Causal inference. In causal inference, we seek to make inferences about the causal relationship of a treatment variable A and an outcome variable Y by counterfactual contrasts of the form $Y(a)$ representing a hypothetical experiment where treatment A is set to a, possibly contrary to the fact. A common setting considers, in addition to A and Y, a vector of baseline variables C, yielding an observed data distribution of the form $p(Y, A, C)$. Under standard assumptions of *consistency*, which states that counterfactual outcome is the same as observed outcome if treatment is set to observed value, *conditional ignorability* which states that $\{Y(a)\}$ is independent of A conditional on C, and *positivity* of $p(A \mid C)$, the counterfactual distribution $p(Y(a))$ is identified as the following function of observed data

$$p(Y(a)) = \sum_c p(Y \mid A = a, C = c) \times p(C = c). \quad (2)$$

The *average causal effect* (ACE) of a binary treatment on an outcome is defined as ACE $= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$. Under the above assumptions, the counterfactual mean $\mathbb{E}[Y(a)]$ is given as the following function of the observed data, called the *adjustment formula* or *g-formula*,

$$\mathbb{E}[Y(a)] = \mathbb{E}\big[\mathbb{E}[Y \mid A = a, C]\big], \quad (3)$$

where the outer expectation is taken with respect to $p(C)$; see Tian and Pearl [2002], Shpitser and Pearl [2006], Bhattacharya et al. [2020] for general identification algorithms in the presence of unmeasured confounders.

There are several different approaches on estimating the adjustment formula such as plug-in, inverse probability weighting (IPW), and semiparametric based estimators such as augmented IPW (AIPW) [Van der Vaart, 2000, Bang and Robins, 2005, Van der Laan et al., 2011]. An alternative class of IPW estimators models the relationship between A and Y via a *marginal structural model (MSM)*, or a causal regression. A simple version of such a model takes the form $\mathbb{E}[Y(a)] = f(a;\beta)$, for finite set of parameters β. Given

such a model, inferences about $\mathbb{E}[Y(a)]$ reduce to inferences about β. For binary treatments, $f(a;\beta)$ can be written as $\beta_0 + \beta_a \times a$ without loss of generality, with ACE $= \beta_a$. An MSM is different from an ordinary regression model, since $\mathbb{E}[Y(a)] \neq \mathbb{E}[Y|A=a]$ given our causal assumptions. Thus, one approach to estimating β is via the following estimating equation, appropriately reweighted by the treatment propensity score model, $W_a(C;\eta_a) := p(A|C;\eta_a)$,

$$\mathbb{P}_n\left[\frac{p^*(a)}{W_a(C;\widehat{\eta}_a)} \times \{Y - f(a;\beta)\}\right] = 0, \qquad (4)$$

where $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n}(.)$, $p^*(a)$ is an arbitrary function of a with the same dimension as β, and $\widehat{\eta}_a$ is the maximum likelihood estimate of η_a. This IPW procedure is known to be inefficient. A more efficient (in fact optimal in a wide class of reasonable estimators) approach is to use influence functions, described in detail in [Robins, 1999]. Our approach to causal SDR stands in the same relation to the semiparametric approach to SDR for regression problems in [Ma and Zhu, 2012] as fitting regression models does to fitting marginal structural models.

3 CAUSAL SUFFICIENT DIMENSION REDUCTION

We are interested in the causal effect of a multidimensional treatment $A \in \mathbb{R}^p$ on outcome Y, assuming all relevant covariates that need to be controlled for are observed and denoted by C. We would like to reduce the dimension of treatment A such that the causal relationship between A and Y is preserved. Let $g(.;\beta)$ be a function parameterized by β that takes values in $\mathbb{R}^p$ and maps them to values in $\mathbb{R}^d$, $d < p$, i.e., $g : A \in \mathbb{R}^p \mapsto g(A;\beta) \in \mathbb{R}^d$. We want to reduce the dimension of A in such a way that the counterfactual response $\mathbb{E}[Y(a)]$ only depends on A via $g(a)$. Specifically, we assume that if $\mathbb{E}[Y(a)]$ is identified, that is if $\mathbb{E}[Y(a)]$ is a mapping f from values a of A to functionals $h_a(p(V))$ of the observed data distribution, where $p(V)$ denotes the joint distribution over the set of observed variables V, then $f(a) = f(g(a;\beta))$. The methodology proposed in this paper does not depend on the choice of $g(.;\beta)$, although we fix a particular $g(.;\beta)$ in our experiments. We assume the three identification assumptions that were discussed in the previous section, namely consistency, conditional ignorability, and positivity, hold in our analysis. Therefore, we fix $h_a\big(p(C,A,Y)\big) = \mathbb{E}\big[\mathbb{E}[Y|A=a,C]\big]$, as shown in (2).

The estimation procedure for MSMs shown in (4) can be viewed as a standard estimating equation for a regression model relating treatment and outcome, but applied to observed data readjusted via inverse weighting in such a way that treatment appear randomly assigned. In other words, MSMs are regressions applied to a version of observed data in such a way that regression parameters can be interpreted causally. Unlike other estimating equations that solve for β by maximizing the feature outcome relationship, the equation in (1) fits β to maintain the identity $\mathbb{E}[Y|X] = \mathbb{E}[Y|X^T\beta]$. As a consequence, semiparametric causal SDR can be viewed as an MSM version of this regression problem, which seeks to find β which maintains $\mathbb{E}[Y(a)] = \mathbb{E}[Y(g(a;\beta))]$. In other words, our aim is to estimate β by maintaining the following identity

$$\mathbb{E}\Big[\mathbb{E}\big[Y \mid a, C\big]\Big] = \mathbb{E}\Big[\mathbb{E}\big[Y \mid g(a;\beta), C\big]\Big], \qquad (5)$$

where the outer expectation is wrt the density $p(C)$.

We note here the different roles that variables play in regression SDR and causal SDR. The goal of regression SDR is to preserve the associative relationship between high dimensional features X and outcome Y. The goal of causal SDR, as we view it here, is to preserve the causal relationship between a multidimensional treatment A and outcome Y, which is made complicated by the presence of spurious associations induced by covariates C. Thus, the goal in causal SDR is *not* to maintain the regression relationship between covariates and outcome by assuming $\mathbb{E}[Y|\{A,C\}] = \mathbb{E}[Y|g(\{A,C\};\beta)]$, but to preserve the relationship as in (5) where C is marginalized (adjusted for). The set of confounders C could still be high dimensional, but they are not of primary interest in our problem. Incorporating baseline covariates into the dimension reduction strategy along with the treatment, as is done in some MSMs, is left as an interesting avenue for future work. Examples of work focusing on dimension reduction of common confounders include Imai and Ratkovic [2014], Hu et al. [2014], Shortreed and Ertefaie [2017], Banijamali et al. [2018], Ma et al. [2019], Luo and Zhu [2020], Cheng et al. [2020].

As stated earlier, our objective is to preserve the causal effect of A on Y, which is of the form shown in (3). However, it suffices to say that if the counterfactual response curve, i.e., $\mathbb{E}[Y(a)]$, is preserved under our dimensionality reduction scheme, then the causal effect is preserved. Hence, we stated our constraint in (5) in terms of the counterfactual mean rather than the counterfactual contrast that would define the effect. Moreover, even though treatment is multidimensional, we emphasize that each unit still receives one treatment session; e.g., a single session of radiation therapy with no followups. Records of radiation treatment are usually stored as monodimentional cumulative dose-volume histograms, and are summarized as amount of radiation on $k\%$ of the organ's volume, where k ranges from 1 to 100.

In a conditionally ignorable causal model, intervention on A corresponds to dropping the term $p(A|C)$ from the observed density $p(Y,A,C)$ yielding (2). Define $q(Y,A,C)$ as the following modified version of (2):

$$q(Y,A,C) := p(Y \mid A,C) \times p^*(A) \times p(C),$$

where $p^*(A)$ is any density with the same support as $p(A)$.

Then (5) can be rewritten as

$$\mathbb{E}_q[Y \mid A = a] = \mathbb{E}_q[Y \mid g(a;\beta)], \tag{6}$$

where $\mathbb{E}_q$ is the expectation taken with respect to the density $q(Y, A, C)$ defined above, and $q(Y|A) = \sum_C q(Y, C|A) = \sum_C p(Y|A, C) \times p(C)$ by definition.

Equations (5) and (6) are equivalent forms of our constraint in the causal SDR problem where the MSM model for $\mathbb{E}[Y(a)] = \mathbb{E}_q[Y|a]$, is assumed to be a function of the multidimensional treatment intervention a only through its lower dimension representation $g(a;\beta)$. We now describe two approaches to estimating β.

3.1 INVERSE PROBABILITY WEIGHTED SDR

Let $\ell(g(A;\beta)) := \mathbb{E}_q[Y|g(A;\beta)]$ and $\nu\big(g(A;\beta)\big) := \mathbb{E}_q[\alpha(A) \mid g(A;\beta)]$ be two unspecified smooth functions of $g(A;\beta)$. A simple estimation strategy for β based on generalizing (4), entails solving

$$\mathbb{E}\left[\frac{p^*(a)}{p(A = a \mid C)} \times \widetilde{U}(\beta)\right] = 0, \tag{7}$$

where $\widetilde{U}(\beta) = \{Y - \ell(g(a;\beta))\} \times \big\{\alpha(A) - \nu(g(a;\beta))\big\}$, $p^*(a)$ is an arbitrary function of a, and $p(A|C)$ is a correctly specified statistical model which governs how the treatment A is assigned based on baseline characteristics C. The above equation may be solved using observed data by evaluating the expectation empirically.

Lemma 1. *An estimator for β based on solving (7) is unbiased under correct specification of $p(A|C)$, and either one of $\ell(g(A;\beta)) := \mathbb{E}_q[Y|g(A;\beta)]$ or $\nu(g(A;\beta)) := \mathbb{E}_q[\alpha(A) \mid g(A;\beta)]$.*

3.2 SEMIPARAMETRIC CAUSAL SDR

A general approach for deriving *regular and asymptotically linear* (RAL) estimators of β is based on deriving $\widetilde{\Lambda}_\eta^\perp$, the orthogonal complement of the nuisance tangent space of a semiparametric model that enforces the constraint (5), but places no other restrictions on the observed data distribution; $\widetilde{\Lambda}_\eta^\perp$ is the class of all influence functions. One approach is to derive this space explicitly, as was done in [Ma and Zhu, 2012]. An alternative is to take advantage of general theory relating orthogonal complements of regression problems, and orthogonal complements of "causal regression problems," or MSMs, developed by Robins [1999]. Given the semiparametric model $\mathcal{M}$ induced by the restriction (5), we take advantage of this theory in the following result.

Theorem 1. *The orthogonal complement of the nuisance tangent space $\widetilde{\Lambda}_\eta^\perp$ for β that satisfies (6) is:*

$$\widetilde{\Lambda}_\eta^\perp = \Big\{\frac{\widetilde{U}(\beta)}{W_a(C)} - \phi(A, C) + \mathbb{E}[\phi(A, C) \mid C]\Big\},$$

where $\phi(A, C)$ is an arbitrary function of A and C, $W_a(C)$ is the IPW weight $p(A = a|C)/p^*(a)$ for a fixed $p^*(a)$, and $\widetilde{U}(\beta)$ is of the form

$$\widetilde{U}(\beta) = \big\{Y - \ell(g(a;\beta))\big\} \times \big\{\alpha(A) - \nu(g(a;\beta))\big\},$$

where $\ell(g(a;\beta)) := \mathbb{E}_q[Y|g(a;\beta)]$ and $\nu(g(a;\beta)) := \mathbb{E}_q[\alpha(A)|g(A;\beta)]$. Moreover, the most efficient estimator in this class, for any fixed $\alpha(A)$, is recovered by setting $\phi^{opt}(A, C) = \mathbb{E}\Big[\frac{\widetilde{U}(\beta)}{W_a(C)} \mid A, C\Big]$.

Lemma 2. *For a fixed choice of $\alpha(A)$ and normalized function $p^*(A)$, the element $\widetilde{U}(\beta^*) \in \widetilde{\Lambda}_\eta^\perp$ corresponding to the optimal choice of $\phi(A, C)$ has the form.*

$$\frac{p^*(A)}{p(A \mid C)} \times \widetilde{U}(\beta) - \frac{p^*(A)}{p(A \mid C)} \times \mathbb{E}\left[\widetilde{U}(\beta)\Big|A, C\right] + \mathbb{E}_q\left[\mathbb{E}\left[\widetilde{U}(\beta)\Big|A, C\right]\Big|C\right], \tag{8}$$

where $\mathbb{E}_q[.]$ is the expectation taken with respect to the density $q(Y, A, C) := p(Y|A, C) \times p^*(A) \times p(C)$.

3.3 ROBUSTNESS PROPERTIES

Just as $\Lambda_\eta^\perp$ in Section 2 entailed double robustness of $U(\beta)$ for semiparametric regression SDR, we now show that the structure of $\widetilde{\Lambda}_\eta^\perp$ yields additional robustness properties.

Lemma 3. *If one of $\big\{p(A|C),\ \mathbb{E}[\widetilde{U}(\beta)|A, C]\big\}$ and one of $\big\{\ell(g(A;\beta)) := \mathbb{E}_q[Y|g(A;\beta)],\ \nu(g(A;\beta)) := \mathbb{E}_q[\alpha(A)|g(A;\beta)]\big\}$ is correctly specified, then the estimator for β based on (8) is consistent and asymptotically normal with mean zero and variance $\tau^{-1} \times Var\big(\widetilde{U}(\beta^*)\big) \times \tau^{-1'}$, where $\widetilde{U}(\beta^*)$ is given in (8) and $\tau = \mathbb{E}[\partial\widetilde{U}(\beta^*)/\partial\beta]$.*

This result implies that the estimating equation in (8) yields a "2 × 2" robustness property. In practice, since we will be dealing with multidimensional problems, correct specification of models is difficult to ensure. However, robustness properties of semiparametric estimators also implies that in regions where sufficient subset of models are approximately correct, the overall bias remains small. If $p(A|C)$ and one of the models in $\widetilde{U}(\beta)$ is correctly specified, the AIPW estimator using (8) remains consistent for any choice of $\mathbb{E}\big[\widetilde{U}(\beta)|A, C\big]$. One promising direction of future work is to consider cases where $p(A|C)$ and $\widetilde{U}(\beta)$ are known and search for $\mathbb{E}[\widetilde{U}(\beta)|A, C]$ which yields good properties of the overall estimator.

4 ESTIMATION AND IMPLEMENTATION

In order to estimate the parameters β in 6, we need to solve the estimating equation $\mathbb{P}_n[\widetilde{U}(\beta^*)] = 0$, where $\widetilde{U}(\beta^*)$ is given in (8). For any $\widetilde{U}(\beta)$ of the form given in Section

3.1, Theorem 1, provides the class of all RAL estimators for β^* along with the most efficient estimator in this class. Under the general form of $\widetilde{U}(\beta) = \big\{Y - \ell(g(A;\beta))\big\} \times \big\{\alpha(A) - \nu(g(A;\beta))\big\}$, the term $\mathbb{E}\big[\widetilde{U}(\beta)|A,C\big]$ in $\widetilde{U}(\beta^*)$ equals $\big\{\mathbb{E}[Y|A,C] - \ell(g(A;\beta))\big\} \times \big\{\alpha(A) - \nu(g(A;\beta))\big\}$. Hence, in the expression in 8, four different models are involved in estimating $\widetilde{U}(\beta^*)$, namely (i) $\ell(g(A;\beta)) := \mathbb{E}_q[Y|g(A;\beta)]$, (ii) $\nu(g(A;\beta)) := \mathbb{E}_q[\alpha(A)|g(A;\beta)]$, (iii) $p(A|C)$, and (iv) $\mathbb{E}[Y|A,C] = \mathbb{E}_q[Y|A,C]$. The last term in (8) is equal to $\mathbb{E}_a\big[\mathbb{E}[U(\beta)|A,C]\big]$, where $\mathbb{E}_a[.]$ is the expectation wrt the marginal distribution of A which can be evaluated empirically without additional modeling.

For a pre-specified functional form of $\ell(g(A;\beta))$, we need to fit three different nuisance models. Given models $\nu(g(A;\beta);\eta_\nu)$, $p(A|C;\eta_a)$, and $\mathbb{E}[Y|A,C;\eta_y]$ for $\nu(g(A;\beta))$, $p(A|C)$, and $\mathbb{E}[Y|A,C]$, respectively, it can be shown that if $n^{\frac{1}{4}+\epsilon}(\widehat{\eta} - \eta_0)$ is bounded in probability for some $\epsilon > 0$, then the estimating equation $\mathbb{P}_n[\widetilde{U}(\beta^*);\widehat{\eta}] = 0$ yields an estimate of β with the same asymptotic properties as if the nuisance models were known. Here $\eta = \{\eta_\nu, \eta_a, \eta_y\}$, and $\widehat{\eta}$, η_0 denote the estimated and the true parameters of the nuisance models, respectively.

Theorem 2. *Let ϕ_0 denote the influence function of the estimator β obtained from the estimating equation $\mathbb{P}_n[\widetilde{U}(\beta^*,\eta_0)] = 0$. If $n^{\frac{1}{4}+\epsilon}(\widehat{\eta} - \eta_0)$ is bounded in probability for some $\epsilon > 0$, then the influence function corresponding to the estimator $\widehat{\beta}$ obtained from the estimating equation $\mathbb{P}_n[\widetilde{U}(\beta^*,\widehat{\eta})] = 0$ is the same as ϕ_0. In other words, $\widehat{\beta}$ follows the same asymptotic properties as if we knew the true nuisance models.*

The condition for the rate of convergence of nuisance models in Theorem 2 is a sufficient condition and is potentially too conservative. In practice, we might be able to use models with the slower convergence rates, see [Fisher and Kennedy, 2018] for more details. [Stone, 1982] provides a detailed analysis of the convergence rates of nonparametric models.

Implementation. We now describe in detail our procedure for estimating β by solving the empirical version of $\mathbb{E}[\widetilde{U}(\beta^*)] = 0$, where $\widetilde{U}(\beta^*)$ is given in (8). In what follows, we assume the structural dimension d, i.e. the cardinality of the range of $g(;\beta)$, is known; we provide a discussion on choosing the structural dimension at the end of this section.

For a given choice of $p^*(A)$ and $\alpha(A)$,

1. First estimate $\widehat{\eta}_a$ and $\widehat{\eta}_y$ in $p(A|C;\eta_a)$ and $\mathbb{E}[Y|A,C;\eta_y]$ by maximum likelihood or nonparametric methods. These two models do not depend on β and are not updated within the iterations below.

2. Pick starting values $\beta^{(1)}$.

3. At j^{th} iteration, given a fixed $\beta^{(j)}$, estimate $\widehat{\ell}(g(A;\beta^{(j)}))$ and $\widehat{\nu}(g(A;\beta^{(j)}))$, and compute:

$$U^q(\beta^{(j)}) = \big\{Y - \widehat{\ell}(g(A;\beta^{(j)}))\big\} \times \big\{\alpha(A) - \widehat{\nu}(g(A;\beta^{(j)}))\big\},$$
$$\mathbb{E}[U^q(\beta^{(j)}) \mid A,C] = \big\{\mathbb{E}[Y \mid A,C;\widehat{\eta}_y] - \widehat{\ell}(g(A;\beta^{(j)}))\big\} \times \big\{\alpha(A) - \widehat{\nu}(g(A;\beta^{(j)}))\big\}.$$

4. Form the sample version of $\mathbb{E}[\widetilde{U}(\beta^*)]$ as follows.

$$\zeta(\beta^{(j)}) = \mathbb{P}_n\Bigg[\frac{p^*(A)}{p(A \mid C;\widehat{\eta}_a)} \times \Big\{U^q(\beta^{(j)}) - \mathbb{E}\big[U^q(\beta^{(j)}) \mid A,C\big]\Big\} + \mathbb{E}_q\Big[\mathbb{E}\big[U^q(\beta^{(j)}) \mid A,C\big] \,\Big|\, C\Big]\Bigg]$$

   where $\mathbb{P}_n[.] := \frac{1}{n}\sum_{i=1}^n [.]_i$.

5. Calculate the first and second derivatives of $\partial\{||\zeta(\beta)||^2\}/\partial\{\beta\}$ numerically and evaluate them at $\beta^{(j)}$, and use the Newton-Raphson update rule to update $\beta^{(j)}$.

6. Repeat steps (3) through (5) until convergence.

The implementation of an empirical evaluation of (7) follows a similar set of steps, except all steps pertaining to second and third terms of (8) are skipped. Moreover, in step 3 of the above implementation, we need to specify individual models for $\ell(g(A;\beta)) := \mathbb{E}_q[Y|g(A;\beta)]$ and $\mathbb{E}[Y|A,C] := \mathbb{E}_q[Y|A,C]$. However, due to variation dependence of these models, it may be difficult to fit these two models in a congenial way in general. We provide an alternative approach in the following section.

In order to deal with the issue of congeniality, we may opt to specify $\mathbb{E}_q[Y|g(A;\beta)]$ and $\widetilde{f}(A,C,\beta) = \mathbb{E}_q[Y|A,C] - \mathbb{E}_q[Y|g(A;\beta)]$, which yield a variationally independent specification of $\mathbb{E}_q[Y|g(A;\beta)]$ and $\mathbb{E}_q[Y|A,C] = \mathbb{E}_q[Y|g(A;\beta)] + \widetilde{f}(A,C,\beta)$. Consequently, the four variationally independent models we need to specify are as follows: $\ell(g(A;\beta))$, $\nu(g(A;\beta))$, $p(A|C)$, and $\widetilde{f}(A,C,\beta)$; the last term in (8) can be evaluated empirically without additional modeling. Thus, we need to specify the additional nuisance model $\widetilde{f}$. We propose to fit $\widetilde{f}$ by borrowing ideas from the theory of structural nested mean models (SNMMs) in [Vansteelandt and Joffe, 2014, Robins, 1999]. We defer the descriptions to the appendix and refer to $\widetilde{f}$ as an "inverted" structural nested mean model.

Choosing the structural dimension. Up until here, we assumed the structural dimension was known a priori. Finding the correct dimension is not an straightforward task and incorrect choices may greatly affect performance. We adapt the technique in [Ma and Zhu, 2012] that was used to select the structural dimension in regression SDR to causal SDR. Specifically, we utilize a resampling procedure to select the structural dimension. This procedure was originally described by [Dong and Li, 2010] and adapts the

idea of [Ye and Weiss, 2003]. We consider a family of functions $g^1(.;\beta^1), \ldots, g^m(.;\beta^m)$ with different structural dimensions, and use the cross-validation procedure we describe below to pick the best dimension.

Let $\widehat{\beta}_\rho$ be the estimate of β from the original sample for the ρth working dimension, where $\rho = 1, \ldots, p-1$, and let $\widehat{\beta}_{\rho,b}$ be the estimate of β from the bth bootstrap sample, for $b = 1, \ldots, B$. The structural dimension can be estimated by finding the dimension ρ to be the cardinality of the range of the function

$$g^* = \arg\max_{g^i} \frac{1}{B} \sum_{b=1}^{B} r^2\big(g^i(A;\widehat{\beta}_\rho), g^i(A;\widehat{\beta}_{\rho,b})\big),$$

where $r^2(u,v) = k^{-1}\sum_{i=1}^{k} \lambda_i$ and λ_is are the non-zero eigenvalues of

$$\{\text{var}(u,v)\}^{-1/2}\,\text{cov}(u,v)\,\{\text{var}(v)\}^{-1}\,\text{cov}(v,u)\,\{\text{var}(u)\}^{-1/2}.$$

This procedure uses resampling to choose β to maximize variability of the reduced set of features given by $g^i(.;\beta^i)$ where $g^i(.;\beta^i)$ is chosen in a way that aims to preserve the causal regression relationship between A and the mean of Y. Exploring other alternatives for choosing the structural dimension is an interesting area for future work.

In the next section, we describe a set of simulations to illustrate the key results presented in this paper and provide a real-data analysis using a cohort of patients with head and neck cancer treated with radiation therapy. All code necessary to reproduce the results is available at https://github.com/raziehna/multidimensional-treatments.

5 SIMULATION STUDY

Causal SDR is not well-solved via standard methods for dimension reduction such as PCA, as they do not take the feature-outcome relationship into account, nor by standard SDR methods, as they do not take the confounding issues into account. We illustrate the utility of our proposal to causal SDR, via simulation studies, and compare them with regression SDR and PCA methods. We also illustrate the consistency of our estimators and illustrate the procedure for selecting the structural dimension. To provide continuity with previous work, our simulation study is similar to that described in [Ma and Zhu, 2012].

We perform 50 replications with fixed sample sizes, where the true response $\mathbb{E}[Y(g(a))]$ is an object of dimension $d = 2$, and the observed data distribution $p(Y, A, C)$ is set as follows. The dimension of the baseline factors C is fixed as 4 and the observed treatment dimension p is set to be 6 and 12. The baseline factors C are generated from a standard multivariate normal distribution. We consider two cases for the treatment vector: one where the linearity and the constant covariance conditions in regular SDR are violated, and one where these assumptions are satisfied.

Case 1. We generated $(A_1, A_2)^T$ (when $p = 6$) and $(A_1, A_2, A_{7:12})^T$ (when $p = 12$) from a multivariate normal distribution where the mean of each component is given as: $\mu_1 = \sum_i C_i$, $\mu_2 = \sum_i (-1)^i C_i$, $\mu_7 = C_1$, $\mu_8 = C_2$, $\mu_9 = C_3$, $\mu_{10} = -C_1 + C_2$, $\mu_{11} = -C_2 + C_3$, $\mu_{12} = -C_3 + C_4$, and the covariance matrix is $(\sigma_{ij})_{(p-4)\times(p-4)}$ where $\sigma_{ij} = 0.5^{|i-j|}$. We generated A_3 from a normal distribution with mean $|A_1 + A_2|$ and variance $|A_1|$. A_4 has a normal distribution with mean $|A_1 + A_2|^{1/2}$ and variance $|A_2|$. A_5 and A_6 were generated from Bernoulli distributions with success probabilities $\exp(A_2)/\{1 + \exp(A_2)\}$, and $\Phi(A_2)$, respectively, where $\Phi(.)$ denotes the standard normal cumulative distribution.

Case 2. The treatment vector is generated from a multivariate normal distribution where the mean of each component is given as follows. $\mu_1 = \sum_i C_i, \mu_2 = \sum_i (-1)^i C_i, \mu_3 = C_1 - C_2 - C_3 + C_4, \mu_4 = -C_1 + C_2 + C_3 - C_4, \mu_5 = \sum_i C_i - 2C_3, \mu_6 = \sum_i C_i - 2C_1$, and $\mu_{6+i} = C_i, \mu_{9+i} = -C_i$ for $i = 1, 2, 3$, and the covariance matrix is $(\sigma_{ij})_{p\times p}$ where $\sigma_{ij} = 0.5^{|i-j|}$.

The response variable is generated using

$$Y = A^T\beta_1 + (A^T)^2\beta_2 + \sum_{i=1}^{4} C_i + \Big\{\sum_{j=1}^{p} A_j\Big\} \times \Big\{\sum_{i=1}^{4} C_i\Big\} + \epsilon,$$

where the error term ϵ is generated from standard normal. For $p = 6$, we set $\beta_1 = (1,1,1,1,1,1)^T/\sqrt{6}$, and $\beta_2 = (1,-1,1,-1,1,-1)^T/\sqrt{6}$. For $p = 12$, the last 6 components of β_1 and β_2 are identically zero.

As mentioned in Section 3.2, Theorem 1 provides the whole class of estimating equations for a given $\widetilde{U}(\beta)$. For simplicity, we assume $\mathbb{E}[\alpha(A) \mid g(A;\beta)] = 0$, and therefore $\widetilde{U}(\beta) = \{Y - \ell(g(A;\beta))\} \times \alpha(A)$ in the following simulations. The performance of the estimates was computed using the distance between true β and $\widehat{\beta}$, defined as the Frobenius norm of the matrix $\widehat{\beta}(\widehat{\beta}^T\widehat{\beta})^{-1}\widehat{\beta}^T - \beta(\beta^T\beta)^{-1}\beta^T$.

Simulation 1. In this set of simulations, we aim for evaluating the performance of different estimation strategies for β and fix the sample size to 200. The results for both Case 1 and Case 2 when $p = 6$ are presented in Fig. 1, and the results for both Case 1 and Case 2 when $p = 12$ is deferred to the appendix. In each case, there are 4 different boxplots. The first one, from the left hand side, labeled as *Reg*, corresponds to semiparametric SDR estimating equation (1). Since regular SDR ignores the influence of confounding variables C, the estimates are not capturing the true causal relationship between A and Y. In the second boxplot, labeled as *IPW*, we use the IPW estimator in (7) with the correct model for $p(A|C)$, by properly adjusting for all the confounders. This recovers a more reasonable β^* estimate than the first one. However, while *IPW* generally performs better than PCA or regression SDR, the improvement is relatively modest. This might be due to the inefficiency of naive IPW estimators at the reported sample size. The third plot,

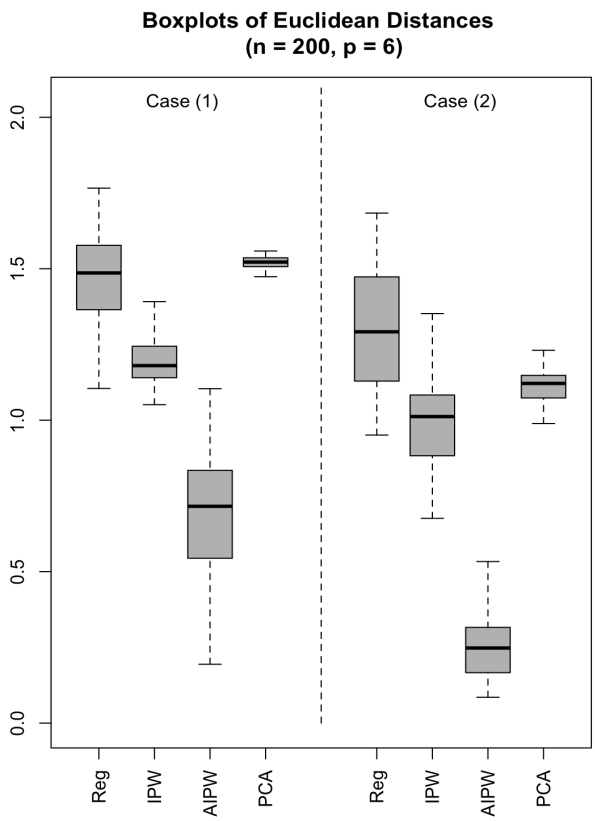


Figure 1: Boxplots of Frobenius norms between true and estimated parameters in simulations ($p = 6$).

labeled *AIPW*, uses the augmented IPW (AIPW) estimator corresponding to (8), which greatly outperforms the other estimators. The last plot corresponds to the classical PCA dimension reduction technique where the treatment-outcome relation is ignored. In this case, the first two principal directions are reported as estimating the basis of the lower dimensional space. As illustrated in the plots, this naive approach does not seek to preserve a causal, nor indeed *any*, relationship to the outcome.

Our main objective was to reduce the dimension of the treatment such that the cause-effect relation between treatment and outcome is preserved. In order to show that our estimating procedures actually preserve this relation, we compute the contrast between $E[Y(g(a_i;\beta))]$ and $\mathbb{E}[Y(g(a_j;\beta))]$ for $i, j = 1, \ldots, n$, given the true parameters and the estimated ones. The $n \times n$ heatmap of effects are provided in Fig. 2 for the true effects and the ones estimated by regular SDR and AIPW. We used 500 sample points generated from Case 2 with $p = 6$ to plot these heatmaps. The plots in 2(a) and (c) demonstrates the significant similarity between the true surface and the one estimated by AIPW. The surface estimated by regression SDR appears to be a very different surface. The root-mean-squared errors between the true causal surface and the ones estimated from AIPW and regular regression SDR are 0.48 and 14.29, respectively.

Simulation 2. We also evaluate the performance of our bootstrap procedure for estimating the structural dimension d, discussed in Section 4. We use the same data generating process as in Simulation 1, with $p = 6$ and $n = 200$. We set the bootstrap size to $B = 50$. The relative frequency of the selected dimension are reported in the appendix, and it reveals that the bootstrap procedure reliably recovers the true structural dimension, namely 2 in both cases (98% of the times in Case 1 and 90% of the times in Case 2.)

Simulation 3. We also demonstrate the effect of sample size on *IPW* and *AIPW* estimators of β in the causal SDR model. Results are revealed in the appendix.

6 DATA APPLICATION

We now illustrate our methods using a cohort of patients treated with radiation therapy for head and neck cancer. The cohort consists of 613 patients who received radiation therapy at the Johns Hopkins hospital prior to 2016. Radiation therapy is one of the most effective modalities for the treatment of head and neck cancers. However, because of the complex shape of target volumes in close proximity to sensitive organs, it may be associated with acute and late radiation morbidities such as xerostomia, mucositis, and dysphagia affecting the patient's quality of life. Such morbidities can lead to severe reduction in food intake and undesirable and possibly dangerous weight loss in patients. There are prospective studies that evaluated risk factors for weight loss in patients who undergo radiation therapy [Johnston et al., 1982, Cacicedo et al., 2014]. However, a proper analysis of whether radiation causes weight loss has not yet been reported likely due to the methodological challenges involved in using high dimensional variables such as radiation therapy as a treatment in causal analysis.

Here, we focus on the parotid glands which are incidentally irradiated by radiation and examine the summary measures of radiation therapy given by the cumulative dose-volume histograms extracted from the raw voxel maps of radiation doses. In particular, we looked at 5 equally spaced percentages of volume to construct a vector of treatment doses. We used weight loss as the outcome of interest, which was defined as the difference between weight measured within 100 to 160 days after the completion of treatment and the weight measured during consultation before the start of treatment. The data has records on demographics such as age, gender, race, and baseline clinical factors such as whether the patient had used feeding tubes and/or received chemotherapy before the initiation of treatment. We assumed these variables are sufficient to control for confounding and thus would ensure the conditional ignorability assumption was met. A copy of this dataset is available on https://github.com/raziehna/multidimensional-treatments.

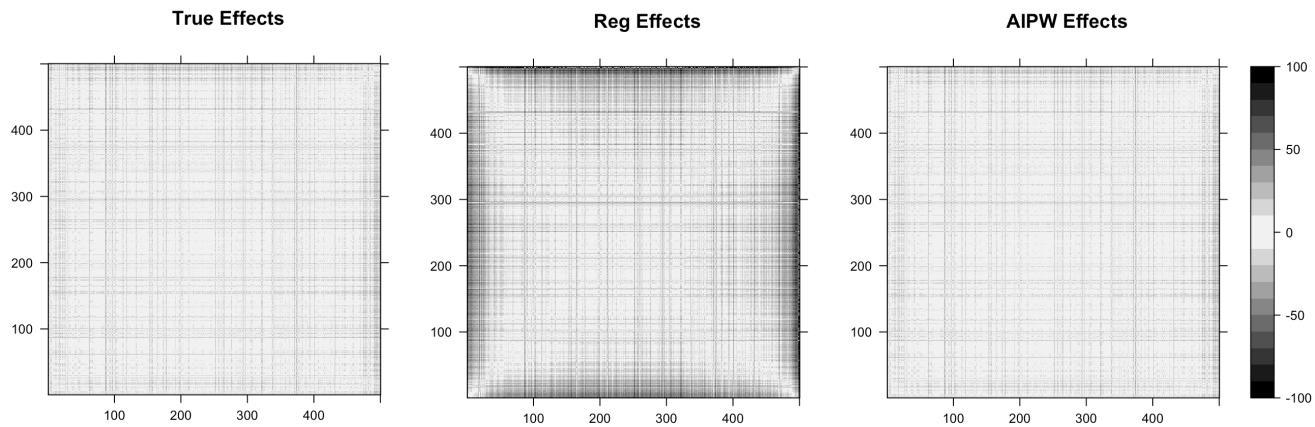


Figure 2: Heatmaps of true causal effects and effects computed by estimating β via the regular SDR and the AIPW estimators. Heatmaps are antidiagonally symmetric.

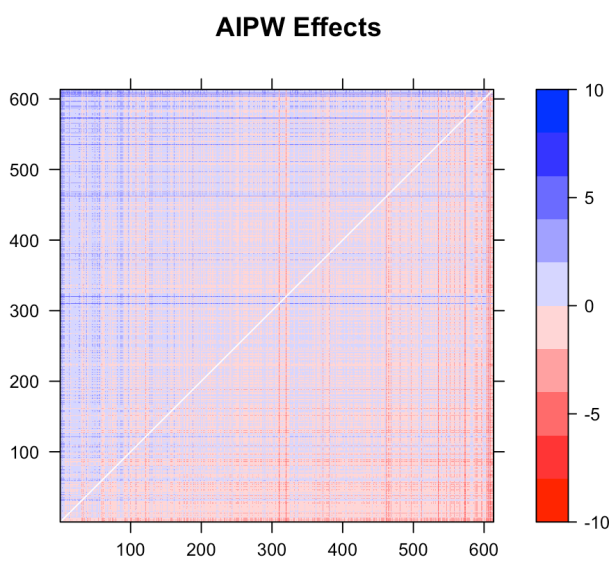


Figure 3: Heatmap to illustrate the causal effect of radiation on weight loss, where effects are computed by estimating β via AIPW estimator. Heatmap is antidiagonally symmetric with opposite color tones.

There exists a rich literature relating parotid dose-volume characteristics to radiotherapy-induced salivary toxicity. It has been shown that the mean dose to the parotid glands correlates strongly with xerostomia and salivary dysfunction which are risk factors of weight loss [Deasy et al., 2010]. In light of such studies, we assume there exists a single dimension in the radiation exposure that captures the relationships between exposure and side effects including weight loss. Therefore, we set the structural dimension d to be one. We set the mapping function $g(.;\beta)$ to be linear in its parameters β, and use Bayesian additive regression trees to fit all nuisance models. The code is provided as part of the supplementary materials. The oncology data were excluded for reasons of patient confidentiality.

We generated $n \times n$ heatmaps in Fig. 3 to illustrate the cause-effect relationship between radiation treatment and weight loss. We use AIPW estimator obtained from Theorem 1. The absolute values on the plots are antidiagonally symmetric. Radiation doses were sorted in increasing values along both axes. We interpret the heatmaps as follows. Consider the $(i,i)^{\text{th}}$ point on the plot and draw a line along the y-coordinate. Since radiation doses were sorted in increasing order, then the radiation value at any point on the line to the right of (i,i) is higher than the radiation value at the $(i,i)^{\text{th}}$ point. For any point to the left of (i,i), the radiation value is lower. The value at the $(k,i)^{\text{th}}$ coordinate corresponds to the contrast $\mathbb{E}\big[Y(g(a_k;\beta)) - Y(g(a_i;\beta))\big]$. Consequently, if $k > i$, then a red dot at (k,i) coordinate implies that an increase in radiation doses leads to an increase in weight loss. On the other hand, a blue dot would imply that an increase in radiation doses would not lead to an increase in weight loss. Similarly, a blue dot at (k,i), for $k < i$, would imply that a decrease in radiation leads to a decrease in weight loss. Reverse is implied when the dot is red. Focusing on the bottom right triangle, we note that most of the area is filled with red color. It implies that as we increase the amount of radiation, the severity of weight loss increases. Thus, radiation therapy is potentially a cause of weight loss among patients who undergo the treatment.

We investigated the relationship between the treatment and outcome as the treatment size increases by selecting larger numbers of equally spaced percentages of volume in the dose-volume histograms. The plots are provided in the supplement. Throughout the experiment, we examined the summary measures of radiation therapy given by the cumulative dose-volume histograms extracted from the raw 3D voxel maps of radiation doses. A more fine-tuned approach is to look at the exact dose localization information in the raw 3D voxel maps. A voxel-based approach would identify the relations between radiation-induced morbidity and local dose release, thus providing a potentially better insight

into spatial signature of radiation sensitivity in composite regions like the head and neck district [Monti et al., 2017]. Given the small cohort of patients that we have access to, a voxel-based approach would fall into $p \gg n$ paradigm, and would require strong sparsity assumptions [Li, 2007] to deal with. This is an interesting and challenging direction for future work.

7 CONCLUSIONS

In this paper, we have described a generalization of the semiparametric sufficient dimension reduction (SDR) approach for regression problems described in [Ma and Zhu, 2012] to causal SDR. Specifically, we developed a method that reduces the dimension of a multidimensional treatment, while preserving the causal relationship between the treatment and the outcome quantified as a counterfactual mean. Using ideas from structural models [Robins, 1999], we provided semiparametric estimators for parameters of the function that maps the multidimensional treatment to a lower dimensional subspace. We have shown our estimator exhibits "2x2 robustness," where the estimator remains consistent if one of two models, for two pairs of models, is chosen correctly.

Even though we use radiation therapy as our main motivation example, our methodology can be used in other applications as well. For instance, in natural language processing, a growing literature is focused on understanding causal effects of high dimensional text data; see [Feder et al., 2021]. Another example is relating neuronal network activity (collected via high dimensional neuroimaging data) to cognitive processing and behavior; see [Ramsey et al., 2010, Mather et al., 2013] for examples. Our proposed framework can also be useful in settings where we are interested in causal effects of multiple treatments simultaneously. In order to scale our methods to high dimensional applied settings, such as fMRI scans, text data, or radiation oncology voxel data, we need to incorporate ideas from parametric modeling, and sparsity within a semiparametric framework. Another natural extension for future work is to apply these methods to classical causal inference in longitudinal studies, where multiple time points render a collection of binary treatments a multidimensional object. Our causal SDR approach would provide an alternative to parametric marginal structural models typically employed in such settings.

Author Contributions

RN and IS contributed to conception and design of the statistical framework. RN performed the statistical analysis. TM provided the radiation database, and was the key in formulating the challenge of multidimensional treatments in oncology and radiation therapy. All authors contributed to the write up and revision of the manuscript.

Acknowledgements

We thank Todd McNutt's research group for their insightful discussions and facilitating our access to the database on head and neck cancer patients. We thank the anonymous reviewers for their comments. Ilya Shpitser is sponsored in part by NSF CAREER: 1942239, ONR N00014-21-1-2820, NIH R01 AI127271-01A1, and NSF 2040804.

References

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

Ershad Banijamali, Amir-Hossein Karimi, and Ali Ghodsi. Deep variational sufficient dimensionality reduction. *arXiv preprint arXiv:1812.07641*, 2018.

Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint arXiv:2003.12659*, 2020.

Jon Cacicedo, Francisco Casquero, Lorea Martinez-Indart, Olga Del Hoyo, Alfonso Gomez de Iturriaga, Arturo Navarro, and Pedro Bilbao. A prospective analysis of factors that influence weight loss in patients undergoing radiotherapy. *Chinese journal of cancer*, 33(4):204, 2014.

Debo Cheng, Jiuyong Li, Lin Liu, and Jixue Liu. Sufficient dimension reduction for average causal effect estimation. *arXiv preprint arXiv:2009.06444*, 2020.

RD. Cook and B. Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30:455–474, 2002.

RD. Cook and S. Weisberg. Discussion of sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86:28–33, 1991.

Joseph O Deasy, Vitali Moiseenko, Lawrence Marks, KS Clifford Chao, Jiho Nam, and Avraham Eisbruch. Radiotherapy dose–volume effects on salivary gland function. *International Journal of Radiation Oncology* Biology* Physics*, 76(3):S58–S63, 2010.

Y. Dong and B. Li. Dimension reduction for non-elliptically distributed predictors: Second-order moments. *Biometrika*, 97:279–294, 2010.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*, 2021.

A. Fisher and E. H. Kennedy. Visually communicating and teaching intuition for influence functions. *arxiv preprint: 1810.03260*, 2018.

M. Gentzkow, B. Kelley, and M. Taddy. Text as data. *Journal of Economic Literature*, 57:535–574, 2019.

W. Hardle and TM. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84:986–995, 1989.

Zonghui Hu, Dean A Follmann, and Naisyin Wang. Estimation of mean response via the effective balancing score. *Biometrika*, 101(3):613–624, 2014.

H. Ichimura. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, 58:71–120, 1993.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

Catherine A Johnston, Thomas J Keane, and Susan M Prudo. Weight loss in patients receiving radical radiation therapy for head and neck cancer: a prospective study. *Journal of Parenteral and Enteral Nutrition*, 6(5):399–402, 1982.

KC. Li. Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86:316–342, 1991.

L. Li. Sparse sufficient dimension reduction. *Biometrika*, 94:603–613, 2007.

Wei Luo and Yeying Zhu. Matching using sufficient dimension reduction for causal inference. *Journal of Business & Economic Statistics*, 38(4):888–900, 2020.

Shujie Ma, Liping Zhu, Zhiwei Zhang, Chih-Ling Tsai, and Raymond J Carroll. A robust and efficient approach to causal inference based on sparse sufficient dimension reduction. *Annals of statistics*, 47(3):1505, 2019.

Y. Ma and L. Zhu. A semiparametric approach to dimension reduction. *Journal of American Statistical Association*, 107:168–179, 2012.

Mara Mather, John T Cacioppo, and Nancy Kanwisher. How fmri can inform cognitive theories. *Perspectives on Psychological Science*, 8(1):108–113, 2013.

Serena Monti, Giuseppe Palma, Vittoria D?Avino, Marianna Gerardi, Giulia Marvaso, Delia Ciardo, Roberto Pacelli, Barbara A Jereczek-Fossa, Daniela Alterio, and Laura Cella. Voxel-based analysis unveils regional dose differences associated with radiation-induced morbidity in head and neck cancer patients. *Scientific Reports*, 7(1): 1–8, 2017.

Joseph D Ramsey, Stephen José Hanson, Catherine Hanson, Yaroslav O Halchenko, Russell A Poldrack, and Clark Glymour. Six problems for causal inference from fmri. *neuroimage*, 49(2):1545–1558, 2010.

S.P. Robertson, H. Quon, A. P. Kiess, J. A. Moore, W. Yang, Z. Cheng, S. Afonso, M. Allen, M. Richardson, A. Choflet, A. Sharabi, and T. R. McNutt. A data-mining framework for large scale analysis of dose-outcome relationships in a database of irradiated head and neck cancer patients. *Med Phys*, pages 4329–4337, 2015.

J. M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology: The Environment and Clinical Trials*. NY: Springer-Verlag, 1999.

James M. Robins, Miguel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

Susan M Shortreed and Ashkan Ertefaie. Outcome-adaptive lasso: variable selection for causal inference. *Biometrics*, 73(4):1111–1122, 2017.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*. AAAI Press, Palo Alto, 2006.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals ofStatistics*, 10:1040–1053, 1982.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Eighteenth National Conference on Artificial Intelligence*, pages 567–573, 2002. ISBN 0-262-51129-0.

Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer Science & Business Media, 2007.

Mark J Van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

S. Vansteelandt and M. Joffe. Structural nested models and g-estimation: The partially realized promise. *Statistical Science*, 29(4):707–731, 2014.

Z. Ye and R. E. Weiss. Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98:968–979, 2003.

APPENDIX

For clearer presentation of materials and equations in this supplement, we switch to a single-column format. Appendix A contains a brief overview of inference in semiparametric models. Appendix B contains additional discussion on how to ensure the nuisance models are trained in a congenial manner. Appendix C contains additional results with simulated data and real data application. Appendix D contains all the proofs.

A BRIEF OVERVIEW OF SEMIPARAMETRIC ESTIMATION

Let $Z_1, \ldots, Z_n$, be iid samples from a general class of probability densities $p(Z;\theta)$ parameterized by $\theta^T = (\beta^T, \eta^T)$, where $\beta \in \mathbb{R}^q$ denotes the set of target parameters, and η denotes a possibly infinite dimensional set of nuisance parameters. The goal of statistical inference in semiparametric models is to find "the best" estimator of β in the model, denoted by $\widehat{\beta}$. We will consider *regular asymptotically linear (RAL)* estimators, which are estimators of the form

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(Z_i) + o_p(1),$$

where $\phi \in \mathbb{R}^q$ with mean zero and finite variance, $o_p(1)$ denotes a term that approaches to zero in probability, and $\phi(Z_i)$ is the *influence function (IF)* of the ith observation for the parameter vector β. RAL estimators are consistent and asymptotically normal, with the variance of the estimator given by its IF,

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \mathbb{E}[\phi\phi^T]\big).$$

There is a bijective correspondence between RAL estimators and IFs. IFs provide a geometric view of the behavior of RAL estimators. Consider a Hilbert space $\mathcal{H}$ of all mean-zero $q-$dimensional functions, equipped with an inner product, defined between two arbitrary elements of the Hilbert space h_1 and h_2 as $\mathbb{E}[h_1^T h_2]$. The *nuisance tangent space* Λ in the semiparametric model is defined to be the mean square closure of elements of the nuisance tangent spaces $\Lambda_\beta = \{B^{q\times r} S_\eta(Z;\theta)\}$ of every *parametric submodel*. A parametric submodel is defined as a subset of densities in the semiparametric model parameterized by $\theta_\beta^T = (\beta^T, \eta_\beta^T)$, where $\eta_\beta^T \in \mathbb{R}^r$, such that the subset contains the density $p(Z;\theta_0)$ in the semiparametric model evaluated at the true parameter values θ_0. The space Λ is important because it is known all influence functions lie in the orthogonal complement $\Lambda^\perp$ of Λ with respect to $\mathcal{H}$. For this reason, recovering $\Lambda^\perp$ is often the first step for constructing RAL estimators in semiparametric models. Out of all IFs in $\Lambda^\perp$ there exists a unique one which lies in the tangent space, and which yields the most efficient RAL estimator by recovering the *semiparametric efficiency bound*, see Tsiatis [2007] for details.

B ADDITIONAL DISCUSSIONS

"Inverted" Structural Nested Mean Model. In order to deal with the issue of congeniality, we may opt to specify $\mathbb{E}_q[Y \mid g(A;\beta)]$ and $\widetilde{f}(A,C,\beta) = \mathbb{E}_q[Y \mid A,C] - \mathbb{E}_q[Y \mid g(A;\beta)]$, which yield a variationally independent specification of $\mathbb{E}_q[Y \mid g(A;\beta)]$ and $\mathbb{E}_q[Y \mid A,C] = \mathbb{E}_q[Y \mid g(A;\beta)] + \widetilde{f}(A,C,\beta)$. Consequently, the four variationally independent models we need to specify are as follows: $\ell(g(A;\beta))$, $\nu(g(A;\beta))$, $p(A|C)$, and $\widetilde{f}(A,C,\beta)$. The last term in (8) can be evaluated empirically without additional modeling. Thus, we need to specify the additional nuisance model $\widetilde{f}$. We propose to fit $\widetilde{f}$ by borrowing ideas from the theory of structural nested mean models (SNMMs) in [Vansteelandt and Joffe, 2014, Robins, 1999]. Unlike MSMs, which are regression models for causal relationships, SNMMs directly model the so called "blip effects," namely counterfactual differences between the response to a particular treatment, and a response to a reference treatment, given a particular observed trajectory. For a single treatment, this difference simplifies to $\gamma(A,C;\psi) = \mathbb{E}[Y(A) \mid A,C] - \mathbb{E}[Y(0) \mid A,C]$. Let $U_{sn}(\psi) := Y - \gamma(A,C;\psi)$. Consequently, $\mathbb{E}[U_{sn}(\psi) \mid A,C] = \mathbb{E}[Y(0) \mid A,C] = \mathbb{E}[Y(0) \mid C] = \mathbb{E}[U_{sn}(\psi)|C]$ (by conditional ignorability). This estimating equation then leads to a consistent estimation of parameters ψ:

$$\mathbb{P}_n\Big[\{d(A,C) - \mathbb{E}[d(A,C) \mid C]\} \times \{U_{sn}(\psi) - \mathbb{E}[U_{sn}(\psi) \mid C]\}\Big] = 0,$$

where $d(A,C)$ is a function of A and C with the same cardinality as ψ [Vansteelandt and Joffe, 2014]. Assuming $\widetilde{f}$ is parameterized by ψ, we now show that estimating ψ can be viewed as an estimation problem for a kind of "inverted SNMM."

Lemma 4. *Let $U_{dim}(\psi) = Y - \widetilde{f}(A, C, \beta; \psi)$, and fix any $d(A, C)$. If either $\mathbb{E}[d(A, C) \mid g(A; \beta)]$ or $\mathbb{E}[U_{dim}(\psi) \mid g(A; \beta)]$ are correctly specified, the following estimating equations yield a consistent estimator of ψ,*

$$\mathbb{P}_n\Big[\big\{d(A, C) - \mathbb{E}[d(A, C) \mid g(A; \beta)]\big\} \times \big\{U_{dim}(\psi) - \mathbb{E}[U_{dim}(\psi) \mid g(A; \beta)]\big\}\Big] = 0.$$

For the purposes of robustness, specifying both $\mathbb{E}[\widetilde{f} \mid g(A; \beta)]$ and $\mathbb{E}[U_{dim}(\psi) \mid g(A; \beta)]$ correctly is part of the correct specification of $\mathbb{E}[U(\beta) \mid A, C]$, given the type of estimation strategy we use.

The implementation provided in Section 4 can be modified to take advantage of modeling congenial models. Right before step (c), we need to estimate $\widehat{\widetilde{f}^{(j)}}\big(A, C, \beta^{(j)}; \widehat{\psi}\big)$ using Lemma 4, and modify step (c) by letting $\mathbb{E}[U^q(\beta^{(j)}) \mid A, C] = \widehat{\widetilde{f}^{(j)}} \times \big\{\alpha(A) - \widehat{\nu}(g(A; \beta^{(j)}))\big\}$. A downside of estimating congenial models is that the overall procedure becomes quite computationally intensive.

C ADDITIONAL EXPERIMENTS

In light of an anonymous reviewer's suggestion, we provide the steps for numerical computation of the first derivative. Computing the second derivative follows similarly.

- Step 1. Define a function that takes as input a vector-valued β and computes $\zeta(\beta)$ as defined in Step 4 of the implementation discussion.
- Step 2. Let K be the length of β, that is (# of rows) $\times$ (# of columns)
- Step 3. Let *deriv* be the vector that collects the derivatives of $\zeta(\beta)$ with respect to $\beta[i, j]$. Thus *deriv* is a vector of size K.
- Step 4. For $k \in 1 : K$ do:
  - (i) Let the kth element in the matrix β be perturbed by δ, as follows
    - (a) Let e be a matrix of all zeros with the same number of rows and columns as β
    - (b) $e[k] = 1$
    - (c) $\beta_k^+ = \beta + \delta \times e$
    - (d) $\beta_k^- = \beta - \delta \times e$
  - (ii) Compute $\zeta(\beta_k^+)$ and $\zeta(\beta_k^-)$
  - (iii) *deriv*$[k] = \{\sum \zeta^2(\beta_k^+) - \sum \zeta^2(\beta_k^-)\}/\{2 \times \delta\}$

C.1 SIMULATIONS

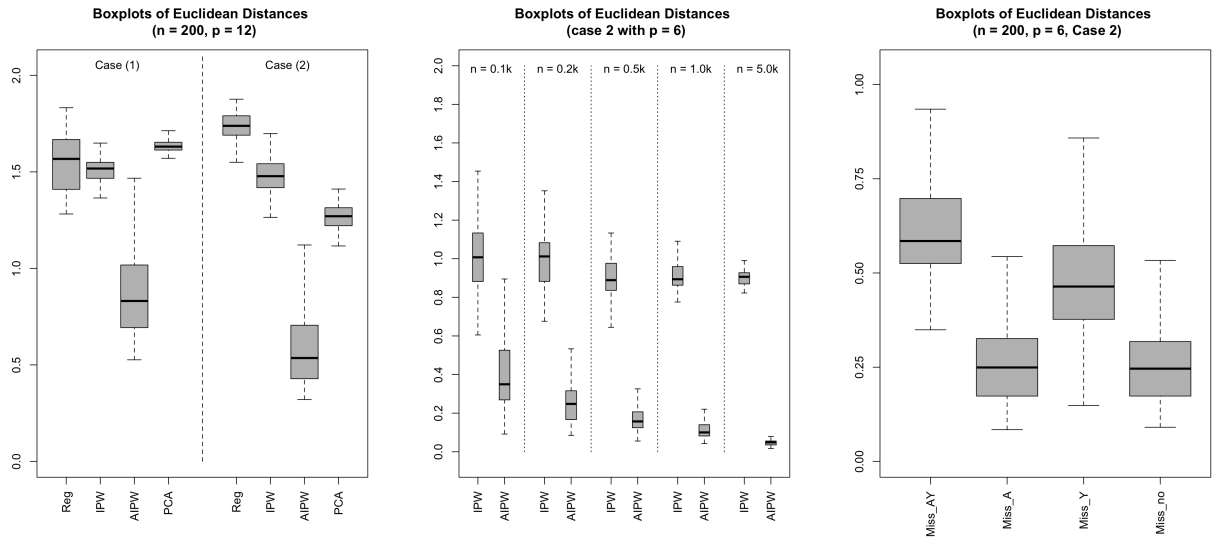
Simulation 1. The boxplots in Fig. 4(a) illustrate the performance of different estimation strategies for β for both Case 1 and Case 2, when sample size is set for 200 and $p = 12$.

Simulation 2. The relative frequency of the selected dimension are reported in the table below, which reveals that the bootstrap procedure reliably recovers the true structural dimension, namely 2.

Model ($p = 6$)	$\hat{d} = 1$	$\hat{d} = 2$	$\hat{d} = 3$	$\hat{d} = 4$	$\hat{d} = 5$
Case 1	0%	98%	2%	0%	0%
Case 2	0%	90%	10%	0%	0%

Simulation 3. In the third set of simulations, we demonstrated the effect of sample size on *IPW* and *AIPW* estimators of β in the causal SDR model. Results are shown in Fig. 4(b). While both estimators are consistent under our model specification, AIPW exhibits favorable convergence rates compared to IPW, as expected.

Simulation 4. In light of an anonymous reviewer's suggestion, we ran simulation to illustrates the robustness property of our AIPW-style estimator. For such illustration, we focused on Case 2 with $p = 6$ and fix the sample size to 200. We computed the Frobenius norms between true and estimated parameters under four different scenarios: (1) *miss-no* where all model are correctly specified, (2) *miss-A* where only the propensity model is misspecified, (3) *miss-Y* where only the



(a) (b) (c)

Figure 4: (a) Boxplots of Frobenius norms between true and estimated parameters in simulations ($p = 12$); (b) Illustration of the effect of sample size on the Frobenius norms between true and estimated parameters using data generated from Case 2 with $p = 6$. (c) Boxplots to illustrate the robustness property of our AIPW-style estimator.

outcome model is misspecified, and (4) *miss-AY* where both models are misspecified. For miss-specification of outcome model, we excluded the dependency of Y on C_3 and C_4. For miss-specification of the propensity model, we changed the mean of the multivariate normal distribution by altering the dependency of A on the baseline factors. We performed 100 iterations and the boxplots are provided in Fig. 4(c).

We assure the reviewer to include demonstrations of the robustness property in the camera ready revision.

C.2 REAL DATA

Assume treatment is collected using p equally spaced percentages of volume. In other words, treatment is assumed to be a vector in $\mathbb{R}^p$ where the i-th element corresponds to the radiation dose on p percentages of the parotid glands. The effect of radiation on weight loss is illustrated in Fig. 5 by allowing p to be 10 and 20, and reducing the size of treatment to one dimension. We use IPW estimators to calculate the effects. Both plots agree with our stated conclusion in the main body of the manuscript, i.e., radiation has a negative effect on weight loss.

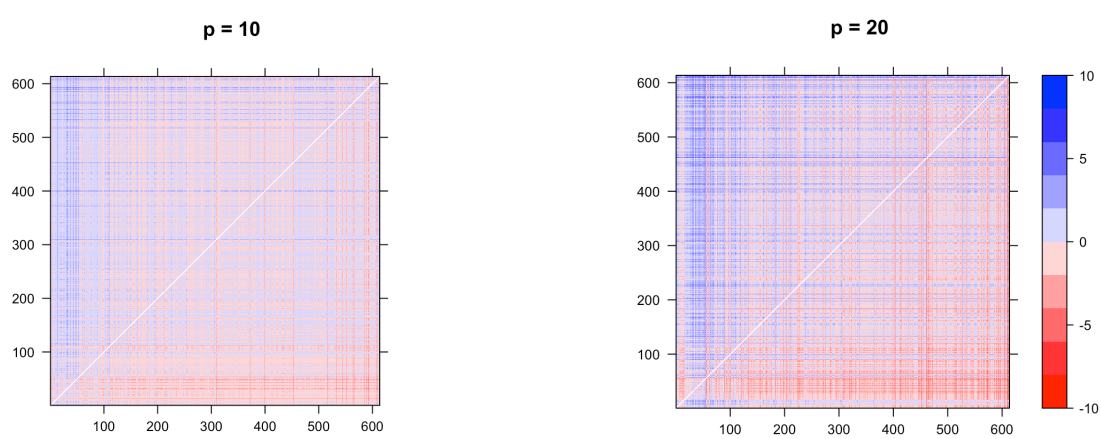


Figure 5: Heatmaps to illustrate the causal effect of radiation on weight loss. Treatment is collected using (right) 10 and (left) 20 equally spaced percentages of volume in parotid glands.

D PROOFS

Lemma 1.

Proof. Choosing $\phi(A, C) = 0$ in Theorem 1 yields (7). All elements of the orthocomplement of the nuisance tangent space are mean zero under the true distribution (we give an argument for elements of $\widetilde{\Lambda}_\eta^\perp$ in Proposition 1). Since $\widetilde{U}(\beta)$ exhibits double robustness, i.e. remaining consistent if either $\ell(g(A;\beta))$ or $\nu(g(A;\beta))$ is correctly specified [Ma and Zhu, 2012], the correct specification of $p(A \mid C)$ yields our conclusion. □

Proposition 1. *For all* $\tilde{U}(\beta^*) \in \widetilde{\Lambda}_\eta^\perp$, $\mathbb{E}[\widetilde{U}(\beta^*)] = 0$.

Proof. The second and third terms of $\widetilde{U}(\beta^*)$ are mean zero by construction. The first term, under truth with the property that $\mathbb{E}_q[Y \mid A] = \mathbb{E}_q[Y \mid g(A;\beta)]$, is

$$\begin{aligned}
\mathbb{E}\Big[\frac{p^*(A)}{p(A \mid C)} \times \widetilde{U}(\beta)\Big] &= \int \widetilde{U}(\beta) \times p(Y \mid A, C) \times p^*(A) \times p(C)\, d\mu_{Y,A,C} \\
&= \int \big\{Y - \ell(g(a;\beta))\big\} \times \big\{\alpha(A) - \nu(g(a;\beta))\big\} \times q(Y, A, C)\, d\mu_{Y,A,C} \\
&= \mathbb{E}_q\Big[\big\{Y - \ell(g(a;\beta))\big\} \times \big\{\alpha(A) - \nu(g(a;\beta))\big\}\Big] \\
&= \mathbb{E}_q\Big[\big\{\alpha(A) - \nu(g(a;\beta))\big\} \times \mathbb{E}_q\big[\big\{Y - \ell(g(a;\beta))\big\} \mid A = a\big]\Big] \\
&= \mathbb{E}_q\Big[\big\{\alpha(A) - \nu(g(a;\beta))\big\} \times \big\{\mathbb{E}_q[Y \mid A = a] - \ell(g(a;\beta))\big\}\Big] \\
&= 0.
\end{aligned}$$

since $\ell(g(a;\beta)) := \mathbb{E}_q[Y \mid A = a]$. Note that even if $\ell(g(a;\beta))$ is miss-specified, the expectation will still be zero if $\nu(g(a;\beta))$ is correctly specified, shown by iterative expectations.

□

Theorem 1.

Proof. This is a direct consequence of Theorems 3.1 and 3.2 in Robins [1999], and results in Appendix 3 of Ma and Zhu [2012]. □

Lemma 2.

Proof. Plugging in the optimal $\phi(A, C)$ yields $\widetilde{U}(\beta^*)$ to be

$$\frac{p^*(A)}{p(A \mid C)} \times \widetilde{U}(\beta) - \mathbb{E}\left[\frac{p^*(A)}{p(A \mid C)}\, \widetilde{U}(\beta)\middle| A, C\right] + \mathbb{E}\left[\mathbb{E}\left[\frac{p^*(A)}{p(A \mid C)}\, \widetilde{U}(\beta)\middle| A, C\right]\middle| C\right].$$

The conclusion follows, since

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left[\frac{p^*(A)}{p(A \mid C)}\, \widetilde{U}(\beta)\middle| A, C\right]\middle| C\right] \\
&\qquad = \mathbb{E}\left[\frac{p^*(A)}{p(A \mid C)} \mathbb{E}\left[\, \widetilde{U}(\beta)\middle| A, C\right]\middle| C\right] \\
&\qquad = \int \frac{p^*(A)}{p(A \mid C)}\, \mathbb{E}[\widetilde{U}(\beta) \mid A, C]\, p(Y, A \mid C)\, d\mu_{Y,A} \\
&\qquad = \int \mathbb{E}[\widetilde{U}(\beta) \mid A, C]\, p(Y \mid A, C)\, p^*(A)\, d\mu_{Y,A} \\
&\qquad = \int \mathbb{E}[\widetilde{U}(\beta) \mid A, C]\, q(Y, A \mid C)\, d\mu_{Y,A}
\end{aligned}$$

$$= \mathbb{E}_q\Big[\mathbb{E}\big[\widetilde{U}(\beta) \mid A, C\big]\Big|C\Big].$$

□

Lemma 3.

Proof. Assume either $\ell(g(A;\beta))$ or $\nu(g(A;\beta))$, and $p(A \mid C)$ are correctly specified. Consequently, the second and third terms in the expression of $\widetilde{U}(\beta^*)$ are both mean zero, even under an incorrect specification of $\mathbb{E}[\widetilde{U}(\beta) \mid A, C]$. Following the same the argument in Proposition 1, the first term is zero if either $\ell(g(A;\beta))$ or $\nu(g(A;\beta))$ is correctly specified.

Assume either $\ell(g(A;\beta))$ or $\nu(g(A;\beta))$, and $\mathbb{E}[\widetilde{U}(\beta) \mid A, C]$ are correctly specified. Consequently, the first two terms in the expression of $\widetilde{U}^*$ are both mean zero, even under an incorrect specification of $p^*(A \mid C)$. For the last term, we have:

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{E}_q\Big[\mathbb{E}\big[\widetilde{U}(\beta) \mid A, C\big] \;\Big|\; C\Big]\Big] \\
&\qquad = \mathbb{E}\Big[\mathbb{E}_q\Big[\int \widetilde{U}(\beta) \times p(Y \mid A, C)\, d\mu_Y \;\Big|\; C\Big]\Big] \\
&\qquad = \mathbb{E}\Big[\int \Big(\int \widetilde{U}(\beta) \times p(Y \mid A, C)\, d\mu_Y\Big) \times p^*(A)\, d\mu_A\Big] \\
&\qquad = \int \Big(\int\int \widetilde{U}(\beta) \times p(Y \mid A, C) \times p^*(A)\, d\mu_Y\, d\mu_A\Big) \times p(C)\, d\mu_C \\
&\qquad = \int \widetilde{U}(\beta) \times p(Y \mid A, C) \times p^*(A) \times p(C)\, d\mu_{Y,A,C} \\
&\qquad = \int \widetilde{U}(\beta) \times q(Y, A, C)\, d\mu_{Y,A,C} \\
&\qquad = \mathbb{E}_q\big[\widetilde{U}(\beta)\big].
\end{aligned}$$

We conclude the proof by noting that $\mathbb{E}\big[\widetilde{U}(\beta)\big]$ is mean zero if either $\ell(g(A;\beta))$ or $\nu(g(A;\beta))$ is correctly specified. Note that the normalized version of $\widetilde{U}(\beta^*)$, that is $\mathbb{E}[\frac{\partial \widetilde{U}(\beta^*)}{\partial \beta}]^{-1} \times \widetilde{U}(\beta^*)$, is an influence function that lives in the orthogonal complement of the tangent space $\widetilde{\Lambda}_\eta^\perp$. Therefore, the estimator obtained by solving $\mathbb{E}[\widetilde{U}(\beta^*)] = 0$ is RAL and is consistent and asymptotically normal with mean zero and variance equal to the variance of the influence function [Van der Vaart, 2000, Tsiatis, 2007]. □

Theorem 2.

Proof. Let $\beta \in \mathbb{R}^q$ and let η be infinite dimensional. We prove this theorem for the parametric submodel in the semiparametric model of $\{p(Z;\beta,\eta)\}$. With a slight abuse of notation, we denote $\eta \in \mathbb{R}^r$ to be the nuisance parameters within the parametric submodel. The Taylor series expansion of $\widetilde{U}(Z;\widehat{\beta}(\widehat{\eta}),\widehat{\eta})$ around β_0 is

$$\begin{aligned}
0 =& \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{U}(z_i;\widehat{\beta}(\widehat{\eta}),\widehat{\eta}) \\
=& \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{U}(z_i;\beta_0,\widehat{\eta})}_{(a)} + \underbrace{\frac{\partial}{\partial \beta}\left\{\frac{1}{n}\sum_{i=1}^n \widetilde{U}(z_i;\beta_0,\widehat{\eta})\right\}}_{(b)} \sqrt{n}(\widehat{\beta} - \beta_0) + o_p(1)
\end{aligned} \tag{9}$$

$$\begin{aligned}
(a) =& \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{U}(z_i;\beta_0,\eta_0) \\
&+ \frac{1}{n}\sum_{i=1}^n \left(\frac{\partial \widetilde{U}(z_i;\beta_0,\eta_0)}{\partial \eta}\right)_{q\times r} \times \sqrt{n}(\widehat{\eta} - \eta_0) \\
&+ \frac{1}{2}\underbrace{n^{1/4}(\widehat{\eta} - \eta_0)'}_{1\times 1\times r} \underbrace{\left(\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 \widetilde{U}(z_i;\beta_0,\eta_0)}{\partial^2 \eta}\right)}_{r\times q\times r \text{ (tensor)}} \underbrace{n^{1/4}(\widehat{\eta} - \eta_0)}_{r\times 1\times 1} + o_p(1)
\end{aligned}$$

$$(b) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \widetilde{U}(z_i;\beta_0,\eta_0)}{\partial \beta}\right)_{q\times q}}_{(b_1)} + \frac{\partial}{\partial \beta}\left\{\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \widetilde{U}(z_i;\beta_0,\eta_0)}{\partial \eta}\right)_{q\times r}}_{(b_2)} \times (\widehat{\eta}-\eta_0)_{r\times 1}\right\}$$

$$(b_1): \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \widetilde{U}}{\partial \beta}\right)_{q\times q} \longrightarrow \mathbb{E}_{\theta_0}\left[\frac{\partial \widetilde{U}}{\partial \beta}\right]_{q\times q} = -\mathbb{E}_{\theta_0}\left[\widetilde{U}(Z;\theta_0)S'_{\beta}(Z;\theta_0)\right]$$

$$(b_2): \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \widetilde{U}}{\partial \eta}\right)_{q\times r} \longrightarrow \mathbb{E}_{\theta_0}\left[\frac{\partial \widetilde{U}}{\partial \eta}\right]_{q\times r} = -\mathbb{E}_{\theta_0}\left[\widetilde{U}(Z;\theta_0)S'_{\eta}(Z;\theta_0)\right] = \mathbf{0}_{q\times r}$$

Since $n^{1/4}(\widehat{\eta}-\eta_0)$ and $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \widetilde{U}(z_i;\beta_0,\eta_0)}{\partial \eta}\right)_{q\times r}$ both converge in probability to zero, then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{U}(z_i;\beta_0,\widehat{\eta}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widetilde{U}(z_i;\beta_0,\eta_0) + o_p(1).$$

Therefore, from equation 9

$$\sqrt{n}(\widehat{\beta}-\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{-\mathbb{E}_{\theta_0}^{-1}\left[\frac{\partial \widetilde{U}(z_i;\beta_0,\eta_0)}{\partial \beta}\right]\widetilde{U}(z_i;\beta_0,\eta_0)\right\} + o_p(1)$$

Which concludes the proof. This procedure carries over to the case where the nuisance parameter is infinite dimensional, Tsiatis [2007]. □

Lemma 4.

Proof. Define $U_{dim}(\psi) = Y - \widetilde{f}(A,C,\beta;\psi)$. Therefore,

$$\mathbb{E}[U_{dim}(\psi) \mid A,C] = \ell(g(A;\beta)) = \mathbb{E}[U_{dim}(\psi) \mid g(A;\beta)].$$

This is a situation precisely isomorphic to single treatment SNMMs above, except with the roles of A and C reversed (hence this is an "inverted SNMM"). Our conclusion will then follow by results in [Robins et al., 2000, Vansteelandt and Joffe, 2014]. We provide a more detailed proof as follows. We have that $\widetilde{f}(A,C,\beta;\psi) = \mathbb{E}[Y \mid A=a, C=C] - \ell(g(a;\beta))$. Therefore,

$$\mathbb{E}[Y \mid A=a, C=C] = \ell(g(a;\beta)) + \widetilde{f}(a,C,\beta;\psi),$$

which we can rewrite as follows,

$$Y = \ell(g(a;\beta)) + \widetilde{f}(a,C,\beta;\psi) + \epsilon, \text{ s.t. } \mathbb{E}[\epsilon \mid C,a] = 0.$$

Observed data are instances of the form $\mathbf{Z} = (C,A,Y)$. The goal is to find semiparametric estimators for ψ in the semiparametric model $\mathcal{P} = \{p(\mathbf{z};\psi,\psi()), \mathbf{z} = (C,a,y)\}$ and the truth is $p_0(\mathbf{z}) = p(\mathbf{z};\psi_0,\eta_0())$. The observed data likelihood can be written as follows,

$$\begin{aligned} p(c,a,y) &= p(C,a)\times p(y \mid a,C) \coloneqq p(C,a)\times p(\epsilon \mid a,C) = \eta_1(C,a)\times\eta_2(\epsilon,a,C) \\ &= \eta_1(C,a)\times\eta_2\Big(y-\ell(g(a;\beta))-\widetilde{f}(a,C,\beta;\psi),a,C\Big), \end{aligned}$$

where $\epsilon = Y - \ell(g(a;\beta)) - \widetilde{f}(a,C,\beta;\psi)$, $\eta_1(C,a)$ denotes the nuisance model for $p(C,a)$, and $\eta_2(\epsilon,a,C)$ denotes the nuisance model for $p(\epsilon \mid a,C)$, which is any density such that $\mathbb{E}[\epsilon \mid a,C] = 0$. ψ is the parameter of interest and the nuisance parameters are $\{\eta_1,\eta_2,\ell(g(a;\beta))\}$.

The nuisance tangent space of this semiparametric model, Λ, is defined as the mean-square closure of parametric submodel nuisance tangent spaces:

$$\mathcal{P}_{\psi,\zeta} = \left\{p(z;\psi,\psi_\zeta) = p(c,a;\zeta_1)\times p(\epsilon \mid a,C;\zeta_2)\right\}$$

$$= \Big\{ p(C,a;\zeta_1) \times p\big(y - \ell(g(a;\beta)) - \widetilde{f}(a,C,\beta;\psi) \mid a,C;\zeta_2\big) \Big\},$$

where ζ_1, ζ_2 are r_1, r_2 dimensional vectors. Thus nuisance parameters in parametric submode are finite dimensional, $\zeta = \{\zeta_1, \zeta_2, \ell(g(a;\beta))\}$.

$$\begin{aligned}
\Lambda_\zeta &= \{B \times S_\zeta, \forall B\}, \\
S_\zeta &= \frac{\partial\{\text{log likelihood of the submodel evaluated at truth}\}}{\partial \zeta} \\
&= \left\{ \Big(\frac{\partial \log p(z;\psi,\zeta)}{\partial \zeta_1}\Big), \Big(\frac{\partial \log p(z;\psi,\zeta)}{\partial \zeta_2}\Big), \Big(\frac{\partial \log p(z;\psi,\zeta)}{\partial \ell(g(a;\beta))}\Big) \right\}\Bigg|_{\psi_0,\zeta_0} \\
&= \Big\{ S_{\zeta_1}(z;\psi_0,\zeta_0), S_{\zeta_2}(z;\psi_0,\zeta_0), S_{\ell(g(a;\beta))}(z;\psi_0,\zeta_0) \Big\}.
\end{aligned}$$

Hence, $\Lambda_\zeta = \Lambda_{\zeta_1} + \Lambda_{\zeta_2} + \Lambda_{\ell(g(a;\beta))}$. S_{ζ_1} should satisfy the density conditions. In addition, S_{ζ_2} should satisfy the condition that $\mathbb{E}[\epsilon \mid a, C] = 0$. We derive each of these subspaces using theorems in [Tsiatis, 2007] as a guideline.

(Theorem 4.6)	$\Lambda_{\zeta_1} = \{f(C,a); \mathbb{E}[f] = 0\}$
(Theorem 4.7)	$\Lambda_{\zeta_2} = \{f(\epsilon,a,C); \mathbb{E}[f \mid a,C] = 0, \mathbb{E}[\epsilon f \mid a,C] = 0\}$
(Lemma 4.3)	$\Lambda_{\zeta_1}^\perp = \{g(\epsilon,a,C); \mathbb{E}[g \mid a,C] = 0\}$
(Theorem 4.8)	$(\Lambda_{\zeta_1} + \Lambda_{\zeta_2})^\perp = \{g(C,a)\epsilon\}$
(Equation 10)	$\Lambda_{\ell(g(a;\beta))} = \{\frac{\psi'_{2\epsilon}(\epsilon,C,a)}{\psi_2(\epsilon,C,a)} f(g(a;\beta))\}$

In order to derive $\Lambda_{\ell(g(a;\beta))}$, we write down the corresponding score function as follows.

$$\begin{aligned}
S_{\ell(g(a;\beta))} &= \frac{\partial \log p(z;\psi,\zeta)}{\partial \ell(g(a;\beta))}\Big|_{\psi_0,\zeta_0} \\
&= \frac{\partial \log \Big(\psi_1(C,a;\zeta_{10}) \times \psi_2\big(y - \ell(g(a;\beta)) - \gamma(C,a;\psi), C, a; \zeta_{20}\big)\Big)}{\partial \ell(g(a;\beta))} \\
&= \frac{\partial \log \psi_2\big(y - \ell(g(a;\beta)) - \gamma(C,a;\psi), l, a; \zeta_{20}\big)}{\partial \ell(g(a;\beta))} \\
&= \frac{\partial \log \psi_2\big(\epsilon, C, a; \zeta_{20}\big)}{\partial \epsilon} \times \frac{\partial \epsilon}{\partial \ell(g(a;\beta))} \quad (\epsilon \text{ is a function of } \ell(g(a;\beta))) \\
&= \frac{\psi'_{2\epsilon}(\epsilon,C,a)}{\psi_2(\epsilon,C,a)} f(g(a;\beta)). \qquad (10)
\end{aligned}$$

In order to derive $\Lambda_\zeta^\perp$, we proceed as follows. Since $\Lambda_\zeta = \Lambda_{\zeta_1} + \Lambda_{\zeta_2} + \Lambda_{\ell(g(a;\beta))}$ and $\Lambda_{\zeta_1} + \Lambda_{\zeta_2} \subset \Lambda_\zeta$, then $\Lambda_\zeta^\perp \subset (\Lambda_{\zeta_1} + \Lambda_{\zeta_2})^\perp = \{g(c,a)\epsilon\}$. Similarly, $\Lambda_\zeta^\perp \subset \Lambda_{\ell(g(a;\beta))}^\perp$, therefore $\Lambda_\zeta^\perp = \{(\Lambda_{\zeta_1} + \Lambda_{\zeta_2})^\perp \cap \Lambda_{\ell(g(a;\beta))}^\perp\}$.

Pick an arbitrary element in $(\Lambda_{\zeta_1} + \Lambda_{\zeta_2})^\perp$, and denote it by $d(C,a)\epsilon$. For $d(C,a)\epsilon$ to be an element in $\Lambda_\zeta^\perp$, it needs to be orthogonal to every element in $\Lambda_{\ell(g(a;\beta))}$. Pick an arbitrary element in $\Lambda_{\ell(g(a;\beta))}$ and denote it by $\frac{\psi'_{2\epsilon}}{\psi_2} h(g(a;\beta))$. We have,

$$\begin{aligned}
\forall h(g(a;\beta)) \quad 0 &=< d(C,a)\epsilon, \frac{\psi'_{2\epsilon}}{\psi_2} h(g(a;\beta)) > \\
&= \mathbb{E}\big[d(C,a)\epsilon \frac{\psi'_{2\epsilon}}{\psi_2} h(g(a;\beta))\big] \\
&= \mathbb{E}\big[d(C,a) h(g(a;\beta))\big].
\end{aligned}$$

Consequently, $\forall h(g(a;\beta))$:

$$0 = \mathbb{E}\big[d(C,a) \times h(g(a;\beta))\big]$$

$$
\begin{aligned}
&= \mathbb{E}\Big[\mathbb{E}\big[d(C,a) \times h(g(a;\beta)) \;\big|\; g(a;\beta)\big]\Big] \\
&= \mathbb{E}\Big[h(g(a;\beta)) \times \mathbb{E}\big[d(C,a) \;\big|\; g(a;\beta)\big]\Big] \\
&= \mathbb{E}\big[h(g(a;\beta))\big] \times \mathbb{E}\big[d(C,a) \;\big|\; g(a;\beta)\big].
\end{aligned}
$$

Therefore, $\mathbb{E}[\; d(C,a) \mid g(a;\beta) \;] = 0$ and

$$
\begin{aligned}
\Lambda_\zeta^\perp &= \Big\{\big(d(C,a) - \mathbb{E}[d(C,a) \mid g(a;\beta)]\big) \times \epsilon\Big\} \\
&= \Big\{\big(d(C,a) - \mathbb{E}[d(C,a) \mid g(a;\beta)]\big) \times \big(Y - \gamma(C,a;\psi) - \ell(g(a;\beta))\big)\Big\} \\
&= \Big\{\big(d(C,a) - \mathbb{E}[d(C,a) \mid g(a;\beta)]\big) \times \big(U(\psi) - \mathbb{E}[U(\psi) \mid C,a]\big)\Big\}.
\end{aligned}
$$

Note that $\mathbb{E}[U(\psi) \mid C,a] = \mathbb{E}_q[Y \mid g(a;\beta)] = \mathbb{E}[U(\psi) \mid g(a;\beta)]$. Hence,

$$
\Lambda_\zeta^\perp = \Big\{\big\{d(C,a) - \mathbb{E}[d(C,a) \mid g(a;\beta)]\big\} \times \big\{U(\psi) - \mathbb{E}[U(\psi) \mid g(a;\beta)]\big\}\Big\}.
$$

$\square$