Semi-Parametric Causal Sufficient Dimension Reduction Of High Dimensional Treatments

Razieh Nabi
Computer Science Department
Johns Hopkins University
rnabiab1@jhu.edu

Ilya Shpitser
Computer Science Department
Johns Hopkins University
ilyas@cs.jhu.edu

Abstract

Cause-effect relationships are typically evaluated by comparing the outcome responses to binary treatment values, representing cases and controls. However, in certain applications, treatments of interest are continuous and high dimensional. For example, in oncology, the causal relationship between severity of radiation therapy, represented by a high dimensional vector of radiation exposure values at different parts of the body, and side effects is of clinical interest. In such circumstances, a more appropriate strategy for making interpretable causal inferences is to reduce the dimension of the treatment. If individual elements of a high dimensional feature vector weakly affect the outcome, but the overall relationship between the feature vector and the outcome is strong, careless approaches to dimension reduction may not preserve this relationship.

The literature on sufficient dimension reduction considers strategies that avoid this issue. Parametric approaches to sufficient dimension reduction in regression problems [6] were generalized to semi-parametric models in [7]. Methods developed for regression problems do not transfer in a straightforward way to causal inference due to complications arising from confounding. In this paper, we use semi-parametric inference theory for structural models [12] to give a general approach to causal sufficient dimension reduction of a high dimensional treatment.

1 Introduction

In causal inference we are interested in finding the causal effect of a treatment on an outcome in a randomized trial or an observation study. Typically, the received treatment is represented by a binary random variable, where 1 corresponds to receiving the treatment itself, and 0 corresponds to receiving a placebo. In some applications, treatments may take on continuous values in $\mathbb{R}$. For example, we might be interested in evaluating the effect of a particular treatment dose on viral load. In such 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 treatments with values that lie in some high dimensional space $\mathbb{R}^p$. An example of such a treatment is radiation exposure in oncology. Radiation dose for each organ or structure of clinical interest is stored as values in a three dimensional voxel map. Of clinical interest in oncology is the effect of radiation therapy represented in this way on tumor size or on post-treatment side effects.

High dimensional problems pose a particular challenge for data science due to sparse sample sizes and the curse of dimensionality. However, it is frequently the case that a high dimensional problem may be viewed without much loss via a lower dimensional representation. A vast literature on dimension reduction has developed an array of methods to take advantage of this observation [1, 3, 4, 5, 6, 8, 15]. In this paper, we consider a particular type of dimension reduction called sufficient dimension reduction (SDR) which seeks to find a low dimensional representation of the feature space $X$ that preserves some relationship with an outcome of interest $Y$ [6].

Given an outcome variable $Y$ and a $p$-dimensional covariate vector $X$, the goal of SDR is to find a known function $f(\cdot)$ with a much smaller range than domain
such that \( Y \) depends on \( X \) only through \( f(X) \). 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 \( X^T \beta \). The set of matrices \( \beta \) such that the conditional distribution of \( Y \) depends only on \( X^T \beta \), i.e. \( F(y \mid X) = F(y \mid X^T \beta) \), where \( F(y \mid X) = Pr(Y \leq y \mid X) \) is called the central subspace, and is denoted by \( S_{Y \mid X} \). If the conditional mean of \( Y \) depends only on \( X^T \beta \), i.e. \( E(Y \mid X) = E(Y \mid X^T \beta) \), the set of matrices \( \beta \) is called the central mean subspace, and is denoted by \( S_{E[Y \mid X]} \).

There exist a rich literature on how to derive the central (mean) subspace. Examples include, but not limited to, sliced inverse regression [15], average derivative estimation [21], inverse regression [1], kernel estimates to, sliced inverse regression [6], sliced average variance estimation [8], and principal Hessian directions [9]. However, all these proposed solutions to SDR suffer from relying on strong parametric assumptions that are unlikely to hold in practical applications, such as linearity condition where \( E[X \mid X^T \beta] \) is a linear function of \( X \), or \( \text{cov}(X \mid X^T \beta) \) being a constant rather than a function of \( X \). Ma et al. [7] took a detour from the existing literature and introduced a completely new approach to SDR by recasting the problem in terms of estimation in a semi-parametric model. Crucially, this approach relies on far weaker assumptions than is typical in SDR, and is thus much more generally applicable.

We are interested in applying SDR ideas to reducing the 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 semi-parametric approach in [7]. In this paper, we use semi-parametric inference theory in structural models [12] to give what we believe is the first approach to causal SDR of a high dimensional treatment.

The paper is organized as follows. We review causal inference in Section 2 and sufficient dimension reduction (SDR) in Section 3. In Section 4, we describe our approach to semi-parametric causal SDR, and derive certain useful properties of our semi-parametric estimators. In Section 5 we describe the estimation and implementation strategy of our estimators in more detail, and report simulation study results in Section 6. Our conclusions are in Section 7.

## 2 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 means of potential outcomes \( Y(1) \) and \( Y(0) \) representing a hypothetical randomized controlled trial where units are assigned randomly to cases (corresponding to \( A = 1 \)), or controls (corresponding to \( A = 0 \)). A common setting considers, in addition to \( A \) and \( Y \), a vector of baseline variables \( C \) (the set of covariates is divided into \( C \) and \( A \)), and an observed data distribution of the form \( p(Y \mid A, C)p(A \mid C)p(C) \). The difficulty in this setting is observed treatment assignments are not properly randomized, but instead were given possibly biased assignments depending on values of \( C \), via the conditional distribution \( p(A \mid C) \).

Under standard assumptions of consistency, which relates observed and counterfactual versions of \( Y \) as \( Y = Y(1)A + Y(0)(1 - A) \), and conditional ignorability which states that \( \{Y(1), Y(0)\} \) is independent of \( A \) conditional on \( C \), and positivity of \( p(A \mid C) \), any counterfactual distribution \( p(Y(a)) \), for any value \( a \) of \( A \) is the following function of the observed data: \( \sum C \) p(Y | a, c)p(c). Comparisons based on counterfactual distributions are often done on the mean difference scale. A common comparison is the average causal effect (ACE), given as the following function of the observed data, called the adjustment formula:

\[
E[Y(1)] - E[Y(0)] = \sum c (E[Y \mid 1, c] - E[Y \mid 0, c])p(c).
\]

One common estimator of the ACE is based on inverse probability weighting (IPW), which seeks to compensate for a biased treatment assignment by reweighting observed outcomes by the inverse of the normalized treatment assignment probability. For every row \(i\), this probability \( p(a \mid c_i) \) is either known by design, or estimated via a statistical model \( p(A \mid C; \alpha) \). In the latter case, the estimator is

\[
\frac{1}{n} \sum_i \left( \frac{1(\hat{A} = 1)Y_i}{W_i(c_i; \hat{\alpha})} - \frac{1(\hat{A} = 0)Y_i}{W_0(c_i; \hat{\alpha})} \right),
\]

where \( 1(.) \) is the indicator function, \( W_a(c_i; \hat{\alpha}) = p(a \mid c_i; \hat{\alpha})/\left( \sum p(a \mid c_i; \hat{\alpha}) \right) \), and \( \hat{\alpha} \) is the maximum likelihood estimate of \( \alpha \). In subsequent sections, we will consider a version of this problem where \( A \) is not a binary variable, but a high dimensional vector.

### 2.1 Semi-Parametric Estimation of the ACE

A semi-parametric model is a set of densities \( p(Z; (\beta, \eta)) \), where \( \beta \) is a finite dimensional set of target parameters, and \( \eta \) is an infinite dimensional set of nuisance parameters. A rich theory of regular asymptotically linear (RAL) estimators, and their corresponding influence functions (IFs) has been developed for semi-parametric models [2] [13]. In particular, IFs for \( \beta \) are closely related to elements of the orthogonal complement of the nuisance tangent space. See the Appendix for
a brief review, and [13] for more details. The counterfactual mean \( E[Y(a)] = \beta(a) \) identified via the adjustment formula can be viewed as a target parameter in a semi-parametric model, yielding the following IPW [2]:

\[
\frac{I(A = a)}{p(A \mid C)} \{ Y - E[Y \mid A, C] \} + E[Y \mid A = a, C] - \beta(a).
\]

This immediately yields the following augmented IPW (AIPW) estimator for the ACE \( \beta(1) - \beta(0) \):

\[
\frac{1}{n} \sum_{i} \frac{I(A = 1) \{ Y - E[Y \mid A, C; \hat{\eta}_b] \}}{p(A \mid C; \hat{\eta}_b)} + E[Y \mid A = 1, C; \hat{\eta}_b] - \frac{I(A = 0) \{ Y - E[Y \mid A, C; \hat{\eta}_b] \}}{p(A \mid C; \hat{\eta}_b)} - E[Y \mid A = 0, C; \hat{\eta}_b],
\]

where \( \hat{\eta}_a, \hat{\eta}_b \) are maximum likelihood estimates for some models of \( p(A \mid C) \) and \( E[Y \mid A, C] \). This estimator exhibits the property of double robustness where the estimate for \( \beta(1) - \beta(0) \) is consistent when either of the two models is correctly specified, even if the other model is arbitrarily misspecified.

### 2.2 Marginal Structural Models

It is possible to generalize IPW estimators to settings with multiple treatments. For example, assume we have \( A = (A^1, \ldots, A^k) \), and \( L^0, \ldots, L^{k-1} \) along with \( Y \) such that the variables obey a temporal ordering \( L^0, A^1, L^1, \ldots, L^{k-1}, A^k, Y \). For every \( a^i \), define \( a^{<i} \equiv \{ a^1, \ldots, a^{i-1} \} \), and \( L^{<i} \equiv \{ L^0, \ldots, L^{i-1} \} \). Assuming sequential ignorability holds, such that \( E[Y(a)] = E[Y(a^1, \ldots, a^k)] \) is identified via the g-computation algorithm [11]:

\[
E[Y \mid a^1, \ldots, a^k, L^0, \ldots, L^{k-1}] \prod_{i=0}^{k-1} p(L^i \mid L^{<i}, a^{<i}),
\]

the following IPW estimator of \( E[Y(a)] \) is consistent:

\[
\frac{1}{n} \sum_{i} \frac{I(A = a^i) Y_i}{W_a(l_i; \hat{\eta}_a)} W_a(l_i; \hat{\eta}_a) \prod_{i=1}^{k-1} p(a^i \mid l^{<i}, a^{<i}; \hat{\eta}_a) \sum_{j} \prod_{i=1}^{k} p(a^i \mid l^{<i}, a^{<i}; \hat{\eta}_a).\]

The difficulty with this estimator in practice is that even for a moderately sized set \( A \), the amount of observed data for any particular value assignment \( a \) is small or zero. The way to “rescue” IPW approaches here is to borrow strength across observed realizations of \( A \) by positing a marginal structural model (MSM), or a causal regression. A simple version of such a model takes the form \( E[Y(a)] = f(a; \beta) \), for some finite \( \beta \). Note: this is not an ordinary regression, since \( E[Y(a)] \neq E[Y \mid a] \), but is instead equal to \( f(a; \beta) \). Given such a model, inferences about \( E[Y(a)] \) reduce to inferences about \( \beta \). One approach to estimating \( \beta \) here is to solve a modified set of estimating equations for regression problems appropriately reweighted by \( W_a(l_i; \hat{\eta}_a) \):

\[
0 = E \left[ \frac{p^*(a) (Y - f(a; \beta))}{W_a(l_i; \hat{\eta}_a)} \right] = \sum_{i} p^*(a) \frac{Y_i - f(a; \beta)}{W_a(l_i; \hat{\eta}_a)},
\]

where \( p^*(a) \) is a user-specified non-degenerate density.

### 3 Sufficient Dimension Reduction

#### 3.1 Sliced Inverse Regression

In his seminal paper [6], Li proposed the sliced inverse regression (SIR) algorithm, as an elegant approach to SDR. He considered the inverse regression curve \( E(X \mid Y) \) to derive the basis for \( S_{E[Y \mid X]} \). \( E(X \mid Y) \) is a \( p \)-dimensional curve in \( \mathbb{R}^p \). Under certain distributional assumption, known as the linearity condition, this curve lies on a \( d \)-dimensional subspace, where \( d \) is the dimension of \( S_{E[Y \mid X]} \). The linearity condition is satisfied when the distribution of \( X \) is elliptic symmetry (e.g., the normal distribution). Li showed under this condition, the eigenvectors associated with the \( d \) largest eigenvalues of \( \text{Cov}(E[X \mid Y]) \), after centralizing \( X \), span the central mean subspace.

Other parametric approaches have been proposed over the last decade on improving SIR, such as sliced variance estimation [5], or directional regression [1]. However, these algorithms are accompanied with more unrealistic assumptions that restrict their use in practice. Ma et al [7] introduced a general class of estimating equations in SDR that gets rid of these distributional assumptions. They illustrated how these previously proposed algorithms are special cases of their semi-parametric approach.

In the remainder of the paper, we focus on describing existing semi-parametric methods for deriving the central mean subspace, and deriving the causal generalization of these methods.

#### 3.2 Semiparametric Approach to SDR

The conditional mean model in sufficient dimension reduction should satisfy \( E[Y \mid x] = E[Y \mid x^T \beta] \). This condition is sometimes written as \( Y = l(x^T \beta + \epsilon) \), where \( l(x^T \beta) = E[x^T \beta] \), is an unspecified smooth function, and \( E[\epsilon \mid x] = 0 \), while the distribution \( p(\epsilon \mid x) \) being otherwise unrestricted. In this semi-parametric model, \( \beta \) are the target parameters, and the infinite dimensional set of parameters for \( p(\epsilon \mid x) \) are nuisance.

Ma et al [7] derived the orthocomplement of the nuisance tangent space for this model as:

\[
A_{\beta}^\perp = \{ Y - E[Y \mid x^T \beta] \} \{ \alpha(x) - E[\alpha(x) \mid x^T \beta] \},
\]

where \( \alpha(x) \) is an arbitrary function of \( x \). A well-known property of semi-parametric models is that all elements of \( A_{\beta}^\perp \) are mean 0 under the true distribution. As a consequence, a consistent estimator of \( \beta \) solves the estimating equation

\[
0 = E[U(\beta)] = E\{ Y - E[Y \mid x^T \beta] \} \{ \alpha(x) - E[\alpha(x) \mid x^T \beta] \},
\]
The estimator in (4) is doubly robust under any choice of models for \(E[Y \mid x^T \beta ]\) and \(E(g(x) \mid x^T \beta ]\), meaning that the estimator remains consistent if either of these two models is correctly specified \([7]\).

4 Causal SDR

In causal inference with high dimensional treatments, what we are interested in is reducing the dimension of \(A\) such that the causal effect of \(A\) on \(Y\) is preserved. Let \(g_A(a; \beta)\), be a function parameterized by \(\beta\) that takes values in \(\mathbb{R}^p\) and map them to values in \(\mathbb{R}^d, d < p\). We want to reduce the dimension of \(A\) in such a way that the counterfactual response \(E[g_A(a; \beta)]\) only depends on \(A\) via \(g_A(a)\). We now describe these types of “reduced” counterfactuals formally.

4.1 Compositional Counterfactuals

For a random variable \(V\), let \(X_V\) be the state space of \(V\). The state space of a set \(V\) is defined as the following Cartesian product \(X_V = \bigotimes_{V \in V}\).

We define counterfactuals \(V(x)\) for any \(x \in X_{\text{pa}(V)}\), where \(\text{pa}(V)\) are direct causes (or parents in a causal diagram) of \(V\) via a structural equation of the form \(f_V: X_{\text{pa}(V) \cup Y_V} \rightarrow X_V\), where \(Y_V\) is a random variable representing exogenous sources of variation in \(V\) not captured explicitly by \(\text{pa}(V)\). The counterfactual random variable \(V(x)\) can be viewed as the output of \(f_V(z, \epsilon_Y)\) as \(\epsilon_Y\) varies. Other counterfactuals are derived via recursive substitution \([10]\):

\[
Y(a) \equiv Y(a_{\text{pa}(Y)}, \{W(a) \mid W \in \text{pa}(Y) \setminus A\}).
\]

Fix \(A\) with a high dimensional state space \(X_A\), another variable \(A^*\) with a lower dimensional state space \(X_{A^*}\), and a function \(g_A: X_A \rightarrow X_{A^*}\). We restrict attention to causal models where structural equations for the outcome \(Y\), namely \(f_Y(\text{pa}(Y), \epsilon_Y)\) can be written as a composition of \(g_A\) and another function \(f_Y^*\), specifically, for any \(x \in \text{pa}(Y)\), \(f_Y(z, \epsilon_Y) = f_Y^*(\{z_{\text{pa}(Y) \setminus \{A\}}, g_A(z_A), \epsilon_Y\}\). In such models, we define \(Y(z_{\text{pa}(Y) \setminus \{A\}}, g_A(z)) \equiv f_Y^*(z_{\text{pa}(Y) \setminus \{A\}}, g_A(z_A), \epsilon_Y)\), and define other counterfactuals by the appropriate generalization of (6). Responses to interventions that involve \(g_A\) are called compositional counterfactuals. Compositional counterfactuals may be defined for an arbitrary causal model, and an arbitrary set of treatments that we wish to reduce the dimension of, but we do not pursue this here in the interests of space.

Causal SDR for \(A\) is possible in models where appropriate compositional counterfactuals exist. For the purposes of this paper we make this assumption, and view it as a causal analogue of the assumption that a lower dimensional manifold exists that preserves the relationship between the feature space and the outcome. The objective in causal SDR for the mean outcome relationship is finding the parameter \(\beta\), for a given class of functions \(g_A(\cdot; \beta)\) such that

\[
E[Y(a)] = E[Y(g_A(a; \beta))].
\]

4.1.1 Identification of Compositional Counterfactuals

Before describing the estimation procedure for \(\beta\), we must show the “reduced” counterfactual is a function of the observed data if the original counterfactual is a function of the observed data. Without this identification result, the estimation problem for \(\beta\) is ill posed.

The identification formula for standard counterfactuals in causal models without hidden variables is given by the g-formula \([11]\):

\[
p(V(a) \mid V \setminus A) = \prod_{V \in V \setminus A} p(V \mid \text{pa}(V) \setminus A, \text{pa}(V) \setminus A).
\]

For every \(V \in V\), let \(\text{pa}^*(V) = \text{pa}(V) \setminus \{A\}\).

**Proposition 1.** In causal models without hidden variables, the distribution \(p(V(g_A(a)) \mid V \in V \setminus \{A\})\) is identified by the following generalization of the g-formula:

\[
\prod_{V \notin A, A \notin \text{pa}(V)} p(V \mid \text{pa}(V)) \prod_{V \notin A, A \notin \text{pa}(V)} p(V \mid \text{pa}^*(V), g_A(a)).
\]

**Proof.** Follows by identification of \(p(V(a) \mid V \in V \setminus A)\) via the g-formula, and the fact that for every \(V\), \(f_V(z_{\text{pa}(V) \setminus \{A\}}, \epsilon_V) = f_V^*(z_{\text{pa}(V) \setminus \{A\}}, g_A(z_A), \epsilon_V)\).

Since \(g_A(\cdot)\) is a function of \(A\), \(p(Y, A, C)\) induces \(p(Y, A, g_A(A), C)\), and this distribution is observed.

**Corollary 1.** Given a known \(g_A\), in causal models where appropriate compositional counterfactuals exist, \(p(Y(g_A(a)))\) is identified from \(p(Y, A, C)\) if it is identified from \(p(Y, A, g_A(A), C)\). Moreover, the identifying functional for \(p(Y(g_A(a)))\) is the identifying functional for \(p(Y(a))\) with \(A\) replaced by \(g_A(a)\).

The methodology proposed in this paper does not depend on the choice of \(g_A\). While from this point on we will assume \(g_A = A^T \beta\), for the sake of simplicity, an interesting direction for future work is to consider dimension reduction strategies developed in the machine learning neural network literature.

4.2 A Semi-Parametric View of Causal SDR

The estimation procedure for MSMs shown in (3) can be viewed as a standard set of estimating equations for a regression model relating treatments and outcome, but applied to observed data readjusted via inverse weighting in such a way that treatment variables appear randomly assigned. That is, MSM are regressions
applied to a version of observed data in such a way that regression parameters can be interpreted causally.

A key observation is that the estimating equation in (4) can be viewed as solving for $\beta$ in a type of regression problem where the function of $\beta$ isn’t to maximize the feature outcome relationship, but maintain the identity $E[Y | X] = E[Y | X^T \beta]$. As a consequence, semi-parametric causal SDR can be viewed as an MSM version of this regression problem, which seeks to find $\beta$ which maintains the following identity

$$\sum C E[Y | a, C]p(C) = \sum C E[Y | a^T \beta, C]p(C).$$

This is equivalent to maintaining (6) by consistency, conditional ignorability, and Proposition 1 under the choice $g_A(a) = a^T \beta$. We now describe two approaches to estimating $\beta$ that maintains the required property, based on combining estimation theory of MSMs [12] and the semi-parametric SDR method in [7].

Inverse Probability Weighted SDR

A simple estimation strategy based on generalizing [2], entails solving

$$0 = E \left[ \frac{p^*(a)U(\beta)}{p(a | C)} \right]$$

$$= \sum \frac{p^*(a)}{p(a | C)} \{Y_i - E[Y | a^T \beta]\} \{\alpha(a_i) - E[\alpha(a_i) | a^T \beta]\},$$

for any function $\alpha(a)$ of $A$, a known density $p^*(a)$, and a correctly specified statistical model $p(A | C; \eta_0)$ which governs how the treatment $A$ is assigned based on baseline characteristics $C$. We show consistency of this estimator in the next section.

A General Semi-Parametric Approach to Causal SDR

A general approach for deriving RAL estimators of $\beta$ is based on deriving $\Lambda_{\beta}^0$. One approach is to derive this space explicitly, as was done in [7]. 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 in [12]. The following results take advantage of this theory.

Our running assumption for these results is a semi-parametric model $p(Y, A, C; (\beta^*, \eta))$, where the appropriate compositional counterfactual exist, $E[Y(a)] = \sum_C E[Y | a, C]p(C)$, and $E[Y(a)] = E[Y(a^T \beta*)]$.  

**Proposition 2.** The orthogonal complement of the nuisance tangent space $\Lambda_{\beta}^0$, contains elements of the form $U(\beta)/W_A(C) + \phi(A, C) - E[\phi(A, C) | C]$, where $\phi(A, C)$ is an arbitrary function of $A, C$, $W_A(C)$ is the usual IPW weight $p^*(a)/p(a | C)$ for a fixed $p^*(a)$, and $U(\beta)$ is $\{Y - E[Y | a^T \beta]\} \{\alpha(a) - E[\alpha(a) | a^T \beta]\}$. Moreover, the most efficient estimator in this class, for any fixed $\alpha(a)$, is recovered by setting $\phi(A, C) = E[g^*(A)U(\beta)/p(A | C) | A, C]$.  

A generalization of this result also exists for arbitrary MSMs with multiple treatments and time-varying confounders, as described in section 2.3 [12]. However, we do not pursue this further in the interests of space.

**Corollary 2.** For a fixed choice of $\alpha(A)$ and $p^*(A)$, the element $U(\beta^*) \in \Lambda_{\beta}^0$, corresponding to the optimal choice of $\phi(A, C)$ has the form

$$g^*(A)U(\beta)/p(A | C) = E_q[E[U(\beta)/A, C] | C],$$

where $E_q[.]$ is the expectation taken with respect to the density $q(Y, A, C) \equiv p(Y | A, C)g^*(A)p(C)$.

**Corollary 3.** An estimator for $\beta$ which solves [7] under the correct specification of $p(A | C)$, and one of $E_q[Y | A]$ and $E_q[\alpha(A) | A]$ is consistent.

Robustness Properties

Just as $\Lambda_{\beta}^0$ ensured double robustness of $U(\beta)$ for semi-parametric SDR, we now show that the structure of $\Lambda^0_\beta$ yields additional robustness properties.

**Proposition 3.** The estimator for $\beta$ based on [8] is consistent if one of $p(A | C)$, $E[U(\beta) | A, C]$, and one of $E_q[Y | A^T \beta]$, $E_q[\alpha(A) | A^T \beta]$ is correctly specified.

This result implies [8] yields a kind of “2 × 2” robustness property. In practice, since we will be dealing with high dimensional problems, correct specification of models is difficult to ensure. However, robustness properties of semi-parametric estimators also implies that in regions where sufficient subset of models are approximately correct, the overall bias remains small.

## 5 Estimation and Implementation

In order to estimate $\beta$, we need to solve the estimating equation $E[U(\beta^*)] = 0$, where $U(\beta^*)$ is given in [8]. The term $E[U(\beta) | A, C]$ in $U(\beta^*)$ is equal to $E[Y | A, C] - E_q[Y | A^T \beta] \{\alpha(A) - E_q[\alpha(A) | A^T \beta]\}$. In this expression, five different models are involved, namely $E_q[Y | A^T \beta]$, $E_q[\alpha(A) | A^T \beta]$, $p(A | C)$, and $E[Y | A, C] = E[Y | A, C]$, to be able to estimate $U(\beta^*)$. Since it is difficult to specify complex models for $E_q[Y | A^T \beta]$ and $E_q[Y | A, C]$ in a congenial way, due to variation dependence of these models, we instead opt to specify $E_q[Y | A^T \beta]$ and $f(A, C, \beta) = E_q[Y | A, C] - E_q[Y | A^T \beta]$, which yield a variationally independent specification of $E_q[Y | A^T \beta]$.
and \( E_q[Y \mid A, C] = E_q[Y \mid A^T \beta] + \tilde{f}(A, C, \beta) \). Consequently, the five variationally independent models we need to specify are: \( E_q[Y \mid A^T \beta] \), \( E_q[\alpha(A) \mid A^T \beta] \), \( p(A \mid C) \), and \( \tilde{f}(A, C, \beta) \). Note that the last term in (8) is equal to \( \int E[U(\beta) \mid A, C]g^*(A)dA \), for a known distribution \( g^*(A) \), which can be evaluated empirically without additional modeling. In addition, we need to specify one more nuisance model to estimate \( \tilde{f} \), which we describe below.

### 5.1 Implementation

In this section, we describe in detail our procedure for estimating \( \beta^* \) by solving the empirical version of the estimating equation \( E(U(\beta^*)) = 0 \), where \( U(\beta^*) \) is given in (8). In these procedures, we chose flexible kernel models based on the Epanechnikov kernel for all models except \( \tilde{f} \), which was chosen to be parametric.

We denote by \( K(.) \) the Epanechnikov kernel and let \( K_h(.) = \frac{1}{h} K(. / h) \) for the choice of bandwidth \( h \). Let \( A \in \mathbb{R}^p \), \( \beta \in \mathbb{R}^{p \times d} \), \( C \) be the baseline vector, and \( Y \) be the outcome of interest. We first estimate \( p(A \mid C) \) by maximum likelihood. This model does not depend on \( \beta \) and is not updated within the iterations below. For a given choice of \( g^*(A), \alpha(A), d(A, C) \):

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

2. At the \( j \)th iteration, given a fixed \( \beta^{(j)} \), nonparametrically estimate the following models:

\[
\hat{E}(Y \mid A^T \beta^{(j)}) = \frac{\sum_{i=1}^{n} Y_i K_h(A^T \beta^{(j)} - A^T \beta^{(j)})}{\sum_{i=1}^{n} K_h(A^T \beta^{(j)} - A^T \beta^{(j)})} \\
\hat{E}(\alpha \mid A^T \beta^{(j)}) = \frac{\sum_{i=1}^{n} \alpha_i K_h(A^T \beta^{(j)} - A^T \beta^{(j)})}{\sum_{i=1}^{n} K_h(A^T \beta^{(j)} - A^T \beta^{(j)})} \\
\hat{E}(d \mid A^T \beta^{(j)}) = \frac{\sum_{i=1}^{n} d_i K_h(A^T \beta^{(j)} - A^T \beta^{(j)})}{\sum_{i=1}^{n} K_h(A^T \beta^{(j)} - A^T \beta^{(j)})} \tag{9}
\]

Let \( U^q(\beta^{(j)}) = \{ Y - \hat{E}(Y \mid A^T \beta^{(j)}) \} \{ \alpha(A) - \hat{E}(\alpha \mid A^T \beta^{(j)}) \} \).

3. Estimate \( \hat{\beta}^{(j)}(A, C, \beta^{(j)}; \hat{\psi}) \). We describe a strategy for doing this below.

4. Let \( E[U^q(\beta^{(j)}) \mid A, C] = \hat{f}^{(j)}(A, C) \{ \alpha - \hat{E}[\alpha \mid A^T \beta^{(j)}] \} \).

5. Form the sample version of \( E[U(\beta^*)] \):

\[
\frac{1}{n} \sum_{i=1}^{n} g^*(A) \{ U_{i}(\beta^{(j)}) + E \{ U_{i}(\beta^{(j)}) \mid A_i, C_i \} \} - E_q \{ E[U_{i}(\beta^{(j)}) \mid A_i, C_i] \mid C_i \}
\]

6. Calculate the first and second derivatives of above expression numerically, and use the Newton-Raphson update rule to update \( \beta^{(j)} \).

7. Repeat until convergence.

Implementing estimating equations (7) follows a similar set of steps, except all steps pertaining to second and third terms of (8) are skipped. The overall procedure is quite computationally intensive, and somewhat sensitive to starting values.

### 5.2 Estimation Of An “Inverted” Structural Nested Mean Model

We fit \( \tilde{f} \) using ideas from the theory of structural nested mean models (SNMMs) in [14]. Unlike MSRs, 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 \( \gamma(A, C; \psi) = E[Y(A) \mid C, A] - E[Y(0) \mid C, A] \). Let \( U_{sn} (\psi) = Y - \gamma(A, C; \psi) \), so \( E[U_{sn} (\psi) \mid C, A] = E[Y(0) \mid C, A] = E[Y(0) \mid C] = E[U_{sn} (\psi) \mid C] \) (by conditional ignorability). Then the following are consistent estimating equations for \( \psi \):

\[
E[(d(A, C) - E[d(A, C) \mid C]) \{ U_{sn} (\psi) - E[U_{sn} (\psi) \mid C] \}] = 0, \tag{10}
\]

where \( d(A, C) \) is a function of the same size as \( \psi \) [14].

We now show that estimating \( \psi \) in \( \tilde{f} \) can be viewed as an estimation problem for a kind of “inverted SNMM.”

**Proposition 4.** Let \( U_{dim} (\psi) = Y - \tilde{f} \), and fix any \( d(A, C) \). Then if \( E[d(A, C) \mid A^T \beta] \) and \( E[U_{dim} (\psi) \mid A^T \beta] \) are correctly specified, the following estimating equations yield a consistent estimator of \( \psi \):

\[
E \left[ \left\{ d(A, C) - E[d(A, C) \mid A^T \beta] \right\} \times \left\{ U_{dim} (\psi) - E[U_{dim} (\psi) \mid A^T \beta] \right\} \right] = 0. \tag{10}
\]

We solve these estimating equations in step 4 of each iteration above, also by numerical Newton Raphson. Inverse weighting in the last term \( E[g^*(A)U(\beta)/p(A \mid C)] \) accounts for the difference between \( E[.] \) and \( E_q[.] \). For the purposes of robustness, specifying both \( E[f \mid A^T \beta] \) and \( E[U_{dim} (\psi) \mid A^T \beta] \) correctly is part of the correct specification of \( E[U_{dim} (\psi) \mid A, C] \), given the type of estimation strategy we use.

### 6 Simulation Study

Causal SDR cannot be performed via standard methods such as PCA, because they do not take the feature/outcome relationship into account, nor by standard SDR methods, because they do not take confounding into account. In this section we illustrate,
via simulation studies, what goes wrong, and demonstrate our methods for causal SDR. To provide continuity with previous work, our study is similar to that described in [7].

We performed 20 replications with sample size $n = 200$, where true $E[Y(g_A(a))]$ is an object of dimension $d = 2$, and the observed data distribution $p(Y, A, C)$ was set as follows. We fix the dimension of the baseline factors $C$ to be 4 and the observed treatment dimension $p$ to be 6 and 12. We generate $C_1$ from chi-squared distribution with two degrees of freedom, $C_2$ from standard normal, $C_3 = |C_1 + C_2|$, and $C_4 = (C_1C_3)^{1/2}$. 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 generate $(A_1, A_2)^T$ (when $p = 6$) and $(A_1, A_2, A_7, \ldots, A_{12})^T$ (when $p = 12$) from a multivariate normal distribution where the mean of each component is given as: $\mu_1 = \sum_{i=1}^4 C_i$, $\mu_2 = \sum_{i=1}^4 (-1)^i C_i$, $\mu_7 = C_1$, $\mu_8 = C_2$, $\mu_9 = C_3$, $\mu_{10} = -C_1 + C_2$, $\mu_{11} = -C_1 + C_3$, $\mu_{12} = -C_3 + C_4$, and the covariance matrix is $(\sigma_{ij})_{(p-4)\times(p-4)}$ where $\sigma_{ij} = 0.5^{i-j}$. We generate $A_3$ 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 $A_6$ are generated from Bernoulli distributions with success probabilities $\exp(A_2)/(1+\exp(A_2))$, and $\Phi(A_2)$ where $\Phi(.)$ denotes the standard normal cumulative distribution, respectively.

Case 2: The treatment vector is generated from a multivariate normal distribution where the mean of each component is given as: $\mu_1 = \sum_{i=1}^4 C_i$, $\mu_2 = \sum_{i=1}^4 (-1)^i C_i$, $\mu_7 = C_1 - C_2 - C_3 + C_4$, $\mu_8 = -C_1 + C_2 + C_3 - C_4$, $\mu_9 = \sum_{i=1}^2 C_i$, and $\mu_{10} = -C_1 + 2C_2$, $\mu_{11} = -C_1 + 2C_3$, $\mu_{12} = -C_1 + \sum_{i=2}^4 C_i$, and the covariance matrix is $(\sigma_{ij})_{p\times p}$ where $\sigma_{ij} = 0.5^{i-j}$.

The response variable is generated using $Y = (A^T \beta_1)^2 + (A^T \beta_2)^2 + \sum_{i=1}^4 C_i + 0.5\epsilon$, where $\beta_1 = (1, 1, 1, 1, 1, 1)^T/\sqrt{6}$ and $\beta_2 = (1, -1, 1, -1, 1, -1)^T/\sqrt{6}$ when $p = 6$. When $p = 12$, the last 6 components of $\beta_1$ and $\beta_2$ are identically zero. The error term $\epsilon$ is generated from standard normal.

The accuracy of the estimates was computed using the distance between the true $\beta$ and $\hat{\beta}$ defined as the Frobenius norm of the matrix $\hat{\beta}(\beta^T \beta)^{-1}\beta^T - \beta(\beta^T \beta)^{-1}\beta^T$. The boxplots of the estimation accuracies are reported in Figure [1]. The results under case 1 are presented in [1](a) and (b) and the results under case 2 are presented in [1](c) and (d).

In each panel, there are 6 different boxplots. The first one, from the left hand side, labeled as regression, corresponds to semi-parametric SDR estimating equations [4], where there exist no baseline confounding factors $C$. The data for this plot is generated from $p(Y \mid A, p(A)$ (the data generation process is the same as what was described at the beginning of section only with ignoring $C$ by setting all $\mu_i$ to 0). The estimation in this case is fairly close to true $\beta$ for the $p = 6$, and less close for the $p = 12$ case. We believe this is due to insufficient sample size relative to the larger dimension of the problem, and the fact that using numerical Newton-Raphson for solving estimating equations in a non-convex problem makes performance quite sensitive to starting values.

The second panel, labeled as noisy.reg, corresponds to a “noisy outcome” version of the first with no confounding. Specifically, the data was generated from $p(Y \mid A, C)p(A)p(C)$, and no modification is applied to the semi-parametric SDR estimating equation. Here in principle we expect [4] to be consistent. However, at the sample sizes we considered in our simulation study, the additional source of noise for the outcome due to $C$ prevents accurate estimation of $\beta$, even if confounding is absent.

In the last four panels, the data is generated from $p(Y \mid A, C)p(A \mid C)p(C)$. In addition to the presence of a “noisy outcome,” the effect of treatment $A$ on outcome $Y$ in this scenario is also confounded by $C$. In the third panel, labelled as confounded.reg, estimating equations [4] are used again. Since this ignores the influence of confounding variables, the estimates are not capturing the true causal relationship of $A$ and $Y$, and are even worse than in the second scenario where the problem was merely noise. In the fourth panel, labeled as ipw.causal, using the IPW estimator in [7] under the correct model for $p(A \mid C)$ properly adjusts for confounding, and recovers a reasonable $\beta^*$ estimate.

The fifth plot, labelled aug.causal, uses the augmented IPW (AIPW) estimator corresponding to [6]. In typical applications of AIPW, as described in Section 2.1 the estimator is used to obtain inferences robust to misspecification of either $E[Y \mid A, C]$ or $p(A \mid C)$. In our case, the modeling problem is much more difficult, due to the need to specify more models, and the high dimensional nature of the problem. In particular, specifying $E[U(\beta) \mid A, C]$ is very difficult, even if we use $\tilde{f}(A, C, \beta)$. Thus, in our simulations, rather than reporting results under the truth for all models, including $E[U(\beta) \mid A, C]$, or fully demonstrate the “2x2 robustness” property, which entails specifying $E[U(\beta) \mid A, C]$ correctly, we instead opted to partially demonstrate robustness by showing that if $p(A \mid C)$ and models in $U(\beta)$ are specified correctly, then the estimator retains the ability to obtain results reason-
ably close to truth even if $E[U(\beta) \mid A, C]$ is not specified correctly, due to a poor choice of $\tilde{f}$.

If $p(A \mid C)$ and one of the models in $U(\beta)$ is correctly specified, the AIPW estimator using $[8]$ remains consistent for any choice of $E[U(\beta) \mid A, C]$. One promising direction of future work is to consider cases where $p(A \mid C)$ and $U(\beta)$ is known, and search for $E[U(\beta) \mid A, C]$ which yields good properties of the overall estimator. This use of the AIPW estimator is similar to that in recent work on randomized trial data, where $p(A \mid C) = p(A)$ is known by design.

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 true dimension reduction directions. As is demonstrated in the plots, this naive approach outputs results far from the true values, since it does not seek to preserve a causal, nor indeed any, relationship to the outcome.

So far we assumed the size of reduced dimension was known. In fact, choosing correct size is not easy, and incorrect choice may greatly affect performance. Advice from standard SDR applies to causal SDR [7]: use a resampling procedure to select the dimension.

7 Discussion and Conclusion

In this paper, we have described a generalization of semi-parametric sufficient dimension reduction (SDR) [7] to causal SDR, where we seek to reduce the dimension of a high dimensional treatment, while preserving the causal treatment-outcome relationship quantified as a counterfactual mean. We have given a semi-parametric estimator which uses ideas from marginal structural model and structural nested mean model literature [12] to reduce the dimension of the treatment while making few parametric assumptions. 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. The generality and robustness of our method comes at a cost – the estimation strategy necessary for our method is a complex nested optimization problem, somewhat sensitive to the choice of a starting point.

An additional methodological obstacle in applied problems where causal dimension reduction is relevant is small sample sizes relative to the dimension of the problem. A promising direction of future work is combining the advantage of our method with those of semi-parametric sparsity methods.
References

[1] B. Li B and S. Wang. On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102:997–1008, 2007.

[2] Heejung Bang and James M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

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

[4] 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.

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

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

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

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

[9] Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs, 2017. Working paper.

[10] Thomas S. Richardson and Jamie M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *preprint: [http://www.csss.washington.edu/Papers/wp128.pdf](http://www.csss.washington.edu/Papers/wp128.pdf)*, 2013.

[11] James M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical Modeling*, 7:1393–1512, 1986.

[12] James 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.

[13] Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer-Verlag New York, 1st edition edition, 2006.

[14] Stijn Vansteelandt and Marshall Joffe. Structural nested models and g-estimation: The partially realized promise. *Statistical Science*, 29(4):707–731, 2014.

[15] LX. Zhu and KT. Fang. Asymptotics for kernel estimation of sliced inverse regression. *The Annals of Statistics*, 3:1053–1068, 1996.
Appendix

Statistical Inference in Semi-parametric Models

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 $\beta \in \mathbb{R}^d$ denotes the set of target parameters, and $\eta$ denotes a possibly infinite dimensional set of nuisance parameters. This type of model is termed semi-parametric, since it has both a parametric and a non-parametric component. The goal of statistical inference in semi-parametric models is to find “the best” estimator of $\beta$. 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 $i$th observation for the parameter vector $\beta$. RAL estimators are consistent and asymptotically normal, with the variance of the estimator given by its IF:

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

Thus, there is a bijective correspondence between RAL estimators and IFs. In fact, 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, and define the inner product of two arbitrary elements of the Hilbert space, $h_1$ and $h_2$, to be equal to $E[h_1 h_2]$. Define a parametric submodel to be a subset of densities in the semi-parametric model parameterized by $\theta^T = (\beta^T, \gamma^T)$, where $\gamma^T \in \mathbb{R}^r$, such that the subset contains the density $p(Z; \theta_0)$ in the semi-parametric model evaluated at the true parameter values $\theta_0$. The nuisance tangent space $\Lambda$ in the semi-parametric model is defined to be the mean square closure of elements of the nuisance tangent spaces $\Lambda_\gamma = \{B^{r \times r} S_\eta(Z; \theta)\}$ of every parametric submodel. The space $\Lambda$ is important because it is known all influence functions lie in the orthogonal complement $\Lambda^\perp$ of $\Lambda$ with respect to $\mathcal{H}$. For this reason, recovering $\Lambda^\perp$ is often the first step for constructing RAL estimators in semi-parametric 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 semi-parametric efficiency bound, see \cite{13} for details.

Proofs

**Proposition 1.** In causal models without hidden variables, the distribution $p(\{V(g_A(a)) \mid V \in V \setminus \{A\})$ is identified by the following generalization of the g-formula:

$$\prod_{V : A \notin \text{pa}(V)} p(V \mid \text{pa}(V)) \prod_{V : A \in \text{pa}(V)} p(V \mid \text{pa}^T(V), g_A(a)).$$

*Proof.* Follows by identification of $p(\{V(a) \mid V \in V \setminus A\})$ via the g-formula, and the fact that for every $V$,

$$f_V(z_{\text{pa}(V)}, \epsilon_V) = f_V(z_{\text{pa}(V)) \setminus \{A\}, g_A(z_A), \epsilon_V).$$

**Corollary 1.** Given a known $g_A$, in causal models where appropriate compositional counterfactuals exist, $p(Y(g_A(a)))$ is identified from $p(Y, A, C)$ if is identified from $p(Y, A, g_A(A), C)$. Moreover, the identifying functional for $p(Y(g_A(a)))$ is the identifying functional for $p(Y(a))$ with a replaced by $g_A(a)$.

*Proof.* This follows by Proposition 1 and Theorem 60 in \cite{9}.

**Proposition 2.** The orthogonal complement of the nuisance tangent space $\Lambda^\perp_{\beta}$ contains elements of the form $U(\beta)/W_a(C) + \phi(A, C) - E[\phi(A, C) \mid C]$, where $\phi(A, C)$ is an arbitrary function of $A, C$, $W_a(C)$ is the usual IPW weight $p^*(a)/p(a \mid C)$ for a fixed $p^*(a)$, and $U(\beta)$ is $\{Y - E[Y \mid a^T \beta]\} \{\alpha(a) - E[\alpha(a) \mid a^T \beta]\}$.
Moreover, the most efficient estimator in this class, for any fixed \( \alpha(a) \), is recovered by setting 
\[
\phi(A, C) = E[g^*(A)U(\beta)/p(A \mid C) \mid A, C].
\]

**Proof.** This is a direct consequence of Theorems 3.1 and 3.2 in [12], and results in Appendix 3 of [7]. \( \square \)

**Corollary 2** For a fixed choice of \( \alpha(A) \) and \( p^*(A) \), the element \( U(\beta^*) \in \Lambda^\perp_{\beta} \) corresponding to the optimal choice of \( \phi(A, C) \) has the form,
\[
g^*(A)U(\beta) \mid p(A \mid C) + E \left[ g^*(A)U(\beta) \mid A, C \right] - E_q[E[U(\beta)A, C | C],
\]
where \( E_q[\cdot] \) is the expectation taken with respect to the density \( q(Y, A, C) \equiv p(Y \mid A, C)g^*(A)p(C) \).

**Proof.** Plugging in the optimal \( \phi(A, C) \) yields \( U(\beta^*) \) equal to
\[
g^*(A)U(\beta) \mid p(A \mid C) + \left\{ E \left[ g^*(A)U(\beta) \mid A, C \right] - E \left[ E \left[ g^*(A)U(\beta) \mid A, C \right] \mid C \right] \right\}.
\]
The conclusion follows, since
\[
E \left[ E \left[ g^*(A)U(\beta) \mid A, C \right] \mid C \right] = \int g^*(A)E[U(\beta) \mid A, C]p(Y, A \mid C) / p(A \mid C)
\]
\[
= \int E[U(\beta) \mid A, C]p(Y \mid A, C)g^*(A)
\]
\[
= \int E[U(\beta) \mid A, C]q(Y, A \mid C)
\]
\[
= E_q[E[U(\beta) \mid A, C] \mid C].
\]
\( \square \)

**Corollary 3** An estimator for \( \beta \) which solves \([7]\) under the correct specification of \( p(A \mid C) \), and one of \( E_q[Y \mid A] \) and \( E_q[\alpha(A) \mid A] \) is consistent.

**Proof.** Choosing \( \phi(A, C) = 0 \) 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 \( \Lambda^\perp_{\beta} \) in the Appendix). Since \( U(\beta) \) exhibits double robustness, the correct specification of \( p(A \mid C) \) yields our conclusion. \( \square \)

**Lemma 1.** For all \( U(\beta^*) \in \Lambda^\perp_{\beta} \), \( E[U(\beta^*)] = 0 \).

**Proof.** The second and third terms of \( U(\beta^*) \) are mean zero by construction. The first term, under truth and \( \beta \) with the property that \( E_q[Y \mid A] = E_q[Y \mid A^T \beta] \), is
\[
E \left[ g^*(A)U(\beta) \mid p(A \mid C) \right] = \int U(\beta)p(Y \mid A, C)g^*(A)p(C)
\]
\[
= \int \{Y - E_q[Y \mid A^T \beta]\} \hat{h}(A)q(Y, A, C)
\]
\[
= E_q\{Y - E_q[Y \mid A^T \beta]\} \hat{h}(A)
\]
\[
= E_q\{E_q[Y \mid A] - E_q[Y \mid A^T \beta]\} \hat{h}(A)
\]
\[
= E_q\{\hat{h}(A)\}E_q\{E_q[Y \mid A] - E_q[Y \mid A^T \beta]\}
\]
\[
= E_q\{\hat{h}(A)\}E_q\{E_q[Y \mid A^T \beta] - E_q[Y \mid A^T \beta]\}
\]
\[
= 0,
\]
where \( \hat{h}(A) = \{\alpha(A) - E_q[\alpha(A) \mid A^T \beta]\} \).
\( \square \)
Proposition \[3\]. The estimator for \( \beta \) based on \[8\] is consistent if one of \( p(A \mid C) \), \( E[U(\beta) \mid A, C] \), and one of \( E_q[Y \mid AT^\beta] \), \( E_q(\alpha(A) \mid AT^\beta) \) is correctly specified.

Proof. Assume either \( E_q[Y \mid AT^\beta] \), or \( E_q(\alpha(A) \mid AT^\beta) \), and \( p(A \mid C) \) are correctly specified. Then the second and third terms of \( U(\beta^*) \) together are mean zero, even under an incorrect model for \( E[U(\beta) \mid A, C] \). For the first term, if \( E_q[Y \mid AT^\beta] \) is correct, it is mean zero by the argument in Proposition \[1\]. If \( E_q[Y \mid AT^\beta] \) is incorrect, but \( E(\alpha(A) \mid AT^\beta) \) is correct, we have:

\[
E \left[ \frac{q^*(A)U(\beta)}{p(A \mid C)} \right] = \int U(\beta)p(Y \mid A, C)q^*(A)p(C)dY, A, C
\]

\[
= \int \left\{ Y - E_q[Y \mid AT^\beta]\right\}\left\{ \alpha(A) - E_q[\alpha(A) \mid AT^\beta]\right\}q(Y, A, C)dY, A, C
\]

\[
= \int \left\{ E_q[Y \mid A] - E_q[Y \mid AT^\beta]\right\}\left\{ \alpha(A) - E_q[\alpha(A) \mid AT^\beta]\right\}q(Y, A, C)dY, A, C
\]

\[
= \int \left\{ E_q[E_q[Y \mid AT^\beta] - E_q[Y \mid AT^\beta]\right\}\left\{ \alpha(A) - E_q[\alpha(A) \mid AT^\beta]\right\}q(A, C)dY, A, C
\]

\[
= \int E_q[E_q[Y \mid AT^\beta] - E_q[Y \mid AT^\beta]\right\}\left\{ \alpha(A) - E_q[\alpha(A) \mid AT^\beta]\right\}q(A, C)dY, A, C
\]

\[
= \int E_q[Y \mid AT^\beta] - E_q[Y \mid AT^\beta]\right\}q(A, C)dY, A, C
\]

\[
= E_q[Y \mid AT^\beta] - E_q[Y \mid AT^\beta]\right\}q(A, C)dY, A, C
\]

Assume either \( E_q[Y \mid AT^\beta] \), or \( E_q[\alpha(A) \mid AT^\beta] \), and \( E[U(\beta) \mid A, C] \) are correctly specified. Then the two terms of \( U^* \) are mean zero, even under an incorrect model \( p^*(A \mid C) \). For the last term, we have:

\[
E[E_q[E[U(\beta)A; C] \mid C]] = E \left[ \int \int U^*p(Y \mid A, C)dY, A, C \right]
\]

\[
= \int \left\{ \int U(\beta)q(Y, A \mid C)dY, A \right\}p(C)dC
\]

\[
= \int U(\beta)q(Y, A, C)dY, A, C = E_q[U(\beta)].
\]

Since this is mean \( U(\beta) \) evaluated under truth, if either \( E_q[Y \mid AT^\beta] \) or \( E_q[\alpha(A) \mid AT^\beta] \) is correctly specified, we get mean zero by above arguments.

\[
\square
\]

Proposition \[4\] Fix any function \( d(A, C) \). Then under the correct specification of \( E[d(A, C) \mid AT^\beta] \) and \( E[U_{dim}(\psi) \mid AT^\beta] \), solving the following set of estimating equations yields a consistent estimator of \( \psi \)

\[
E \left\{ d(A, C) - E[d(A, C) \mid AT^\beta] \right\} \times \left\{ U_{dim}(\psi) - E[U_{dim}(\psi) \mid AT^\beta] \right\} = 0.
\]

Proof. Define \( U_{dim}(\psi) = Y - \tilde{f}(A, C, \beta; \psi) \). Then \( E[U_{dim}(\psi) \mid A, C] = E[Y \mid AT^\beta] = E[U_{dim}(\psi) \mid AT^\beta] \). 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 follows immediately.

\[
\square
\]