Semiparametric Causal Sufficient Dimension Reduction of Multidimensional Treatments (Supplementary Material)

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

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 $Z_1, \dots, Z_n$, 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 1. *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,

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

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 1, and modify step (c) by letting $\mathbb{E}[U^q(\beta^{(j)}) \mid A,C] = \widehat{\widetilde{f}^{(j)}} \times \{\alpha(A) - \widehat{\nu}(g(A;\beta^{(j)}))\}$. 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. 1(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 $p = 12$.
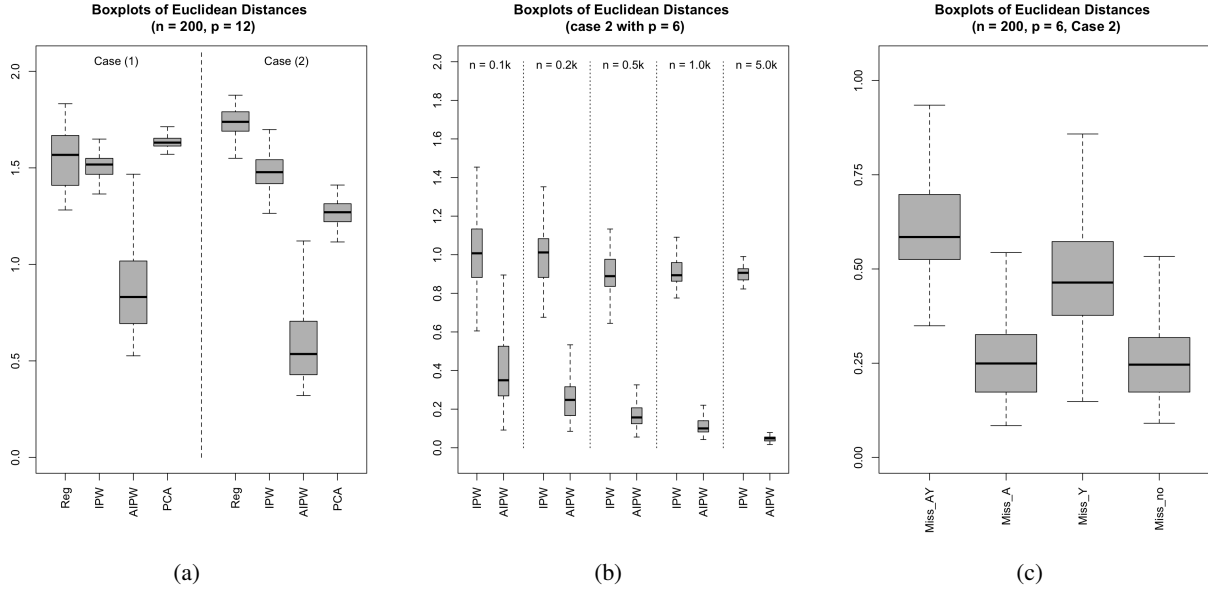
Figure 1: (a) Boxplots of Frobenius norms between true and estimated parameters in simulations ($p = 12$); (b) Illustration of the effect of sample size on the Frobenius norms between true and estimated parameters using data generated from Case 2 with $p = 6$. (c) Boxplots to illustrate the robustness property of our AIPW-style estimator.

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. 1(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 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. 1(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. 2 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.
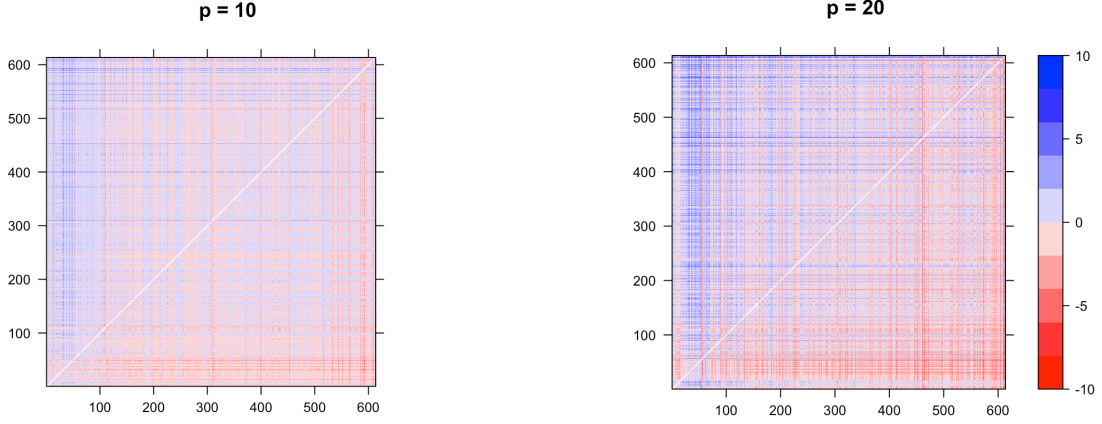
Figure 2: Heatmaps to illustrate the causal effect of radiation on weight loss. Treatment is collected using (right) 10 and (left) 20 equally spaced percentages of volume in parotid glands.

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

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

since $\ell(g(a; \beta)) \coloneqq \mathbb{E}_q[Y \mid A = a]$. 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

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

The conclusion follows, since

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

$$= \mathbb{E}\left[\frac{p^*(A)}{p(A \mid C)}\mathbb{E}\left[\widetilde{U}(\beta)\Big| A, C\right]\Big| C\right]$$

$$= \int \frac{p^*(A)}{p(A \mid C)}\, \mathbb{E}[\widetilde{U}(\beta) \mid A, C]\, p(Y, A \mid C)\, d\mu_{Y,A}$$

$$= \int \mathbb{E}[\widetilde{U}(\beta) \mid A, C]\, p(Y \mid A, C)\, p^*(A)\, d\mu_{Y,A}$$

$$= \int \mathbb{E}[\widetilde{U}(\beta) \mid A, C]\, q(Y, A \mid C)\, d\mu_{Y,A}$$

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

$\square$

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:

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

$$= \mathbb{E}\left[\mathbb{E}_q\left[\int \widetilde{U}(\beta) \times p(Y \mid A, C)\, d\mu_Y\,\Big|\, C\right]\right]$$

$$= \mathbb{E}\left[\int \left(\int \widetilde{U}(\beta) \times p(Y \mid A, C)\, d\mu_Y\right) \times p^*(A)\, d\mu_A\right]$$

$$= \int \left(\int \int \widetilde{U}(\beta) \times p(Y \mid A, C) \times p^*(A)\, d\mu_Y\, d\mu_A\right) \times p(C)\, d\mu_C$$

$$= \int \widetilde{U}(\beta) \times p(Y \mid A, C) \times p^*(A) \times p(C)\, d\mu_{Y,A,C}$$

$$= \int \widetilde{U}(\beta) \times q(Y, A, C)\, d\mu_{Y,A,C}$$

$$= \mathbb{E}_q[\widetilde{U}(\beta)].$$

We conclude the proof by noting that $\mathbb{E}[\widetilde{U}(\beta)]$ 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]. $\square$

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

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

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

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

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

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

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 1

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

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

Lemma 1.

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,

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

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,

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

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:

$$\mathcal{P}_{\psi, \zeta} = \Big\{p(z; \psi, \psi_\zeta) = p(c, a; \zeta_1) \times p(\epsilon \mid a, C; \zeta_2)\Big\}$$
$$= \Big\{p(C, a; \zeta_1) \times p\big(y - \ell(g(a; \beta)) - \widetilde{f}(a, C, \beta; \psi) \mid a, C; \zeta_2\big)\Big\},$$

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

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

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 2)	$\Lambda_{\ell(g(a;\beta))} = \{\frac{\psi_{2\epsilon}'(\epsilon, C, a)}{\psi_2(\epsilon, C, a)} f(g(a; \beta))\}$

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

$$S_{\ell(g(a;\beta))} = \frac{\partial \log p(z; \psi, \zeta)}{\partial \ell(g(a; \beta))}\Bigg|_{\psi_0, \zeta_0}$$
$$= \frac{\partial \log \Big(\psi_1(C, a; \zeta_{10}) \times \psi_2\big(y - \ell(g(a; \beta)) - \gamma(C, a; \psi), C, a; \zeta_{20}\big)\Big)}{\partial \ell(g(a; \beta))}$$

$$= \frac{\partial \log \psi_2 \big(y - \ell(g(a; \beta)) - \gamma(C, a; \psi), l, a; \zeta_{20}\big)}{\partial \ell(g(a; \beta))}$$

$$= \frac{\partial \log \psi_2 \big(\epsilon, C, a; \zeta_{20}\big)}{\partial \epsilon} \times \frac{\partial \epsilon}{\partial \ell(g(a; \beta))} \qquad (\epsilon \text{ is a function of } \ell(g(a; \beta)))$$

$$= \frac{\psi'_{2\epsilon}(\epsilon, C, a)}{\psi_2(\epsilon, C, a)} f(g(a; \beta)). \tag{2}$$

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,

$$\forall h(g(a; \beta)) \quad 0 = <d(C, a)\epsilon, \frac{\psi'_{2\epsilon}}{\psi_2} h(g(a; \beta)) >$$

$$= \mathbb{E}\big[d(C, a)\epsilon \frac{\psi'_{2\epsilon}}{\psi_2} h(g(a; \beta))\big]$$

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

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

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

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

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

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

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

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

$$= \Big\{\big(d(C, a) - \mathbb{E}[d(C, a) \mid g(a; \beta)]\big) \times \big(Y - \gamma(C, a; \psi) - \ell(g(a; \beta))\big)\Big\}$$

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

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,

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

$\square$

References

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

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.

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

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.