Submitted to the Annals of Applied Statistics
0000 Vol. 0, No. 0,
arXiv: arXiv:0000.0000

SUPPLEMENTARY MATERIALS FOR PROPENSITY SCORE WEIGHTING FOR CAUSAL INFERENCE WITH MULTIPLE TREATMENTS

By Fan Li and Fan Li

Yale University and Duke University

SUPPLEMENT A: ON TRANSITIVITY

With multiple treatments, a desirable property of a given class of estimands is *transitivity*. For pairwise comparisons, lack of transitivity often implies that comparisons of treatments are based on different populations. As a result, non-transitivity may lead to incompatible pairwise contrasts; for example, it is possible that treatment A is favored over treatment B, treatment B is favored over treatment C, but treatment C is found to better than treatment A at the same time. Below we provide a formal definition of transitivity and offer two related remarks.

Definition 1. *The class of causal estimands* $\mathbb{T}(h,\mathbb{A}) = \{\tau^h(\boldsymbol{a}) : \boldsymbol{a} \in \mathbb{A} \subset \mathbb{R}^J\}$ *is transitive if the following equivariance relationship holds:* $\tau^h(\boldsymbol{a})+\tau^h(\boldsymbol{a}') = \tau^h(\boldsymbol{a}'')$ *whenever* $\boldsymbol{a},\boldsymbol{a}',\boldsymbol{a}'' \in \mathbb{A}$ *and* $\boldsymbol{a}+\boldsymbol{a}' = \boldsymbol{a}''$.

Remark 1. *Fixing a tilting function* h*, the class of estimands specifying all pairwise contrasts, namely,* $\mathbb{T}(h,\mathbb{S})$ *is transitive. For example, with* $h(\boldsymbol{X}) = 1$*, the class of pairwise ATE estimands is transitive; with* $h(\boldsymbol{X}) = e_j(\boldsymbol{X})\prod_{k=1}^J E_k(\boldsymbol{X})$*, the class of ATT estimands in Lopez and Gutman (2017) is also transitive.*

Remark 2. *The union of* $\mathbb{T}(h_1,\mathbb{S})$ *and* $\mathbb{T}(h_2,\mathbb{S})$ *or that of their subsets is generally non-transitive for* $h_1 \neq h_2$*. This explains why several existing classes of estimands are non-transitive, including the class of ATT estimands of Lechner (2001),* $\{\mathbb{E}[Y_i(j)-Y_i(j')|Z_i = j \text{ or } Z_i = j'] : j < j'\}$*. The reason is that each individual estimand corresponds to a distinct tilting function* $h_{j,j'}(\boldsymbol{X}) = (e_j(\boldsymbol{X})+e_{j'}(\boldsymbol{X}))/e_1(\boldsymbol{X})$*, and therefore this class of estimands is the union of* $\binom{J}{2}$ *elements, each of which is contained in* $\mathbb{T}(h_{j,j'},\mathbb{S})$ *for some* $j < j'$.

SUPPLEMENT B: PROOF OF PROPOSITIONS

For proving the Propositions, we assume regularity conditions on $m_j(\boldsymbol{X}) = \mathbb{E}[Y(j)|\boldsymbol{X}]$ and $v_j(\boldsymbol{X}) = \mathbb{V}(Y(j)|\boldsymbol{X})$ necessary to ensure that the integrals

are well defined.

Proof of Proposition 1. By definition of the generalized propensity score, we must have $\mathbb{E}[\mathbb{1}(Z = j)/e_j(\boldsymbol{X})|\boldsymbol{X}] = 1$ for all $j \in \mathbb{Z}$. Then the average of the potential outcomes in target population h

$$\begin{aligned} m_j^h &= \frac{\int_{\mathbb{X}} m_j(\boldsymbol{X}) f(\boldsymbol{X}) h(\boldsymbol{X}) \mu(d\boldsymbol{X})}{\int_{\mathbb{X}} f(\boldsymbol{X}) h(\boldsymbol{X}) \mu(d\boldsymbol{X})} \\ &= \frac{\int_{\mathbb{X}} \mathbb{E}[\mathbb{1}\{Z = j\} Y(j) (h(\boldsymbol{X})/e_j(\boldsymbol{X}))|\boldsymbol{X}] f(\boldsymbol{X}) \mu(d\boldsymbol{X})}{\int_{\mathbb{X}} \mathbb{E}[\mathbb{1}\{Z = j\} (h(\boldsymbol{X})/e_j(\boldsymbol{X}))|\boldsymbol{X}] f(\boldsymbol{X}) \mu(d\boldsymbol{X})} \\ &= \frac{\int_{\mathbb{X}} \mathbb{E}[\mathbb{1}\{Z = j\} Y(j) w_j(\boldsymbol{X})|\boldsymbol{X}] f(\boldsymbol{X}) \mu(d\boldsymbol{X})}{\int_{\mathbb{X}} \mathbb{E}[\mathbb{1}\{Z = j\} w_j(\boldsymbol{X})|\boldsymbol{X}] f(\boldsymbol{X}) \mu(d\boldsymbol{X})} \end{aligned} \tag{A.1}$$

where the second equation holds due to the weak unconfoundedness assumption, $Y(j) \perp \mathbb{1}\{Z = j\}|\boldsymbol{X}$ (Imbens, 2000). Because $D_{ij} = \mathbb{1}\{Z_i = j\}$, it follows that the estimators, $n^{-1} \sum_{i=1}^n D_{ij} Y_i w_j(\boldsymbol{X}_i)$ and $n^{-1} \sum_{i=1}^n D_{ij} w_j(\boldsymbol{X}_i)$, consistently estimate the numerator and denominator of (A.1). Therefore, $\hat{m}_j^h = \sum_{i=1}^n D_{ij} Y_i w_j(\boldsymbol{X}_i) \Big/ \sum_{i=1}^n D_{ij} w_j(\boldsymbol{X}_i)$ is consistent for m_j^h, and $\hat{\tau}^h(\boldsymbol{a}) = \sum_{j=1}^J a_j \hat{m}_j^h$ must be consistent for $\tau^h(\boldsymbol{a}) = \sum_{j=1}^J a_j m_j^h$.

Proof of Proposition 2. By SUTVA (Imbens and Rubin, 2015), we write

$$\hat{\tau}_{\boldsymbol{a}}^h = \sum_{j=1}^J a_j \frac{\sum_{i=1}^n D_{ij} Y_i w_j(\boldsymbol{X}_i)}{\sum_{i=1}^n D_{ij} w_j(\boldsymbol{X}_i)} = \sum_{j=1}^J a_j \frac{\sum_{i=1}^n D_{ij} Y_i(j) w_j(\boldsymbol{X}_i)}{\sum_{i=1}^n D_{ij} w_j(\boldsymbol{X}_i)}.$$

Conditional on the assignment $\underline{\boldsymbol{Z}}$ and sample design $\underline{\boldsymbol{X}}$, only the potential outcomes are random. Therefore the residual variance of $\hat{\tau}^h(\boldsymbol{a})$ is

$$\begin{aligned} \mathbb{V}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}}, \underline{\boldsymbol{X}}] &= \sum_{j=1}^J a_j^2 \frac{\sum_{i=1}^n v_j(\boldsymbol{X}_i) D_{ij} w_j^2(\boldsymbol{X}_i)}{[\sum_{i=1}^n D_{ij} w_j(\boldsymbol{X}_i)]^2} \\ &= \sum_{j=1}^J a_j^2 \frac{\sum_{i=1}^n \{v_j(\boldsymbol{X}_i)/e_j(\boldsymbol{X}_i)\} \{D_{ij}/e_j(\boldsymbol{X}_i)\} h^2(\boldsymbol{X}_i)}{[\sum_{i=1}^n \{D_{ij}/e_j(\boldsymbol{X}_i)\} h(\boldsymbol{X}_i)]^2}. \end{aligned}$$

Averaging over the joint distribution of $\underline{\boldsymbol{Z}}$ and $\underline{\boldsymbol{X}}$, we observe by the Weak Law of Large Numbers that

$$\frac{1}{n} \sum_{i=1}^n \{D_{ij}/e_j(\boldsymbol{X}_i)\} h(\boldsymbol{X}_i) \xrightarrow{p} \int_{\mathbb{X}} \mathbb{E}[\mathbb{1}\{Z = j\}/e_j(\boldsymbol{X})|\boldsymbol{X}] h(\boldsymbol{X}) f(\boldsymbol{X}) \mu(d\boldsymbol{X}) = C_h,$$

and

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\{v_j(\boldsymbol{X}_i)/e_j(\boldsymbol{X}_i)\}\{D_{ij}/e_j(\boldsymbol{X}_i)\}h^2(\boldsymbol{X}_i) \\ \xrightarrow{p} & \int_{\mathbb{X}} v_j(\boldsymbol{X})/e_j(\boldsymbol{X})\,\mathbb{E}[\mathbb{1}\{Z=j\}/e_j(\boldsymbol{X})|\boldsymbol{X}]h^2(\boldsymbol{X})f(\boldsymbol{X})\mu(d\boldsymbol{X}) \\ = & \int_{\mathbb{X}}\{v_j(\boldsymbol{X})/e_j(\boldsymbol{X})\}h^2(\boldsymbol{X})f(\boldsymbol{X})\mu(d\boldsymbol{X}) \end{aligned}$$

An application of the Slutsky's Theorem shows $n\cdot\mathbb{V}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}] \xrightarrow{p} Q(\boldsymbol{a},h)$, where $Q(\boldsymbol{a},h)$ is a constant defined in Proposition 2. The uniform integrability assumption for the family of random variables $\{\mathbb{V}[\hat{\tau}^h(\boldsymbol{a})|\underline{\boldsymbol{Z}},\underline{\boldsymbol{X}}], n\geq 1\}$ then gives the desired L_1 convergence result.

Proof of Proposition 3. For notational simplicity, we use the $\mathbb{E}[\cdot]$ operator to represent $\int_{\mathbb{X}}\cdot f(\boldsymbol{X})\mu(d\boldsymbol{X})$. Under homoscedasticity, $v_j(\boldsymbol{X})=v$,

$$\begin{aligned} Q(\boldsymbol{a},h) &= (v/C_h^2)\int_{\mathbb{X}}\Big(\sum_{j=1}^{J}a_j^2/e_j(\boldsymbol{X})\Big)h^2(\boldsymbol{X})f(\boldsymbol{X})\mu(d\boldsymbol{X}) \\ &= (v/C_h^2)\,\mathbb{E}\Big\{h^2(\boldsymbol{X})\Big(\sum_{j=1}^{J}a_j^2/e_j(\boldsymbol{X})\Big)\Big\}. \end{aligned}$$

Applying the Cauchy-Schwarz inequality, we have

$$\begin{aligned} C_h^2=[\mathbb{E}\{h(\boldsymbol{X})\}]^2 &= \Big[\mathbb{E}\Big\{h(\boldsymbol{X})\Big(\sum_{j=1}^{J}a_j^2/e_j(\boldsymbol{X})\Big)^{1/2}\Big(\sum_{j=1}^{J}a_j^2/e_j(\boldsymbol{X})\Big)^{-1/2}\Big\}\Big]^2 \\ &\leq \mathbb{E}\Big\{h^2(\boldsymbol{X})\Big(\sum_{j=1}^{J}a_j^2/e_j(\boldsymbol{X})\Big)\Big\}\,\mathbb{E}\Big\{\Big(\sum_{j=1}^{J}a_j^2/e_j(\boldsymbol{X})\Big)^{-1}\Big\}, \end{aligned}$$

and the equality is attained when $h=\tilde{h}(\boldsymbol{X})\propto\Big(\sum_{j=1}^{J}a_j^2/e_j(\boldsymbol{X})\Big)^{-1}$. This implies that

$$\mathbb{E}\Big\{h^2(\boldsymbol{X})\Big(\sum_{j=1}^{J}a_j^2/e_j(\boldsymbol{X})\Big)\Big\}\Big/C_h^2 \geq \Big[\mathbb{E}\Big\{\Big(\sum_{j=1}^{J}a_j^2/e_j(\boldsymbol{X})\Big)^{-1}\Big\}\Big]^{-1}=C_{\tilde{h}}^{-1},$$

which gives $Q(\boldsymbol{a},\tilde{h})=v/C_{\tilde{h}}$.

SUPPLEMENT C: PROOF OF THEOREM 1

From the multinomial logistic model, we have for $i = 1, \ldots, n$,

$$\begin{aligned} e_1(\boldsymbol{X}_i) &= Pr(Z_i = 1|\boldsymbol{X}_i) = \frac{1}{1 + \sum_{k=2}^{J} \exp(\alpha_k + \boldsymbol{X}_i^T \boldsymbol{\beta}_k)}. \\ e_j(\boldsymbol{X}_i) &= Pr(Z_i = j|\boldsymbol{X}_i) = \frac{\exp(\alpha_j + \boldsymbol{X}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{k=2}^{J} \exp(\alpha_k + \boldsymbol{X}_i^T \boldsymbol{\beta}_k)}, \quad j = 2, \ldots, J. \end{aligned}$$

Since $D_{ij} = \mathbb{1}\{Z_i = j\}$, it is straightforward to show that the log likelihood function

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \Big[\sum_{j=2}^{J} \Big\{ D_{ij}(\alpha_j + \boldsymbol{X}_i^T \boldsymbol{\beta}_j) \Big\} - \log \Big\{ 1 + \sum_{k=2}^{J} \exp(\alpha_k + \boldsymbol{X}_i^T \boldsymbol{\beta}_k) \Big\} \Big]$$

When the estimation of model parameters is carried out by maximum likelihood, the first-order condition is obtained by differentiating the log likelihood with respect to $\boldsymbol{\theta}$,

$$\begin{aligned} \mathbf{0} &= \boldsymbol{S}_{\boldsymbol{\theta}} = \sum_{i=1}^{n} \boldsymbol{S}_{\boldsymbol{\theta},i} = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} l_i(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \left(\frac{\partial}{\partial \alpha_2} l_i(\boldsymbol{\theta}), \ldots, \frac{\partial}{\partial \alpha_J} l_i(\boldsymbol{\theta}), \frac{\partial}{\partial \boldsymbol{\beta}_2^T} l_i(\boldsymbol{\theta}), \ldots, \frac{\partial}{\partial \boldsymbol{\beta}_J^T} l_i(\boldsymbol{\theta}) \right)^T, \end{aligned} \tag{A.2}$$

where for $l = 2, \ldots, J$,

$$\frac{\partial}{\partial \boldsymbol{\beta}_l} l_i(\boldsymbol{\theta}) = \boldsymbol{X}_i \frac{\partial}{\partial \alpha_l} l_i(\boldsymbol{\theta}) = \boldsymbol{X}_i \{ D_{il} - e_l(\boldsymbol{X}_i) \}.$$

We further let $\boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}} = -\mathbb{E}[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} l_i(\boldsymbol{\theta})]$ be the information matrix, whose exact form can be expressed in a similar fashion but is omitted here for brevity. We denote a consistent estimator for this information by $\hat{\boldsymbol{I}}_{\boldsymbol{\theta}\boldsymbol{\theta}}$. Under standard regularity conditions (Lehmann, 1983), the stochastic expansion for the maximum likelihood estimator is

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{S}_{\boldsymbol{\theta},i} + o_p(1),$$

where $o_p(1)$ is asymptotically negligible as $n \to \infty$.

With the multinomial logistic model, the generalized overlap weights are expressed as functions of $\boldsymbol{\theta}$:

$$\begin{aligned} w_1(\boldsymbol{X}_i) &= w_1(\boldsymbol{X}_i; \boldsymbol{\theta}) = \frac{1}{1 + \sum_{k=2}^{J} \exp(-\alpha_k - \boldsymbol{X}_i^T \boldsymbol{\beta}_k)} \\ w_j(\boldsymbol{X}_i) &= w_j(\boldsymbol{X}_i; \boldsymbol{\theta}) = \frac{\exp(-\alpha_j - \boldsymbol{X}_i^T \boldsymbol{\beta}_j)}{1 + \sum_{k=2}^{J} \exp(-\alpha_k - \boldsymbol{X}_i^T \boldsymbol{\beta}_k)}, \quad j = 2, \ldots, J, \end{aligned}$$

and the derivative of the weights takes the form

$$\dot{w}_j(\boldsymbol{X}_i) \equiv \frac{\partial}{\partial \boldsymbol{\theta}} w_j(\boldsymbol{X}_i) = \left(\frac{\partial}{\partial \alpha_2} w_j(\boldsymbol{X}_i), \ldots, \frac{\partial}{\partial \alpha_J} w_j(\boldsymbol{X}_i), \frac{\partial}{\partial \boldsymbol{\beta}_2^T} w_j(\boldsymbol{X}_i), \ldots, \frac{\partial}{\partial \boldsymbol{\beta}_J^T} w_j(\boldsymbol{X}_i) \right)^T,$$

where for $j = 1, \ldots, J$ and $l = 2, \ldots, J$,

$$\frac{\partial}{\partial \boldsymbol{\beta}_l} w_j(\boldsymbol{X}_i) = \boldsymbol{X}_i \frac{\partial}{\partial \alpha_l} w_j(\boldsymbol{X}_i) = \boldsymbol{X}_i \{ w_j(\boldsymbol{X}_i) w_l(\boldsymbol{X}_i) - \delta_{jl} w_l(\boldsymbol{X}_i) \},$$

and $\delta_{jl} = \mathbb{1}\{j = l\}$.

For $j = 1, \ldots, J$, the plug-in weighting estimator $\hat{m}_j^h$ can be regarded as the solution of the following estimating equation

$$\sum_{i=1}^{n} \boldsymbol{U}(\hat{m}_j^h, \hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} D_{ij}(Y_i - \hat{m}_j^h) w_j(\boldsymbol{X}_i; \hat{\boldsymbol{\theta}}) = \mathbf{0}.$$

Under standard regularity conditions (van der Vaart, 1998), a first-order Taylor expansion of the unbiased estimating equations around the truth leads to

$$\begin{aligned} \sqrt{n}(\hat{m}_j^h - m_j^h) &= \varpi^{-1} \Big\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{U}(m_j^h, \boldsymbol{\theta}) + \boldsymbol{H}_j^T \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \Big\} + o_p(1) \\ &= \varpi^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big\{ D_{ij}(Y_i - m_j^h) w_j(\boldsymbol{X}_i) + \boldsymbol{H}_j^T \boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \boldsymbol{S}_{\boldsymbol{\theta},i} \Big\} + o_p(1), \end{aligned} \tag{A.3}$$

where $\varpi = \mathbb{E}[D_{ij} w_j(\boldsymbol{X}_i)] = \mathbb{E}[h(\boldsymbol{X}_i)]$, and $\boldsymbol{H}_j = \mathbb{E}[D_{ij}(Y_i - m_j^h)\dot{w}_j(\boldsymbol{X}_i)] = \mathbb{E}[(Y_i - m_j^h) e_j(\boldsymbol{X}_i) \dot{w}_j(\boldsymbol{X}_i)]$. Therefore, given any fixed coefficient $\boldsymbol{a} = (a_1, \ldots, a_J)'$, we have

$$\sqrt{n}\{\hat{\tau}^h(\boldsymbol{a}) - \tau^h(\boldsymbol{a})\} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big\{ \frac{1}{\varpi} \sum_{j=1}^{J} a_j \psi_{ij} \Big\} + o_p(1),$$

where we define $\psi_{ij} = D_{ij}(Y_i - m_j^h) w_j(\boldsymbol{X}_i) + \boldsymbol{H}_j^T \boldsymbol{I}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1} \boldsymbol{S}_{\boldsymbol{\theta},i}$. Since the triplets $\{Y_i, \boldsymbol{X}_i, Z_i\}'s$ are assumed i.i.d., an application of the standard Central Limit Theorem gives,

$$\sqrt{n}\{\hat{\tau}^h(\boldsymbol{a}) - \tau^h(\boldsymbol{a})\} \xrightarrow{d} \mathcal{N}\left(0, \varpi^{-2} \, \mathbb{E}\Big\{ \sum_{j=1}^{J} a_j \psi_{ij} \Big\}^2 \right).$$

In practice, we use the empirical sandwich estimator to consistently estimate the large-sample variance (Stefanski and Boos, 2002); the variance of $\hat{\tau}^h(\boldsymbol{a})$ is estimated by

$$\frac{1}{(n\hat{\varpi})^2}\sum_{i=1}^{n}\left\{\sum_{j=1}^{J}a_j\hat{\psi}_{ij}\right\}^2,$$

where

$$\begin{aligned}\hat{\psi}_{ij} &= D_{ij}(Y_i-\hat{m}_j^h)w_j(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}})+\hat{\boldsymbol{H}}_j^T\hat{\boldsymbol{I}}_{\boldsymbol{\theta}\boldsymbol{\theta}}^{-1}\hat{\boldsymbol{S}}_{\boldsymbol{\theta},i},\\ \hat{\varpi} &= \frac{1}{n}\sum_{i=1}^{n}\Big\{\sum_{k=1}^{J}1/\hat{e}_k(\boldsymbol{X}_i)\Big\}^{-1},\\ \hat{\boldsymbol{H}}_j &= \frac{1}{n}\sum_{i=1}^{n}D_{ij}(Y_i-\hat{m}_j^h)\dot{w}_j(\boldsymbol{X}_i;\hat{\boldsymbol{\theta}}),\end{aligned}$$

and $\hat{\boldsymbol{S}}_{\boldsymbol{\theta},i}$ is the estimated individual score function (A.2) from the propensity model. For pairwise comparisons, we substitute $\boldsymbol{a}$ with $\boldsymbol{\lambda}_{j,j'}$ to obtain the results in Theorem 1.

For completeness, we next offer three remarks regarding variance estimation.

REMARK 3. *One could similarly characterize the asymptotic distribution of a collection of estimators specified by different contrast coefficients. Briefly, let the coefficient matrix* $\boldsymbol{A}_{J\times R}=(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_R)$*, where the vector* $\boldsymbol{a}$*'s are distinct from one another. For pairwise comparisons, each vector* $\boldsymbol{a}$ *is a distinct element in the set* $\mathbb{S}$*. Write* $\boldsymbol{\tau}=(\tau^h(\boldsymbol{a}_1),\ldots,\tau^h(\boldsymbol{a}_R))'$*, and* $\hat{\boldsymbol{\tau}}=(\hat{\tau}^h(\boldsymbol{a}_1),\ldots,\hat{\tau}^h(\boldsymbol{a}_R))'$ *as the corresponding weighting estimators. Further denote* $\boldsymbol{\psi}_i=(\psi_{i1},\ldots,\psi_{iJ})'$*, and it can be shown using similar arguments that*

$$\sqrt{n}(\hat{\boldsymbol{\tau}}^h-\boldsymbol{\tau}^h)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varpi^{-1}\boldsymbol{A}^T\boldsymbol{\psi}_i+o_p(1)\xrightarrow{d}\mathcal{N}\left(\boldsymbol{0},\varpi^{-2}\boldsymbol{A}^T\,\mathbb{E}\{\boldsymbol{\psi}_i\boldsymbol{\psi}_i^T\}\boldsymbol{A}\right).$$

The covariance for $\hat{\boldsymbol{\tau}}^h$ *can then be estimated by the empirical sandwich estimator*

$$\hat{\mathbb{V}}(\hat{\boldsymbol{\tau}}^h)=(n\hat{\varpi})^{-2}\boldsymbol{A}^T\left\{\sum_{i=1}^{n}\hat{\boldsymbol{\psi}}_i\hat{\boldsymbol{\psi}}_i^T\right\}\boldsymbol{A}.$$

REMARK 4. *Although the above derivation focuses on the generalized overlap weights, a more general presentation for other members of the balancing weights is possible, provided that the balancing weights is a differentiable function in the generalized propensity scores. This differentiability condition rules out the generalized matching weights, which is smooth but not everywhere differentiable and so closed-form variance requires parametric smooth approximation (Li and Greene, 2013) (such approximations could be challenging with multiple treatments since the weight function have infinite-many non-differentiable points). In particular, if we choose the balancing weights as the inverse probability weights, in which case* $h(\boldsymbol{X}) = 1$ *and the target population is the combined population from all groups, the above derivation can be repeated by substituting the correct forms of* $w_j(\boldsymbol{X}_i)$ *and* $\dot{w}_j(\boldsymbol{X}_i)$*. For example, the inverse probability weights are*

$$\begin{aligned} w_1(\boldsymbol{X}_i) &= 1/e_1(\boldsymbol{X}_i) = 1 + \sum_{k=2}^{J} \exp(\alpha_k + \boldsymbol{X}_i^T \boldsymbol{\beta}_k) \\ w_j(\boldsymbol{X}_i) &= 1/e_j(\boldsymbol{X}_i) = \frac{1 + \sum_{k=2}^{J} \exp(\alpha_k + \boldsymbol{X}_i^T \boldsymbol{\beta}_k)}{\exp(\alpha_j + \boldsymbol{X}_i^T \boldsymbol{\beta}_j)}, \qquad j = 2, \ldots, J, \end{aligned}$$

and the derivative of the weights takes the form

$$\frac{\partial}{\partial \boldsymbol{\beta}_l} w_j(\boldsymbol{X}_i) = \boldsymbol{X}_i \frac{\partial}{\partial \alpha_l} w_j(\boldsymbol{X}_i) = \boldsymbol{X}_i \{ w_j(\boldsymbol{X}_i)/w_l(\boldsymbol{X}_i) - \delta_{jl} w_l(\boldsymbol{X}_i) \},$$

for $j = 1, \ldots, J$ *and* $l = 2, \ldots, J$*. Of note, theis empirical sandwich variance for* $h(\boldsymbol{X}) = 1$ *extends the one proposed by Lunceford and Davidian (2004) for binary treatments, and is used to obtain the interval estimates for IPW in the main manuscript.*

REMARK 5. *We have focused on the case with a multinomial logistic propensity score model, but in fact the derivation can be made more general to accommodate other propensity score models that admit a regular and asymptotically linear estimator for the model parameters (Tsiatis, 2006). This condition permits a stochastic expansion for* $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$*, which can then be substituted into (A.3) to obtain the corresponding sandwich variance estimator. In particular, one could replace the multinomial logistic model with a multinomial Probit model, which is another commonly used regression model to accommodate categorical responses.*

SUPPLEMENT D: ADDITIONAL SIMULATION RESULTS

In the second simulation with $J = 6$ groups, we specify the parameters to simulate both adequate and lack of overlap. Specifically, we spec-

ify $\boldsymbol{\beta}_2^T = \kappa_2 \times (1,1,2,1,1,1)$, $\boldsymbol{\beta}_3^T = \kappa_3 \times (1,1,1,1,1,-5)$, $\boldsymbol{\beta}_4^T = \kappa_4 \times (1,1,1,1,1,5)$, $\boldsymbol{\beta}_5^T = \kappa_5 \times (1,1,1,-2,1,1)$ and $\boldsymbol{\beta}_6^T = \kappa_6 \times (1,1,1,-2,-1,1)$. We use $(\kappa_2, \kappa_3, \kappa_4, \kappa_5, \kappa_6) = (0.1, 0.15, 0.2, 0.25, 0.3)$ to simulate a scenario with adequate overlap and $(\kappa_2, \kappa_3, \kappa_4, \kappa_5, \kappa_6) = (0.4, 0.6, 0.8, 1, 1.2)$ to represent a challenging scenario with strong propensity tails. The intercepts are chosen so that the marginal treatment proportions are fixed around $(0.12, 0.16, 0.12, 0.25, 0.2, 0.15)$. Finally, the coefficients for the outcome model is specified as $\boldsymbol{\gamma}_1^T = (-1.5, 1, 1, 1, 1, 1, 1)$, $\boldsymbol{\gamma}_2^T = (-4, 2, 3, 1, 2, 2, 2)$, $\boldsymbol{\gamma}_3^T = (4, 3, 1, 2, -1, -1, -4)$, $\boldsymbol{\gamma}_4^T = (1, 4, 1, 2, -1, -1, -3)$, $\boldsymbol{\gamma}_5^T = (3.5, 5, 1, 2, -1, -1, -2)$ and $\boldsymbol{\gamma}_6^T = (3.5, 6, 1, 2, -1, -1, -1)$. The total sample size is fixed at $n = 6000$ for $J = 6$. Visual inspections of the overlap in each simulation scenario are provided in Supplementary Figures 1-4. Simulation results for $J = 6$ are presented in Supplementary Figures 5 and 6.

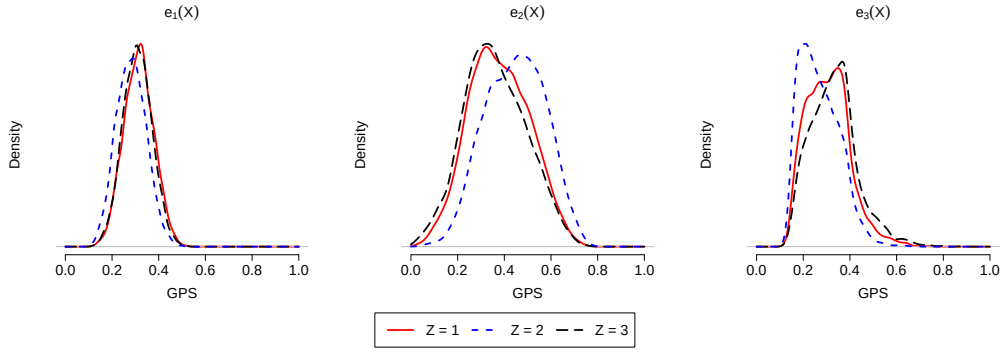

FIG 1. *Distribution of the generalized propensity scores in the simulation with $J = 3$ groups and adequate overlap.*

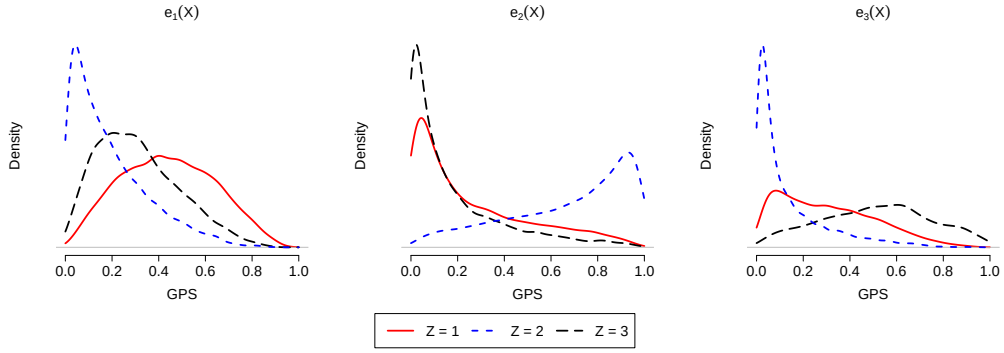

FIG 2. *Distribution of the generalized propensity scores in the simulation with $J = 3$ groups and lack of overlap.*

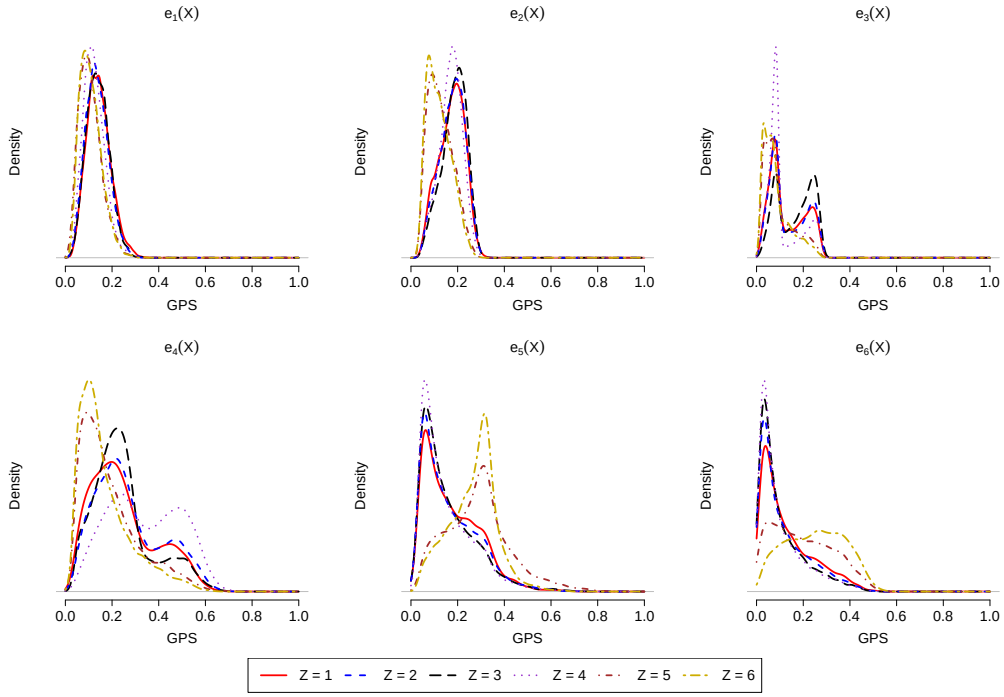

FIG 3. *Distribution of the generalized propensity scores in the simulation with $J = 6$ groups and adequate overlap.*

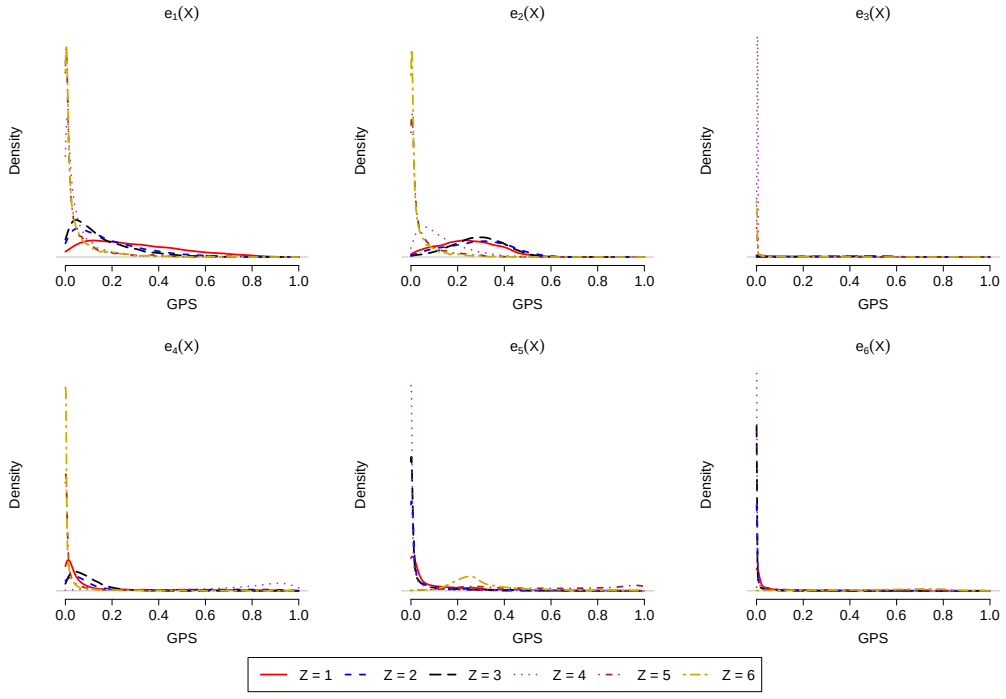

FIG 4. *Distribution of the generalized propensity scores in the simulation with* $J = 6$ *groups and lack of overlap.*

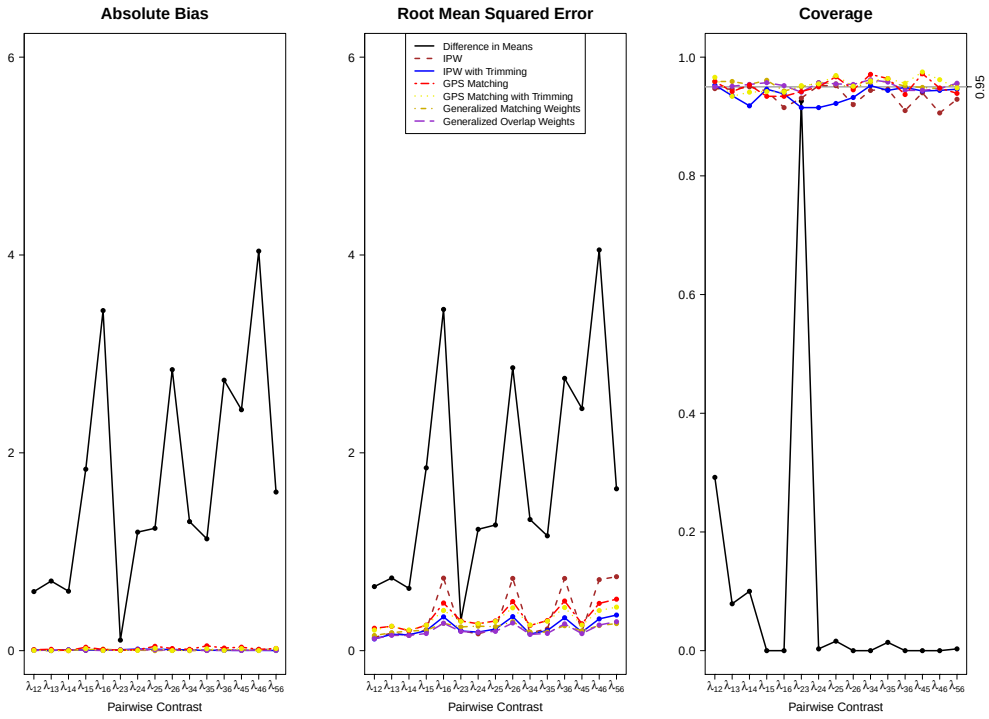

FIG 5. *Simulation results with* $J = 6$ *treatment groups and* $(\kappa_2, \kappa_3, \kappa_4, \kappa_5, \kappa_6) = (0.1, 0.15, 0.2, 0.25, 0.3)$*, i.e., with adequate overlap. Optimal trimming excludes* $3\% \sim 7\%$ *of the total sample. For a given approach, each one of the* 15 *causal comparisons is represented by the contrast* $\boldsymbol{\lambda}_{j,j'}$ *for notational simplicity.*

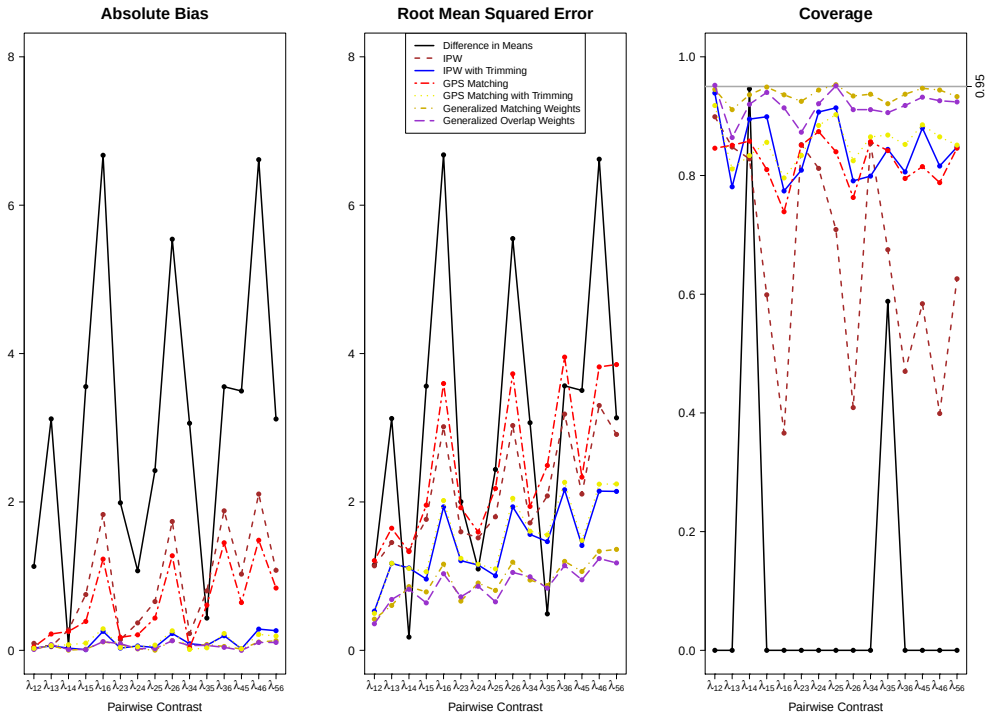

FIG 6. *Simulation results with* $J = 6$ *treatment groups and* $(\kappa_2, \kappa_3, \kappa_4, \kappa_5, \kappa_6) = (0.4, 0.6, 0.8, 1, 1.2)$*, i.e., with strong propensity tails. Optimal trimming excludes* $52\% \sim 74\%$ *of the total sample. For a given approach, each one of the* 15 *causal comparisons is represented by the contrast* $\boldsymbol{\lambda}_{j,j'}$ *for notational simplicity.*

SUPPLEMENT E: SAMPLE R CODE FOR IMPLEMENTING GENERALIZED OVERLAP WEIGHTS

We provide sample R code to illustrate the application of the generalized overlap weighting scheme along with the implementation of the empirical sandwich variance, based on a simulated observational data set with $J = 3$ treatments. We use the following code to simulate a data set for an illustrative analysis. The data generating process is described in Section 5 of the main manuscript.

```
# Consider three treatments and sample size 1500
set.seed(2019)
require(MASS)
J = 3
n = 1500

# Simulate pre-treatment covariates
# The simulation setup is similar to Yang et al. (2016)
# X1-X3 are multivariate normal covariates
# X4 is uniform variable
# X5 is a Chi-squared variable
# X6 is binary
vars = c(2,1,1)
covars = c(1,-1,-.5)
mu = c(0,0,0)
tau = 1
Sigma = diag(vars)
Sigma[2,1] = Sigma[1,2] = covars[1]
Sigma[3,1] = Sigma[1,3] = covars[2]
Sigma[3,2] = Sigma[2,3] = covars[3]

X13 = mvrnorm(n, mu=mu, Sigma=Sigma, empirical = FALSE)
X4 = runif(n,-3,3)
X5 = rchisq(n, df=1)
X6 = rbinom(n, size=1, prob=.5)
X16 = cbind(X13, X4, X5, X6)
X06 = cbind(1, X13, X4, X5, X6)

# Assignment mechanism
beta1 = c(0,0,0,0,0,0,0)
beta2 = c(0.344, 0.2*c(1,1,1,-1,-1,1))
beta3 = c(-0.178, 0.1*c(1,1,1,1,1,1))
xb2 = c(X06 %*% beta2)
xb3 = c(X06 %*% beta3)
exb2 = exp(xb2)
exb3 = exp(xb3)
e1 = 1 / (1+exp(xb2)+exp(xb3))
e2 = exp(xb2)/(1+exp(xb2)+exp(xb3))
```

```
e3 = exp(xb3)/(1+exp(xb2)+exp(xb3))
e = cbind(e1,e2,e3)    # true propensity scores

# Simulate observed treatment
D = matrix(NA, n, J)
colnames(D) = c("D1", "D2", "D3")
for(k in 1:n){
  D[k,] = rmultinom(1, 1, prob = e[k,])
}
Z = D[,"D1"] + 2*D[,"D2"] + 3*D[,"D3"]

# True potential outcome models
u  = rnorm(n)
gamma1 = c(-1.5,1,1,1,1,1,1)
gamma2 = c(-4,2,3,1,2,2,2)
gamma3 = c(3,3,1,2,-1,-1,-1)
EY1 = c(X06 %*% gamma1)
EY2 = c(X06 %*% gamma2)
EY3 = c(X06 %*% gamma3)
EY = cbind(EY1, EY2, EY3)
Y = rowSums(EY*D) + u
```

We now create a data set with only observed outcomes, treatments and pre-treatment covariates, and examine the first few rows to get a sense of the data structure.

```
# Create the analysis datasets
analdata = data.frame(Y=Y, Z=Z, X=X16)
colnames(analdata) = c("Y", "Z", "X1", "X2", "X3", "X4", "X5",
    "X6")

# Peek at the data
round(head(analdata),3)

#          Y Z      X1      X2      X3      X4     X5 X6
# 1    7.673 2   0.869   1.696   0.261   2.310  0.156  0
# 2   -3.701 2  -1.142   0.370   0.467  -1.628  0.041  1
# 3 -11.345 2  -1.975  -1.866   1.247  -0.468  0.819  1
# 4    3.534 2   1.857   0.463  -0.003   2.195  0.015  0
# 5    0.524 1  -2.555  -0.790  -0.120   2.398  2.391  0
# 6    7.198 3   1.472   0.767   0.349   1.229  0.819  0
```

We could repeat the above data generating process for a large number of times, and numerically approximate the true values of the pairwise ATO (by averaging out the Monte Carlo errors in repeated simulations). We will omit the details here but indicate that the true pairwise ATO quantities are $\tau^h(\boldsymbol{\lambda}_{1,2}) = 1.03$, $\tau^h(\boldsymbol{\lambda}_{1,3}) = -1.36$ and $\tau^h(\boldsymbol{\lambda}_{2,3}) = -2.39$. From now on, we

will remain agnostic to the true data generating process, and estimate the pairwise ATO using only the observed data object `analdata`. We estimate the generalized propensity scores using a correctly specified multinomial logistic regression.

```
require(nnet)
require(numDeriv)

# Create a n x J matrix of treatment indicators
# Each row corresponds to an observation
# Each column represents a treatment category
D = cbind(as.numeric(analdata$Z == 1),
          as.numeric(analdata$Z == 2),
          as.numeric(analdata$Z == 3))

# Estimate the GPS model
n = length(analdata$Y)
gps.fit = multinom(Z ~ X1 + X2 + X3 + X4 + X5 + X6, data =
    analdata, maxit = 500, Hess = TRUE, trace = FALSE)
Thetah = t(coef(gps.fit))          # matrix coefficient
thetah = c(Thetah)                 # vec operator
IthetaInv = n*vcov(gps.fit)        # extract the covariance
    matrix
e = gps.fit$fitted.values          # estimated propensity scores
eInv = 1/e                         # inverse probability weights
h = as.numeric(1/rowSums(eInv))    # obtain the optimal tilting
    function
w = eInv * h                       # create the generalized
    overlap weights
```

In practice, we need to check whether the estimated generalized overlap weights adequately balance the weighted covariates across groups (therefore assess the adequacy of the generalized propensity score model). If the weighted balance is unsatisfactory, we may need to revise the propensity score model by including additional higher-order terms or interactions to make improvements. We defined the following two functions to calculate the *population standardized difference* (PSD) and the pairwise *absolute standardized difference* (ASD), defined in Section 3.2.

```
# Population standardized difference
PSDfun = function(Z, D, covM, h, w){
  # Z: treatment value
  # D: matrix of treatment indicators
  # covM: covariate matrix, n by p
  # h: value of the h function, n by 1
  # w: matrix of weights, n by J

```

```
  p = ncol(covM)
  J = ncol(D)
  psdM = matrix(NA, J, p)
  for(k in 1:p){
    x = as.numeric(covM[,k])
    xbar.p = sum(rowSums(D*x*h))/sum(rowSums(D*h))
    xbar.j = colSums(D*x*w)/colSums(D*w)
    s2x = mean(tapply(x, Z, var))
    psdM[,k] = abs(xbar.j - xbar.p)/sqrt(s2x)
  }
  return(t(psdM))
}

# create contrast matrix
J = 3
Jc = choose(J, 2)                # number of contrasts
loc = t(combn(J,2))
C = matrix(0, Jc, J)
for(i in 1:Jc){
  C[i,loc[i,1]] = 1
  C[i,loc[i,2]] = -1
}

# Absolute standardized difference
ASDfun = function(Z, D, C, covM, h, w){
  # Z: treatment value
  # D: matrix of treatment indicators
  # C: contrast matrix
  # covM: covariate matrix, n by p
  # h: value of the h function
  # w: matrix of weights, n by J

  p = ncol(covM)
  Jc = nrow(C)
  asdM = matrix(NA, Jc, p)

  # loop through each pair
  for(k in 1:p){
    x = as.numeric(covM[,k])
    xbar.j = colSums(D*x*w)/colSums(D*w)
    s2x = mean(tapply(x, Z, var))
    asdM[,k] = abs(C %*% xbar.j)/sqrt(s2x)
  }
  return(t(asdM))
}
```

We compared the covariate balance before and after weighting using the following code. The covariate balance is satisfactory after weighting, which assures the adequacy of the propensity score model.

```
# Before weighting: PSD and ASD may exceed 10%
round(PSDfun(Z=Z, D=D, covM=X, h=rep(1,n), w=rep(1,n)), 3)
#       [,1]  [,2]  [,3]
# [1,] 0.182 0.122 0.006
# [2,] 0.240 0.144 0.033
# [3,] 0.085 0.042 0.026
# [4,] 0.083 0.254 0.294
# [5,] 0.003 0.141 0.207
# [6,] 0.058 0.034 0.008
round(ASDfun(Z=Z, D=D, C=C, covM=X, h=rep(1,n), w=rep(1,n)), 3)
#       [,1]  [,2]  [,3]
# [1,] 0.304 0.188 0.116
# [2,] 0.384 0.273 0.111
# [3,] 0.127 0.111 0.016
# [4,] 0.338 0.210 0.548
# [5,] 0.144 0.204 0.348
# [6,] 0.092 0.066 0.026

# After weighting: PSD and ASD all below 10%
round(PSDfun(Z=Z, D=D, covM=X, h=h, w=w), 3)
#       [,1]  [,2]  [,3]
# [1,] 0.018 0.016 0.024
# [2,] 0.019 0.002 0.005
# [3,] 0.007 0.004 0.013
# [4,] 0.006 0.004 0.007
# [5,] 0.002 0.034 0.011
# [6,] 0.013 0.007 0.008

round(ASDfun(Z=Z, D=D, C=C, covM=X, h=h, w=w), 3)
#       [,1]  [,2]  [,3]
# [1,] 0.002 0.006 0.008
# [2,] 0.021 0.014 0.006
# [3,] 0.004 0.006 0.010
# [4,] 0.003 0.001 0.004
# [5,] 0.035 0.009 0.044
# [6,] 0.006 0.005 0.001
```

The following code is used to estimate the pairwise ATO using the generalized overlap weights. The variance is obtained using the empirical sandwich estimator, which is also used to construct the 95% confidence interval. Recall that the true pairwise ATO quantities are $\tau^h(\boldsymbol{\lambda}_{1,2}) = 1.03$, $\tau^h(\boldsymbol{\lambda}_{1,3}) = -1.36$ and $\tau^h(\boldsymbol{\lambda}_{2,3}) = -2.39$; the estimated pairwise ATO, $\hat{\tau}^h(\boldsymbol{\lambda}_{1,2}) = 1.08$, $\hat{\tau}^h(\boldsymbol{\lambda}_{1,3}) = -1.19$ and $\hat{\tau}^h(\boldsymbol{\lambda}_{2,3}) = -2.27$, are close to the true quantities in this example.

```
# Point estimation
mhat = as.numeric(colSums(D*Y*w)/colSums(D*w))
```

```
tau = as.matrix(C%*%mhat)

# Variance and interval estimation
omega = mean(h)

# Calculate gradient of weights
Hmat = NULL
for(j in 1:J){
  wj = function(theta){
    Theta = matrix(theta, ncol(X)+1, ncol(D)-1)
    Eta = cbind(1,X) %*% cbind(0,Theta)
    return(as.numeric(exp(-Eta[,j])) / as.numeric(rowSums(exp(-Eta))))
  }
  wdotj = jacobian(wj,thetah)
  Hmat = rbind(Hmat, c(colMeans(D[,j] * (Y - mhat[j]) * wdotj)))
}

# Multinomial logistic score function
loglik = function(theta){
  Theta = matrix(theta, ncol(X)+1, ncol(D)-1)
  Eta = cbind(1,X) %*% Theta
  ltheta = as.numeric(rowSums(D[,-1]*Eta)-log(1 + rowSums(exp(Eta))))
  return(ltheta)
}
Sthetah = jacobian(loglik, thetah)

# Covariance matrix of pairwise ATO estimates
# This a multivariate version of Theorem 1
# and relevant details are provided in Remark 3 (Supplement C)
YMat = matrix(rep(Y, J), n, J)
mhatMat = matrix(rep(mhat, each=n), n, J)
Psi = t(D*(YMat - mhatMat)*w) + Hmat %*% IthetaInv %*% t(Sthetah)
Sigmah = diag(C %*% tcrossprod(Psi) %*% t(C) / (n * omega)^2)
se = sqrt(Sigmah)
lcl = tau - qnorm(0.975) * se
ucl = tau + qnorm(0.975) * se
results = cbind(tau, lcl, ucl)
colnames(results) = c("Point estimates", "95% Lower limit", "95% Upper limit")
rownames(results) = c("1-2", "1-3", "2-3")
round(results, 3)
#     Point estimates 95% Lower limit 95% Upper limit
# 1-2           1.080           0.758           1.402
# 1-3          -1.188          -1.515          -0.861
# 2-3          -2.268          -2.756          -1.779
```

REFERENCES

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika* **87** 706–710.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press, New York, NY.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluations of Active Labor Market Policies in Europe* (M. Lechner and F. Pfeiffer, eds.) 43–58. Heidelberg: Physica.

Lehmann, E. (1983). *Theory of Point Estimation*. Spinger, New York.

Li, L. and Greene, T. (2013). A weighting analogue to pair matching in propensity score analysis. *International Journal of Biostatistics* **9** 1-20.

Lopez, M. J. and Gutman, R. (2017). Estimation of causal effects with multiple treatments: A review and new ideas. *Statistical Science* **32** 432–454.

Lunceford, J. K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* **23** 2937–2960.

Stefanski, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *American Statistician* **56** 29–38.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York, NY.

van der Vaart, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

F. Li
Department of Biostatistics
Yale University
135 College St
New Haven, Connecticut 06510
USA
E-mail: fan.f.li@yale.edu

F. Li
Department of Statistical Science
Duke University
122 Old Chemistry Building
Durham, North Carolina 27708
USA
E-mail: fli@stat.duke.edu