STA 640 — Causal Inference

Chapter 3.6: Non-binary Treatments

Fan Li

Department of Statistical Science
Duke University

Non-binary Treatments

Non-binary treatments include

- Nominal categorical treatments: no ordering of different categories
  - Example 1: compare the patient satisfaction between several physicians ("treatment" is physician)
  - Example 2: compare the treatment effect between three different medications
- Ordinal categorical treatments: categories are ordered
  - Scales: e.g. never - sometimes - always
  - Doses: e.g. low - medium - high
- Continuous treatments: can always be discretized into ordinal treatments with doses

Nominal Treatments: Notations

- Units: $1, ..., N$
- Treatments: $j \in \mathbb{Z}$, where $\mathbb{Z} = \{1, \cdots, J\}$. Example: for binary treatments, $J = 2$ and $\mathbb{Z} = \{0, 1\}$
- $\mathbf{T}_i = (T_{i1}, \cdots, T_{iJ})$: the multinomial trial of Z_i, where $T_{ij} = 1$ when $Z_i = j$ and $T_{ij} = 0$ otherwise
- Potential outcomes: $Y_i(j)$, for all $j \in \mathbb{Z}$
- Observed data: pre-treatment variables (covariates) X_i, treatment status Z_i, and the observed outcome $Y_i = Y_i(Z_i)$.

Nominal Treatments: Estimands

For nominal treatments, the most common causal estimands are pairwise comparisons (Lechner, 2001): for any $j, j' \in \mathbb{Z}$

- The ATE between treatment j and j' for all units:

$$ATE_{j,j'} \equiv \mathbb{E}[Y_i(j) - Y_i(j')]$$

  Here the equivalent estimand is $\mu(j) = \mathbb{E}[Y_i(j)]$, for all $j \in \mathbb{Z}$

- The ATT between treatment j and j' for units in treatments j and j':

$$ATT_{j,j'} \equiv \mathbb{E}[Y_i(j) - Y_i(j')|Z_i = j, Z_i = j']$$

- The ATT between treatment j and j' for units in trt j:

$$ATT_{j|j,j'} \equiv \mathbb{E}[Y_i(j) - Y_i(j')|Z_i = j]$$

Transitive Effects

- $ATE_{j,j'}$ is transitive, i.e. ATE_{j_1,j_2} and ATE_{j_1,j_3} can determine ATE_{j_2,j_3}, e.g.

$$ATE_{1,3} = ATE_{1,2} + ATE_{2,3}$$

- $ATT_{j|j,j'}$ with the same referent group j is also transitive:

$$ATE_{1|1,3} = ATE_{1|1,2} + ATE_{1|2,3}$$

- $ATT_{j,j'}$ is not transitive because of the change of target population in different pairwise comparisons

Assumptions

- SUTVA
- Overlap: $0 < \Pr(Z_i = j|X) < 1$ for all $j \in \mathbb{Z}$ – much harder to satisfy in multiple treatments
- Unconfoundedness: versions
  - Strong unconfoundedness: $(Y(1), \ldots, Y(J)) \perp Z|X$.
    - Interpretation: within each strata of X, the assignment to any treatment $Z = j$ is randomized
    - Standard joint independence – think about $J = 2$
    - Not necessary for identification of causal quantities, can be weakened
  - Weak unconfoundedness (next page)

Weak Unconfoundedness

(Imbens, 2000, Biometrika)

- Weak unconfoundedness: for all $j \in \mathbb{Z}$,

$$Y(j) \perp \mathbf{1}\{Z = j\}|X$$

  - Interpretation: For any treatment level j, within the strata of X, each unit being assigned to treatment j vs. not is randomized
  - Weak unconfoundedness is weaker than the strong unconfoundedness assumption that requires the joint independence of all potential outcomes.
  - Different implications wrt balancing scores (more later)

Generalized Propensity Score (GPS)

(Imbens, 2000, Biometrika)

Definition: Generalized Propensity Score (GPS) – the conditional probability of being assigned to a treatment level given the covariates:

$$e_j(x) \equiv \Pr(Z_i = j | X_i = x)$$

- Each unit has J GPSs: $\boldsymbol{e} = \{e_1, ..., e_J\}$, and $\sum_{j=1}^{J} e_j(x) = 1$. In practice, $J - 1$ scores are adequate to characterize each unit, but not fewer
- Example: $J = 3$, three units with $\boldsymbol{e} = (.3, .6, .1)$, $(.3, .25, .45)$, $(.3, .1, .6)$.
- Individual matching less suited to multiple treatments (Imbens, 2000)

Properties of GPS

- Under strong unconfoundedness,

$$(Y(1), \ldots, Y(J)) \perp Z | (e_1(X), \ldots, e_{J-1}(X)))$$

Interpretation: if the treatment strongly unconfounded given X, it is so given the vector of GPSs

  - Dimension of balancing scores $\uparrow$ with J (Imbens, 2000)
  - Without additional assumptions, there is in general no scalar function $b(X)$ s.t.

$$(Y(1), \ldots, Y(J)) \perp Z | b(X)$$

suggesting that the advantages of the propensity score approach do not carry over to the multiple treatments

Properties of GPS

- GPS has similar dimension-reduction properties of the propensity score in binary treatments
  - Balancing property: for all $j \in \mathbb{Z}$

$$\mathbf{1}\{Z = j\} \perp X | e_j(X)$$

    Interpretation: for each treatment level j, within the strata of GPS for that j, the covariates X are balanced between the group $Z = j$ and all other groups combined

  - Weak unconfoundedness: if $Y(j) \perp \mathbf{1}\{Z = j\}|X$ for all $j \in \mathbb{Z}$, then

$$Y(j) \perp \mathbf{1}\{Z = j\}|e_j(X)$$

    Interpretation: If the treatment is weakly unconfounded given X, it is also weakly unconfounded given the scalar score $e_j(X)$, for all j levels

Nominal Treatments: Estimate GPS

- To estimate the GPS for nominal treatments
  - Option 1: Use multinomial logistic or probit model
  - Option 2: Conduct a series pairwise comparisons (e.g., using the units only in two groups) using the standard binary propensity score
  - Option 3: generalized boosted models (McCaffrey et al. 2013, SIM) – R package `twang`
- Lechner (2002): correlation coefficients of the conditional trt assignment probabilities estimated from option 1 and 2 (with probit link) is 0.99
- Check overlap and balance is important: leverage the balancing property of GPS, similar to binary settings, check for each treatment level

Nominal Treatments: Inverse Probability Weighting

(Robins et al., 2000, Epidemiology; Feng et al. 2011, SIM)

- Similar to binary treatments, one can use (inverse probability) weighting for categorical nominal and ordinal treatments
- Under weak unconfoundedness,

$$\mathbb{E}_x\{\mathbb{E}[Y\mathbf{1}\{Z_i = j\}/e_j(x)]\} = \mathbb{E}[Y(j)]$$

- Define the inverse probability weight as $1/e_j(x)$ for each unit under each treatment j
- Estimate the group mean $\mathbb{E}[Y(j)]$ by the weighted average outcome in group j:

$$\hat{\mathbb{E}}[Y(j)] = \frac{\sum_{i:Z_i=j} Y_i/e_j(X_i)}{\sum_{i:Z_i=j_k} 1/e_j(X_i)}$$

- This only applies to the estimation of ATE because the target population is combined from all groups

Nominal Treatments: Inverse Probability Weighting

- IPW is sensitive to extreme weights, exacerbate in the setting of multiple treatments
- Trimming is used to remove extreme weights, but
  - can end up losing a large portion of sample
  - rule of thumbs on threshold are difficult to specify; optimal trimming depends on data and results in ambiguous target population
- Possible remedy: extend the class of balancing weights to multiple treatments

Nominal Treatments: General Estimands

(Li and Li, 2019, AOAS)

- ATE is a special case of the general class of WATE estimands
- Define $m_j(X) = \mathbb{E}[Y(j)|X]$ as the conditional expectation of potential outcome
- Assume the population density corresponding to the *observed* sample, $f(X)$, exists
- Consider a *target population* with a different density $g(X) \propto f(X)h(X)$, for a pre-specified $h(\cdot)$ – *tilting function*
- Average potential outcome in the target population

$$m_j^h \equiv \mathbb{E}_g[Y(j)] = \frac{\mathbb{E}[h(X)m_j(X)]}{\mathbb{E}[h(X)]}.$$

- Estimand: $\tau^h(a) = \sum_{j=1}^J a_j m_j^h$ for coefficient $a = (a_1, \ldots, a_J)$

Nominal Treatments: Balancing Weights

- Recall that

$$f_j(X) = f(X|Z = j) \propto f(X)e_j(X), \quad \forall\ j \in \mathbb{Z}$$

- For a given $h(X)$, to estimate m_j^h, we can weight $f_j(X)$ to the target population using analytical weights

$$w_j(X) \propto \frac{f(X)h(X)}{f(X)e_j(X)} = \frac{h(X)}{e_j(X)}, \quad \forall\ j \in \mathbb{Z}$$

- The class of weights $\{w_j,\ j \in \mathbb{Z}\}$ is called the *balancing weights* – balancing the weighted distributions of covariates across J comparison groups:

$$f_j(X)w_j(X) = f(X)h(X), \quad \forall\ j \in \mathbb{Z}$$

Examples: target population and balancing weights

- Choice of coefficient a determines the causal contrast; often choose a to define pairwise comparisons
  - $a \in \mathcal{S} = \{\lambda_j - \lambda_{j'} : j < j'\}$, where λ_j is the $J \times 1$ vector with one at the jth position and zero everywhere else
- Choice of tilting function h determines the target population and weights

Target population	Tilting function $h(X)$	Weights $\{w_j(X),\ j \in \mathbb{Z}\}$
Combined	1	$\{1/e_j(X),\ j \in \mathbb{Z}\}$
Treated (j'th group)	$e_{j'}(X)$	$\{e_{j'}(X)/e_j(X),\ j \in \mathbb{Z}\}$
Trimmed combined	$\mathbf{1}\{X \in C\}$	$\{\mathbf{1}\{X \in C\}/e_j(X),\ j \in \mathbb{Z}\}$
Matching	$\min_{1 \le l \le J}\{e_l(X)\}$	$\{\min_l\{e_l(X)\}/e_j(X),\ j \in \mathbb{Z}\}$
Overlap	$\dfrac{1}{\sum_{l=1}^{J} 1/e_l(X)}$	$\left\{\dfrac{1/e_j(X)}{\sum_{l=1}^{J} 1/e_l(X)},\ j \in \mathbb{Z}\right\}$

Aside: Transitivity

- Example of transitivity: $ATE_{1,3} = ATE_{1,2} + ATE_{2,3}$, or $a'' = a + a' \Rightarrow \tau(a'') = \tau(a) + \tau(a')$
- Non-transitivity leads to incompatible pairwise contrasts (different populations)
- For pairwise comparisons, fixing a common tilting function $h(X)$ guarantees transitivity
- The non-transitive estimands (Lechner, 2001)

$$\{ATT_{j,j'} = \mathbb{E}[Y(j) - Y(j')|Z = j, Z = j'] : j < j'\}$$

correspond to multiple distinct tilting functions – essentially different target populations

Nonparametric Weighting Estimator

- General principle: estimating the average of the potential outcomes separately for each treatment level, which requires adjusting only a scalar GPS
- Sample weighting estimator (with normalized weights)

$$\hat{m}_j^h = \frac{\sum_{i:Z_i=j} Y_i w_j(X_i)}{\sum_{i:Z_i=j} w_j(X_i)}, \qquad \hat{\tau}^h(a) = \sum_{z=1}^{J} a_j \hat{m}_j^h$$

- **Theorem 1.** Under weak unconfoundedness, for any h and a, $\hat{\tau}^h(a)$ is a consistent estimator of $\tau^h(a)$

Large-sample Properties

- **Theorem 2.** Given a and under regularity conditions, the expectation of the conditional variance converges

$$n \cdot \mathbb{E}\{\mathbb{V}[\hat{\tau}^h(a)|Z, X]\} \rightarrow$$
$$Q(a, h) \equiv \int \left(\sum_j a_j^2 v_j(X)/e_j(X) \right) h^2(X) f(X) \mu(dX)/C_h^2,$$

where $v_j(X) = \mathbb{V}[Y(j)|X]$ and $C_h \equiv \int h(X) f(X) \mu(dX)$ is a constant.

Large-sample Properties - Cont'd

- **Corollary 1.** Under homoscedasticity of residual variance, the following tilting function (the harmonic mean of the propensities to each group)

$$\tilde{h}(X) \propto \frac{1}{\sum_{l=1}^{K} a_l^2 / e_l(X)}$$

gives the smallest asymptotic variance of the weighted estimator $\hat{\tau}^h(a)$ among all h's, and $\min_h Q(a, h) = v/C_{\tilde{h}}$.

Generalized Overlap Weights

(Li and Li, 2019)

- Nominal treatments: pairwise comparison
- Choose $a \in \mathcal{S} = \{\lambda_j - \lambda_{j'} : j < j'\}$, where λ_j is the $J \times 1$ unit vector with one at the jth position and zero everywhere else
- Choose h to minimize the total asymptotic variance of the weighting estimators for all pairwise comparisons

$$\tilde{h}(X) = \arg\min_h \sum_{j<j'} Q(\lambda_j - \lambda_{j'}, h) \propto \frac{1}{\sum_{l=1}^{J} 1/e_l(X)}$$

- For $J = 2$, $\tilde{h}(X) \propto e(X)\{1 - e(X)\}$
- Generalized overlap weights

$$w_j(X) \propto \frac{1}{e_j(X)} \times \frac{1}{\sum_{l=1}^{J} 1/e_l(X)}, \quad j \in \mathbb{Z}$$

Generalized Overlap Weights

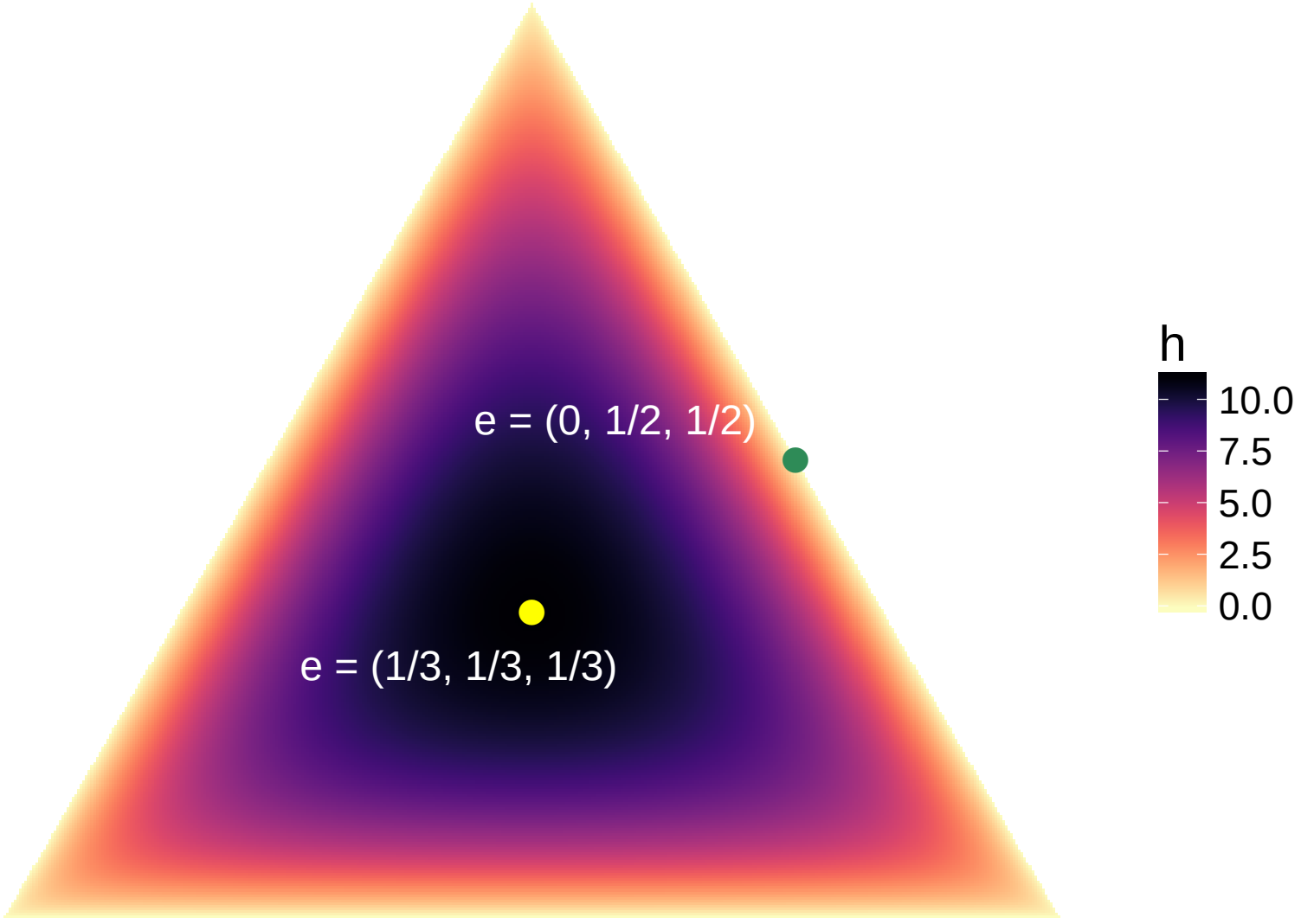
- Recall

$$\tilde{h}(X) \propto \frac{1}{\sum_{l=1}^{K} 1/e_l(X)}, \quad w_j(X) \propto \frac{\tilde{h}(X)}{e_j(X)}$$

- Maximum h is attained when $e_j(X) = 1/J$ for all j – substantial probability to receive each treatment
- Target population: subpopulation with the most overlap in covariates among all groups
- Target estimand: pairwise average treatment effect among the overlap population (ATO)
- Generalized overlap weights are also bounded and robust to extreme propensities
- Pairwise ATO is also adaptive – close to pairwise ATE for "well-designed" studies (randomized studies, $h \propto 1$)

Optimal Tilting Function: Ternary Plot

- For $J = 3$, plot $h(X)$ over two-dimensional probability simplex



Generalized Overlap Weights: Advantages

Conceptual:

- Emphasizes units at clinical equipoise, i.e., with substantial probability of receiving both treatments
- Exemplifies "observational studies analyzed like randomized trials"

Statistical:

- Minimum total variance of the sample weighted estimators among all balancing weights
- Weights are strictly bounded between 0 and 1, bypassing the bias and excessive variance of IPW (common $J \geq 3$)
- Continuously down-weights units in the tails, avoids ad hoc trimming decisions (loss of sample size)

Balance Check

- The adequacy of GPS model should be informed by the resulting covariate balance
- Recall that balancing weights ensure

$$f_j(X)w_j(X) = f(X)h(X) = f_{j'}(X)w_{j'}(X)$$

- Population standardized difference (PSD)

$$\text{PSD}_j = |\bar{X}_j - \bar{X}_p|/S_X; \quad \max_j\{\text{PSD}_j\}$$

- Absolute standardized differences (ASD)

$$\text{ASD}_{j,j'} = |\bar{X}_j - \bar{X}_{j'}|/S_X; \quad \max_{j,j'}\{\text{ASD}_{j,j'}\}$$

where $\bar{X}_j$ is the weighted covariate mean, $\bar{X}_p$ is the covariate mean in the target population, S_X is the pooled standard deviation

Empirical Sandwich Variance

- **Theorem 2.** Under standard regularity conditions, when the GPS is estimated by multinomial logistic regression, the resulting ATO is asymptotically normal

$$\sqrt{n}\{\hat{\tau}^h_{j,j'} - \tau^h_{j,j'}\} \xrightarrow{p} \mathcal{N}\left(0, \mathbb{E}\left\{\psi_{ij} - \psi_{ij'}\right\}^2 / [\mathbb{E}\{h(X)\}]^2\right),$$

where

$$\begin{aligned}\psi_{ij} \quad = \quad & \mathbf{1}\{Z_i = j\}(Y_i - m^h_j)w_j(X_i) + \\ & \mathbb{E}\left\{\mathbf{1}\{Z_i = j\}(Y_i - m^h_j)\frac{\partial}{\partial \theta^T} w_j(X_i)\right\} \mathcal{I}^{-1}_{\theta\theta} S_{\theta,i},\end{aligned}$$

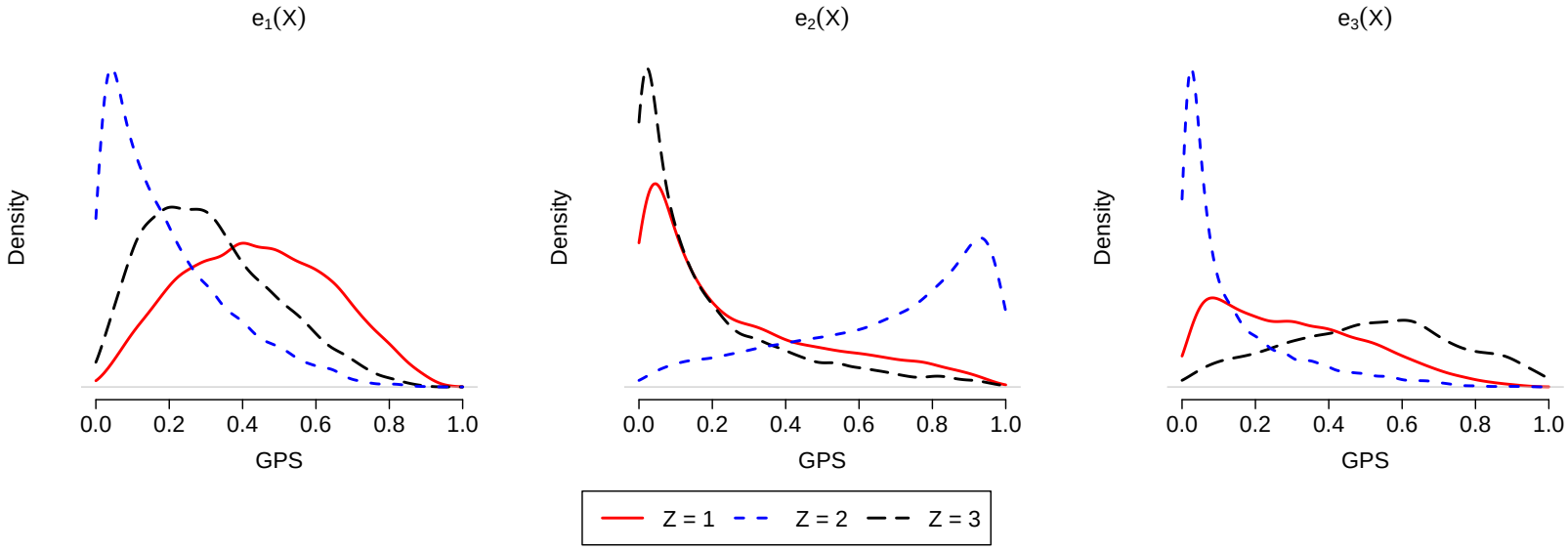
$S_{\theta,i}$, $\mathcal{I}_{\theta\theta}$ individual score and information matrix wrt GPS model.

- Consistent variance estimator

$$\hat{\mathbb{V}}[\hat{\tau}^h_{j,j'}] = \frac{\sum_i \left(\hat{\psi}_{ij} - \hat{\psi}_{ij'}\right)^2}{\left[\sum_i \{\sum_{l=1}^K 1/\hat{e}_l(X_i)\}^{-1}\right]^2},$$

Simulated Example

- Consider $J = 3$ groups with total sample size $N = 1500$
- Specify true GPS via a multinomial logistic regression, and generate Z
- Simulate $Y(j)$ from a linear model
- Consider both adequate overlap and lack of overlap



Simulated Example

	Absolute Bias			RMSE		
	$\tau_{1,2}$	$\tau_{1,3}$	$\tau_{2,3}$	$\tau_{1,2}$	$\tau_{1,3}$	$\tau_{2,3}$
IPW	0.19	0.02	0.17	1.04	0.61	1.16
IPW (Trim)	0.03	0.01	0.01	0.38	0.28	0.47
Overlap	0.01	0.01	0.003	0.28	0.23	0.35

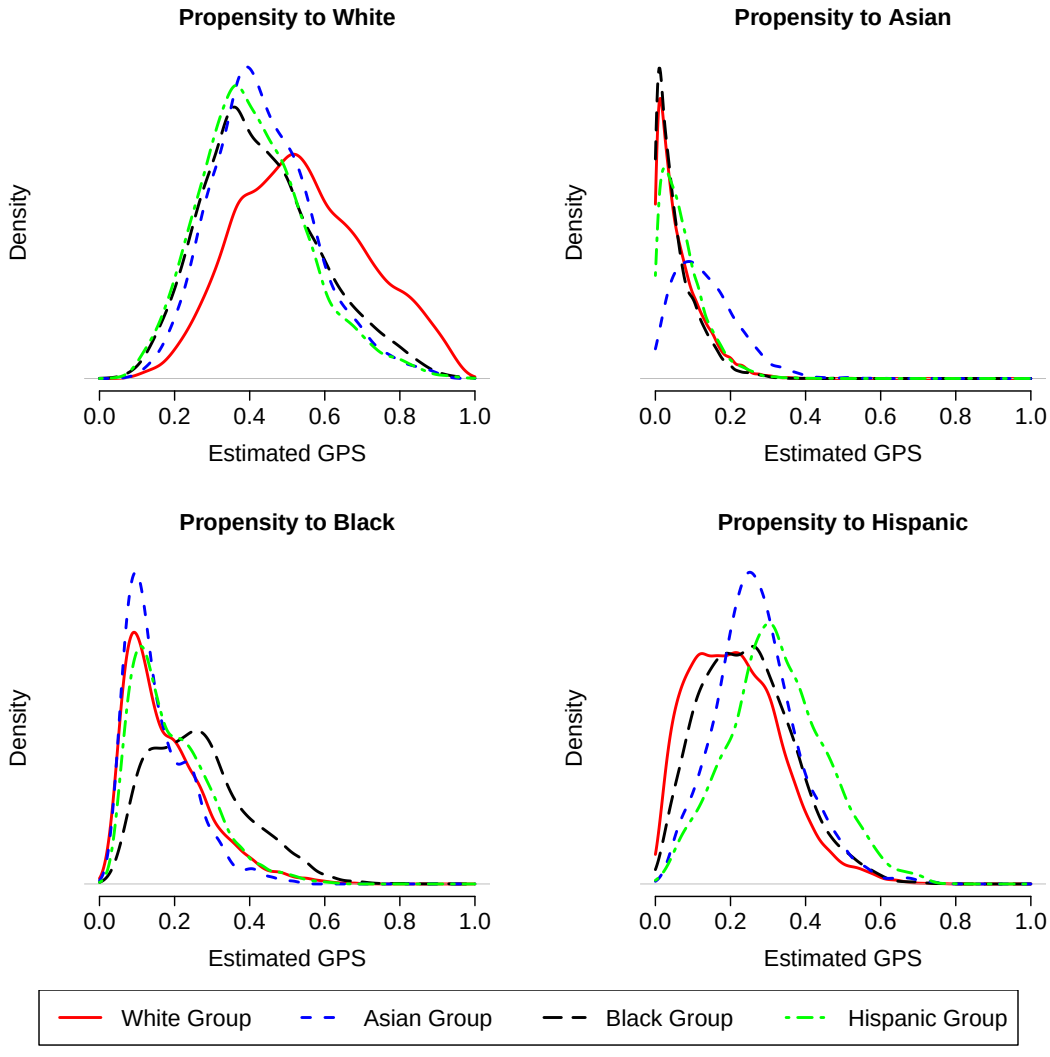
	95% Coverage		
	$\tau_{1,2}$	$\tau_{1,3}$	$\tau_{2,3}$
IPW	0.79	0.88	0.91
IPW (Trim)	0.93	0.90	0.91
Overlap	0.95	0.94	0.94

- Generalized overlap weights – small bias, largest efficiency and nominal coverage – even under lack of overlap

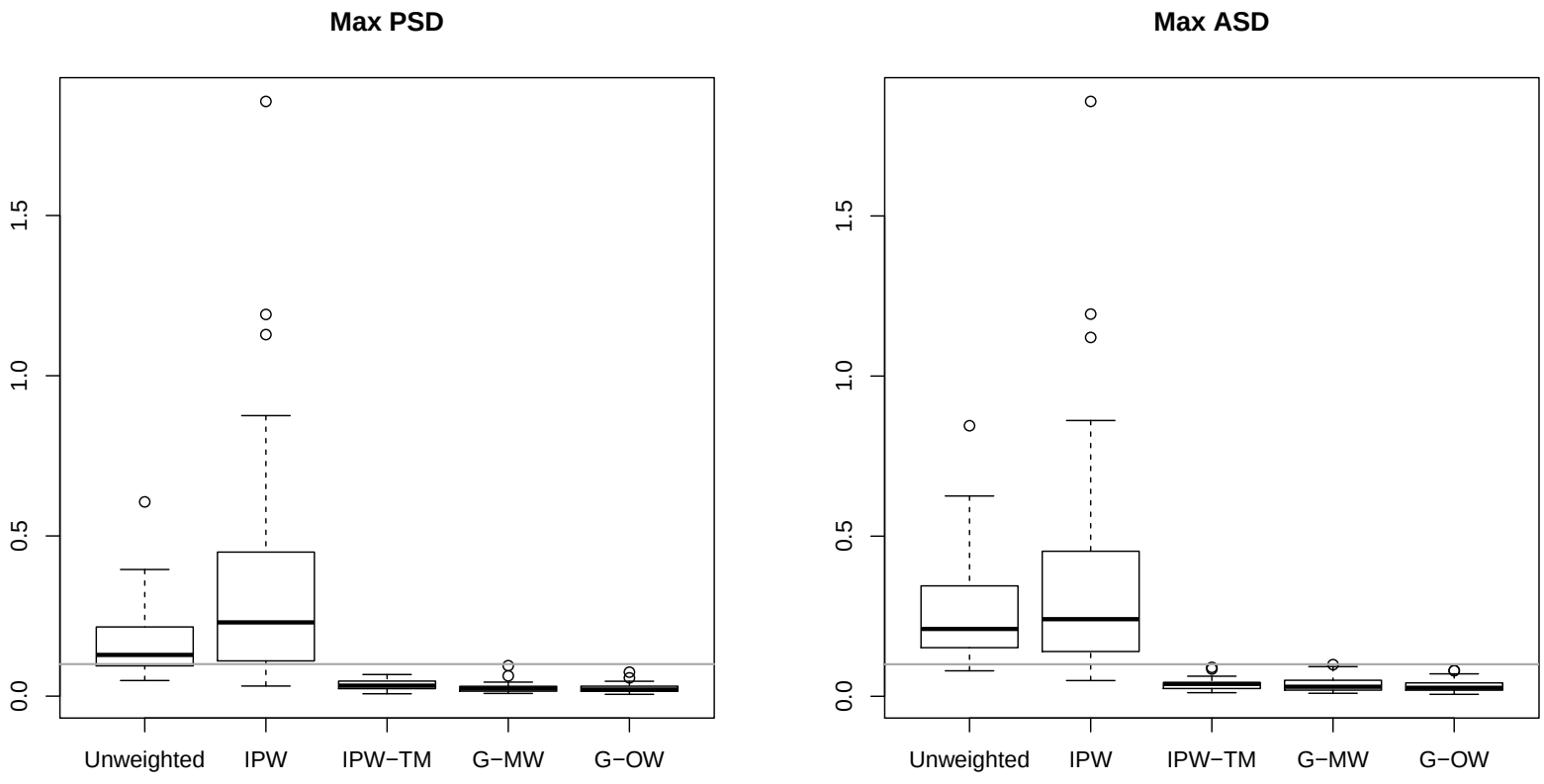
Case Study: Racial Disparity in Medical Expenditure

- Goal: estimate racial disparity in medical expenditures after balancing covariates (Le Cook et al., 2010)
- Race not manipulable so comparisons are descriptive, not causal
- Data: 2009 Medical Expenditure Panel Survey: 9830 non-Hispanic Whites, 4020 Blacks, 1446 Asians, 5150 Hispanics
- Separate binary comparison may result in non-transitivity; interested in simultaneous multiple-group comparisons
- Multinomial logistic regression to estimate GPS, 25 covariates (5 continuous, 20 categorical)
- Ignore survey weights here, but weighting allows easy incorporation of survey weights

Racial Disparity in Medical Expenditure



Racial Disparity in Medical Expenditure



- Generalized overlap weights do not lead to exact balance when $K \geq 3$ under a multinomial logistic GPS model

Racial Disparity in Medical Expenditure

- Estimates (CIs) for difference in yearly medical expenditure (in dollars)

	White-Asian	White-Black	White-Hispanic
IPW	2402 (530, 4274)	908 (505, 1311)	719 (129, 1309)
IPW (Trim)	1335 (671, 1999)	1148 (781, 1515)	1257 (804, 1711)
Overlap	1160 (660, 1661)	886 (518, 1253)	1221 (849, 1593)

- One Asian subject has over 30% of the weight (out of 1446 Asians) – a 78 year old Asian woman with a BMI of 55.4: $e_{\text{Asian}}(X) \approx 0$

- Optimal trimming excludes 2125 Whites, 44 Asians, 1001

Multiple Treatments: R package `PSweight`

- All the above propensity score weighting methods, with balance check, diagnostic plots, closed-form variance, etc. are implemented in the R package `PSweight` (Zhou et al., 2020)
  `https://cran.r-project.org/web/packages/PSweight/index.html`

Nominal Treatments: Matching

- Similar as in binary treatment, one can also use matching and subclassification of GPS, but not as convenient as weighting
- Matching on the full set of p covariates is NOT attractive with multiple treatments when p large (Imai and Van Dyke, 2004; Abadie and Imbens, 2006)
- GPS as a balancing score $\Rightarrow$ match on the vector of GPS (e.g., Lopez and Gutman, 2017)
- Curse of dimensionality even for moderate J moderate to large... But the point of PS is dim reduction, matching seems to get you right back into the hole.

Nominal Treatments: Matching

(Yang et al., 2016, Biometrics)

- Recall that under weak unconfoundedness

$$Y(j) \perp \mathbf{1}\{Z = j\} | e_j(X) \quad \forall\, z \in \mathbb{Z}$$

- Sufficient to estimate $\mathbb{E}[Y(j)]$ by matching on scalar $e_j(X)$
  - estimate GPS via multinomial logistic model
  - for unit i receiving $Z_i = j$, need to impute $Y_i(j')$
  - search from $\{l : Z_l = j'\}$ for a unit j with smallest $||e_{j'}(X_j) - e_{j'}(X_i)||$
  - impute $Y_i(j') = Y_j(j')$, repeat $\forall\, i, j'$
  - average to compute $\hat{\mathbb{E}}[Y(j)]$, $\hat{\mathbb{E}}[Y(j')]$ etc.
- Dimension reduction; closed-form variance also available
- In general, matching is not robust to extreme GPSs, and less efficient v.s. generalized overlap weights

Nominal Treatments: Subclassification

(Huang et al., 2005, Heath Service Research)

Subclassification Procedure

1. For each unit i, calculate the $J-1$ propensity scores, one for each treatment level j
2. For each each category j:
   - 2.1 Balance check: subclass units into quintiles of the propensity scores to that category (e_j) and remove all the subjects that are in non-overlapping region
   - 2.2 Estimate the average outcome $Y(j)$ by combining the estimates in the five subclasses
3. Improve the estimates using shrinkage methods (Stein's estimator) to shrink across treatment levels

Nominal Treatments: Subclassification

(Huang et al., 2005, Heath Service Research)

- The same unit's subclass may be different for different treatment levels j
- When the number of treatment levels increases, the overlap region often quickly diminish
- Subclassification is less prone to extreme GPS
- Advisable to combine with additional outcome regression
- Achy-Brou et al. (Biometrics, 2010) applied a similar method in a longitudinal setting

Case study: PS subclassification in nominal treatments

Huang et al. 2005, Heath Service Research

- Huang et al. (2005): Application of a Propensity Score Approach for Risk Adjustment in Profiling Multiple Physician Groups on Asthma Care. *Health Services Research* 40(1): 253-278.
- Binary outcome, 20 physicians – each physician is a “treatment level"
- Subclassfication of GPS
- Details of overlap and balance check
- Standard errors: Morris’s method (1983, JASA)

Ordinal Treatments

- Ordinal treatments can be viewed as “in-between” nominal and continuous treatments
- Estimands:
  - pairwise comparisons, particularly between adjacent levels
  - dose-response functions
- All the previously discussed GPS-based methods for nominal treatments (including balance check) are applicable to ordinal treatments
- Ignoring the ordered structure leads to loss of information

Ordinal Treatments: Estimate the GPS

- A standard model to estimate the GPS for ordinal treatments is the proportional odds model (i.e. prop odds form of a cumulative logit model): for $j = 1, \cdots, J-1$

$$\text{logit}(\Pr(Z_i < j|X_i) = \log\left(\frac{\Pr(Z_i < j|X_i)}{\Pr(Z_i \geq j|X_i)}\right) = \alpha_j - \beta' X_i \qquad (1)$$

Ordinal Treatments: Balancing Weights using GPS

- The framework of balancing weight directly applies to ordinal treatment
- The key is to specify the estimand by specifying the a vector. Examples:
  - Example 1: one may be interested in the quadratic contrasts between unit increases in the treatment level, namely

$$\tau^h = (m^h_{j+1} - m^h_j) - (m^h_j - m^h_{j-1})$$

  - Example 2: one may be interested in the weighted average of unit increase in the treatment level

$$\tau^h = \sum_{j=1}^{J-1} \xi_j (m^h_{j+1} - m^h_j)$$

  - Example 3: the accumulative effect of the maximum treatment, $\tau^h = m^h_J - m^h_1$

Ordinal Treatments: p-function

- An alternative method to GPS: propensity-function
- Given model (1), differences in outcomes between units with different Z but equal $\beta' X$ is unbiased for causal effects at that $\beta' X$ (Joffe and Rosenbaum, 1999, AJE)
- $\beta' X$ (not the propensity score) is a balancing score – an example of the p-function (Imai and wan Dyk, 2004)
- Balancing $\beta' X$ would balance all X – subclassification, weighting, matching
- Strong unconfoundedness + GPS model assumption $\Rightarrow$ balancing the scalar $\beta' X$ instead of all vectors of GPS

Ordinal Treatments: p-function

- Estimation strategy using the p-function:
  1. Subclass units by the value of $\beta' X$
  2. Estimate the treatment effect within each class
  3. Calculate a weighted average of the subclass-specific treatment effects to estimate ATE or ATT
- Standard errors: bootstrap (Feng at al., 2011); sandwich estimator (McCaffrey et al. 2013). Still an open area
- Alternative strategy: dichotomize the treatment using a cutoff point, but subject to loss of information, sensitive to the choice of cutoff, violation of SUTVA. Not recommended unless for particular reasons

Continuous Treatments: Notation and Estimand

- Units: indexed by $i = 1, \ldots, N$.
- Continuous treatment: $T_i \in \mathcal{T} = [t_0, t_1]$, an interval (Note: I changed from the usual notation of treatment Z to T to emphasize it is continuous)
- Potential outcome: $Y_i(t)$, for $t \in \mathcal{T}$ (implicitly assuming SUTVA)
- Observed outcome: $Y_i(T_i)$
- A vector of pre-treatment covariates: X_i
- Estimand: dose-response function (DRF)

$$\mu(t) = \mathbb{E}[Y_i(t)] = \mathbb{E}_x[Y(t)|X = x], \forall t \in \mathcal{T}$$

Continuous Treatments: Generalize Propensity Score

(Hirano and Imbens, 2004)

- Generalized Propensity Score (GPS): let $r(t, x)$ be the conditional density (not probability) of the treatment given the covariates:

$$r(t, x) = f_{T|X}(t|x),$$

- The GPS for continuous treatment is $R = r(T, X)$, that is, $r(t, x)$ evaluated at the observed treatment T and covariates X
- Balancing property of GPS: $X_i \perp 1\{T_i = t\}|r(t, X_i)$ for all $t \in \mathcal{T}$

Assumptions and Properties of GPS

- SUTVA
- Overlap: $r(t, x) > 0$ for all $t \in \mathcal{T}$ and x (non-zero density)
- Weak unconfoundedness: $Y_i(t) \perp T_i | X_i$ for all $t \in \mathcal{T}$
  - it is weak in the sense that it only requires conditional independence for each potential outcome $Y_i(t)$ rather than joint independence of all potential outcomes.

Bias removal using GPS

(Hirano and Imbens, 2004)

- Theorem: If the assignment mechanism is weakly unconfounded given X, then it is also weakly unconfounded given the GPS $r(t, X)$:

$$Y_i(t) \perp T_i | r(t, X_i), \forall t \in \mathcal{T}$$

- Consequently, we have
  - i $\beta(t, r) \equiv \mathbb{E}[Y(t)|r(t, X) = r] = \mathbb{E}[Y|T = t, R = r]$
  - ii $\mu(t) = \mathbb{E}_X[\beta(t, r(t, X))]$

- Thus one can use the GPS as the only predictor in a regression adjustment

- This property applies to both categorical and continuous treatments, but mostly used in continuous cases

Estimation via Regression

(Hirano and Imbens, 2004)

- To use GPS to estimate DRF $\mu(t) = \mathbb{E}[Y(t)]$, three steps:
  1. Estimate the GPS, check overlap and balance
  2. Estimate $\beta(t, r) = \mathbb{E}[Y|T = t, R = r]$ for each t
  3. Estimate the outcome $\mu(t)$ by averaging $\beta(t, r)$ over X: $\mu(t) = \mathbb{E}[\beta(t, r(t, X))]$
- The key in implementing regression adjustment with GPS is to specify the density function $\beta(t, r) = \mathbb{E}[Y|T = t, R = r]$.

Step 1: Estimate GPS - Balance Check

- Leverage the balancing property of GPS: within strata of GPS, the "probability" of $T = t$ does not depend on the value of X
- One simple approach is to coarsen the continuous treatment into ordinal treatments, and then apply the subclassification type of balance check as in nominal treatments (Hirano and Imbens, 2004) – a bit cumbersome

Step 1: Estimate GPS - Balance Check

- Another approach to check balance (Imai and van Dyk, 2004): compare the balance of each covariate, separately, before and after accounting for the GPS

    1. Run a regression (or logit models for binary covariates) of each covariate X on T with the GPS: $X_i \sim T_i + R(T_i, X_i)$
    2. Test the significance of the coefficient for T in the above regression, if GPS gives good balance, coef of T should be insignificant. For complex models, use a likelihood ratio test to compare the two models $X \sim T$ vs. $X \sim T + R$ (Flores et al. 2012):

Step 2-3: Estimate Dose-Response Function

(Hirano and Imbens, 2004)

- Procedure:
  1. Estimate GPS: e.g. specify $T_i|X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$, and obtain
  $$\hat{R}_i \approx \exp\{-(T_i - \hat{\beta}_0 + \hat{\beta}_1' X_i)^2/(2\hat{\sigma}^2)\}$$
  2. Check Balance to ensure the GPS estimates (previous page)
  3. Specify an outcome model for $\mathbb{E}[Y|T = t, R = r]$, e.g.
  $$\mathbb{E}[Y_i|T_i, R_i] = \alpha_0 + \alpha_1 \cdot T_i + \alpha_2 \cdot T_i^2 + \alpha_3 \cdot R_i + \alpha_4 \cdot R_i^2 + \alpha_5 \cdot T_i \cdot R_i$$
  and fit to the data and estimate the parameters $\hat{\alpha}$
  4. Finally obtain the dose-response function: for each t, average over all of units (plug in X_i and $\hat{\alpha}$):
  $$\widehat{\mathbb{E}[Y(t)]} = \frac{1}{N}\sum_{i=1}^{N} \hat{\alpha}_0 + \hat{\alpha}_1 \cdot t + \hat{\alpha}_2 \cdot t^2 + \hat{\alpha}_3 \cdot \hat{r}(t, X_i) + \hat{\alpha}_4 \cdot \hat{r}(t, X_i)^2 + \hat{\alpha}_5 \cdot t \cdot \hat{r}(t, X_i)$$
- Standard errors: bootstrap (refit GPS and outcome model for each bootstrap sample

Flexible Outcome Models and Software

- Hirano and Imbens (2004): a parametric outcome model of Y with GPS and quadratic term of GPS as the only predictors
- The parametric specification is often not flexible enough, results are sensitive to the specification (Bia et al. 2013)
- Flores et al. (2012) proposed a semiparametric local linear regression for outcome model (Y on T and R) with weighted kernel function (Newey, 1994)
- Bia et al. (2013): provide a STATA package for the kernel methods of Flores et al.

Estimation via Inverse Probability Weighting

- Flores et al. (2005):

$$\mu(t) = \mathbb{E}\left[\frac{r(T_i, X_i|t)Y_i}{\mathbb{E}(r(T_i, X_i|t)|X_i)}\right]$$

- Based on the above, Flores et al. (2012, RESTAT) proposed a GPS-based weighting method for continuous treatment. Plug in sample estimates

- Zhao et al. (2020): using normalized (stablized) weights improves substantially over non-normalized weights

Continuous Treatments: Case Study

- Flores, Flores-Lagunes, Gonzalez, Neumann (2012): Estimating the Effects of Length of Exposure to Instruction in a Training Program: The Case of Job Corps. *Review of Economics and Statistics*. 94(1): 153-171.
- Treatment: Length of exposure to instruction in a job training program
- Outcome: earnings (continuous)

Continuous Treatments: Propensity Function

- Alternative: propensity function (Imai and van Dyk, 2004)
  - Key assumption: for every value of X, there exists a unique parameter $\boldsymbol{\theta}$, such that treatment assignment depends on X only through $\boldsymbol{\theta}$.
  - Example of $\boldsymbol{\theta}$: if we assume $T_i|X_i \sim N(\beta X_i, \sigma^2)$, then βX_i is the unique $\boldsymbol{\theta}$
  - Then one can do subclassification or outcome model based on $\hat{\boldsymbol{\theta}}$ instead of the GPS
  - Flexible outcome model is still the key, e.g. smooth coefficient model($f(\cdot), g(\cdot)$ are smooth functions):

$$E(Y|T, \hat{\boldsymbol{\theta}}) = f(\hat{\boldsymbol{\theta}}) + g(\hat{\boldsymbol{\theta}}) \cdot T$$

- Zhao et al (2020) use simulations to show: all PS-based (GPS or P-function) methods, despite efforts in checking balance, is generally biased in estimating causal DRF

Continuous Treatments: Final Words

- Zhao et al. (2020): researchers should be cautious when using any method to estimate the full DRF with a continuous treatment in an observational study
- Not surprising:
  - Estimating (causal) dose-response function is much tougher than causal effects for binary or multiple treatments: much fewer observations $Y_i(T_i)$ compared to estimand $\{Y(t) : \forall t \in \mathcal{T}\}$, and thus much more reliant on extrapolation
  - Overlap is much more difficult to satisfy: require all units having non-zero probability being assigned to the whole dose range – practically impossible
- What does causal DRF mean? Do we really need a global function? Or just change of local increment is enough?
- My preference: coarsen the continuous to ordinal treatment.

References

Lechner, M. (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. Review of Economics and Statistics. 84, 205-220.

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

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R. and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Statistics in Medicine. 32, 3388-3414.

Feng, P., Zhou, X. H., Zou, Q. M., Fan, M. Y. and Li, X. S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. Statistics in Medicine. 31, 681-697.

Robins, J. M., Hernan, M. A., Brumback, B. (2000). Marginal structural models and causal inference in epidemiology.

Li F, Li F. (2019). Propensity score weighting for causal inference with multiple treatments. Annals of Applied Statistics. 13(4), 2389-2415.

References

Le Cook, B., McGuire, T. G., Lock, K., Zaslavsky, A. M. (2010). Comparing methods of racial and ethnic disparities measurement across different settings of mental health care. Health services research, 45(3), 825-847.

Zhou, T., Tong, G., Li, F., Thomas, L., Li, F (2020). R Package 'PSweight'.

Imai, K., Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. Journal of the American Statistical Association, 99(467), 854-866.

Abadie, A., Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. econometrica, 74(1), 235-267.

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

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E. and Kadziola, Z. (2016). Propensity score matching and subclassification in observational studies with multi-level treatments. Biometrics. 72, 1055-1065.

References

Huang, I. C., Frangakis, C., Dominici, F., Diette, G. B., Wu, A. W. (2005). Application of a propensity score approach for risk adjustment in profiling multiple physician groups on asthma care. Health services research, 40(1), 253-278.

Achy-Brou, A. C., Frangakis, C. E., Griswold, M. (2010). Estimating treatment effects of longitudinal designs using regression models on propensity scores. Biometrics, 66(3), 824-833.

Hansen, M. H., Madow, W. G., Tepping, B. J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. Journal of the American Statistical Association, 78(384), 776-793.

Joffe, M. M., Rosenbaum, P. R. (1999). Invited commentary: propensity scores. American journal of epidemiology, 150(4), 327-333.

Hirano, K., Imbens, G. W. (2004). The propensity score with continuous treatments. Applied Bayesian modeling and causal inference from incomplete-data perspectives, 226-164, 73-84.

Bia, M., Mattei, A. (2008). A Stata package for the estimation of the dose-response function through adjustment for the generalized propensity score. The Stata Journal, 8(3), 354-373.

References

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., Neumann, T. C. (2012). Estimating the effects of length of exposure to instruction in a training program: the case of job corps. Review of Economics and Statistics, 94(1), 153-171.

Newey, W. K. (1994). Kernel estimation of partial means and a general variance estimator. Econometric Theory, 233-253.

Flores, C. A. (2007). Estimation of dose-response functions and optimal doses with a continuous treatment. University of Miami, Department of Economics, November.

Zhao, S., van Dyk, D. A., Imai, K. (2020). Propensity score-based methods for causal inference in observational studies with non-binary treatments. Statistical Methods in Medical Research, 29(3), 709-727.