DOES REGRESSION PRODUCE REPRESENTATIVE CAUSAL RANKINGS?

APOORVA LAL

ABSTRACT. We examine the challenges in ranking multiple treatments based on their estimated effects when using linear regression or its popular double-machine-learning variant, the Partially Linear Model (PLM), in the presence of treatment effect heterogeneity. We demonstrate by example that overlap-weighting performed by linear models like PLM can produce Weighted Average Treatment Effects (WATE) that have rankings that are inconsistent with the rankings of the underlying Average Treatment Effects (ATE). We define this as ranking reversals and derive a necessary and sufficient condition for ranking reversals under the PLM. We conclude with several simulation studies conditions under which ranking reversals occur.

1. Introduction

In both the public and private sector, ranking treatments based on their causal effects is crucial for decision-making. In commercial applications, it is common to rank user actions by estimating their effect on a target metric, and subsequently seeking to encourage actions with large estimated effects, which are deemed 'high value'. An increasingly popular approach is to use Partially Linear Models (PLM) to flexibly condition on a large set of confounders as part of estimating causal effects of treatments while relaxing the stringent form assumptions (Chernozhukov, Chetverikov, et al., 2018). This estimator is rooted in the seminal Frisch-Waugh-Lovell theorem and is extremely popular in practice, and is commonly viewed as *the* Double Machine Learning (DML) estimator by applied users[1].

However, under treatment effect heterogeneity, it is well known from that linear regression performs overlap-weighting (Angrist, 1998). As a result, it is is biased for the Average Treatment Effect (ATE), but instead estimates a conditional-variance Weighted average of treatment effects (WATE). Estimating treatment effects via regression adjustment is numerically simple and may be ideal from an bias-variance tradeoff perspective in many applications. In contrast, when unbiased estimation of the ATE is the goal, practitioners opt for direct

NETFLIX

Date: December 19, 2024.

[1]This is not strictly correct, since DML is in fact a recipe for constructing Neyman-orthogonal estimators for a wide variety of causal and structural parameters. However, due to the prevalence of conditional-ignorability-based identification assumptions and the popularity of linear regression, the PLM has become synonymous with DML. Chernozhukov, Newey, and Singh (2022) study Neyman-orthogonal estimators for a wide variety of causal and structural parameters.

estimation methods such as IPW (Inverse Propensity Weighting) or its Augmented variety (AIPW), or regression imputation / g-modelling.

In many public policy or commercial applications, practioners seek to rank treatment effects instead of merely estimating them, and performance of common estimators for ranking purposes is less well-understood. To this end, we first construct an example with two treatments where the ranking of Weighted Average Treatment Effects (WATEs) produced by the PLM is the opposite of the true ranking of underlying Average treatment Effects (ATEs), which we formalize as a 'ranking reversal' property that is undesirable for downstream decision-making. This implies that decision-makers that seek to rank treatments based on the treatment effects may therefore form incorrect rankings if they use PLM coefficients to form these rankings. We then derive a decomposition relating the WATE and ATE, which gives rise to a necessary and sufficient condition for ranking reversals, and provide economic intuition for it. We find that ranking reversals require substantial treatment effect heterogeneity and covariances between regression weights and treatment effects to be of opposite signs across the treatments being ranked. We conclude with an array of simulation designs that mimic realistic DGPs that comport with our theoretical findings about the likelihood of rank reversals under different heterogeneity patterns.

2. **A Simple Numerical Example**

Consider a binary covariate $X \sim \text{Bernoulli}(0.5)$ and two binary treatments W_1, W_2 with the following propensity scores:

	$W_1 = 0$	$W_2 = 1$
$X = 0$	0.01	0.5
$X = 1$	0.5	0.01

The true treatment effects are:

	τ_1	τ_2
$X = 0$	-3	-2
$X = 1$	3	3
ATE	0	0.5

With linear propensity scores, we can plug the above two sets of numbers into 3.4 to construct PLM regression coefficients

$$\tilde{\tau}_1 = \frac{-3 \cdot 0.01 \cdot 0.99 + 3 \cdot 0.5 \cdot 0.5}{0.01 \cdot 0.99 + 0.5 \cdot 0.5} = 2.7714$$
$$\tilde{\tau}_2 = \frac{-2 \cdot 0.5 \cdot 0.5 + 3 \cdot 0.01 \cdot 0.99}{0.01 \cdot 0.99 + 0.5 \cdot 0.5} = -1.8095$$

The partially linear model recovers the incorrect ranking (treatment 1 $\succ$ treatment 2) when in fact in the DGP, treatment 2 $\succ$ treatment 1. In contrast, IPW or AIPW correctly recovers

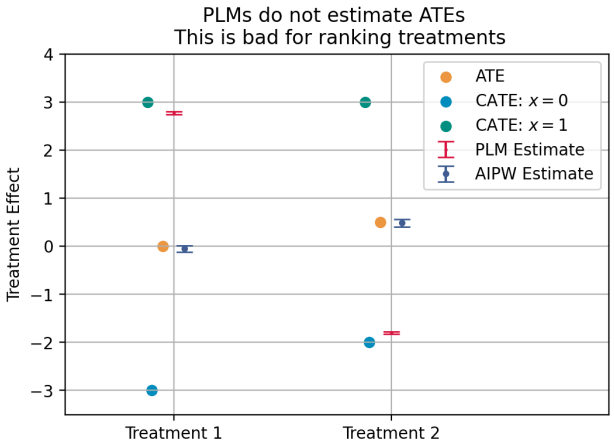

FIGURE 1. Strata-level and overall true effects, and estimated effects from PLM and AIPW

the ATEs and therefore also recovers the correct rankings. We report true and estimated parameters from PLM and AIPW in fig 1. These results demonstrate that PLM leads to incorrect ranking of treatments, while AIPW provides the correct ranking based on ATEs. This is an admittedly contrived example; in the next section, we formalize the properties of this example that yielded the poor ranking performance of PLM.

3. Methodology

We consider a setting with multiple binary treatments where for each unit i, we observe an outcome $Y_i \in \mathbb{R}$, treatment assignment $W_i \in \{1, ..., K\}$ indicating which of K treatments was received (with $W_i = 0$ denoting control), and pre-treatment covariates $\mathbf{X}_i \in \mathbb{R}^d$. Our goal is to rank treatments according to their average treatment effects relative to control, defined as $\tau_j := \mathbb{E}[Y_i(j) - Y_i(0)]$ for each treatment j. We seek to form a poset ordering $(\leq, \boldsymbol{\tau}_j)$, and want to estimate $\boldsymbol{\tau}_j$s using standard techniques under selection-on-observables assumptions [Unconfoundedness and Overlap (Imbens, 2004)].

Defn 3.1 (Partially Linear Model).
For each treatment W_i, the PLM approach models the outcome as:

$$Y_i = \tau W_i + g(\mathbf{X}_i) + \varepsilon_i \tag{3.1}$$

Estimation typically involves a residuals-on-residuals regression:

$$Y_i - \mathbb{E}[Y_i|\mathbf{X}_i] = \widehat{\tau}(W_i - \mathbb{E}[W_i|\mathbf{X}_i]) + \eta_i \tag{3.2}$$

Where the conditional expectations $\mathbb{E}[Y \mid \mathbf{X}] =: \mu(\mathbf{X})$ and $\mathbb{E}[W \mid \mathbf{X}] =: p(\mathbf{X})$ are estimated using flexible non-parametric regression methods and cross-fit to avoid over-fitting to satisfy the technical requirements in Chernozhukov et al (2018).

Theorem 3.2 (Conditional Variance weighting property of linear regression). Under treatment effect heterogeneity, PLM estimates a weighted average treatment effect:

$$\widehat{\tau} = \frac{\mathbb{E}[\omega_i \tau_i]}{\mathbb{E}[\omega_i]}$$

where $\omega_i := (W_i - \mathbb{E}[W_i \mid X_i])^2$ (Angrist, 1998; Angrist and Krueger, 1999; Aronow and Samii, 2016).Defining normalized weights $\gamma_i = \omega_i / \mathbb{E}[\omega_i]$ and working (without loss of generality) with discrete $\mathbf{X}$ lets us rewrite the above as

$$\text{plim } \hat{\tau} = \mathbb{E}[\gamma(\mathbf{X})\tau(\mathbf{X})] =: \text{WATE} \tag{3.3}$$

where $\gamma(\mathbf{X})$ are (normalized) weights that depend on the propensity scores. The weights take the following form

$$\gamma(\mathbf{X}) = \frac{\mathbb{V}[D \mid \mathbf{X}]}{\mathbb{E}[\mathbb{V}[D \mid \mathbf{X}]]} = \frac{p(\mathbf{X})(1 - p(\mathbf{X}))}{\mathbb{E}[p(\mathbf{X})(1 - p(\mathbf{X}))]} \tag{3.4}$$

where the second equality uses the fact that each treatment is binary and substitutes in the expression for binomial variance. Proof in A.1.

This means that in the presence of treatment effect heterogeneity (i.e. $\tau(\mathbf{X})$ is not a constant function $= \tau$), the probability limit of the regression coefficient is no longer the Average Treatment Effect (ATE $:= \mathbb{E}[\tau(\mathbf{X})]$) but is instead the above Weighted Average Treatment Effect (WATE), with weights γ implicitly chosen by the regression specification. These weights are largest for propensity scores close to 0.5, which results in OLS performing 'overlap-weighting' where it down-weights strata with extreme propensity scores, and discards strata with no overlap (with propensity scores equal to 0 or 1).

An interesting alternative but complementary decomposition is studied by Słoczyński (2022), who shows that the regression coefficient $\hat{\tau}$ can also be decomposed into the ATT (Average Treatment Effect on the Treated) and ATU (Average Treatment Effect on the Untreated), with weights that are inversely proportional to group sizes. In other words, the larger the share of the treated group, the lower weight it receives, and vice versa.

3.1. **Rank Reversal: definition and conditions.**

With these weights in hand, we can define the property observed in the previous section.

Defn 3.3 (Rank Reversal).
For any two treatments j and k, a ranking reversal implies that we have $\text{ATE}_j > \text{ATE}_k$ but $\text{WATE}_j < \text{WATE}_k$. This occurs when

$$\overbrace{\mathbb{E}\left[\tau_j(\mathbf{X})\right]}^{\text{ATE}_j} > \overbrace{\mathbb{E}\left[\tau_k(\mathbf{X})\right]}^{\text{ATE}_k} \tag{3.5}$$

$$\underbrace{\mathbb{E}\left[\gamma_j(\mathbf{X})\tau_j(\mathbf{X})\right]}_{\text{WATE}_j} < \underbrace{\mathbb{E}\left[\gamma_k(\mathbf{X})\tau_k(\mathbf{X})\right]}_{\text{WATE}_k} \tag{3.6}$$

We first derive an expression relating the ATE and WATE. For any treatment g, we can decompose the WATE using the definition of covariance ($\text{Cov}\left[a, b\right] = \mathbb{E}\left[ab\right] - \mathbb{E}\left[a\right]\mathbb{E}\left[b\right]$)

$$\mathbb{E}[\gamma_g(\mathbf{X})\tau_g(\mathbf{X})] = \mathbb{E}[\gamma_g(\mathbf{X})]\mathbb{E}[\tau_g(\mathbf{X})] + \text{Cov}(\tau_g(\mathbf{X}), \gamma_g(\mathbf{X}))$$

Note that by construction of regression weights $\gamma_g(\mathbf{X}) := \mathbb{V}\left[W \mid \mathbf{X}\right]/\mathbb{E}\left[\mathbb{V}\left[W \mid \mathbf{X}\right]\right]$ have an expected value of 1. So, we arrive at the following decomposition

$$\underbrace{\mathbb{E}[\gamma_g(\mathbf{X})\tau_g(\mathbf{X})]}_{\text{WATE}_g} = \underbrace{\mathbb{E}[\tau_g(\mathbf{X})]}_{\text{ATE}_g} + \text{Cov}(\tau_g(\mathbf{X}), \gamma_g(\mathbf{X})) \tag{3.7}$$

We provide three simple examples numerically illustrating the above decomposition with negative, zero, and positive covariance between the regression weights and treatment functions in figure 2.

This decomposition immediately illustrates how rank reversals may arise in practice: when the second term in 3.7 is large enough to offset the first, rank-reversals may occur.

Proposition 3.4 (Necessary and Sufficient Condition for Rank Reversal).
The following condition yields rank-reversal between treatments j and k

$$\mathbb{E}\left[\tau_j(\mathbf{X}\right] + \text{Cov}\left[\tau_j(\mathbf{X}, \gamma_j(\mathbf{X})\right] < \mathbb{E}\left[\tau_k(\mathbf{X})\right] + \text{Cov}\left[\tau_k(X), \gamma_k(\mathbf{X})\right] \tag{3.8}$$

This is an immediate implication of the decomposition 3.7. Proof in A.1. We also provide slightly more transparent sufficient conditions that parametrises the magnitudes of the two covariances in 3.8 in appdx A.2.

When can we expect PLM coefficients to yield correct rankings?

(1) Constant treatment effects ($\tau(\mathbf{X}) = \tau$): Here, PLM, IPW, and AIPW all estimate the same quantity. This is rare in practice but serves as a useful benchmark.

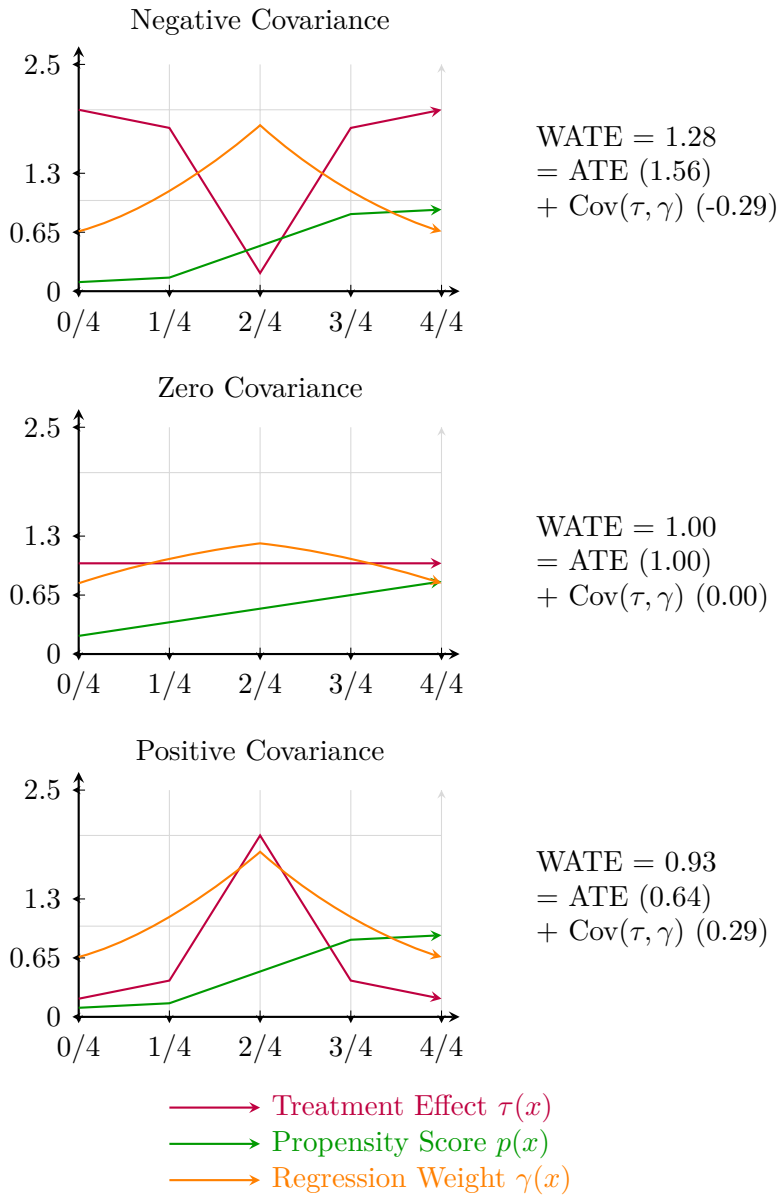

FIGURE 2. Treatment effect heterogeneity and regression weights under negative, zero, and positive scenarios for the $\text{Cov}\left[\tau_g(\mathbf{X}), \gamma_g(\mathbf{X}\right]$ term in 3.7. We have a single covariate X with 5 discrete strata with equal probability, and vary propensity scores and treatment effects according to the green and red functions specified above, which gives rise to the orange regression weights function. The right panel for each scenario shows how the weighted average treatment effect (WATE) estimated using regression decomposed into the true average treatment effect (ATE) and the covariance between treatment effects and regression weights.

(2) Uncorrelated weights and effects ($\text{Cov}\left[\gamma(\mathbf{X}), \tau(\mathbf{X})\right] \approx 0$): This can happen when:
   (a) Treatment assignment is relatively balanced ($p(\mathbf{X}) \approx 0.5$)
   (b) Treatment effects vary independently of variables that predict treatment

   (c) As-good-as-random assignment: if units don't have the opportunity to sort into treatment based on private information about their own treatment effects $\tau(\mathbf{X})$, this covariance will be more likely to be small.
(3) Uniform selection on gains: If units sort into treatments j and k based on private information about their expected gains $\tau_j(\mathbf{x}), \tau_k(\mathbf{x})$, the covariance $\text{Cov}\left[\gamma_g(\mathbf{X}), \tau_g(\mathbf{X})\right]$ will be of the same sign for $g \in \{j, k\}$, which would not flip the rankings between the ATEs.
(4) Similar propensity score distributions: When $p_j(\mathbf{X})$ and $p_k(\mathbf{X})$ have similar distributions, $\gamma_j(\mathbf{X})$ and $\gamma_k(\mathbf{X})$ will be similar, reducing the chance of rank reversals. This suggests observational studies with very different propensity scores across treatments are more prone to rank reversals
(5) Moderate treatment effect heterogeneity: If heterogeneity in treatment effects is modest, and this is known to agents, it is less likely that they actively seek or avoid treatments (which pushes $p_g(\mathbf{X})$ towards 0 or 1) based on this information, which weakens the magnitude of $\text{Cov}\left[\tau(\cdot), \gamma(\cdot)\right]$, which in turn makes it less likely that the covariances for different treatments are of contrasting signs to result in rank reversals.

A practical implication of the above is that when treatment effects are suspected to be highly heterogeneous with units selecting into treatments, researchers should prefer AIPW over PLM for ranking.

Defn 3.5 (Augmented Inverse-Propensity Weighting (AIPW) Estimators).
An alternative to the PLM that does not fall prey to the ranking reversal property is the AIPW estimator, which involves construction of a 'pseudo-outcome' Γ_i^j that is the estimated potential outcome under treatment j (Cattaneo, 2010; Chernozhukov, Chetverikov, et al., 2018)

$$\widehat{\Gamma}_i^j = \widehat{\mu}^{j,-k}(\mathbf{X}_i) + \frac{\mathbb{1}\{W_i = j\}}{\widehat{p}^{j,-k_i}(\mathbf{X}_i)} \left(Y_i - \widehat{\mu}^{j,-k_i}(\mathbf{X}_i)\right)$$

$$\widehat{\tau}^{\text{AIPW},a,b} = \frac{1}{n} \sum_i^n \left(\widehat{\Gamma}_i^a - \widehat{\Gamma}_i^b\right)$$

where we first partition data by assigning each observation into $k_i \in \mathsf{U}\left[K\right]$ folds, and cross-fit nuisance functions $\widehat{\mu}(\cdot)$ (an outcome regression within treatment level j) and $\widehat{p}$ (a multi-class propensity score that models the probability of treatment level j) so that their predictions for unit i are produced from models that were not trained on the k_i-th fold. The above estimator is consistent for the ATE regardless of the level of heterogeneity in the underlying treatment effect function $\tau(\mathbf{X})$, which implies that it does not exhibit rank-reversal properties, but conversely may have poor empirical performance in the presence of extreme propensity scores.

4. Numerical Experiments

4.1. Simulation Design. We conduct Monte Carlo simulations to evaluate the performance of PLM and AIPW estimators under various data generating processes (DGPs). Each DGP is characterized by:

- A binary covariate $X \sim \text{Bernoulli}(0.5)$
- Two binary treatments W_1, W_2 with stratum-specific propensity scores $p_j(X)$
- Heterogeneous treatment effects $\tau_j(X)$ for each treatment

We consider five scenarios that vary in their degree of effect heterogeneity and propensity score distributions:

(1) **Extreme Heterogeneity**: Large differences in treatment effects across strata with extreme propensity scores
(2) **Constant Effects**: Homogeneous effects within treatments but different across treatments
(3) **Uncorrelated**: Moderate heterogeneity with balanced propensity scores
(4) **Selection on Gains**: Treatment probability correlated with treatment effects
(5) **Balanced**: Equal propensity scores across strata with heterogeneous effects

For each scenario, we simulate 1,000 datasets with 10,000 observations each. We evaluate the estimators on three dimensions:

- Distribution of point estimates
- Bias relative to true effects
- Proportion of correct rankings between treatments

We report figures for each of these settings in appendix A.3. We find that with the exception of the extreme heterogeneity setting that expands upon the example in section 2 (fig 7), the rankings produced by the PLM are largely consistent with the AIPW estimator, and conform with the sufficient conditions derived in the previous section.

5. Conclusion

This note highlights the importance of using appropriate methods for estimating and ranking treatment effects in the presence of heterogeneity. We show using an example that commonly used Partially Linear Models can lead to biased estimates and incorrect rankings. We then define a notion of ranking reversals and derive a decomposition relating the WATE and ATE, which gives rise to a necessary and sufficient condition for ranking reversals in linear regression. Finally, we propose interpretations for these conditions and recommend using Augmented Inverse Probability Weighting estimator as a general solution for ranking in the presence of substantial heterogeneity.

Our findings have important implications for decision-making in various fields, including digital platforms and policy evaluation, where accurate ranking of treatments is crucial. Future work could explore the performance of these methods in more complex settings with multiple treatments and high-dimensional covariates.

References

Angrist, Joshua D (1998). "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants". In: *Econometrica: journal of the Econometric Society* 66 (2), pp. 249–288 (cit. on pp. 1, 4).

Angrist, Joshua D and Alan B Krueger (Jan. 1, 1999). "Chapter 23 - Empirical Strategies in Labor Economics". In: *Handbook of Labor Economics*. Ed. by Orley C Ashenfelter and David Card. Vol. 3. Elsevier, pp. 1277–1366 (cit. on p. 4).

Aronow, Peter M and Cyrus Samii (Jan. 28, 2016). "Does Regression Produce Representative Estimates of Causal Effects?" In: *American journal of political science* 60 (1), pp. 250–267 (cit. on p. 4).

Cattaneo, Matias D (Apr. 1, 2010). "Efficient semiparametric estimation of multi-valued treatment effects under ignorability". In: *Journal of econometrics* 155 (2), pp. 138–154 (cit. on p. 7).

Chernozhukov, Victor, Denis Chetverikov, et al. (Feb. 16, 2018). "Double/debiased machine learning for treatment and structural parameters". In: *The econometrics journal* 21 (1), pp. C1–C68 (cit. on pp. 1, 7).

Chernozhukov, Victor, Whitney K Newey, and Rahul Singh (May 1, 2022). "Automatic Debiased Machine Learning of Causal and Structural Effects". In: *Econometrica: journal of the Econometric Society* 90 (3). https://doi.org/10.3982/ECTA18515, pp. 967–1027 (cit. on p. 1).

Imbens, Guido W (Feb. 1, 2004). "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review". In: *The review of economics and statistics* 86 (1). doi: 10.1162/003465304323023651, pp. 4–29 (cit. on p. 3).

Słoczyński, Tymon (2022). "Interpreting OLS Estimands when Treatment Effects are Heterogeneous". In: *Review of Economics and Statistics* (cit. on p. 4).

Appendix A. Proofs

A.1. Conditional Variance Weighting.

We observe $(Y_i, W_i, \mathbf{X}_i)_{i=1}^N \in \mathbb{R} \times \{0,1\} \times \mathbb{R}^d$. We project the covariate vector X_i into some basis Φ, which approximates the flexible function $g(\mathbf{X})$.

(1) Unconfoundedness: $Y_i^0, Y_i^1 \perp\!\!\!\perp W_i \mid X_i$
(2) Linearity of propensitye score $\mathbb{E}[W_i \mid X_i] = \phi_i'\psi$

Define $Z_i = (1 : W_i : \phi_i)$. We run the following regression

$$Y_i \sim Z_i = \alpha + \tau W_i + \underbrace{\phi_i'\zeta}_{g(x)} + \varepsilon_i$$

By FWL, we can write the coefficient $\hat{\tau}$ as

$$\widetilde{W}_i = W_i - \phi_i'\psi, \ \widehat{\psi} = (\phi'\phi)^{-1}\phi' w$$

$$\begin{aligned}
\hat{\tau} &= \frac{\sum_i \widetilde{W}_i Y_i}{\sum_i \widetilde{W}_i^2} \\
&= \frac{\sum_i \widetilde{W}_i \left(Y_i^0 + \tau_i W_i\right)}{\sum_i \widetilde{W}_i^2} = \frac{\sum_i \widetilde{W}_i Y_i^0}{\sum_i \widetilde{W}_i^2} + \frac{\sum_i \widetilde{W}_i \tau_i W_i}{\sum_i \widetilde{W}_i^2} \\
&= \underbrace{\frac{\sum_i \widetilde{W}_i Y_i^0}{\sum_i \widetilde{W}_i^2}}_{\to 0 \text{ by A1}} + \underbrace{\frac{\sum_i \widetilde{W}_i \tau_i \phi_i'\psi}{\sum_i \widetilde{W}_i}}_{\to 0 \text{ by orthogonality bw } \widetilde{W}_i \text{ and } \phi_i'\psi} + \frac{\sum_i \widetilde{W}_i^2 \tau_i}{\sum_i \widetilde{W}_i^2} \qquad \text{Expand out } W_i = \phi_i'\psi + \widetilde{W}_i \\
&= \frac{\sum_i \widetilde{W}_i^2 \tau_i}{\sum_i \widetilde{W}_i^2} = \frac{\sum_i (W_i - \phi_i'\psi)^2 \tau_i}{\sum_i (W_i - \phi_i'\psi)^2}
\end{aligned}$$

Proof [Proof of necessity and sufficiency of 3.8 for rank reversal]

We need it to be the case that 3.8, combined with the definitional assumption that $\text{ATE}_j > \text{ATE}_k$ ($\mathbb{E}\left[[]\, \tau_j(\mathbf{X})\right] > \mathbb{E}\left[\tau_k(\mathbf{X})\right]$) $\Leftrightarrow$ rank reversal $\text{WATE}_j < \text{WATE}_k$.

$\Leftarrow$ Using the decomposition 3.7, we note that LHS of 3.8 is equal to WATE_j, and the right hand side is WATE_k, so this is rank reversal by definition.

$\Rightarrow$ By the same token, since the LHS and RHS of 3.8 are the definition of WATE_j and WATE_k respectively by 3.7, this immediately implies the conclusion.

□

A.2. **Interpretable Sufficient Conditions.**

(1) $\text{Cov}(\tau_j(\mathbf{X}), \gamma_j(\mathbf{X})) < -\delta$ for some $\delta > 0$
(2) $\text{Cov}(\tau_k(\mathbf{X}), \gamma_k(\mathbf{X})) > \delta$
(3) The difference in ATEs is smaller than the combined covariance in effects: $\mathbb{E}[\tau_j(\mathbf{X})] - \mathbb{E}[\tau_k(\mathbf{X})] < 2\delta$

We proceed by showing that conditions (1)-(3) together imply rank reversal as defined in Definition 3.3. We need to show that for treatment effect functions $\tau_j(\mathbf{X}), \tau_k(\mathbf{X})$ that satisfy 3.5 and conditions (1-3), 3.6 holds.

Next, use this definition for j and k and plug in conditions (1) and (2)

$$\begin{aligned}
\mathbb{E}[\gamma_j(\mathbf{X})\tau_j(\mathbf{X})] &= \mathbb{E}[\tau_j(\mathbf{X})] + \text{Cov}(\tau_j(\mathbf{X}), \gamma_j(\mathbf{X})) \\
&< \mathbb{E}[\tau_j(\mathbf{X})] - \delta && \text{condition (1)} \\
\mathbb{E}[\gamma_k(\mathbf{X})\tau_k(\mathbf{X})] &= \mathbb{E}[\tau_k(\mathbf{X})] + \text{Cov}(\tau_k(\mathbf{X}), \gamma_k(\mathbf{X})) \\
&> \mathbb{E}[\tau_k(\mathbf{X})] + \delta && \text{condition (2)}
\end{aligned}$$

From condition (3): $\mathbb{E}[\tau_j(\mathbf{X})] - \mathbb{E}[\tau_k(\mathbf{X})] < 2\delta$. Therefore:

$$\begin{aligned}
\mathbb{E}[\gamma_j(\mathbf{X})\tau_j(\mathbf{X})] &< \mathbb{E}[\tau_j(\mathbf{X})] - \delta \\
&< \mathbb{E}[\tau_k(\mathbf{X})] + \delta && \text{plug in cond (3)} \\
&< \mathbb{E}[\gamma_k(\mathbf{X})\tau_k(\mathbf{X})] && \square
\end{aligned}$$

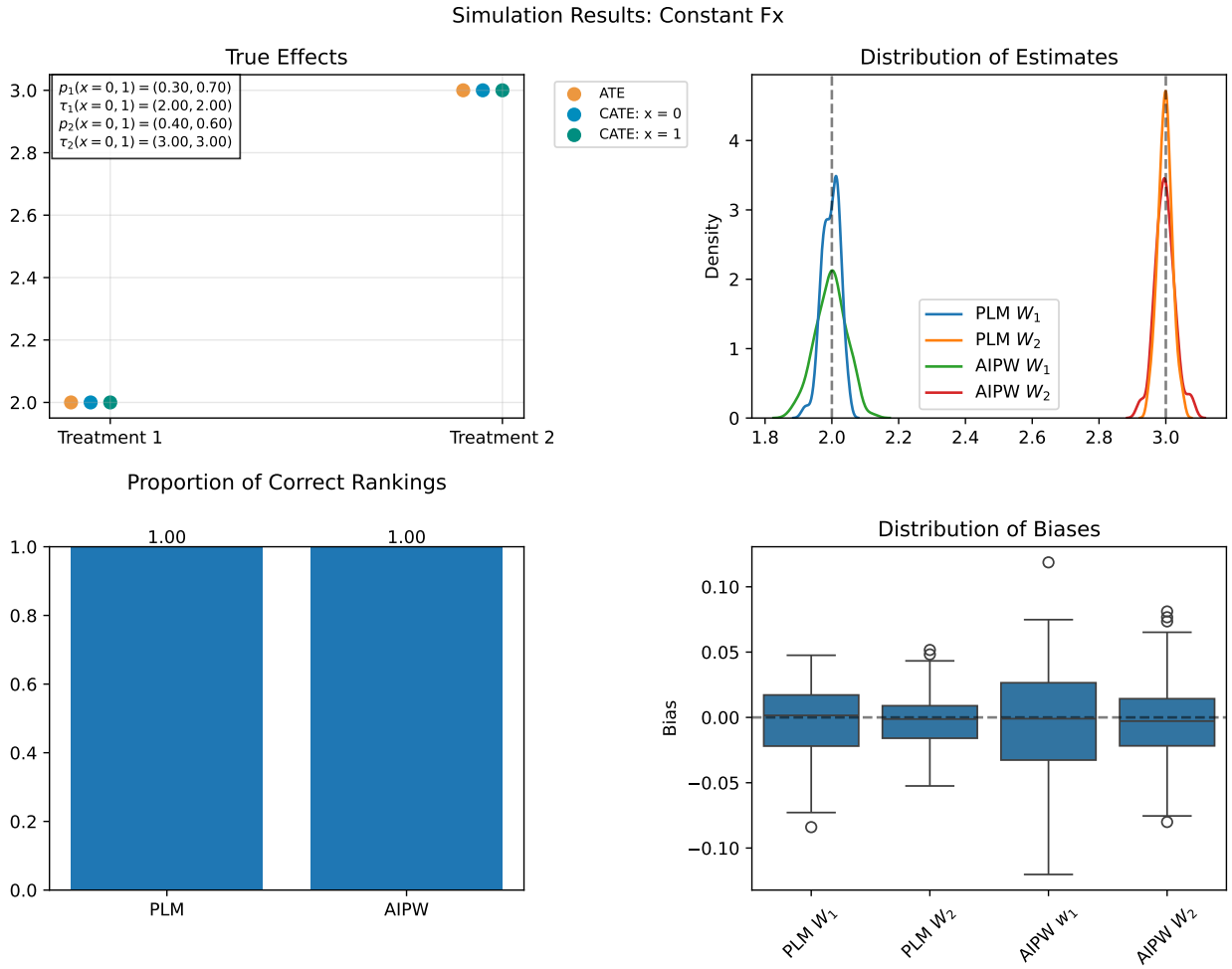

FIGURE 3. Results for Constant Effects

A.3. **Simulation Study Results.**

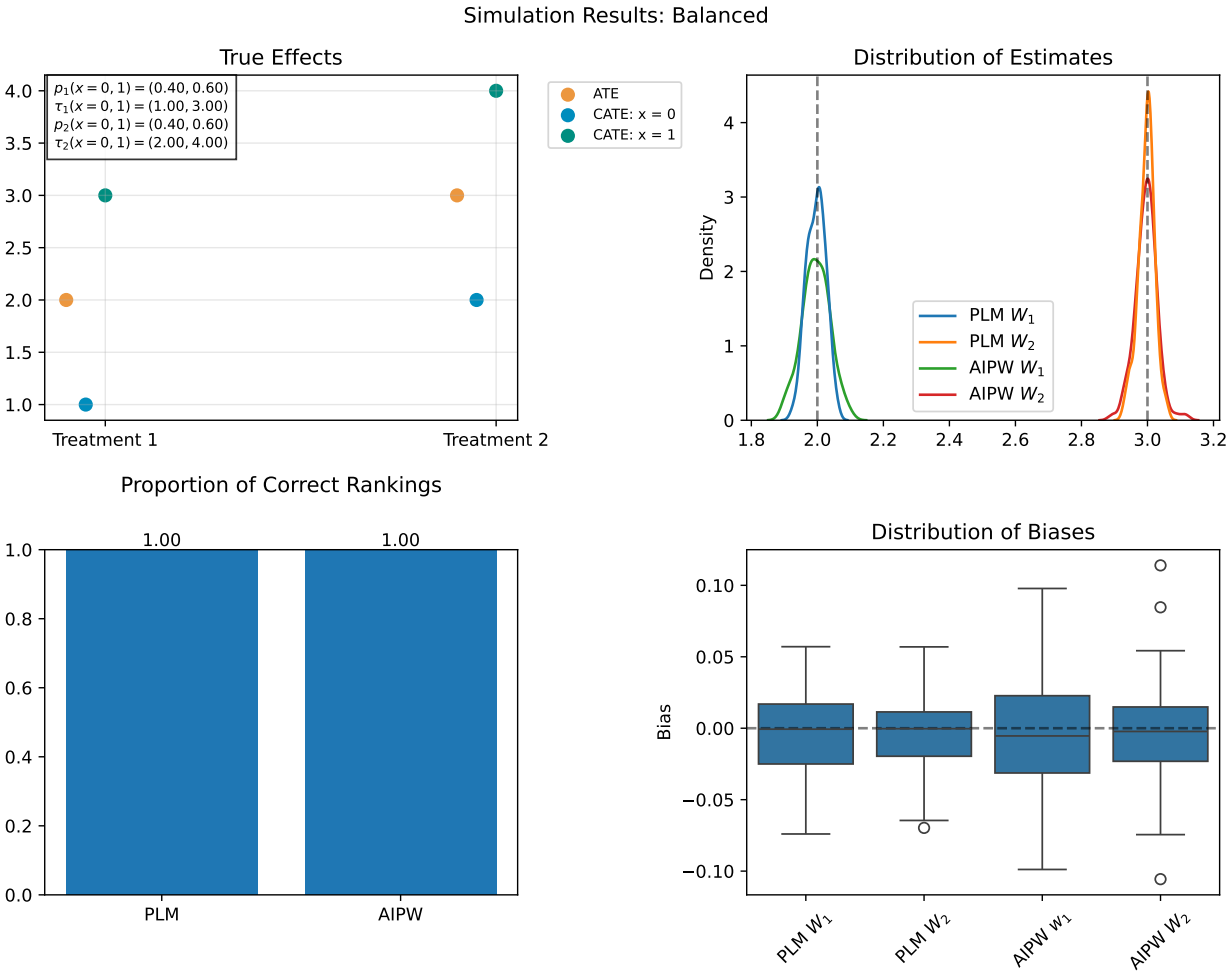

FIGURE 4. results for balanced assignment

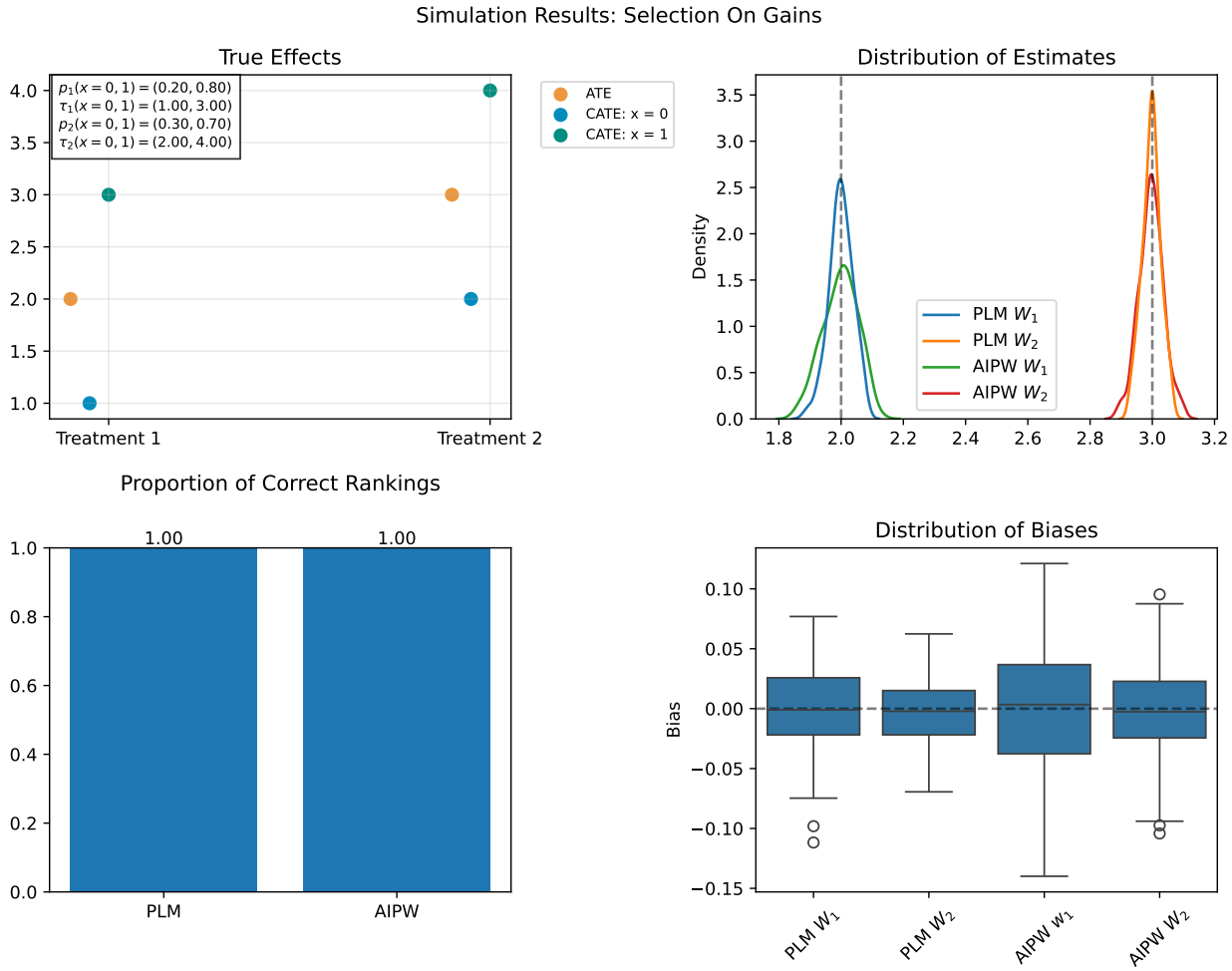

FIGURE 5. Results for Selection on Gains

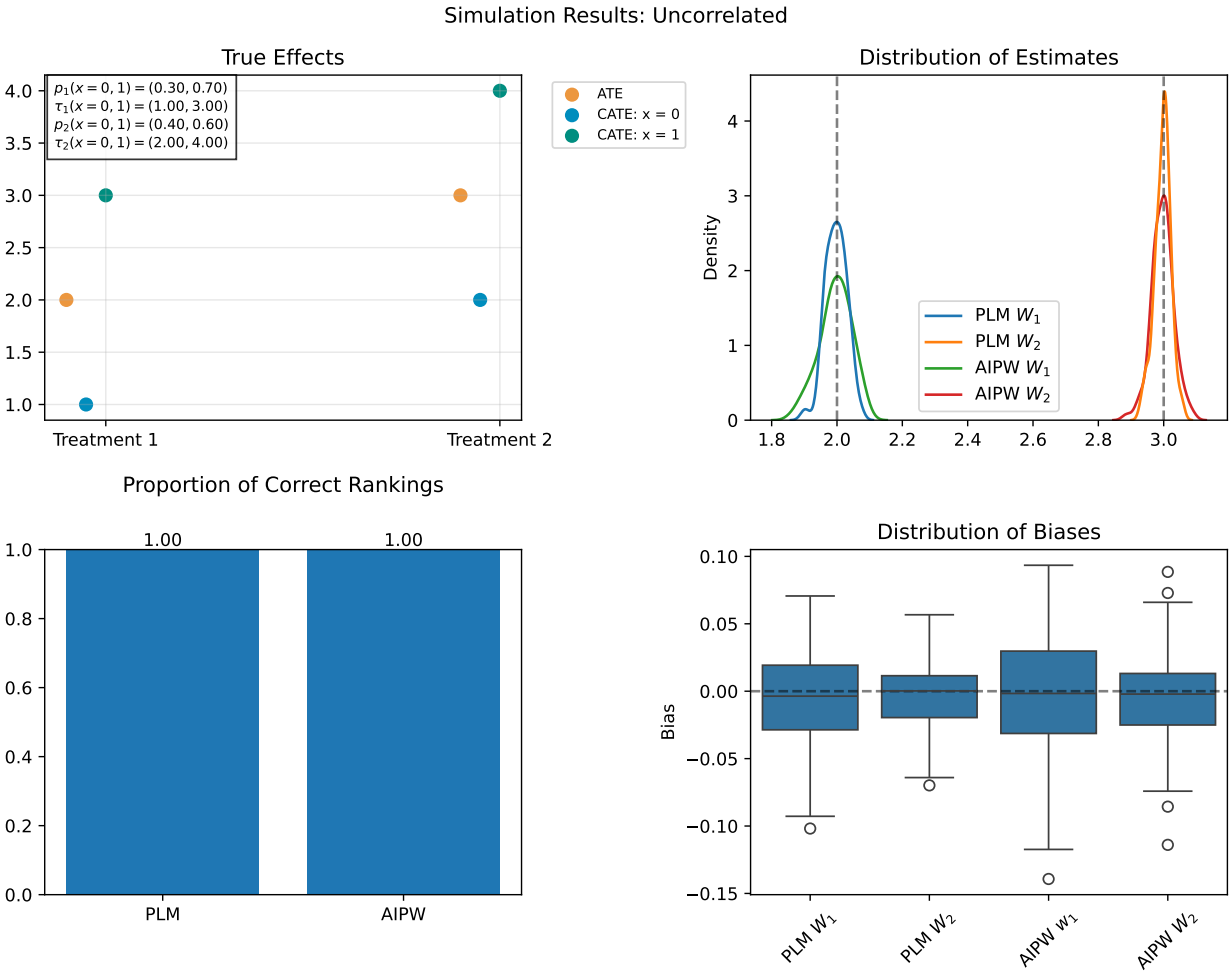

FIGURE 6. Results for Uncorrelated propensity and treatment effects

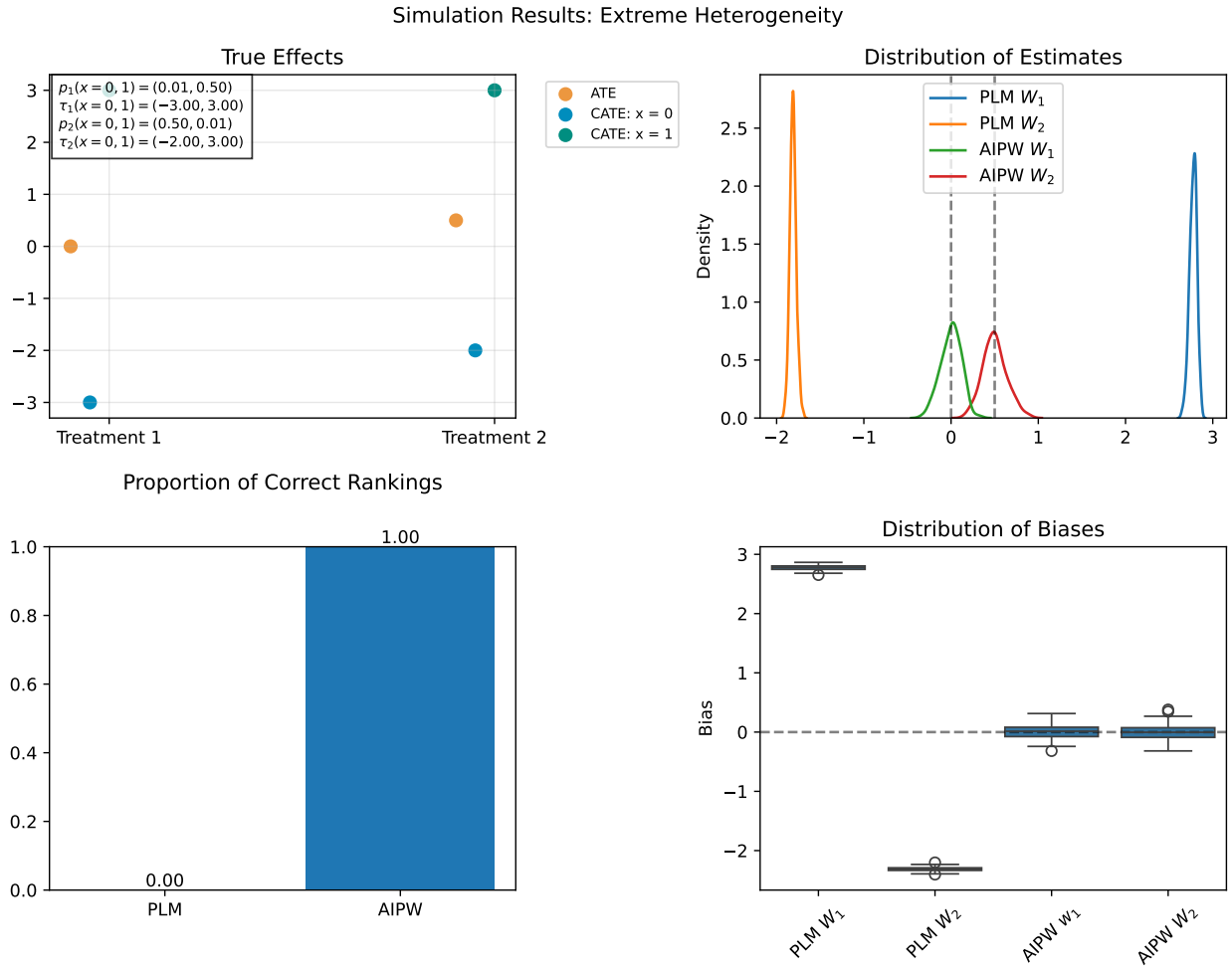

FIGURE 7. Results for Extreme heterogeneity