Cluster-Specific Propensity Score Weighting To Stabilize Treatment Effect Estimation

Hardev Ranglani[1]

EXL Service Inc.

This paper introduces a novel method for estimating the Average Treatment Effect (ATE) in observational causal inference studies by calculating cluster-specific ATEs and taking a weighted average across clusters. Instead of directly applying Inverse Propensity Weighting (IPW), this approach leverages clustering to address issues such as positivity violations and extreme weights, which often arise when propensity scores are near 0 or 1. Each cluster is formed based on covariates, and the ATE is estimated for each cluster using propensity score weighting. The overall ATE is then calculated as a weighted average of the cluster-specific ATEs, where the weights are based on the average propensity score within each cluster. This method effectively captures treatment effect heterogeneity and mitigates the instability caused by extreme individual weights. Simulations on synthetic data and real-world datasets demonstrate the superiority of this method in producing more stable and reliable treatment effect estimates compared to traditional IPW.

Keywords: Inverse Propensity Weighting, Clustering, Average treatment effects

1 Introduction

Causal inference is the process of determining the causal effect of a treatment, policy, or intervention on an outcome of interest. Unlike traditional correlational analyses, causal inference seeks to establish cause-and-effect relationships by adjusting for confounding variables that might influence both the treatment and the outcome. This is critical in observational studies where randomized control trials (RCTs)—the gold standard for establishing causality—are not possible.

Inverse Propensity Weighting (IPW) is a widely used method in causal inference for estimating treatment effects in observational studies, where randomization is not possible. In such settings, individuals self-select into treatment or control groups, and confounding variables may influence both the treatment assignment and the outcome. IPW helps to address this bias by weighting individuals according to the probability of receiving treatment, known as the propensity score. The goal of IPW is to create a

DOI:10.5121/mlaij.2025.12103

pseudo-population where the distribution of covariates is balanced between treated and untreated individuals, making the treatment assignment independent of the covariates. For example, in an observational study comparing two drug treatments for heart disease, where researchers want to estimate the effect of a new drug or treatment on patient outcomes, patients may not be randomly assigned to treatments and thus the overall results of the study may not be reliable. IPW can adjust for differences in patient characteristics (such as age, severity of illness, or prior health conditions, etc.) by reweighting the data, ensuring that the treatment groups are comparable, and allowing for an unbiased estimate of the treatment effect.

The propensity score is defined as the probability of receiving the treatment(such as a new drug) given the observed covariates, often estimated using logistic regression. The propensity score $e(X)$ for an individual with covariates X, where the treatment is denoted by T is

$$e(X) = P(T = 1|X)$$

.

In many applications of IPW, logistic regression is employed to estimate the propensity score. Logistic regression models the relationship between the covariates X and the binary treatment T, and it provides a flexible and interpretable way to calculate the probability of treatment assignment. The estimated propensity scores from the logistic regression model can then be used to compute weights for each individual in the dataset.

The weights are computed as follows:

$$w_i = \begin{cases} \frac{1}{e(X_i)} & \text{if } T_i = 1 \\ \frac{1}{1-e(X_i)} & \text{if } T_i = 0 \end{cases}$$

Thus:

- Treated individuals $(T = 1)$ are weighted by $\frac{1}{e(X_i)}$, which gives more weight to those who were unlikely to be treated.
- Control individuals $(T = 0)$ are weighted by $\frac{1}{1-e(X_i)}$ which gives more weight to those who were unlikely to be in the control group.

The ATE is then estimated as using the weighted outcome means as :

$$\widehat{ATE}_{IPW} = \frac{1}{n} \sum_{i=1}^{n} w_i \left(T_i Y_i - (1 - T_i) Y_i\right)$$

where :

- T_i is the treatment indicator(1 if treated, 0 if control)
- Y_i is the observed outcome.
- $e(X_i)$ is the estimated propensity score for individual i.

While IPW thus reweights observations based on their propensity scores, (i.e. —the probability of receiving treatment given a set of covariates), thereby mimicing the conditions of a randomized trial. However, in practical applications, traditional IPW often suffers from positivity violations and high variance in the weights. Positivity violations occur when certain subgroups are almost always treated or almost never treated, leading to extreme propensity scores near 0 or 1, which result in excessively large weights and unstable Average Treatment Effect (ATE) estimates [2][3].

Several methods have been proposed to address these issues, such as trimming extreme weights [4] or using machine learning methods to estimate propensity scores more flexibly [5]. While these approaches are promising, they either discard observations or fail to fully stabilize the variance in estimates, especially in datasets where treatment groups are small or covariate distributions are highly skewed.

To address these challenges, we propose a new method that builds upon traditional IPW by addressing these limitations through cluster-specific ATE estimation. Instead of applying IPW directly to the entire population, we first cluster individuals based on their covariates using a suitable clustering algorithm (e.g., k-means or hierarchical clustering). Within each cluster, the ATE is calculated by comparing the weighted outcomes between treated and untreated individuals in the traditional IPW manner. The overall ATE is then obtained by taking a weighted average of the cluster-specific ATEs, where the weights are derived from the average propensity score within each cluster. This stabilizes extreme weights, mitigates positivity violations, and reduces variance.

This method has several advantages:

- Captures treatment effect heterogeneity: By estimating ATEs within clusters, this approach captures variations in treatment effects across subgroups with similar covariate profiles.
- Mitigates extreme weights: Using average propensity scores at the cluster level reduces the influence of extreme individual weights, leading to more stable ATE estimates.
- Improved stability: The weighting at the cluster level mitigates the impact of extreme weights caused by positivity violations, where some individuals have very low or very high likelihoods of receiving treatment.

By aggregating the cluster-specific ATEs into an overall ATE, we not only stabilize the estimates but also capture treatment effect heterogeneity across different subgroups. This technique is especially useful in datasets with high levels of heterogeneity, where a single ATE estimate may not accurately reflect the variation in treatment effects.

Through a series of simulations on synthetic datasets, as well as applications to real-world datasets, we demonstrate that this method provides more robust ATE estimates with lower bias and variance compared to traditional IPW. Additionally, the use of cluster-specific ATEs offers insights into how treatment effects vary across different subgroups, providing a more nuanced understanding of the data. This makes the method particularly valuable in fields like healthcare, economics, and social sciences, where treatment effects often differ across populations.

For the synthetic datasets, we demonstrate that the cluster-specific IPW technique is more accurate in estimating the average treatment effect(which is already known) as compared to the traditional IPW technique, and for real-life datasets, we compare

various metrics relating to stability and robustness of the ATE estimates to highlight the superiority of the cluster-specific IPW technique over the traditional IPW technique.

The rest of the paper is structured as follows: The literature review section presents the existing techniques that address this problem, the methodology section describes in detail the cluster-specific hybrid weighting technique, the results section describes the results of using this technique on a synthetic dataset and a real-life dataset, and the conclusion and future work section mentions about how this technique can be explored further.

2 Literature Review

Causal inference methods, particularly Inverse Propensity Weighting (IPW), have become fundamental tools in observational studies where randomization is absent.

In observational studies, estimating the Average Treatment Effect (ATE) often involves adjusting for confounding variables to account for biases in treatment assignment. A common approach is Inverse Propensity Weighting (IPW), which creates a pseudo-population where treatment assignment is independent of covariates by weighting individuals according to their propensity scores. Introduced by Rosenbaum and Rubin [1], propensity score weighting has become a fundamental technique for addressing confounding. However, traditional IPW methods are prone to instability due to extreme weights, which occur when propensity scores are close to 0 or 1, leading to high variance in the ATE estimates [2].

The challenge of extreme weights has been widely documented. Kang and Schafer [3] highlighted the sensitivity of IPW to model misspecification, especially when propensity scores are poorly estimated, causing extreme weights to disproportionately influence the results. Similarly, Robins et al. [4] noted that when positivity violations (i.e., individuals with covariates that make treatment assignment highly unlikely) occur, the ATE estimates become unstable, reducing the reliability of the findings. To mitigate these issues, several extensions of IPW have been proposed, including trimming extreme weights [5] and stabilized weights [6], where propensity scores are modified to reduce the influence of extreme values. However, these methods do not fully resolve the problem of treatment effect heterogeneity across subgroups.

The growing recognition of treatment effect heterogeneity has led to the development of clustering-based techniques, which aim to estimate cluster-specific treatment effects and capture variations in treatment response across different subpopulations. Crump et al. [7] explored the idea of restricting the analysis to a subset of the population where treatment effect estimation is more reliable, but this does not generalize to estimating the ATE for the entire population. More recently, Athey and Imbens [8] introduced methods such as Causal Trees, which partition the population into subgroups and estimate treatment effects for each subgroup. Although this approach improves the estimation of heterogeneous effects, it does not address the instability caused by extreme weights in IPW.

In response to these challenges, the Cluster-Specific ATE Estimation method builds on the idea of subgroup analysis by clustering individuals based on their covariates and calculating treatment effects within each cluster. This technique leverages the strengths of traditional IPW while mitigating the impact of extreme weights by

estimating cluster-level propensity scores and taking a weighted average of the cluster-specific ATEs. This method reduces the sensitivity to extreme propensity scores, leading to more stable estimates [9]. By focusing on cluster-specific treatment effect heterogeneity, the method also addresses limitations associated with traditional IPW approaches that assume a homogeneous treatment effect across all individuals.

Recent advances in causal inference, such as Bayesian causal forests [10] and meta-learners [11], have emphasized the importance of improving model robustness in the presence of complex data structures. The Cluster-Specific ATE Estimation method complements these innovations by providing a framework that combines propensity score weighting with clustering to improve ATE estimation, particularly when positivity violations and high variance are present.

Thus, the proposed method fills an important gap in the literature by offering a solution that addresses both extreme weight sensitivity and treatment effect heterogeneity. As observational studies continue to be a cornerstone of causal inference in fields such as healthcare and economics, developing more robust approaches like the Cluster-Specific ATE Estimation will enhance the reliability of treatment effect estimates in practical applications.

This paper contributes to literature on causal inference by introducing a hybrid approach that combines the strengths of individual-level and cluster-level weighting methods. By addressing the limitations of traditional IPW, particularly in cases of positivity violations and high variance, this technique provides a more robust framework for estimating treatment effects in observational studies. Its applicability across different domains, including healthcare, social sciences, and economics, makes it a valuable tool for researchers and practitioners alike.

3 Methodology

In this section, we describe the methodology for evaluating the Cluster-Specific ATE Estimation with Propensity Score Weighting method and compares its performance with the traditional Inverse Probability Weighting (IPW) method. This comparison is based on a series of experiments using synthetic datasets with varying true Average Treatment Effects (ATEs) and introducing positivity violations. The overall goal is to demonstrate the robustness of the clustering-based method in scenarios where traditional IPW may fail due to extreme propensity score weights.

3.1 Varying the True ATE:

To test the performance of both methods across different treatment effect magnitudes, we generate a list of true ATE values ranging from 1 to 100. Each of these ATE values represents the actual treatment effect for a different synthetic dataset. By systematically increasing the true ATE, we can observe how each method performs in terms of bias, variance, and stability across a broad range of scenarios. This ensures that the results are not biased toward a specific ATE value and that the conclusions are robust. For each true ATE value, $\tau \in 1, 2, 3, ...100$ we generate a synthetic dataset and calculate the ATE using both methods.

3.2 Synthetic Data Generation with Positivity Violations:

Each synthetic dataset consists of 15,000 observations and 8 covariates $X_1, X_2, X_3, ...X_8$. The covariates are generated from normal distributions, and the treatment assignment is determined based on a logistic model. To introduce positivity violations and simulate real-world challenges, we deliberately manipulate the treatment assignments for some observations, creating situations where individuals with very low propensity scores receive treatment, and individuals with very high propensity scores do not. This ensures that extreme weights will be present in the traditional IPW method.

The data generation process is as follows:

- Covariates $X_1, X_2, X_3, ...X_8$ are drawn from normal distributions.
- We choose 3 of these covariates- X_1, X_2, X_3 which affect both the outcome and the treatment. The propensity score $P(T = 1|X_1, X_2, X_3)$ is calculated using a logistic function of X_1, X_2 and X_3.
- Treatment assignment T is based on the propensity score, with the probability of treatment equal to the propensity score, but positivity violations are introduced by forcing treatment for individuals with low propensity scores and forcing control for individuals with high propensity scores.
- The outcome Y is generated as a linear combination of the covariates and the treatment effect, with the true ATE being systematically varied from 1 to 100 across datasets.

3.3 Propensity Score Estimation and IPW Weights

Once the synthetic data is generated, the propensity scores are estimated using a logistic regression model that predicts treatment assignment based on the covariates. The propensity scores $e(X_i)$ are then used to calculate the weights for both the traditional IPW method and the cluster-specific method.

For the traditional IPW, the weights w_i are calculated as follows:

$$w_i = \begin{cases} \frac{1}{e(X_i)} & \text{if } T_i = 1 \\ \frac{1}{1-e(X_i)} & \text{if } T_i = 0 \end{cases}$$

These weights are then used to estimate the ATE using the traditional IPW formula:

$$\widehat{\tau}_{IPW} = \frac{\sum_{i=1}^{N} w_i Y_i T_i}{\sum_{i=1}^{N} w_i T_i} - \frac{\sum_{i=1}^{N} w_i Y_i (1 - T_i)}{\sum_{i=1}^{N} w_i (1 - T_i)}$$

3.4 Cluster-Specific ATE Estimation

For the Cluster-Specific ATE Estimation, we first apply k-means clustering to group individuals based on their covariates. The number of cluster k is set to 10 based on exploratory analysis, and each cluster contains individuals with similar covariate values. Within each cluster, the ATE is estimated using the traditional IPW formula

but applied to the individuals within that cluster. The overall ATE is then calculated as a weighted average of the cluster-specific ATEs, where the weights are based on the average propensity score within each cluster:

$$OverallATE = \frac{\sum_{c=1}^{C}(ATE_c * avg\_propensity_c * n_c)}{\sum_{c=1}^{C}(avg\_propensity_c * n_c)}$$

where:

- ATE_c is the ATE for cluster c.
- $avg\_propensity_c$ is the average of all propensity scores in cluster c.
- C is the total number of clusters.

This method accounts for treatment effect heterogeneity across clusters and stabilizes the ATE estimation by reducing the impact of extreme weights. For the synthetic dataset, where the true ATE is known, we calculate the ATE estimate using the traditional IPW method and compare it with the ATE calculated using the cluster-specific ATE method.

The results of this methodology provide insights into the robustness and stability of the clustering-based method compared to traditional IPW, particularly in cases with positivity violations and extreme weights. By systematically varying the true ATE and generating multiple synthetic datasets, we demonstrate that the Cluster-Specific ATE Estimation consistently outperforms traditional IPW in terms of lower variance, reduced bias, and fewer extreme weights, making it a more reliable method for estimating treatment effects in complex observational studies.

3.5 Application on Real-life datasets:

In addition to testing the Cluster-Specific ATE Estimation technique on synthetic data, the methodology extends to real-life datasets to assess its effectiveness in practical scenarios where the true ATE is unknown. This section describes how the method is applied to real-world data and what metrics are compared for the traditional Inverse Propensity Weighting (IPW) technique and the cluster-specific IPW technique.

The following metrics are evaluated and compared for the traditional IPW technique to the cluster-specific technique:

- Variance of the ATE Estimates: The variance of the ATE estimates is calculated using bootstrapping to assess the stability of the estimates. A lower variance indicates a more stable and reliable method. This is calculated in a similar way to how the variance was calculated for the synthetic datasets
- 95 % confidence interval for the estimted ATE: While estimating the ATE using both the traditional IPW method and the cluster-specific method, we also calculate the 95 % C.I. around the estimated ATE by taking 1000 bootstrap samples of data, calculating the estimated ATE in each sample, and then calculating the 2.5th and the 97.5th percentiles. Narrower confidence intervals suggest a higher precision for the ATE estimates.

- Effective Sample Size (ESS): The effective sample size reflects how much information is retained after applying the inverse propensity weights. A smaller ESS indicates that the estimator is relying on a small subset of the data, which could lead to greater variability and less reliable estimates. The ESS is calculated as :

$$ESS = \frac{(\sum_{i=1}^{n} w_i)^2}{\sum_{i=1}^{n} w_i^2} \tag{1}$$

  A higher ESS indicates that the method makes better use of the available data. The hybrid method should typically result in a higher ESS than traditional IPW.

- Number of Extreme Weights: The number of extreme weights (defined as weights greater than 10 times the mean weight) is counted to determine how prone each method is to instability caused by extreme propensity scores. The Cluster-Specific ATE method is expected to produce fewer extreme weights compared to traditional IPW.

3.6 Expected Outcomes on Real-life datasets:

It is hypothesized that the Cluster-Specific ATE Estimation technique will outperform traditional IPW in real-life datasets by providing

- More stable ATE estimates, with lower variance, narrower 95 % confidence interval band, and a higher effective sample sizes.
- Fewer extreme weights, leading to more reliable estimates of treatment effects, especially in datasets with positivity violations.

By applying this method to a diverse set of real-life datasets and comparing the results with traditional IPW, the paper demonstrates the generalizability and robustness of the Cluster-Specific ATE Estimation approach in observational studies.

4 Results

4.1 Evaluation on Synthetic Datasets

For evaluation of this technique on a synthetic dataset, we create a list of true ATE values ranging from 1 to 100. Traditional IPW estimates are computed as a benchmark for comparison. For each ATE, value, we create a synthetic dataset with this ATE value and compare the performance of traditional IPW technique to the cluster-specific technique. For a given value of true ATE, The synthetic data for this technique is created as follows:

We generate 15000 observations of 8 covariates $X_1, X_2, X_3, ..., X_8$, 3 out of which are influential observations, X_1, X_2, X_3, which affect both the treatment probability and the outcome. The probability of receiving the treatment is modeled as a logistic function of the covariates:

$$P(T = 1|X) = \frac{1}{1 + exp(-(2.5 * X_1 - 0.25 * X_2 + 0.5 * X_3 - 3))}$$

The treatment assignment variable, T is assigned as 1 or 0 where the probability of treatment being 1 is the probability calculated in the above equation. The true effect is The outcome Y is then simulated as a linear combination of the covariates, treatment and the True ATE as :

$$Y = (TrueATE) * T + 0.5 * X_1 * X_2 + 0.3 * X_3 + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, 1)$

3. Propensity Score Estimation The propensity scores are estimated using a logistic regression model:

$$P(T = 1|X) = \frac{1}{(1 + exp(-(\beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + ... + \beta_8 * X_8))))}$$

where β is a vector of of coefficients estimated by maximum likelihood. For each scenario, the propensity scores are used to calculate inverse probability weights:

$$w_i = \frac{1}{(P(T = 1|X_i) * T_i + (1 - P(T = 1|X_i)) * (1 - T_i))}$$

These weights are then used in the calculation of the IPW ATE estimate.

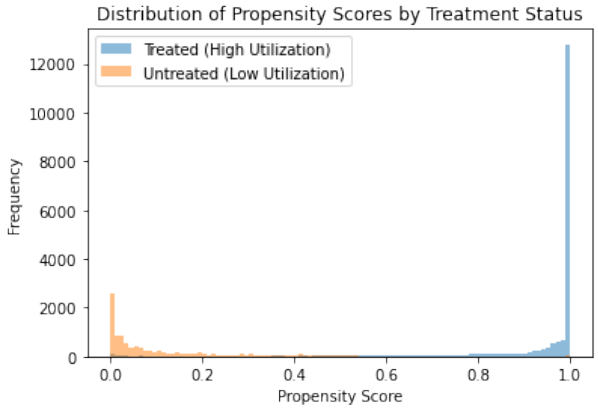

Figure 1: Distribution of propensity scores for a synthetic dataset where actual ATE = 10. We can clearly see that a large number of propensity scores are close to 0 or close to 1, leading to positivity violation.

As an illustration for positivity violation for synthetic datasets, for one of the true ATE values, $ATE = 10$, figure 1 shows that majority of the propensity scores are concentrated around 0 or 1, clearly showing positivity violation for this dataset.

For each synthetic dataset created using a different true ATE value from 1 to 75, we calculate the ATE estimate using the traditional IPW method and the cluster-specific IPW method, and compare the results. As illustrated in figure 2, we clearly see that the ATE estimate calculated using the cluster-specific method is much closer to the actual ATE as compared to the traditional IPW method. For lower values of true ATE, even though the ATE estimated using traditional method is somewhat closer to

the true ATE as compared to the cluster-specific method(this can be due to effect of noise outweighing the ATE effect for lower values of true ATE), for higher values, the cluster-specific method clearly outperforms. Thus, we are clearly able to demonstrate the superiority of the cluster-specific method over the traditional IPW method.

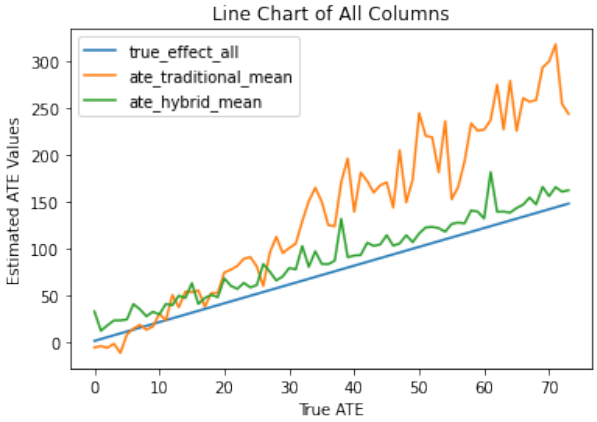

Figure 2: Comparison of True ATE to the estimated ATE using the traditional IPW technique and the Cluster-Specific IPW technique. We can clearly see that the Cluster-Specific technique is much closer than the traditional IPW technique in estimating the true ATE.

4.2 Evaluation on Real life Datasets

We use the Default of Credit Card Clients dataset from the UCI ML repository to test the effectiveness of this technique on a real life dataset. This dataset is regarding the case of customers' default payments on credit card in Taiwan. We have data of 30,000 customers across 25 columns, with demographic information such as Gender, Age, Education level, marital status, credit card balance limit, billed amount and amount paid, and payment status for each of the last 6 months, and the outcome variable is the binary flag of whether the customer defaulted on the payment for the next month.

Here, the treatment variable is defined as whether the customer's total billed amount for the last 6 months is more than 50% of the customer's balance limit for the credit card. The goal here for this dataset is to estimate the effect of high credit limit utilization on the probability of defaulting on the next month. We apply the Cluster-Specific ATE Estimation technique to estimate the effect of high utilization on default rates and compare the results with the traditional Inverse Probability Weighting (IPW) method.

4.2.1 Defining Treatment and Outcome Variables:

Treatment Variable: The treatment variable represents high credit utilization, which is computed as the sum of the billing amounts for the first six months (BILL_AMT1 to BILL_AMT6) divided by the credit limit (LIMIT_BAL). Customers with a utilization rate exceeding 50% are categorized as receiving the treatment (T = 1), indicating high utilization, while those below the threshold are assigned as control (T = 0).

Outcome Variable: The outcome of interest is whether the customer defaults on their credit card payment in the next month (default.payment.next.month), with 1 indicating default and 0 indicating no default.

4.2.2 Covariates and Propensity Score Estimation

We use a range of demographic and financial covariates to estimate the propensity scores and cluster the individuals. The covariates include:

Demographics: Age, Sex, Education, and Marital Status. Credit and Payment History: Past payment behavior (PAY_0 to PAY_6), credit limit (LIMIT_BAL), and previous billing amounts (BILL_AMT1 to BILL_AMT6). A logistic regression model is used to estimate the propensity score, which is the probability that a customer falls into the treatment group (high credit utilization) based on their covariates.

We plot the distribution of propensity scores for the default of credit card clients dataset and we can clearly see in figure 3 that majority of the propensity scores are very close to 0 or very close to 1, thus indicating positivity violation, which makes this dataset a suitable candiate for comparing the performance of cluster-specific IPW technique to traditional IPW technique.

4.2.3 Clustering

To account for heterogeneity in treatment effects, we apply k-means clustering on the covariates. The goal of clustering is to group individuals with similar credit and demographic profiles into distinct clusters. Each cluster represents a subgroup of customers who share similar risk factors and credit behavior patterns.

4.2.4 Cluster-Specific ATE Estimation

After clustering the customers, we compute the Average Treatment Effect (ATE) for each cluster. The ATE is calculated by comparing the weighted outcomes between the treated and untreated individuals within each cluster. The weights are derived from the individual propensity scores, adjusting for confounding factors between the treated and untreated groups.

Once the cluster-specific ATEs are computed, the overall ATE is obtained by taking a weighted average of the cluster-specific ATEs, where the weights are the number of individuals in each cluster.

4.2.5 Comparison with Traditional IPW

For comparison, we also estimate the ATE using the traditional IPW method, where the propensity score is used to compute individual-level weights. The traditional IPW method directly compares treated and untreated customers without accounting for potential heterogeneity in treatment effects across different clusters.

4.2.6 Performance Metrics

We evaluate the performance of the Cluster-Specific ATE Estimation and Traditional IPW methods using the following metrics:

- **Estimated ATE:** The estimated treatment effect of high credit utilization on default rates.
- **Variance of ATE:** Variance is assessed through bootstrapping to evaluate the stability of the ATE estimates.

- **95 % confidence interval of the ATE estimates:** The 95 % C.I. around the ATE estimates using both methods is calculated by taking 1000 bootstrap samples and calculating the 2.5 and the 97.5 percentiles of the ATE estimates.
- **Effective Sample Size (ESS):** The ESS is computed to measure the efficiency of the weighting schemes in both methods.
- **Number of Extreme Weights :** We calculate the number of extreme weights (weights greater than 10 times the mean) to assess the stability of the methods.

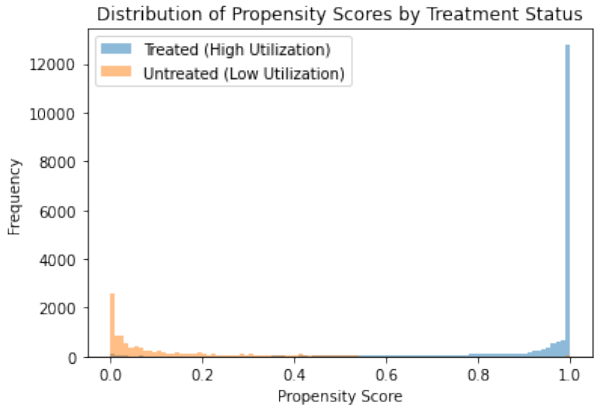

Figure 3: Distribution of propensity scores for a Credit card default dataset. We can clearly see that a large number of propensity scores are close to 0 or close to 1, leading to positivity violation.

Using the IPW method to estimate the propensity scores, we can see in the graph below that a majority of the propensity scores are near 0 or 1, leading to positivity assumption violation and high variance violation, making this dataset a suitable candidate for the cluster-specific IPW technique. We perform clustering and ATE estimation in a similar method, and below are the results of the analysis:

Metric	**Traditional IPW technique**	**Cluster-specific IPW technique**	**Column 4**
ATE(Actual ATE unknown)	0.2099	-0.0319	Actual ATE unknown
95 % C.I. for ATE	[0.1303, 0.2663]	[-0.0404, -0.0229]	Lower is better
Effective Sample Size	2321.98	18275.38	Higher is better
Variance of ATE	0.0014	2.001E-05	Lower is better
Number of extreme weights	6969	0	Lower is better

Table 1: Comparison of metrics for Credit Card default dataset

For these metrics, as we do not know the actual ATE for this dataset, we cannot comment on which estimated ATE is correct, however, for all the other metrics, we can

clearly see that for each of the performance metric, the cluster-specific IPW technique outperforms the traditional IPW technique. This clearly demonstrates the superiority of the cluster-specific IPW technique over traditional IPW technique, while applied on a real-life dataset. The code for the analysis on both the real-life dataset and the synthetic dataset can be found here

5 Conclusion & Future Work

This paper introduces a novel Cluster-Specific Inverse Probability Weighting (IPW) method to address the challenges associated with traditional IPW, particularly in situations involving positivity violations, high variance, and treatment effect heterogeneity. By integrating individual-level propensity score weights with cluster-level weights, the cluster-specific method stabilizes weight estimates while capturing the nuanced heterogeneity of treatment effects across subgroups. The performance of the method was demonstrated on a set of large, complex synthetic datasets where the Cluster-Specific method significantly outperformed traditional IPW in terms of more accurate ATE estimation and for the real-life dataset, the Cluster-Specific method outperformed the traditional IPW method on various metrics.

Future Work:

While the results demonstrate the effectiveness of the Hybrid Cluster-Weighted IPW method, several avenues remain for future research and optimization. This section outlines potential directions for enhancing the method further and addressing some of its current limitations.

- Optimal Number of Clusters: Currently, the number of clusters is selected arbitrarily (e.g., 10 or 12 clusters), but determining the optimal number of clusters is crucial for maximizing the performance of the method. Clustering plays a key role in reducing extreme weights and improving the balance between treated and untreated groups. Future work could focus on developing data-driven methods to determine the optimal number of clusters based on: Cross-validation techniques that minimize the variance of the ATE across different cluster counts. Cluster quality measures, such as the silhouette score, gap statistic, or inertia, that assess the separation and compactness of clusters based on covariate distributions. BIC/AIC-based selection to penalize overfitting, especially when using more flexible clustering methods like Gaussian mixture models. Additionally, adaptive clustering methods could be explored, where clusters are formed dynamically based on the underlying propensity score distribution or treatment effect heterogeneity across the dataset.

- Incorporating More Complex Data Structures: Future research could expand the Hybrid Cluster-Weighted IPW method to handle more complex data structures, such as:

  Longitudinal data: Where treatment assignment and outcomes evolve over time, requiring dynamic adjustments of weights across time points. Multilevel/hierarchical data: Where individuals are nested within groups (e.g., students in schools, patients in hospitals). Adapting the hybrid weighting approach to multilevel settings would enable more accurate causal inference in clustered or hierarchical datasets. High-dimensional data: With a large number of covariates,

regularization methods or dimensionality reduction techniques (e.g., principal component analysis) could be integrated into the hybrid method to prevent overfitting while maintaining covariate balance.

6 References

1. Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika.
2. Hirano, K., & Imbens, G. W. (2001). Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. Health Services and Outcomes Research Methodology.
3. Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. Econometrica.
4. Robins, J. M., Hernán, M. A., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. Epidemiology.
5. Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. Statistics in Medicine.
6. McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological Methods.
7. Zubizarreta, J. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. Journal of the American Statistical Association.
8. Athey, S., Imbens, G. W., & Wager, S. (2018). Estimating treatment effects with causal forests. Annals of Statistics.
9. Imai, K., & Ratkovic, M. (2013). Covariate balancing propensity score. Journal of the Royal Statistical Society: Series B (Statistical Methodology).
10. Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science.
11. Cattaneo, M. D. (2010). Efficient semiparametric estimation of multivalued treatment effects under ignorability. Journal of Econometrics.
12. Sant'Anna, P. H., & Zhao, J. (2020). Doubly robust difference-in-differences estimators. Journal of Econometrics.
13. Pearl, J. (2009). Causality: Models, reasoning, and inference. Cambridge University Press.
14. Rubin, D. B. (2004). Direct and indirect causal effects via potential outcomes. Scandinavian Journal of Statistics.