

Article

Robust Interval Estimation for Treatment Effects with Few Treated and Many Control Observations

Chong Ding ^{1,*} , Linzi Sun ² and Zheng Li ³

¹ Key Laboratory for Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun 130024, China

² School of Mathematics and Statistics, Peking University, Beijing 100871, China

³ School of Mathematics and Systems Science, Liaocheng University, Liaocheng 252059, China

* Correspondence: cdnenu@gmail.com

Abstract

When some treatment labels are missing, the number of the observed treated group is much smaller than the control group. The imbalanced dataset can make treatment effect estimation more challenging. To address this issue, we develop an inference framework that integrates an Absent Data Generating (ADG) algorithm with kernel matching. The proposed method incorporates a modified Kernel Fisher Discriminant (KFD) into the causal inference framework to recover missing treatment assignments. SMOTE and Borderline-SMOTE are mainly designed for imbalanced classification problems. These methods generate synthetic observations through local interpolation. In contrast, the ADG algorithm learns the main patterns in the observed treated units. The algorithm then generates the unobserved samples by updating the data structure step by step and drawing samples based on the learned relationships. Consequently, it provides a distribution-based mechanism for recovering missing treatment information rather than relying solely on neighborhood-based interpolation. The recovered treatment labels are subsequently incorporated into kernel matching to improve ATT estimation and interval inference. Monte Carlo simulations and a real-data application demonstrate that the proposed method provides more accurate ATT estimates and more reliable confidence intervals under severe treatment-group imbalance.

Keywords: missing treatment labels; synthetic data generation; imbalanced data; kernel matching; causal inference; confidence intervals

1. Introduction

In practical research, missing treatment labels constitute a common yet often overlooked data issue, widely arising in medical studies, policy evaluations, and industrial experiments. In medical studies and electronic health record applications, treatment information is often partially missing because of incomplete documentation, irregular follow-up, or inconsistencies across data sources. Discarding observations with missing treatment labels may substantially reduce the effective sample size and introduce estimation bias. When the treatment is relatively rare, missing labels can further aggravate sample imbalance, leading to underestimated uncertainty and overly narrow confidence intervals. To address these challenges, we propose a framework that recovers missing treatment labels from observed covariates and subsequently estimates treatment effects in observational studies with incomplete treatment information. This type of imbalanced data can reduce the performance of treatment effect estimation methods. This problem can affect



Academic Editor: George Karabatsos

Received: 5 May 2026

Revised: 22 June 2026

Accepted: 29 June 2026

Published: 2 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

the reliability of the results. Some common methods include inverse probability weighting (IPW) [1], doubly robust (DR) estimation [2], and targeted maximum likelihood estimation (TMLE) [3,4]. These methods are widely used in practice. However, IPW is sensitive to extreme propensity scores. It can lead to unstable results [5]. DR and TMLE also depend on the quality of the propensity score model and the outcome model. This issue becomes more serious in small samples. These limitations motivate the use of kernel matching. Kernel matching uses smooth weights and uses information from all observations. It reduces the impact of poor matches and provides a flexible way to estimate the ATT.

As an estimator that is theoretically asymptotically unbiased and possesses desirable statistical properties, kernel matching assigns weights to control samples through a kernel function to construct counterfactual outcomes for each treated individual. Ding et al. [6] established the consistency and asymptotic normality of the kernel-matching estimator for both the average treatment effect (ATE) and the ATT, and further extended the theoretical results to the setting where the propensity score is unknown, providing a solid theoretical foundation for empirical applications. However, the asymptotic normality of these estimators, particularly when confidence intervals are constructed using bootstrap methods, typically relies on the assumption that the sample size tends to infinity.

To address the above issue, one possible strategy is to first apply binary classification algorithms to identify and impute the missing treatment labels, thereby obtaining a complete dataset on which confidence interval estimation can subsequently be performed. However, when the identification task focuses on a rare treated subgroup, directly applying conventional binary classifiers, including the support vector machine (SVM) [7] and logistic regression [8], often becomes unstable and yields suboptimal performance. The fundamental reasons can be summarized in three aspects. First, when treated samples are extremely scarce, prediction bias is likely to occur, with the model tending to classify most test samples into the majority control group [9]. Second, an imbalanced training set may mislead the model in learning the true classification boundary [10]. Finally, in regions where the two classes substantially overlap, standard learning objectives may place insufficient emphasis on difficult minority class observations, making it challenging for the classifier to accurately distinguish these critical regions and thereby exacerbating classification bias [11,12].

To overcome the limitations of existing machine learning classification methods under such settings, this paper introduces data resampling and data generation techniques from the machine learning literature into the causal inference framework to address the data imbalance problem. The Synthetic Minority Oversampling Technique (SMOTE) [13] introduced the core idea of generating new synthetic samples through interpolation between minority class observations, establishing a research paradigm for synthetic oversampling. Unlike traditional approaches that rely solely on resampling the original data, SMOTE innovatively achieves data balance by effectively “creating” new samples. In addition to SMOTE, a refined method proposed by Han et al. [14] concentrates the generation of synthetic minority instances in regions close to the class separation boundary, while Batista et al. [15] further studied frameworks that combine multiple heuristic strategies with synthetic oversampling to enhance the representation and classification performance of minority samples. Overall, these synthetic oversampling methods generally outperform simple resampling of the original data because the additional information utilized by the model does not come from repeated sampling but is generated within the feature space through synthesis mechanisms. This further helps reduce the lack of representation caused by too few minority samples. Other studies including ADASYN [16], CTGAN [17], and diffusion-based models [18] show good results for creating synthetic tabular data to solve problems about the imbalanced dataset. For example, Sauber-Cole and Khoshgoftaar [19] showed that synthetic data can improve prediction performance when the minority class has very

few samples. Majeed and Hwang [20] studied CTGAN for tabular data generation. They found that CTGAN can produce more realistic data than many traditional oversampling methods. The model also keeps more structure in the data. In addition, Chen et al. [21] reviewed several modern oversampling methods. They showed that these methods can help solve class imbalance problems in real applications. These studies show that generative oversampling methods are becoming more important. At the same time, they also show that generating high-quality synthetic data is still a difficult and open problem.

Motivated by these developments, we further combine the ADG method with kernel matching to improve the accuracy of treatment effect estimation. Since kernel matching leverages similarity measures in a reproducing kernel Hilbert space to construct weighted comparisons between treated and control units, it captures complex nonlinear relationships in covariates. However, kernel matching does not always work well when the data is highly unbalanced. It also performs poorly when overlap is weak. The weights can become unstable. The treatment effect estimation may then have high variance. Then, we use ADG to address this problem because it increases the number of treated samples and improves the overlap between groups. It also makes the weighting process more stable. The generated samples are built in the same feature space used by the kernel. This makes them consistent with the matching process. The kernel method can then use the added structure in the data. As a result, the combined method reduces variance and improves balance between covariates. It produces more stable and reliable estimates of treatment effects.

In this paper, we aim to recover missing treatment labels by integrating the ADG into a causal inference framework and propose an ADG-based kernel-matching approach to address severe treatment-group imbalance arising from missing treatment assignments. We show that incorporating synthetic oversampling into the label recovery process can mitigate the imbalance induced by missing labels and consequently provide more reliable treatment effect estimates than existing approaches while maintaining negligible bias.

The remainder of this paper is organized as follows. Section 2 introduces the notation framework and outlines the core ideas of kernel matching for estimating treatment effects. In Section 3, we propose a kernel-matching estimator combined with the ADG algorithm to handle missing treatment assignments and imbalanced data structures. Section 4 reports Monte Carlo simulation results to assess the finite-sample performance of the proposed approach. An empirical application is presented in Section 5 to demonstrate its practical usefulness. Finally, Section 6 concludes the paper with closing remarks and discusses possible avenues for future research.

2. Notation and Fundamental Concepts of Kernel Matching

In observational studies, we adopt the potential outcome framework proposed by Robins et al. [22]. Under this framework, each individual i is associated with a pair of potential outcomes $\{Y_i(1), Y_i(0)\}$, where $Y_i(1)$ denotes the potential outcome if the individual receives the treatment, and $Y_i(0)$ denotes the potential outcome if the individual does not receive the treatment. We observe the treatment indicator $W_i \in \{0, 1\}$, along with a set of pre-treatment covariates X_i . The observed outcome Y_i can be written as $Y_i = W_i Y_i(1) + (1 - W_i) Y_i(0)$. The primary causal estimator of interest in this paper is ATT, defined as

$$\tau_{ATT} = E[Y(1) - Y(0) \mid W = 1].$$

A dataset of size N is assumed to contain N_1 treated observations and N_0 control observations, denoted by $\{Y_i, X_i, W_i\}_{i=1}^N$, where all observations are assumed to be independently drawn from their respective populations. Let $\mathcal{X}$ denote the covariate space. Define $\mathcal{T}_w$ as the index set of observations satisfying $W_i = w$, so that $\mathcal{T}_1$ and $\mathcal{T}_0$ correspond to the treated and control groups, respectively. Let $\mathcal{X}_{\mathcal{T}_0} = \{x_i : i \in \mathcal{T}_0\} \subset \mathcal{X}$ denote the

set of covariates for the control group, and $\mathcal{X}_{\mathcal{T}_1} = \{x_i : i \in \mathcal{T}_1\} \subset \mathcal{X}$ denote the set of covariates for the treated group. In a well-specified setting with adequately large and properly balanced treated and control groups, and free from missing or misclassified treatment labels, kernel-matching estimators for ATT typically yield stable point estimates and reliable confidence intervals.

The following assumptions are crucial for deriving the asymptotic properties and will be maintained throughout the analysis. They are standard in the causal inference literature and widely adopted in existing studies; see, for instance, refs. [23,24]. Let $X \in \mathbb{R}^k$ be a vector of covariates, and denote by $\mathbb{X}$ its support. For any $x \in \mathbb{X}$, define the propensity score as $p(x) = \Pr(W = 1 \mid X = x)$.

Assumption 1. (i) (Conditional Independence) The treatment assignment is independent of the potential outcomes given covariates, i.e., $W \perp \{Y(1), Y(0)\} \mid X$, where $\perp$ denotes statistical independence [25]; (ii) (Positivity) the propensity score is strictly bounded away from 0 and 1, namely, $p(x) \in (0, 1)$.

Assumption 1 (i), the conditional independence assumption [26–28], restricts the way in which potential outcomes may differ across treatment status by ensuring that selection into treatment is fully explained by observed covariates. While Assumption 1 implies that the conditional distribution of $Y(0)$ given X is the same for treated and control units, the distribution of X conditional on $W = w \in \{0, 1\}$ may still depend on w . This imbalance in the covariate distributions is precisely what can lead to bias in simple comparisons of outcomes between treated and control groups when covariates are not properly accounted for. Since the focus of this paper is on the ATT, it is not necessary to impose conditional independence for $Y(1)$ in the same way as in the ATE setting.

Next, we present kernel-based matching estimators designed for estimating the ATT. Under the notation introduced above, we assume N to be a natural number, i.e., $N \in \mathbb{N}$, and assume that $K(\cdot)$ denotes a kernel function while h represents a bandwidth parameter satisfying $h \rightarrow 0$ as N increases. For convenience, without loss of generality, we let

$$\frac{X_i - x}{h} = \left(\frac{X_{i1} - x_1}{h_1}, \frac{X_{i2} - x_2}{h_2}, \dots, \frac{X_{id} - x_d}{h_d} \right),$$

and

$$K\left(\frac{X_i - x}{h}\right) = \prod_{s=1}^q K\left(\frac{X_{is} - x_s}{h_s}\right).$$

The kernel-based matching estimators that rely on covariates X to impute the unobserved potential outcomes are constructed as follows:

$$\hat{Y}_i(0) = \begin{cases} Y_i & \text{if } W_i = 0, \\ \frac{\sum_{j:W_j=0}^N Y_j K\left(\frac{X_j - X_i}{h}\right)}{\sum_{j:W_j=0}^N K\left(\frac{X_j - X_i}{h}\right)} & \text{if } W_i = 1, \end{cases}$$

and

$$\hat{Y}_i(1) = \begin{cases} \frac{\sum_{j:W_j=1}^N Y_j K\left(\frac{X_j - X_i}{h}\right)}{\sum_{j:W_j=1}^N K\left(\frac{X_j - X_i}{h}\right)} & \text{if } W_i = 0, \\ Y_i & \text{if } W_i = 1. \end{cases}$$

Then, the proposed kernel-matching estimator for the ATT is given by

$$\hat{\tau}_t = \frac{1}{N_1} \sum_{i=1}^N W_i \left(Y_i - \frac{\sum_{j=1}^N (1 - W_j) Y_j K\left(\frac{X_j - X_i}{h}\right)}{\sum_{j=1}^N (1 - W_j) K\left(\frac{X_j - X_i}{h}\right)} \right), \quad (1)$$

where $N_1 = \sum_{i=1}^N W_i$ gives the sample size of the treated subgroup, obtained by summing the treatment indicators over all units. The asymptotic results for kernel-matching estimators are established based on the following assumptions concerning the kernel function.

Assumption 2. The kernel function $K(\cdot)$ is assumed to be symmetric and to satisfy a set of standard regularity conditions, namely:

- (1) $\int K(u) du = 1$;
- (2) $\int u^2 K(u) du < \infty$;
- (3) $\int K^2(u) du < \infty$;
- (4) $\int |u|^3 K(u) du < \infty$.

The regularity conditions on the kernel function form the basis for deriving the estimator's asymptotic behavior, providing the foundation for both its asymptotic unbiasedness and a precise description of its limiting distribution.

Under Assumptions 1 and 2, Theorem 3.2 in [6] establishes that the kernel-matching estimator possesses desirable large-sample properties: it is consistent and asymptotically normal. However, these theoretical guarantees fundamentally rely on asymptotic arguments, namely that the sample sizes of both the treated and control groups tend to infinity. When the sample size is sufficiently large, the asymptotic normality established in Theorem 3.2 provides a theoretical justification for constructing confidence intervals via bootstrap methods. In small-sample settings, however, particularly when the treated group is extremely sparse, these guarantees no longer hold. In such cases, bootstrap resampling may fail to adequately capture the distributional characteristics of the original sample, leading to confidence intervals with unsatisfactory performance, such as under-coverage or excessive conservativeness, manifested by reduced coverage probability or abnormal fluctuations in interval length. Therefore, we integrate the ADG algorithm with kernel matching to recover missing treated samples and restore data balance before treatment effect estimation.

3. Confidence Interval Estimation for ATT

In this section, we first introduce the ADG algorithm and then develop its integration with kernel matching for estimating ATT and constructing reliable confidence intervals under imbalanced data settings.

In practical applications, due to survey errors, data entry mistakes, and other factors, the treatment indicator W is often missing for a subset of individuals, leading to a severe shortage of treated samples. This issue directly affects the performance of kernel-matching estimation: when the treated group is sparse, the matching weights tend to concentrate on a small number of control observations, resulting in increased variance of the ATT estimator, wider confidence intervals, and reduced coverage probability.

To address this issue, we incorporate the ADG algorithm into the causal inference framework. By rebalancing the original data distribution, the proposed approach improves the estimation accuracy of kernel matching and enhances the reliability of confidence interval coverage under imbalanced samples. Specifically, ADG first identifies and imputes missing treatment labels, and then kernel matching is applied to the rebalanced data to obtain more stable and accurate confidence intervals for ATT.

Suppose we observe a sample of size N , but the treatment assignment is partially missing. To characterize this missingness, we introduce an observation indicator R_i :

$$R_i = \begin{cases} 1, & \text{if } W_i \text{ is observed,} \\ 0, & \text{if } W_i \text{ is missing.} \end{cases}$$

The observed dataset can therefore be written as $\mathcal{D}_{\text{obs}} = \{(X_i, Y_i, W_i, R_i) : i = 1, \dots, N\}$, where W_i is only observed when $R_i = 1$. In the presence of severe class imbalance, we define the training set without loss of generality as follows. Let the covariates of the control group, denoted by $\mathcal{X}^-$, be defined as $\{\mathbf{x}_i^-\}_{i=1}^{l_-}$, representing the set of observed samples satisfying $W_i = 0$ and $R_i = 1$. Similarly, let the covariates of the treated group, denoted by $\mathcal{X}^+$, be defined as $\{\mathbf{x}_i^+\}_{i=1}^{l_+}$, representing the set of observed samples satisfying $W_i = 1$ and $R_i = 1$. Both $\mathcal{X}^-$ and $\mathcal{X}^+$ are subsets of the full covariate set $\mathcal{X}$. Together, they form the training set. Under class imbalance, we have $l_+ \ll l_-$. For notational simplicity, l_+ is written as l . To incorporate the above data generation mechanism into the kernel-matching framework, we apply an artificial data generation approach tailored for settings with scarce treated samples. The main idea is to make new treated data points near the border between the two groups. The method creates these new points inside the existing groups and selects each group for new data. At the same time, a group with fewer samples has a higher chance of being selected. This increases the size of the smaller class and keeps the original data pattern as much as possible. Therefore, this gives a more balanced dataset for later treatment effect analysis.

First, within the causal inference framework, we define “missing treated data” as those observations that should belong to the treated group but are unobserved, leading to class imbalance. In other words, these missing observations can be understood as “treated samples that ought to exist but were not recorded.” By introducing this concept, we can leverage synthetic methods to regenerate these missing observations, thereby alleviating the imbalance caused by the scarcity of treated samples and providing a more reliable data foundation for subsequent treatment effect estimation. The available observed data may provide sufficient information to infer and synthesize additional within-class observations. However, the existence of missing data does not imply prior knowledge of their exact quantity or precise locations in the feature space. Nevertheless, in imbalanced classification settings, particularly in causal inference settings where treated samples are scarce, this assumption provides a theoretical foundation for synthetic oversampling of the minority class. The set of missing treated data is represented by $\mathcal{A} = \{\mathbf{x}_i^+\}_{i=l+1}^{l+k} \subset \mathcal{X}^+$, assumed to be independent and identically distributed. Furthermore, for $j = 1, 2, \dots, k$, denote $\mathbf{x}_{l+j}^+ \in \mathcal{A}$ as $\mathbf{a}_j$.

Based on Fisher linear discriminant analysis, for the binary classification problem of distinguishing treated from control samples, we introduce Kernel Fisher Discriminant (KFD). As noted by Hofmann et al. [29], when a properly selected kernel function is employed for the data transformation, and all computations rely solely on kernel evaluations, explicit operations in the high-dimensional feature space are unnecessary. The so-called kernel trick allows this property to be applied to Fisher discriminant analysis, resulting in KFD [30]. Specifically, KFD extends Fisher linear discriminant analysis to the feature space, with the objective

$$\max_w J(w) = \frac{w^T S_B^\phi w}{w^T S_W^\phi w}, \quad (2)$$

where S_B^ϕ represents the between-class scatter matrix and S_W^ϕ the within-class scatter matrix in the feature space, whose forms are given by

$$\begin{aligned} S_B^\phi &= dd^T, \\ S_W^\phi &= \sum_{i \in \{-, +\}} S_i^\phi, \end{aligned} \quad (3)$$

where $d = m_-^\phi - m_+^\phi$, $S_i^\phi = \sum_{x \in \mathcal{X}^i} (\phi(x) - m_i^\phi)(\phi(x) - m_i^\phi)^T$ with $m_i^\phi = \frac{1}{l_i} \sum_{j=1}^{l_i} \phi(x_j^i)$ and a nonlinear mapping $\phi: \mathcal{X} \rightarrow \mathcal{F}$, which embeds data into a Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle$. The corresponding kernel function K is defined on $\mathcal{X} \times \mathcal{X}$ such that $K(x, x') = \langle \phi(x), \phi(x') \rangle$, thus enabling inner products in the feature space to be computed implicitly via K .

To address data imbalance, we extend KFD by generating synthetic treated samples in the projected feature space. In most settings, informative samples lie near the boundary of the control group. The treated group often generate multiple clusters in high-dimensional settings. We leverage projection-region similarity to guide data generation, which preserves the clustering structure of the data while also reducing model bias.

Assume that the treated samples in the training set can be divided into C distinct clusters, where C is an integer greater than one. The positive dataset $\mathcal{X}^+$ is organized as a collection of C mutually exclusive subsets. Specifically, it can be expressed as the union of cluster-wise groups, $\mathcal{X}^+ = \bigcup_{c=1}^C \mathcal{X}_c^+$, with the property that any two distinct clusters share no common elements, i.e., $\mathcal{X}_c^+ \cap \mathcal{X}_{c'}^+ = \emptyset$, $c \neq c'$. Each subset $\mathcal{X}_c^+$ corresponds to the samples of the c -th minority cluster and is defined as $\mathcal{X}_c^+ = \{x_{1,c}^+, x_{2,c}^+, \dots, x_{l_c,c}^+\}$. The size of each cluster is denoted by $l_c = |\mathcal{X}_c^+|$, and the total number of positive samples across all clusters is given by $\sum_{c=1}^C l_c = l$. Correspondingly, the missing treated dataset $\mathcal{A}$ is partitioned into C disjoint clusters, denoted by $\mathcal{A}_c$, where $c = 1, 2, \dots, C$, and each $\mathcal{A}_c$ corresponds to a minority cluster $\mathcal{X}_c^+$. Specifically, $\mathcal{A}_c = \{x_{l_c+i,c}^+\}_{i=1}^{k_c}$, with $\bigcup_{c=1}^C \mathcal{A}_c = \mathcal{A}$, $\mathcal{A}_{c'} \cap \mathcal{A}_c = \emptyset$, $c \neq c'$, where $|\mathcal{A}_c| = k_c$ and $\sum_{c=1}^C k_c = k$. For each c , let $a_{j,c} = x_{l_c+j,c}^+$, $j = 1, 2, \dots, k_c$, which serves as an extension of the notation a_j defined earlier.

To achieve the two desired properties, we introduce the following constraints. First, newly generated samples are required to lie close to the projection region of existing minority samples. Second, the synthesized samples are expected to lie in close proximity to the decision boundary of the majority class:

$$\left(w^T \phi(a_{j,c}) - w^T m_{+,c}^\phi \right)^2 \leq \delta, \quad (4)$$

$$\left(\phi(a_{j,c}) - m_-^\phi \right)^T \left(\phi(a_{j,c}) - m_-^\phi \right) \leq \Lambda, \quad (5)$$

where each class index $c \in \{1, 2, \dots, C\}$, the subscript j ranges over $j = 1, 2, \dots, k_c$. $\delta > 0$ is a constant, and $\Lambda > 0$ is a tuning parameter controlling the proximity of generated samples to the majority class boundary. Furthermore, $m_{+,c}^\phi = \frac{1}{l_c} \sum_{j=1}^{l_c} \phi(x_j^+)$ is the feature-space mean of the c -th minority cluster, while m_-^ϕ denotes the mean of the majority class.

According to the Representer's Theorem [29], we assume that w and each $\phi(a_{j,c})$ can be represented by the feature-space images of $\mathcal{X}^- \cup \mathcal{X}^+$. For notational simplicity, denote all training samples by $\{x_p\}_{p=1}^n$, where $n = l + l_-$. To be more specific, $w = \sum_{p=1}^n \alpha_p \phi(x_p)$, and $\phi(a_{j,c}) - m_-^\phi = \sum_{p=1}^n \beta_p^{j,c} \phi(x_p)$, where α_p and $\beta_p^{j,c}$ are scalar-valued parameters in $\mathbb{R}$.

Following Mika et al. [30], we define $M := (M_- - M_+)(M_- - M_+)^T$, $N := \sum_{i \in \{-, +\}} K_i(I - \mathbf{1}_{l_i})K_i^T$. Here, $M_+ \in \mathbb{R}^{n \times 1}$ is a vector whose j -th component is given by $(M_+)_j = \frac{1}{l} \sum_{\ell=1}^l K(x_j, x_\ell^+)$. For each class $i \in \{-, +\}$, $K_i \in \mathbb{R}^{n \times l_i}$ denotes the kernel matrix whose entries are defined as $(K_i)_{r,s} = K(x_r, x_s^i)$, $r \in \{1, \dots, n\}$, $s \in \{1, \dots, l_i\}$. Moreover, I represents the identity matrix of compatible dimension, while $\mathbf{1}_{l_i}$ denotes the

$l_i \times l_i$ matrix whose elements are all equal to $1/l_i$. So for $j \in \{1, 2, \dots, k_c\}$, $c \in \{1, 2, \dots, C\}$, the matrix representation of Equations (4) and (5) are given by

$$\left[\alpha^T K \beta^{j,c} + \alpha^T (M_- - M_+^c) \right]^2 \leq \delta, \quad (6)$$

and

$$(\beta^{j,c})^T K (\beta^{j,c}) \leq \Lambda. \quad (7)$$

Accordingly, the corresponding learning task under class-imbalanced settings can be formulated as the following optimization problem:

$$\max_{\alpha} J(\alpha) = \frac{\alpha^T M \alpha}{\alpha^T N \alpha}, \quad (8)$$

s.t.

$$\left[\alpha^T K \beta^{j,c} + \alpha^T (M_- - M_+^c) \right]^2 \leq \delta, \quad (9)$$

and

$$(\beta^{j,c})^T K (\beta^{j,c}) \leq \Lambda, \quad (10)$$

First, in order to solve the optimization problem (8)–(10), we let $\delta = 0$. The setting indicates that the generated samples $a_{j,c}$ are projected to the same position as the projected mean of the corresponding minority cluster. Under this condition, Equation (8) is equivalently rewritten as

$$\alpha^T K \beta^{j,c} + \alpha^T (M_- - M_+^c) = 0, \quad (11)$$

Although an additional constraint is introduced, it does not impose any essential restriction on the optimization problem, since a relaxed version of the original problem is solved subsequently. By applying the Lagrangian relaxation method [31], the optimization problem (8)–(10) can be equivalently reformulated by incorporating the denominator of the objective function (8) into the constraints, while retaining only the numerator as the new objective function. Consequently, the optimization problem associated with imbalanced classification in (8) can be reformulated as

$$\max_{\alpha} J(\alpha) = \alpha^T M \alpha, \quad (12)$$

s.t.

$$\alpha^T N \alpha \leq R, R > 0 \quad (13)$$

where α denotes an n -dimensional coefficient vector composed of entries $\alpha_1, \alpha_2, \dots, \alpha_n$, i.e., $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n]^T$, and $\beta^{j,c}$ represents another n -dimensional parameter vector defined as $\beta^{j,c} = [\beta_1^{j,c}, \beta_2^{j,c}, \dots, \beta_n^{j,c}]^T$, $(M_-)_j = \frac{1}{l_-} \sum_{\ell=1}^{l_-} K(x_j, x_{\ell}^-)$, $(M_+^c)_j = \frac{1}{l_c} \sum_{\ell=1}^{l_c} K(x_j, x_{\ell,c}^+)$, and K is the $n \times n$ kernel matrix with entries $(K)_{r,s} = K(x_r, x_s)$. Based on the above reformulation, the corresponding Lagrangian function is given by

$$\begin{aligned} Y(\alpha, \beta) = & \alpha^T M \alpha - \sum_{c=1}^C \sum_{j=1}^{k_c} \lambda_j^c \alpha^T K \beta^{j,c} + \lambda_j^c \alpha^T (M_- - M_+^c) \\ & - \sum_{c=1}^C \sum_{j=1}^{k_c} \mu_j^c (\beta^{j,c})^T K (\beta^{j,c}) - \mu_j^c \Lambda \\ & - \gamma \alpha^T N \alpha - \gamma R, \end{aligned} \quad (14)$$

where all parameters satisfy positivity constraints, i.e., $\gamma > 0$, $\lambda_j^c > 0$, and $\mu_j^c > 0$. Next, the optimality conditions are derived by equating the partial derivatives of the Lagrangian with respect to α and $\beta^{j,c}$ to zero:

$$\frac{\partial}{\partial \alpha} Y = 2M\alpha - 2\gamma N\alpha - \sum_{c=1}^C \sum_{j=1}^{k_c} \lambda_j^c (K\beta^{j,c} + (M_- - M_+^c)) = 0, \quad (15)$$

and

$$\frac{\partial}{\partial \beta^{j,c}} Y = \lambda_j^c (K\alpha) + 2\mu_j^c K\beta^{j,c} = 0, \quad (16)$$

By substituting $\beta^{j,c} = -\frac{\lambda_j^c}{2\mu_j^c} \alpha$ derived from Equation (16), into Equation (15), we obtain

$$(M - \gamma N)\alpha = -K\alpha \sum_{c=1}^C \sum_{j=1}^{k_c} \frac{(\lambda_j^c)^2}{4\mu_j^c} - \sum_{c=1}^C \left\{ (M_- - M_+^c) \sum_{j=1}^{k_c} \frac{\lambda_j^c}{2} \right\}. \quad (17)$$

Define $\omega = -\sum_{c=1}^C \sum_{j=1}^{k_c} \frac{(\lambda_j^c)^2}{4\mu_j^c}$, and $\nu^c = -\sum_{j=1}^{k_c} \frac{\lambda_j^c}{2}$. It follows that

$$(M - \gamma N)\alpha - \omega K\alpha = \sum_{c=1}^C \{(M_- - M_+^c)\nu^c\}. \quad (18)$$

Therefore, the vector α satisfies

$$((M - \gamma N) - \omega K)\alpha = \sum_{c=1}^C (M_- - M_+^c)\nu^c. \quad (19)$$

Assume that the coefficient matrix $((M - \gamma N) - \omega K)$ is nonsingular. Then Equation (19) admits a unique solution, and solving the resulting linear system leads to the determination of the optimal coefficients $\alpha^* = [\alpha_1^*, \alpha_2^*, \dots, \alpha_n^*]$. The nonsingularity condition is satisfied when the involved kernel-related matrices are full rank and the regularization term guarantees the positive definiteness of the coefficient matrix. Under these conditions, the optimization problem is well-posed and the resulting estimator is uniquely determined.

Thus, the projection of a test sample x_t onto the direction w can be expressed as follows:

$$\langle w, \phi(x_t) \rangle = \sum_{\ell=1}^l \alpha_\ell^* K(x_\ell, x_t). \quad (20)$$

The solution to (19) yields the coefficients α^* , which define a more compact decision boundary that better encloses the dominant class distribution. Although $\beta^{j,c}$ does not explicitly appear in (19), it influences α implicitly through the Lagrange multipliers λ_j^c and μ_j^c . Therefore, while $\beta^{j,c}$ is not directly involved in prediction, it plays an implicit role in locating missing samples. This mechanism explains how missing data contribute to shaping a more compact majority class boundary.

Next, we present the process of generating synthetic observations according to the newly determined decision boundary and updating the corresponding scatter matrices using these generated samples. Independent observations are drawn according to the estimated density of the existing minority class, with the sampling region restricted according to the separating boundary defined by the optimal coefficient α_* . Let $\hat{F}_{\alpha_*}^u$ be the estimated distribution for minority samples, obtained by modeling $\mathcal{X}^+$ using a Gaussian mixture model composed of u components and applying truncation determined by α_* , i.e.,

$$\hat{F}_{\alpha_*}^u = \frac{1}{u} \sum_{b=1}^u a_b \Psi_b, \quad (21)$$

where $\Psi_b \sim \mathcal{N}(\mu_b, \Sigma_b^2)$, truncated on $\mathcal{X}^+$, $a_b \in [0, 1]$, the set of coefficients $\{a_b\}_{b=1}^u$ indexed by $b = 1, 2, \dots, u$, which satisfy the constraint that their sum equals unity, i.e., $\sum_{b=1}^u a_b = 1$. A set of q independent samples is generated from $\hat{F}_{\alpha_*}^u$ denoted by $\tilde{x}_\ell^+ \sim \hat{F}_{\alpha_*}^u$, for $\ell = 1, 2, \dots, q$. Let $\tilde{\mathcal{A}}$ collect these synthetic samples. The augmented set of minority class samples is constructed by combining the original and synthetic samples as $\tilde{\mathcal{X}}^+ = \mathcal{X}^+ \cup \tilde{\mathcal{A}} = \{x_1^+, \dots, x_l^+, \tilde{x}_1^+, \dots, \tilde{x}_q^+\}$.

Subsequently, the scatter matrices (3) are re-evaluated by using the augmented minority set. A modified optimization problem is obtained, which follows the same form as (8) and (10). In this formulation, the matrices K , N , and M , together with the vectors M_- and M_+^c for $c = 1, 2, \dots, C$, are evaluated using the updated sample sets $\mathcal{X}^-$ and $\tilde{\mathcal{X}}^+$. Solving this problem results in a new optimal coefficient vector α_* . Based on this updated vector, we re-estimate the divergence matrix by refitting the Gaussian mixture model and generating $q \leftarrow \lfloor q/2 \rfloor$ synthetic missing points. This iterative process continues until the condition $q < 1$ is satisfied, at which point the resulting α_* is adopted as the optimal solution.

At every iteration of the process, we adopt the X-means algorithm [32] because the appropriate number of clusters is unknown a priori and can therefore be estimated directly from the data. Nevertheless, other clustering approaches, such as model-based clustering [33], could also be employed. Specifically, X-means is an extension of the k -means algorithm, where the number of clusters denoted by C is determined automatically according to the Bayesian Information Criterion (BIC) [34].

The BIC is also employed to determine the number of Gaussian components in the estimated distribution of minority samples. At each iteration, the corresponding number of mixture components is set as follows:

$$\arg \min_{u \in \mathbb{N}} \text{BIC}(\hat{F}_{\alpha_*}^u), \quad (22)$$

where $\mathbb{N}$ denotes the set of natural numbers and

$$\text{BIC}(\hat{F}_{\alpha_*}^u) = u \log(q) - 2 \log(L) \quad (23)$$

where L represents the likelihood associated with the minority class samples, under the assumption that they are random samples drawn from $\hat{F}_{\alpha_*}^u$.

The observations in $\tilde{\mathcal{A}}$ are put into different groups $c = 1, 2, \dots, C$. Each observation has a group center in the original feature space. The method measures the straight-line distance from each observation to these centers. Each observation is then assigned to the group with the closest center. More precisely, for each $\tilde{x}_\ell^+ \in \tilde{\mathcal{Z}}'$, define a discriminant function $d(c') = \|\tilde{x}_\ell^+ - \bar{x}_{c'}^+\|$; its cluster label is determined as follows:

$$c = \arg \min_{c'} d(c'), \quad (24)$$

where $\bar{x}_{c'}^+ = \frac{1}{l_{c'}} \sum_{\ell=1}^{l_{c'}} x_{\ell, c'}^+$ represents the centroid of cluster c' in the original feature space. Consequently, for each cluster $c = 1, 2, \dots, C$, we obtain q_c newly constructed observations, with the total number of generated samples satisfying $\sum_{c=1}^C q_c = q'$.

The coefficients λ_c^j and μ_c^j in the Lagrangian formulation are adaptively assigned such that their values decrease as the number of minority samples within cluster c increases at the present stage, i.e., $\lambda_c^j = \lambda/l_c$ and $\mu_c^j = \mu/l_c$. This implies that clusters containing only a small number of observations incur a stronger penalty when constraints (6) and (7)

are violated, whereas such violations are penalized less severely in clusters with a larger number of observations. Since clusters containing fewer minority samples are considered to be more severely underrepresented, they are assigned a larger number of perceived absent observations. Accordingly, in cluster c , the number of inferred missing observations, k_c , is set to be inversely proportional to the cluster's current sample size [35].

We apply the ADG algorithm to deal with missing treatment labels. The algorithm classifies and predicts the missing labels. This gives us a complete dataset. In this dataset, all treatment assignments are fully observed. We use a kernel-matching estimator to estimate the ATT. This estimator has a stable normal behavior when the sample size is large. We then build confidence intervals for the treatment effect. These intervals are more accurate and more stable. We present the detailed procedure of the proposed ADG-based kernel-matching method in Algorithm 1. We show the flowchart in Figure 1.

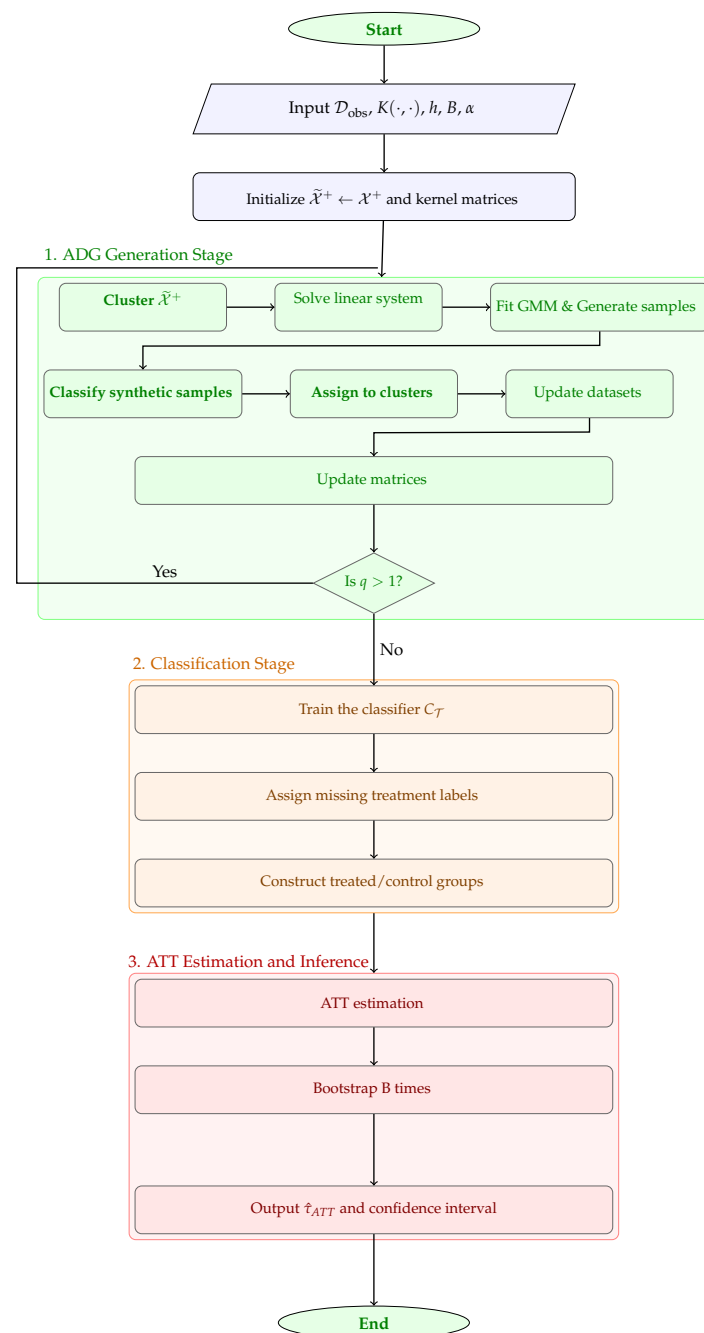


Figure 1. Flowchart of the proposed ADG-based kernel-matching procedure.

Algorithm 1 ADG-Based Kernel Matching for ATT Confidence Interval Estimation

Input: $\mathcal{D}_{\text{obs}}$, kernel function $K(\cdot, \cdot)$, bandwidth h , number of Bootstrap replications B , significance level α

Output: ATT estimator $\hat{\tau}_{ATT}$ and confidence interval $[L, U]$

- 1: Initialize $\tilde{\mathcal{X}}^+ \leftarrow \mathcal{X}^+$ and compute initial matrices $\mathbf{K}, \mathbf{N}, \mathbf{M}, \mathbf{K}_i, \mathbf{M}_i$ for $i \in \{-, +\}$.
- 2: **while** $q > 1$ **do**
- 3: Cluster $\tilde{\mathcal{X}}^+$ into C groups and determine C by minimizing the BIC
- 4: Let $\mu_c^j = \frac{\mu}{l_c}$ and $\lambda_c^j = \frac{\lambda}{l_c}$ for $j = 1, 2, \dots, k_c, c = 1 \dots C$, where k_c is chosen proportionally to $1/l_c$ subject to the constraint $\sum_{c=1}^C k_c = k$.
- 5: Compute $(\mathbf{M}_+^c)_j = \frac{1}{l_c} \sum_{\ell=1}^{l_c} K(\mathbf{x}_j, \mathbf{x}_{\ell,c}^+)$, $\omega = -\sum_{c=1}^C \sum_{j=1}^{k_c} \frac{(\lambda_j^c)^2}{4\mu_j^c}$ and $\nu^c = -\sum_{j=1}^{k_c} \frac{\lambda_j^c}{2}$.
- 6: Solve linear system (19) to obtain α_*
- 7: Fit a Gaussian mixture model to $\mathcal{X}^+$, where the number of components u is chosen to minimize the BIC in (22).
- 8: Generate q synthetic samples $\tilde{\mathcal{A}} = \{\tilde{\mathbf{x}}_1^+, \dots, \tilde{\mathbf{x}}_q^+\}$, based on the resulting Gaussian mixture model obtained above.
- 9: Use the α_* to determine, based on (20), whether each synthetic sample $\tilde{\mathbf{x}}_\ell^+$, for $\ell = 1, 2, \dots, q$, should be assigned to class +1 or not. Define $\tilde{\mathcal{A}}'$ as a subset of $\tilde{\mathcal{A}}$, consisting of observations incorporated into the minority class, i.e., $\tilde{\mathcal{A}}' = \{\mathbf{x}_{l+1}^+, \mathbf{x}_{l+2}^+, \dots, \mathbf{x}_{l+q'}^+\} \subset \tilde{\mathcal{A}}$.
- 10: Assign the newly generated observations to clusters according to (24). For each cluster $c \in \{1, 2, \dots, C\}$, let q_c denote the number of samples in $\tilde{\mathcal{A}}$ that are assigned to cluster c .
- 11: Update $\tilde{\mathcal{X}}^+ \leftarrow \tilde{\mathcal{X}}^+ \cup \tilde{\mathcal{A}}'$ and $\tilde{\mathcal{X}} \leftarrow \mathcal{X}^- \cup \tilde{\mathcal{X}}^+$.
- 12: Update $(\mathbf{M}_-)_j \leftarrow \frac{1}{l_-} \sum_{\ell=1}^{l_-} K(\mathbf{x}_j, \mathbf{x}_\ell^-)$ and $(\mathbf{M}_+)_j \leftarrow \frac{1}{l_+q'} \sum_{\ell=1}^{l_++q'} K(\mathbf{x}_j, \mathbf{x}_\ell^+)$ for $\mathbf{x}_j \in \tilde{\mathcal{X}}$.
- 13: Update $(\mathbf{M}_+^c)_j = \frac{1}{l_c+q_c} \sum_{\ell=1}^{l_c+q_c} K(\mathbf{x}_j, \mathbf{x}_{\ell,c}^+)$ for $\mathbf{x}_j \in \tilde{\mathcal{X}}$.
- 14: Update $(\mathbf{K})_{r,s} \leftarrow K(\mathbf{x}_r, \mathbf{x}_s)$, for $r, s \in \{1, \dots, n+q'\}$.
- 15: $(\mathbf{K}_+)^{r,s} = K(\mathbf{x}_r, \mathbf{x}_s^+)$ for $r \in \{1, 2, \dots, n+q\}, s \in \{1, 2, \dots, l+q\}$,
 $(\mathbf{K}_-)^{r,s} = K(\mathbf{x}_r, \mathbf{x}_s^-)$ for $r \in \{1, 2, \dots, n+q\}, s \in \{1, 2, \dots, l\}$.
- 16: Update $\mathbf{M} \leftarrow (\mathbf{M}_- - \mathbf{M}_+)(\mathbf{M}_- - \mathbf{M}_+)$.
- 17: Update $\mathbf{N} \leftarrow \sum_{i \in \{-, +\}} \mathbf{K}_i (\mathbf{I} - \mathbf{1}_i) \mathbf{K}_i^T$.
- 18: Update $q \leftarrow \lfloor q/2 \rfloor, l \leftarrow |\mathcal{X}^+|, n \leftarrow |\tilde{\mathcal{X}}|$.
- 19: **end while**
- 20: Obtain the optimal projection coefficients α_* corresponding to the discriminant direction that optimizes the Fisher criterion in the kernel-induced feature space.
- 21: Construct the $n \times 1$ kernel vector κ_x by evaluating the kernel function between $\mathbf{x}$ and all elements in the training set augmented with the synthetic minority samples $\tilde{\mathcal{X}}$, where the ℓ -th entry is given by $(\kappa_x)_\ell = K(\mathbf{x}_\ell, \mathbf{Sx})$ for $\mathbf{x}_\ell \in \tilde{\mathcal{X}}$.
- 22: Train a one-dimensional binary classifier $C_{\mathcal{T}}(\text{SVM})$ on $\mathcal{T} = \{(h(\mathbf{x}_\ell; \alpha_*), y_\ell) : \mathbf{x}_\ell \in \tilde{\mathcal{X}}, \ell = 1, \dots, n\}$, where $h(\mathbf{x}_\ell; \alpha_*) = \alpha_*^T \kappa_{\mathbf{x}_\ell}$.
- 23: Obtain the decision threshold v_* . Classify a new sample $\mathbf{x}$ as +1 if $h(\mathbf{x}; \alpha_*) > v_*$, and as -1 otherwise.
- 24: Use the ADG procedure to assign treatment status for units with missing treatment indicators. Construct the treated group $\mathcal{I}_1 = \{i : W_i = 1\}$ and control group $\mathcal{I}_0 = \{i : W_i = 0\}$
- 25: Compute the ATT estimate according to (1).
- 26: **for** $b = 1$ to B **do**
- 27: Draw a bootstrap sample with replacement
- 28: Compute $\hat{\tau}_{ATT}^{(b)}$
- 29: **end for**
- 30: Compute the empirical quantiles $L = Q_{\alpha/2}, U = Q_{1-\alpha/2}$
return $\hat{\tau}_{ATT}$ and confidence interval $[L, U]$

4. Monte Carlo Simulation Experiments

4.1. Simulation Setup

This section presents a Monte Carlo simulation study. The study evaluates the finite-sample performance of the proposed method. It also compares the proposed method with several alternative approaches. The results show that the ADG-based kernel-matching estimator performs well in estimating the ATT. It achieves high accuracy in most settings. For comparison, we also include several commonly used ATT estimation methods:

- Kernel matching [6] (denoted as “KM”).
- Kernel matching + SVM (denoted as “KM + SVM”).
- Kernel matching + SMOTE [13] (denoted as “KM + SMOTE”).
- Kernel matching + Borderline-SMOTE [14] (denoted as “KM + BSMOTE”).
- The ADG-based kernel-matching estimator for the ATT, introduced in Section 3, is referred to as the “proposed” method. Since kernel-based estimators are known to be sensitive to the choice of bandwidth and kernel function, the specification of these tuning parameters plays an important role in practice. A detailed discussion of the corresponding selection criteria is provided in the following subsection.

We consider four different settings. These settings are listed below.

- Setting I:
 - For $w \in \{0, 1\}$, the covariate $X \in \mathbb{R}^2$ follows a bivariate standard normal distribution conditional on $W = w$:

$$X \mid W = w \sim \mathcal{N}(\mathbf{0}, I_2);$$

- Let $\mu_1(x) = x_1$, where x_1 denotes the first component of x . Given $X = x$ and $W = 0$, the potential outcome $Y(0)$ follows a standard normal distribution:

$$(Y(0) - \mu_1(x)) \sim \mathcal{N}(0, 1).$$

- Given $X = x$ and $W = 1$,

$$(Y(1) - \mu_1(x)) \sim \mathcal{N}(0, 1).$$

- Setting II:
 - For $w \in \{0, 1\}$, the covariate $X \in \mathbb{R}^2$ follows a bivariate standard normal distribution conditional on $W = w$:

$$X \mid W = w \sim U([-1, 1]^2);$$

- Given $X = x$ and $W = 0$, the potential outcome $Y(0)$ follows a standard normal distribution:

$$Y(0) \mid X = x \sim \mathcal{N}(0, 1);$$

- Let $\mu_1(x) = x_1$, where x_1 denotes the first component of x . Given $X = x$ and $W = 1$,

$$(Y(1) - \mu_1(x)) \sim \mathcal{N}(0, 1).$$

- Setting III:
 - For $w \in \{0, 1\}$, the covariate $X \in \mathbb{R}^2$ follows a bivariate standard normal distribution conditional on $W = w$:

$$X \mid W = w \sim \mathcal{N}(\mathbf{0}, I_2);$$

- Given $X = x$ and $W = 0$, the potential outcome $Y(0)$ follows a standard normal distribution:

$$Y(0) \mid X = x \sim \mathcal{N}(0, 1);$$

- Let $\mu_1(x) = 0$, where x_1 denotes the first component of x . Given $X = x$ and $W = 1$,

$$(Y(1) - \mu_1(x)) \sim \mathcal{N}(0, 1).$$

- Setting IV:

- For $w \in \{0, 1\}$, the covariate X given $W = w$ follows a standard normal distribution:

$$X \mid W = w \sim \mathcal{N}(\mathbf{0}, I_2);$$

- Given $X = x$ and $W = 0$, the potential outcome under control satisfies

$$Y(0) \mid X = x \sim \mathcal{N}(0, 1).$$

- Let

$$\mu_1(x) = \frac{1}{4}(Q_8(\Phi(x_1)) - 8),$$

where x_1 denotes the first component of x , $\Phi(x)$ denotes the cumulative distribution function of the standard normal distribution, and $Q_8(\cdot)$ is the quantile function of a chi-squared distribution with eight degrees of freedom. Then, given $X = x$ and $W = 1$,

$$(Y(1) - \mu_1(x)) \sim \mathcal{N}(0, 1).$$

For Settings I–IV, the simulated observations are produced from the following data generating mechanisms. These settings are constructed based on the framework developed by Ferman [24], with minor adaptations introduced for the present study. The four settings are designed to progressively increase the level of difficulty in order to comprehensively evaluate the robustness of the proposed method. Setting I serves as a baseline with a well-specified linear structure and identical covariate distributions across treatment groups. Setting II introduces distributional changes and heterogeneous treatment effects, assessing robustness to covariate shift and structural asymmetry. Setting III considers the case where covariates are irrelevant to the outcome, examining whether the method avoids spurious treatment effects. Finally, Setting IV imposes a highly nonlinear and heavy-tailed transformation, creating a challenging scenario to evaluate performance under severe model misspecification. We generate datasets with initial sample sizes of $N_0 = 1000$ and $N_1 = 500$. To better reflect the scarcity of treated units commonly encountered in real-world applications, treatment labels are randomly missing, resulting in an observed control sample size of $N_0 = 500$ and an observed treated sample size of $N_1 \in \{25, 50, 100\}$. For each scenario, we conducted 1000 Monte Carlo simulations to estimate the ATT.

4.2. A Study of Bandwidth and Kernel Function Selection for Kernel-Based Matching

Since the performance of the kernel-matching estimator for the ATT depends critically on the choice of kernel function $K(\cdot)$ and bandwidth h , investigating appropriate selections for both is of substantial theoretical and practical importance. In this section, we conduct a systematic analysis of kernel and bandwidth choices through Monte Carlo simulations.

In the simulation study, we evaluate the performance of the kernel-based ATT estimator in terms of bias and RMSE. The data generating process follows Setting I, with a fixed sample size of $N_0 = 1000$ for the control group and $N_1 = 500$ for the treated group. In this study, two representative kernel functions, namely the Gaussian kernel and the

Epanechnikov kernel, are employed for analysis. The bandwidth parameter is set to $h \in \{0.01, 0.03, 0.05, 0.07, 0.09\}$. All reported results are based on 1000 Monte Carlo replications. The simulation results for bias and RMSE of the ATT estimator are presented in Figures A1 and A2 in Appendix B.

When comparing the Gaussian and Epanechnikov kernels, Figures A1 and A2 show that the estimator based on the Epanechnikov kernel consistently achieves lower bias and RMSE than that based on the Gaussian kernel. This advantage is particularly pronounced when the bandwidth is small. As further illustrated in Figure A1, this pattern holds across most bandwidth values $h \in \{0.01, 0.03, 0.05, 0.07, 0.09\}$, with the difference becoming more evident at smaller bandwidths. These findings indicate that the Epanechnikov kernel yields higher efficiency, especially in settings where control of rejection rates is important. This superior performance can be attributed to its favorable weighting structure and its widespread practical applicability, which together contribute to more stable and reliable estimation in kernel matching. Moreover, the Epanechnikov kernel demonstrates robustness across a relatively wide range of bandwidth values, making it particularly suitable for applications emphasizing estimation accuracy.

Regarding bandwidth selection, as shown in Figures A1 and A2, the estimator is sensitive to the choice of bandwidth for both kernel functions. A small bandwidth leads to high variance, while a large bandwidth induces excessive smoothing and reduces estimation accuracy. Based on Theorem 4 and Corollary 2 of Ding et al. [6], which establish the asymptotic normality of the kernel-matching ATT estimator, the estimation error converges to zero as $N \rightarrow \infty$ and $h \rightarrow 0$, provided that $Nh^4 \rightarrow 0$. Consistent with this theoretical result, and supported by the simulation evidence reported in Appendix B (Figures A1 and A2) regarding the choice of kernel function and bandwidth, the condition $Nh^4 \rightarrow 0$ implies the guideline $h \leq N^{-1/4}$. Motivated by this requirement, and further supported by the simulation results, we adopted the bandwidth choice $h = N^{-2/3}$ for the final implementation. Given the sample size $N = 1500$, we therefore selected the Epanechnikov kernel with bandwidth $h = 0.01$ for the subsequent numerical studies and empirical analysis.

4.3. Performance Analysis of the Proposed Estimation

As shown by the results in Appendix B, we employ the Epanechnikov kernel with a bandwidth of $h = 0.01$ for the settings with sample sizes $N_0 = 1000$ and $N_1 = 500$. To reflect realistic scenarios with missing treatment labels, we further impose random missingness on the treatment indicators, resulting in observable treated group sizes $N_1 \in \{25, 50, 100\}$, while the control group size remains fixed at $N_0 = 1000$. This simulation design is conducted under Settings I–IV. We evaluate and compare the point estimation performance of the ATT across five methods: KM, KM + SVM, proposed, KM + SMOTE, and KM + BSMOTE, in terms of bias, standard deviation (SD), and root mean square error (RMSE). For a more comprehensive evaluation of interval estimation, 95% confidence intervals are constructed for the estimators under consideration using bootstrap techniques. In particular, the percentile bootstrap approach [36] is employed, with 1000 bootstrap samples generated to obtain the corresponding interval estimates. The interval estimation properties of the ATT are further investigated by examining the empirical coverage probabilities (CPs) and average confidence interval widths (AWs) associated with the four competing estimators. Complete simulation findings for Settings I–IV are documented in Appendix C (Tables A1–A4).

In the estimation of ATT, the proposed ADG-kernel-matching method yields bias and RMSE values that are generally comparable to those of competing approaches across most scenarios, despite not consistently attaining the minimum values. From Tables A1–A4, across all four Settings I–IV, the proposed method consistently exhibits the best or near-best

overall performance. It outperforms or significantly surpasses other methods in terms of RMSE, CP, and AW. Specifically, under Setting I, the proposed method performs optimally across all sample sizes: its bias is close to zero, and both SD and RMSE are the lowest. The CP remains stable between 0.96 and 0.97, closely aligning with the nominal level of 0.95. In contrast, KM + SMOTE and KM + BSMOTE exhibit extremely poor coverage, indicating complete failure. Under Setting II, the proposed method continues to perform strongly, with RMSE substantially lower than KM and KM + SVM. For example, when $N_1 = 100$, the RMSE of the proposed method is 0.1107, compared to 0.1271 for KM and 0.1289 for KM + SVM. The CP converges smoothly from 0.960 to 0.950, effectively matching the nominal level. Under Settings III and IV, the proposed method maintains the lowest RMSE and SD, while its coverage probability consistently ranges between 0.93 and 0.97. In contrast, KM and KM + SVM exhibit slightly lower or less stable coverage, whereas KM + SMOTE and KM + BSMOTE show severely deficient coverage, rendering them unreliable.

It is worth noting that in some small-sample scenarios (e.g., $N_1 = 25$ in Settings II–IV), KM + SMOTE and KM + BSMOTE may yield slightly lower SD and RMSE than the proposed method. However, their coverage probabilities are severely inadequate, indicating substantial distortion in confidence intervals and making them unsuitable for valid statistical inference. These seemingly improved error metrics come at the cost of unreliable interval estimation and thus lack practical relevance. In contrast, the proposed method achieves strong accuracy while consistently providing reliable coverage. Moreover, the AW of the proposed method decreases appropriately as the sample size increases across all settings (e.g., from 0.7203 to 0.4242 in Setting I, and from 0.6520 to 0.3017 in Setting IV), reflecting effective control of model complexity and avoiding both overfitting and underfitting. SMOTE and BSMOTE generate synthetic samples by linear interpolation in the feature space. However, they do not consider the cluster structure of the data. They also ignore the uncertainty in the data generation process. Therefore, the generated samples may not accurately reflect the original data distribution, especially near the decision boundary. This can produce biased variance estimates and lead to lower coverage probabilities for the resulting confidence intervals.

In contrast, ADG constructs synthetic samples in a cluster-aware manner by explicitly accounting for the local structure of minority data through the quantities k_c and q_c . This design helps preserve the intrinsic distributional characteristics of each cluster and yields more stable resampling behavior, which in turn improves the accuracy of variance estimation and leads to better coverage probability.

In summary, the proposed method is the most reliable approach for ATT estimation across all four simulation settings. It achieves low estimation error, high stability, coverage probabilities close to the nominal level, and well-controlled model complexity. While KM and KM + SVM provide reasonable benchmarks, they are inferior in both accuracy and reliability, and KM + SMOTE and KM + BSMOTE should be avoided due to severely distorted confidence intervals. In more complicated scenarios, such as Setting IV, both point and interval estimation can be substantially improved through the use of the proposed estimator.

5. Empirical Application

In this section, we apply the proposed method to a real dataset that investigates the effect of maternal smoking during pregnancy on infant birth weight (MSBW) [37,38].

Analysis of Data from the MSBW Dataset

To assess the performance of the proposed estimator, the ATT is estimated using both treated and control groups. The dataset contains ($N = 4642$) observations. The variables

considered in the analysis include: bweight (the infant's birth weight), mb smoke (equal to 1 if the mother smoked during pregnancy and 0 otherwise), mhis p (equal to one if the mother is Hispanic and 0 otherwise), deadkids (number of previous child deaths), alcohol (equal to one if the mother consumed alcohol during pregnancy and 0 otherwise), mrace (equal to one if the mother is White and 0 otherwise), fedu (father's education level), and prenatal (whether the mother received her first prenatal care visit within the first three months of pregnancy).

In this dataset, the outcome variable bweight denotes birth weight, and the treatment variable mb smoke indicates whether the mother smoked during pregnancy. The set of potential confounders includes mhis p, deadkids, alcohol, mrace, fedu, and prenatal. The sample consists of $N_1 = 864$ treated individuals and $N_0 = 3778$ control individuals. Table 1 summarizes the descriptive characteristics of the MSBW dataset and provides the p -values obtained from t -tests assessing differences in covariate means between the treatment and control groups. The results suggest that the covariate balance between the two groups is generally inadequate across the variables. For all variables, the null hypothesis of equal means is rejected, indicating that the treated and control groups have statistically significant differences in their average characteristics.

Table 1. Descriptive statistics for the MSBW dataset variables.

Variable	Experimental Data		p -Value of Two-Sample t -Test
	Treated (864) Mean (SD)	Control (3778) Mean (SD)	Treated/Control Experiments
mhis p	0.024 (0.154)	0.036 (0.187)	0.04
deadkids	0.318 (0.466)	0.246 (0.431)	0.00
alcohol	0.091 (0.288)	0.019 (0.136)	0.01
mrace	0.809 (0.393)	0.847 (0.359)	0.01
fedu	10.703 (4.097)	12.673 (3.481)	0.01
prenatal	1.307 (0.630)	1.178 (0.473)	0.00

We impose random missingness on treatment labels, yielding observed sample sizes of $N_1 \in \{30, 50, 100\}$ for the treated group and $N_0 = 300$ for the control group. Table 2 presents the estimated ATT of maternal smoking during pregnancy on infant birth weight obtained using different methods under varying treated sample sizes. The performance of the proposed ADG-kernel-matching estimator (denoted as “proposed”) is evaluated against three competing approaches: the kernel-matching-based SVM estimator (“KM + SVM”), the SMOTE-based estimator (“KM + SMOTE”), and the BSMOTE-based estimator (“KM + BSMOTE”). The proposed method employs the Epanechnikov kernel with a bandwidth parameter of ($h = 0.01$). For all methods under consideration, estimation uncertainty is quantified by reporting standard errors together with 95% bootstrap confidence intervals obtained from ($B = 1000$) bootstrap samples.

Across all settings, the proposed method consistently yields ATT estimates ranging from -265.9 g to -274.5 g, suggesting that maternal smoking has a markedly adverse impact on birth weight. These estimates are remarkably close to those reported by Almond et al. [39] and Cattaneo [37], which show that maternal smoking reduces infant birth weight by approximately 200–250 g using propensity score subclassification and regression adjustment methods. The similarity between our findings and those of previous studies provides empirical support for the validity of the proposed approach. In comparison, competing methods often yield confidence intervals that include zero or display greater overlap with zero, indicating less stable and statistically weaker estimates. Notably, the KM + SMOTE and KM + BSMOTE approaches tend to substantially underestimate

the magnitude of the negative effect of maternal smoking. These findings suggest that conventional oversampling techniques are insufficient to recover the true treatment effect under severe class imbalance. Importantly, the 95% confidence intervals obtained with the proposed method remain entirely below zero across all sample sizes, offering strong evidence that maternal smoking during pregnancy has a detrimental impact on infant birth weight. Overall, these results demonstrate that the proposed method not only produces estimates consistent with established empirical evidence but also provides superior robustness and reliability, particularly in settings with limited treated samples and imbalanced data.

Table 2. Estimates of the ATT, together with their SE and associated 95% CI, are reported for the MSBW dataset under different estimation methods.

Sample Size	Method	Estimate (SE)	95% CI
$N_1 = 30$	KM	−174.6 (110.5)	(−383.3, 49.1)
	KM + SVM	−295.3 (80.5)	(−283.6, 29.9)
	Proposed	−265.9 (100.2)	(−452.4, −70.1)
	KM + SMOTE	−48.0 (51.2)	(−139.9, 58.76)
	KM + BSMOTE	−47.2 (43.6)	(−454.1, 81.5)
$N_1 = 50$	KM	−175.6 (88.1)	(−337.0, 8.0)
	KM + SVM	−182.6 (56.3)	(−213.5, 1.8)
	Proposed	−272.6 (87.7)	(−389.5, −108.9)
	KM + SMOTE	−48.8 (50.12)	(−142.3, 54.5)
	KM + BSMOTE	−47.4 (54.3)	(−147.6, 59.8)
$N_1 = 100$	KM	−179.3 (65.8)	(−296.2, −39.1)
	KM + SVM	−122.9 (44.0)	(−143.9, 6.8)
	Proposed	−274.5 (63.6)	(−368.4, −130.7)
	KM + SMOTE	−49.8 (43.5)	(−128.9, 46.6)
	KM + BSMOTE	−45.4 (47.6)	(−135.4, 54.4)

6. Concluding Remarks

This paper investigates the problem of treatment effect estimation in observational studies with missing treatment labels. By integrating the ADG with kernel matching, we propose a unified framework that combines synthetic data generation with label recovery to alleviate the imbalance induced by missing treatment information.

Extensive comparisons with classical methods in both simulation studies and real data analyses demonstrate the superior performance of the proposed method. Beyond the specific numerical results reported in this study, our findings suggest that generative oversampling techniques can serve as a useful tool for improving the precision of causal effect estimation when treatment information is incomplete. In this sense, the proposed approach contributes to the growing literature on causal inference under data imperfections and highlights the potential of combining modern data synthesis methods with matching-based estimators.

Several limitations should be acknowledged. First, the performance of the proposed procedure may be sensitive to the tuning parameter λ , which in turn affects the accuracy of label recovery. Although satisfactory performance is observed in our empirical studies, developing data-driven methods for selecting λ remains an important direction for future research. Moreover, further methodological improvements are needed to extend the proposed algorithm to a broader class of observational data structures, including high-dimensional covariates, heterogeneous populations, and more complex missing-data mechanisms, thereby enhancing its applicability in practical causal inference problems. Recent advances in generative oversampling methods, such as GAN-based and diffusion-based approaches, have demonstrated strong performance in imbalanced learning. While

this study focuses on the ADG framework, future work may consider integrating these modern generative models into causal inference procedures to further improve treatment effect estimation under severe imbalance and incomplete treatment information. In addition, the computational cost of the proposed ADG algorithm grows as the sample size increases. This is mainly because the method needs to create synthetic samples and train a classification model. As a result, the method may have difficulty dealing with very large datasets. At the same time, the algorithm can be easily run in parallel. The synthetic data generation step can be split into smaller parts. The model training step can also be done in batches or on different machines. Modern machine learning tools can further reduce the running time in practice.

Author Contributions: Conceptualization, C.D.; methodology, C.D.; software, C.D. and Z.L.; validation, C.D. and Z.L.; formal analysis, C.D.; investigation, C.D.; resources, C.D. and Z.L.; data curation, C.D. and Z.L.; writing—original draft preparation, C.D. and L.S.; writing—review and editing, C.D., L.S. and Z.L.; visualization, C.D., L.S. and Z.L.; supervision, C.D. and Z.L.; project administration, C.D. and Z.L.; funding acquisition, C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author due to the terms and conditions of the original data providers, as well as institutional policies.

Acknowledgments: We confirm that an AI tool (ChatGPT GPT-5.5) was used during the preparation of the manuscript. Specifically, it was used solely for English-language editing and grammatical corrections. No AI tools were used for content generation, data analysis, or drawing scientific conclusions.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Generalization Error Bounds for ADG

In this section, we derive bounds on the generalization error of the ADG algorithm to ensure reliable estimation of the ATT and its confidence intervals.

The generalization error [40] is defined as the expected loss on unseen data drawn from the same distribution $\mathcal{G}$. For a decision function h and loss function $\mathcal{L}$, it is given by $\mathbb{E}_x\{\mathcal{L}(h)\}$. We denote by $B_{\mathcal{D}}$ the one-dimensional binary classifier adopted in ADG. This classifier is trained on the dataset $\mathcal{D} = \{(h(\mathbf{x}_\ell; \boldsymbol{\alpha}_F), y_\ell) \mid \mathbf{x}_\ell \in \mathcal{X}^- \cup \mathcal{X}^+, \ell = 1, \dots, n\}$ where each sample consists of a transformed representation $h(\mathbf{x}_\ell; \boldsymbol{\alpha}_F)$ paired with its corresponding label y_ℓ and $\boldsymbol{\alpha}_F$ is the optimal solution obtained from the KFD optimization problem, and $h(\mathbf{x}_\ell; \boldsymbol{\alpha}_F) = \boldsymbol{\alpha}_F^\top \boldsymbol{\kappa}_{\mathbf{x}_\ell}$. Let v_F denote the decision threshold of $B_{\mathcal{D}}$. For a test point $\mathbf{x}_t$, the KFD classifier assigns a label to the test sample $\mathbf{x}_t$ according to the value of the decision function $h(\mathbf{x}_t; \boldsymbol{\alpha}_F)$. Specifically, the output is given by

$$O_{\text{KFD}}(\mathbf{x}_t) = \begin{cases} 1, & \text{if } h(\mathbf{x}_t; \boldsymbol{\alpha}_F) > v_F, \\ -1, & \text{otherwise.} \end{cases}$$

In other words, the instance is classified into the positive class when the decision score exceeds the threshold v_F , and into the negative class otherwise. Using the law of total probability, the generalization error of KFD can be decomposed into class-weighted conditional error terms. Define the classifier as $\hat{y}_t = \text{sign}(v_F - h(\mathbf{x}_t; \boldsymbol{\alpha}_F))$; then, the overall misclassification rate can be written as

$$\text{err}_K = \pi_- \mathbb{P}(\hat{y}_t = 1 \mid y_t = -1) + \pi_+ \mathbb{P}(\hat{y}_t = -1 \mid y_t = 1),$$

where the two terms correspond respectively to the false positive and false negative probabilities weighted by the class priors where π_i denotes the prior probability of class $i \in \{-1, 1\}$.

Lemma A1 (Generalization Error Bound for KFD [41]). *Suppose that, in the feature space, data from each class follow a Gaussian distribution with means μ_i and a common covariance matrix Σ for $i \in \{-1, 1\}$. Let the training sample size be $n = l_+ + l_-$, where l_i denotes the number of samples from class i , and let π_i be the corresponding class prior. For every $\rho \in (0, 1)$, the generalization error associated with KFD is bounded by*

$$\mathbb{P}\left(\text{err}_K \leq \sum_{i \in \{-1, 1\}} \pi_i \Phi\left(-2\left[g(\bar{\tau}(\epsilon))\Pi - \sqrt{\frac{n}{l_i}}\left(1 + \sqrt{\frac{2}{n} \log \frac{4}{\rho}}\right)\right]\right)\right) \geq 1 - \rho, \quad (\text{A1})$$

where $\Pi = \left[\sqrt{\frac{\|\mu_+ - \mu_-\|^2}{\lambda_{\max}(\Sigma)} + \frac{n}{l_- l_+} \frac{\text{tr}(\Sigma)}{\lambda_{\max}(\Sigma)}} - \sqrt{\frac{2n}{l_- l_+} \log \frac{4}{\rho}}\right]_+$. Here, $g(r) = \frac{\sqrt{r}}{1+r}$, $[x]_+ = \max(0, x)$, $\lambda_{\max}(\Sigma)$ denotes the largest eigenvalue of Σ , and Φ is the cumulative distribution function of the standard normal distribution. Moreover, $\bar{\tau}(\epsilon) = \frac{\lambda_{\max}(\Sigma)}{\eta} \left(1 + \sqrt{\frac{n-2}{n}} + \frac{\epsilon}{\sqrt{n}}\right)^2 + \tau(\Sigma_n)$, where $\epsilon = \sqrt{2 \log \frac{4}{\rho}}$, and the quantity $\tau(\Sigma_n)$ refers to the condition number corresponding to Σ_n , the sample covariance matrix, and $\eta > 0$ is a regularization constant ensuring nonsingularity.

From Lemma A1, $g(\cdot)$ is monotonically decreasing for $r \geq 1$, implying that a smaller $\bar{\tau}(\epsilon)$ yields a tighter upper bound. With the regularization parameter η fixed, $\bar{\tau}(\epsilon)$ is mainly determined by $\tau(\Sigma_n)$ and the sample size n . As n increases, (A1) becomes tighter when other quantities are held constant. This behavior naturally arises in synthetic data generation, particularly in ADG, where additional observations are generated at each iteration. From Theorem A1, as the sample size grows, the condition number of the covariance matrix in the feature space tends to decrease. Consequently, under the ADG framework, better numerical conditioning results in a tighter upper bound for the generalization error. An important theoretical problem is to characterize the behavior of the covariance operator projected onto the observation-induced Hilbert space as new observations are incorporated into the dataset.

Let $\Sigma_n = P_n \Sigma P_n^T$, where P represents the orthogonal projection operator onto the subspace of the Hilbert space generated by the observed samples. It is further assumed that the feature representations corresponding to the observations are linearly independent. Then the projection operator can be expressed as $P_n = \left(X_n^{\phi T} X_n^{\phi}\right)^{-\frac{1}{2}} X_n^{\phi T}$, in which X_n^{ϕ} is constructed by concatenating feature vectors $\phi(x_\ell)$ as column entries, where $x_\ell \in \mathcal{X}^- \cup \mathcal{X}^+$. When a new observed data point x_{n+1} is incorporated into the training set, the projection space is expanded accordingly, leading to a new projection matrix P_{n+1}^{ϕ} . A natural question is how the covariance operator projected onto this data-induced Hilbert space evolves as additional observations are included. Theorem A1 addresses this issue by characterizing the relationship between the sample size and the condition number of the projected covariance operator. Specifically, Theorem A1 shows that, as the number of training samples increases, the condition number of the covariance operator associated with the feature subspace decreases. Consequently, the projected feature space becomes better conditioned, which leads to a tighter generalization error bound for the proposed ADG method.

Theorem A1 (Condition Number Decreases with Sample Size). *Assume that $\mathcal{H}$ be a separable Hilbert space, and $\Sigma : \mathcal{H} \rightarrow \mathcal{H}$ be a self-adjoint positive semidefinite covariance operator. Let $\{x_\ell\}_{\ell=1}^n \subset \mathcal{X}$ and define $\zeta_\ell = \phi(x_\ell)$, $\ell = 1, \dots, n$, where the feature mappings $\{\zeta_\ell\}$ are linearly*

independent. Define the finite-dimensional subspace $\mathcal{H}_n = \text{span}\{\zeta_1, \dots, \zeta_n\}$, and let P_n denote the orthogonal projection onto $\mathcal{H}_n$. The projected covariance operator is defined as

$$\Sigma_n = P_n \Sigma P_n.$$

Then, the condition number

$$\tau(\Sigma_n) = \frac{\lambda_{\max}(\Sigma_n)}{\lambda_{\min}(\Sigma_n)}$$

satisfies

$$\tau(\Sigma_{n+1}) < \tau(\Sigma_n), \quad \forall n \in \mathbb{N}.$$

Proof. Following [35], let $x_\ell \in \mathcal{X}$, $\ell = 1, \dots, n$, and define $\zeta_\ell = \phi(x_\ell) \in \mathcal{H}$, where $\mathcal{H}$ is a separable Hilbert space induced by feature map ϕ . Let $\mathcal{H}_n = \text{span}\{\zeta_1, \dots, \zeta_n\}$, and assume $\zeta_\ell \sim \mathcal{N}(\mu, \Sigma)$ in $\mathcal{H}$. Define the projected covariance operator $\Sigma_n = P_n \Sigma P_n^T$, where P_n is the orthogonal projection onto $\mathcal{H}_n$, given by $P_n = \left(X_n^{\phi T} X_n^\phi \right)^{-1/2} X_n^{\phi T}$, $X_n^\phi = [\zeta_1, \dots, \zeta_n]$.

After appropriate rotation and scaling, assume $X_p^{\phi T} X_p^\phi = I$. Then

$$P_{n+1} = \left(X_{n+1}^\phi \right)^T = \left[\left(X_n^\phi \right)^T \mid \zeta_{n+1}^T \right]$$

and hence

$$\Sigma_{n+1} = \begin{bmatrix} \Sigma_n & \left(X_n^\phi \right)^T \Sigma \zeta_{n+1} \\ \zeta_{n+1}^T \Sigma X_n^\phi & \zeta_{n+1}^T \Sigma \zeta_{n+1} \end{bmatrix}$$

By defining $\|\zeta_{n+1}\|^2 = \zeta_{n+1}^T \Sigma \zeta_{n+1}$ and invoking the result established in Appendix A of Pourhabib et al. [35], one can bound the largest and smallest eigenvalues of Σ_{n+1} as follows:

$$\lambda_{\max}(\Sigma_{n+1}) \leq \lambda_{\max}(\Sigma_n) + \|\zeta_{n+1}\|^2, \quad \lambda_{\min}(\Sigma_{n+1}) \geq \lambda_{\min}(\Sigma_n) + \|\zeta_{n+1}\|^2$$

Therefore, for the condition number $\tau(A) = \lambda_{\max}(A) / \lambda_{\min}(A)$,

$$\tau(\Sigma_{n+1}) \leq \frac{\lambda_{\max}(\Sigma_n) + \|\zeta_{n+1}\|^2}{\lambda_{\min}(\Sigma_n) + \|\zeta_{n+1}\|^2} < \frac{\lambda_{\max}(\Sigma_n)}{\lambda_{\min}(\Sigma_n)} = \tau(\Sigma_n).$$

Thus, the condition number decreases as n increases. $\square$

In summary, the ADG algorithm incorporates a synthetic data generation mechanism into the KFD framework. Based on the aforementioned theoretical analysis characterizing the decrease in generalization error, the introduction of artificially generated samples enriches the structure of the feature space, improves the spectral properties of the projected covariance operator, and ultimately leads to a tighter generalization error bound. In contrast, although methods such as SMOTE and BSMOTE also rely on synthetic sample generation, their construction differs from the KFD framework and therefore cannot be naturally embedded into its kernel-based operator formulation. As a consequence, the covariance operator in the reproducing kernel Hilbert space is not adjusted in a way that allows the above error bound to be directly applied. Furthermore, the proposed method enhances the representation of the treated group in the reproducing kernel Hilbert space. This improvement leads to a more stable ATT estimation. The estimation of ATT becomes less sensitive to noise. The method also produces more robust confidence intervals.

Appendix B. Simulation Study on Bandwidth and Kernel Function Selection

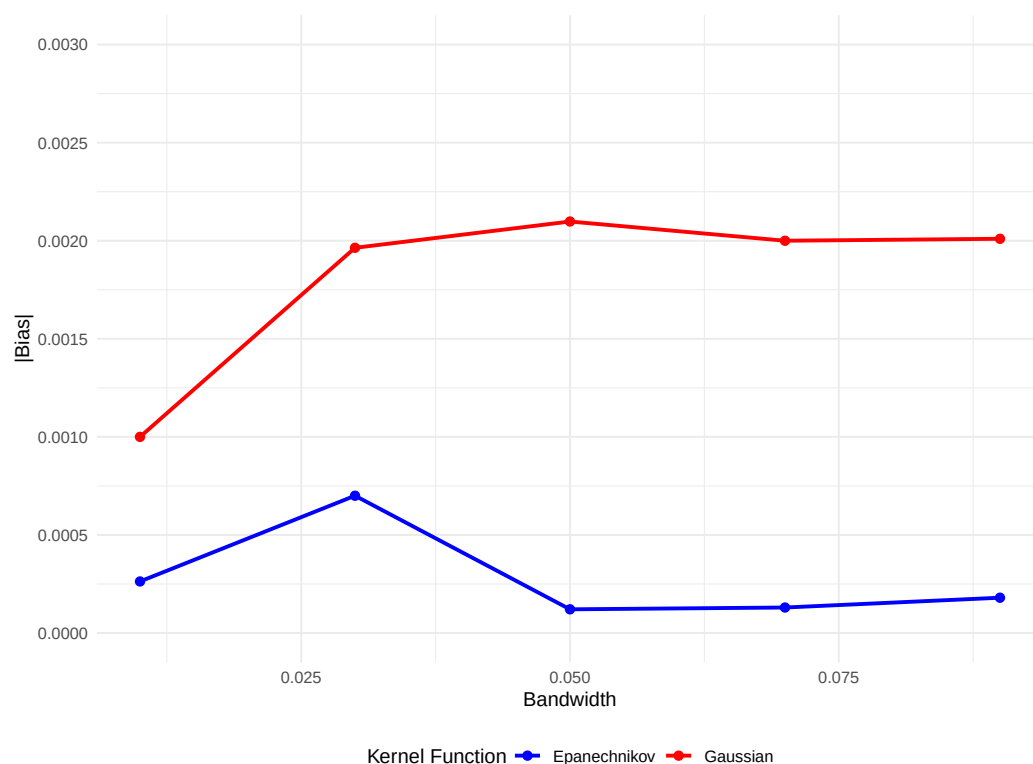


Figure A1. Absolute bias of the kernel-matching-based ATT estimator under Setting I for Gaussian and Epanechnikov kernels across different bandwidths.

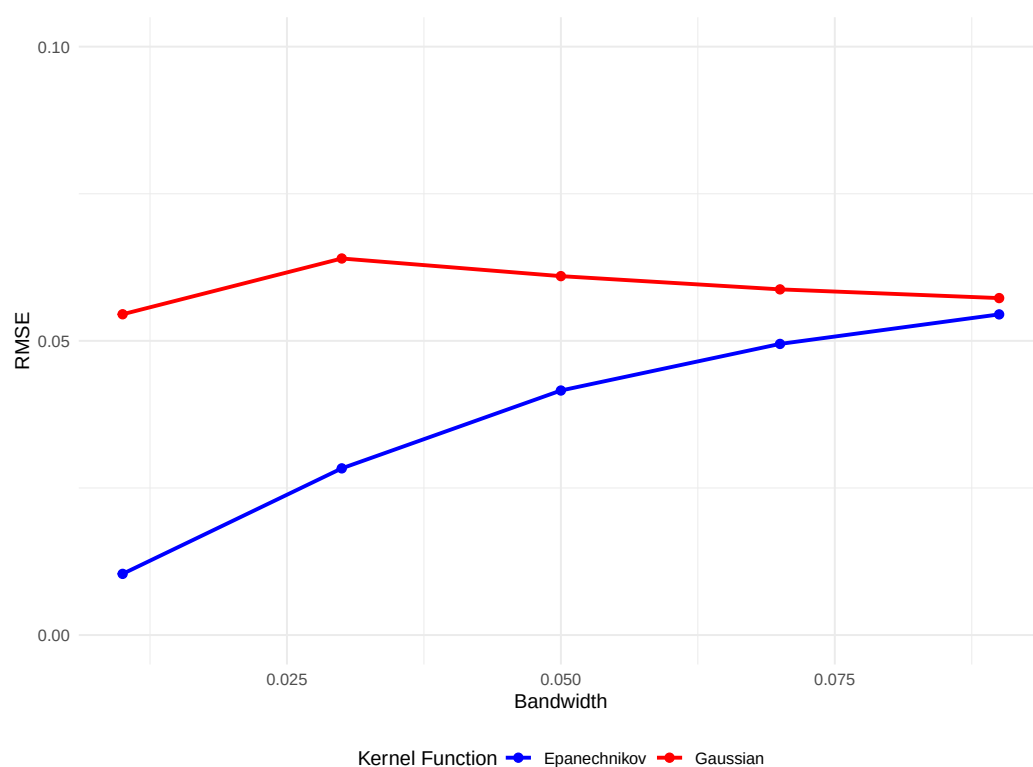


Figure A2. RMSE of the kernel-matching-based ATT estimator under Setting I for Gaussian and Epanechnikov kernels across different bandwidths.

Appendix C. Monte Carlo Results Associated with Section 4.3

Table A1. Estimated ATT performance from simulations under Setting I.

Sample Size	Method	Bias	SD	RMSE	AW	CP
$N_1 = 25$	KM	−0.0042	0.2931	0.2930	1.1289	0.941
	KM + SVM	0.0011	0.2832	0.2831	1.1103	0.945
	Proposed	0.0007	0.1574	0.1571	0.7203	0.967
	KM + SMOTE	0.0295	0.3591	0.3601	0.3046	0.373
	KM + BSMOTE	0.0087	0.4312	0.4312	0.3231	0.380
$N_1 = 50$	KM	−0.0041	0.2050	0.2051	0.8202	0.945
	KM + SVM	0.0145	0.1997	0.2002	0.8001	0.944
	Proposed	−0.0083	0.1545	0.1544	0.5267	0.954
	KM + SMOTE	0.0047	0.3432	0.3431	0.3045	0.320
	KM + BSMOTE	−0.0008	0.3620	0.3620	0.2887	0.313
$N_1 = 100$	KM	−0.0045	0.1553	0.1553	0.6061	0.938
	KM + SVM	0.0119	0.1545	0.1549	0.5751	0.936
	Proposed	−0.0026	0.1083	0.1083	0.4242	0.953
	KM + SMOTE	−0.0269	0.3385	0.3394	0.2987	0.333
	KM + BSMOTE	0.0065	0.3458	0.3458	0.3135	0.320

Table A2. Estimated ATT performance from simulations under Setting II.

Sample Size	Method	Bias	SD	RMSE	AW	CP
$N_1 = 25$	KM	−0.0008	0.2331	0.2329	0.9268	0.941
	KM + SVM	−0.0120	0.2391	0.2393	0.9340	0.938
	Proposed	0.0052	0.1822	0.1816	0.6125	0.956
	KM + SMOTE	−0.0270	0.1819	0.1820	0.2485	0.621
	KM + BSMOTE	0.0085	0.1564	0.1563	0.2258	0.573
$N_1 = 50$	KM	−0.0111	0.1722	0.1725	0.6550	0.935
	KM + SVM	−0.0065	0.1742	0.1742	0.6669	0.946
	Proposed	0.0089	0.1392	0.1390	0.4167	0.953
	KM + SMOTE	−0.0042	0.1171	0.1171	0.2355	0.671
	KM + BSMOTE	0.0028	0.1341	0.1341	0.2358	0.623
$N_1 = 100$	KM	−0.0040	0.1271	0.1271	0.4874	0.943
	KM + SVM	−0.0006	0.1289	0.1289	0.4729	0.953
	Proposed	−0.0010	0.1105	0.1107	0.3945	0.952
	KM + SMOTE	0.0076	0.1181	0.1182	0.2730	0.653
	KM + BSMOTE	0.0065	0.3458	0.3457	0.2490	0.656

Table A3. Estimated ATT performance from simulations under Setting III.

Sample Size	Method	Bias	SD	RMSE	AW	CP
$N_1 = 25$	KM	0.0025	0.2812	0.2811	0.9202	0.940
	KM + SVM	−0.0005	0.2807	0.2806	0.9724	0.955
	Proposed	0.0076	0.2083	0.2019	0.4406	0.958
	KM + SMOTE	0.0101	0.1857	0.1858	0.2892	0.581
	KM + BSMOTE	0.0105	0.1813	0.1815	0.2827	0.611
$N_1 = 50$	KM	0.0101	0.2096	0.2097	0.8668	0.944
	KM + SVM	0.0072	0.2086	0.2086	0.7958	0.941
	Proposed	0.0047	0.1870	0.1852	0.4717	0.968
	KM + SMOTE	0.0025	0.1755	0.1755	0.2266	0.587
	KM + BSMOTE	0.0051	0.1873	0.1873	0.2463	0.476

Table A3. Cont.

Sample Size	Method	Bias	SD	RMSE	AW	CP
$N_1 = 100$	KM	0.0009	0.1518	0.1518	0.4956	0.946
	KM + SVM	−0.0004	0.1478	0.1478	0.5729	0.940
	Proposed	0.0042	0.1135	0.1130	0.4189	0.931
	KM + SMOTE	0.0001	0.1699	0.1698	0.2672	0.524
	KM + BSMOTE	0.0015	0.1829	0.1828	0.2508	0.598

Table A4. Estimated ATT performance from simulations under Setting IV.

Sample Size	Method	Bias	SD	RMSE	AW	CP
$N_1 = 25$	KM	0.0023	0.2807	0.2806	1.1112	0.947
	KM + SVM	0.0003	0.2801	0.2800	1.0700	0.939
	Proposed	0.0053	0.1995	0.1994	0.6520	0.969
	KM + SMOTE	−0.0211	0.1833	0.1844	0.2666	0.468
	KM + BSMOTE	−0.0171	0.1942	0.1972	0.2523	0.667
$N_1 = 50$	KM	0.0100	0.2109	0.2111	0.7994	0.947
	KM + SVM	0.0081	0.2097	0.2097	0.7962	0.938
	Proposed	0.0028	0.1957	0.1948	0.4886	0.965
	KM + SMOTE	−0.0195	0.1741	0.1750	0.2525	0.581
	KM + BSMOTE	−0.0160	0.1895	0.1900	0.2514	0.591
$N_1 = 100$	KM	0.0016	0.1521	0.1520	0.5807	0.945
	KM + SVM	0.0012	0.1484	0.1484	0.5646	0.962
	Proposed	0.0065	0.1111	0.1111	0.3017	0.950
	KM + SMOTE	−0.0132	0.1684	0.1688	0.2801	0.573
	KM + BSMOTE	−0.0063	0.1770	0.1771	0.2436	0.684

Appendix D. Simulation Results for Different Values of δ

To assess the sensitivity of the proposed method to the tuning parameter δ , we conduct a Monte Carlo study over a range of values $\delta \in \{0, 0.01, 0.05, 0.1\}$ under Setting I with $N_1 = 25$. The corresponding results are summarized in terms of absolute bias, SE, and RMSE. As shown in the table, the bias exhibits a mild increasing trend as δ increases, rising from 0.0007 when $\delta = 0$ to 0.00436 when $\delta = 0.1$. However, this increase remains numerically small, suggesting that the estimator is relatively stable across a broad range of δ . In contrast, both the standard error and RMSE remain highly stable across all configurations. The SE varies only slightly between 0.2676 and 0.2838, while the RMSE shows similarly negligible variation, ranging from 0.2665 to 0.2831. These results indicate that although δ may introduce a slight bias–variance trade-off, its overall impact on estimation accuracy is limited.

Overall, the Monte Carlo evidence suggests that the proposed estimator is robust to moderate changes in δ , and the choice of δ does not materially affect its finite-sample performance within the considered range. In particular, setting $\delta = 0$ yields slightly smaller bias and has the additional advantage of simplifying the computation of α , thereby making the overall procedure more computationally convenient.

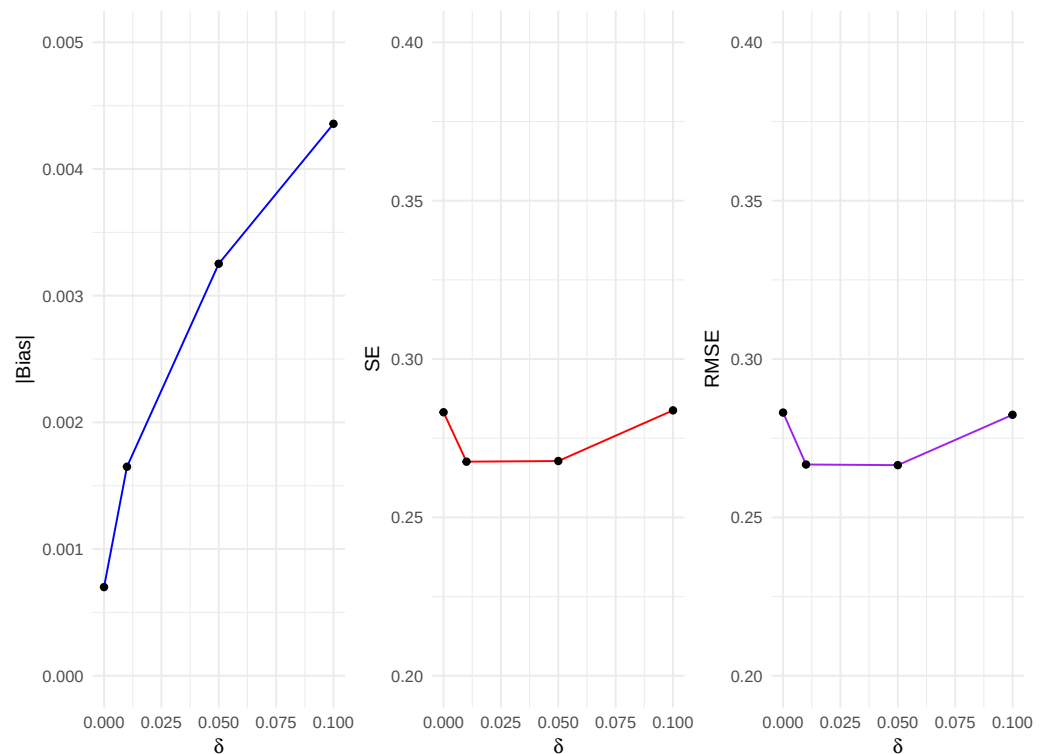


Figure A3. Comprehensive performance evaluation in terms of bias, SE, and RMSE across different values of δ .

References

- Hirano, K.; Imbens, G.W.; Ridder, G. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **2003**, *71*, 1161–1189. [\[CrossRef\]](#)
- Bang, H.; Robins, J.M. Doubly robust estimation in missing data and causal inference models. *Biometrics* **2005**, *61*, 962–973. [\[CrossRef\]](#) [\[PubMed\]](#)
- van der Laan, M.J.; Rubin, D. Targeted maximum likelihood learning. *Int. J. Biostat.* **2006**, *2*, 1–38. [\[CrossRef\]](#)
- Gruber, S.; van der Laan, M.J. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *Int. J. Biostat.* **2010**, *6*, 18. [\[CrossRef\]](#) [\[PubMed\]](#)
- Imai, K.; Ratkovic, M. Covariate balancing propensity score. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2014**, *76*, 243–263.
- Ding, C.; Li, Z.; Ng, H.K.T.; Gao, W. Estimation of Treatment Effects based on Kernel Matching. *arXiv* **2025**, arXiv:2502.10958.
- Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
- Hosmer, D.W.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; Wiley: New York, NY, USA, 2000.
- Chen, J.J.; Tsai, C.A.; Young, J.F.; Kodell, R.L. Classification ensembles for unbalanced class sizes in predictive toxicology. *SAR QSAR Environ. Res.* **2005**, *16*, 517–529. [\[CrossRef\]](#) [\[PubMed\]](#)
- He, H.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. [\[CrossRef\]](#)
- Prati, R.C.; Batista, G.E.A.P.A.; Monard, M.C. Class Imbalances versus Class Overlapping: An Analysis of a Learning System Behavior. In *Proceedings of the MICAI 2004: Advances in Artificial Intelligence*; Monroy, R., Arroyo-Figueroa, G., Sucar, L.E., Sossa, H., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; pp. 312–321.
- Vuttipittayamongkol, P.; Elyan, E.; Petrovski, A. On the class overlap problem in imbalanced data classification. *Knowl.-Based Syst.* **2021**, *212*, 106631. [\[CrossRef\]](#)
- Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [\[CrossRef\]](#)
- Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *Proceedings of the Advances in Intelligent Computing*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 878–887.
- Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *Acm Sigkdd Explor. Newsl.* **2004**, *6*, 20–29. [\[CrossRef\]](#)

16. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China 1–8 June 2008; pp. 1322–1328.
17. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular data using Conditional GAN. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
18. Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. In *Proceedings of the Advances in Neural Information Processing Systems*; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 6840–6851.
19. Sauber-Cole, R.; Khoshgoftaar, T.M. The use of generative adversarial networks to alleviate class imbalance in tabular data: A survey. *J. Big Data* **2022**, *9*, 98. [[CrossRef](#)]
20. Majeed, A.; Hwang, S.O. CTGAN-MOS: Conditional generative adversarial network based minority-class-augmented oversampling scheme for imbalanced problems. *IEEE Access* **2023**, *11*, 85878–85899.
21. Chen, Y.; Pedrycz, W.; Zhang, C.; Wang, J.; Yang, J. Oversampling with GAN via meta-learning for imbalanced data. *IEEE Trans. Multimed.* **2025**, *27*, 8819–8834. [[CrossRef](#)]
22. Robins, J.M.; Rotnitzky, A.; Zhao, L.P. Estimation of regression coefficients when some regressors are not always observed. *J. Am. Stat. Assoc.* **1994**, *89*, 846–866. [[CrossRef](#)]
23. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55. [[CrossRef](#)]
24. Ferman, B. Matching estimators with few treated and many control observations. *J. Econom.* **2021**, *225*, 295–307. [[CrossRef](#)]
25. Dawid, A.P. Conditional independence in statistical theory. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1979**, *41*, 1–15. [[CrossRef](#)]
26. Imbens, G.W. Nonparametric estimation of average treatment effects under exogeneity: A review. *Rev. Econ. Stat.* **2004**, *86*, 4–29. [[CrossRef](#)]
27. Imbens, G.W.; Wooldridge, J.M. Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* **2009**, *47*, 5–86. [[CrossRef](#)]
28. Imbens, G.W. Matching Methods in Practice: Three Examples. *J. Hum. Resour.* **2015**, *50*, 373–419. [[CrossRef](#)]
29. Hofmann, T.; Schölkopf, B.; Smola, A.J. Kernel Methods in Machine Learning. *Ann. Stat.* **2008**, *36*, 1171–1220. [[CrossRef](#)]
30. Mika, S.; Ratsch, G.; Weston, J.; Scholkopf, B.; Mullers, K. Fisher discriminant analysis with kernels. In Proceedings of the Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468), Madison, WI, USA, 25 August 1999; pp. 41–48. [[CrossRef](#)]
31. Anstreicher, K.; Wolkowicz, H. On Lagrangian Relaxation of Quadratic Matrix Constraints. *SIAM J. Matrix Anal. Appl.* **2000**, *22*, 41–55. [[CrossRef](#)]
32. Pelleg, D.; Moore, A.W. X-means: Extending k-means with efficient estimation of the number of clusters. In Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford, CA, USA, 29 June–2 July 2000; Volume 1, pp. 727–734.
33. Fraley, C.; Raftery, A.E. How many clusters? Which clustering method? Answers via model-based cluster analysis. *Comput. J.* **1998**, *41*, 578–588. [[CrossRef](#)]
34. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer Series in Statistics; Springer: New York, NY, USA, 2009.
35. Pourhabib, A.; Mallick, B.K.; Ding, Y. Absent Data Generating Classifier for Imbalanced Class Sizes. *J. Mach. Learn. Res.* **2015**, *16*, 2695–2724.
36. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman and Hall/CRC: Boca Raton, FL, USA, 1993.
37. Cattaneo, M.D. Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *J. Econom.* **2010**, *155*, 138–154. [[CrossRef](#)]
38. Wang, J.; Gao, W.; Tang, M.L. Estimation of treatment effects for heterogeneous matched-pairs data with probit models. *Scand. J. Stat.* **2019**, *46*, 575–594. [[CrossRef](#)]
39. Almond, D.; Chay, K.Y.; Lee, D.S. The costs of low birth weight. *Q. J. Econ.* **2005**, *120*, 1031–1083. [[CrossRef](#)]
40. Williams, C.K.; Rasmussen, C.E. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
41. Durrant, R.; Kabán, A. Error bounds for kernel Fisher linear discriminant in Gaussian Hilbert space. In Proceedings of the Artificial Intelligence and Statistics, PMLR, La Palma, Spain, 21–23 April 2012; pp. 337–345.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.