CAD-DA: Controllable Anomaly Detection after Domain Adaptation by Statistical Inference

Vo Nguyen Le Duy[1,2,6]   **Hsuan-Tien Lin**[3,4]   **Ichiro Takeuchi**[5,6,*]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]Dept. of Computer Science and Information Engineering, National Taiwan University
[4]Center for Data Intelligence, National Taiwan University
[5]Nagoya University
[6]RIKEN

Abstract

We propose a novel statistical method for testing the results of anomaly detection (AD) under domain adaptation (DA), which we call CAD-DA—controllable AD under DA. The distinct advantage of the CAD-DA lies in its ability to control the probability of misidentifying anomalies under a pre-specified level α (e.g., 0.05). The challenge within this DA setting is the necessity to account for the influence of DA to ensure the validity of the inference results. We overcome the challenge by leveraging the concept of Selective Inference to handle the impact of DA. To our knowledge, this is the first work capable of conducting a valid statistical inference within the context of DA. We evaluate the performance of the CAD-DA method on both synthetic and real-world datasets.

1 INTRODUCTION

Anomaly detection (AD) is a fundamental problem in machine learning and statistics, as the vast body of literature surveyed by Aggarwal (2017) suggests. The goal of AD is to identify rare and unusual observations that deviate significantly from the norm within a given dataset. AD plays a critical role in several applications and has been widely applied in many areas such as medical (Wong et al., 2002; Aggarwal, 2005), fraud detection (Pourhabibi et al., 2020), and damage detection (Avci et al., 2021; Du et al., 2020).

In numerous real-world scenarios, the availability of limited data can lead to poor performance of AD. To overcome this challenge, a strategy of increasing the sample size by transferring data points from a readily accessible source domain to the present target domain can be employed. This type of problem is known as Domain Adaptation (DA). By leveraging data-rich source domains to bolster the data pool in the target domain, DA aims to enhance the efficacy of AD in scenarios where limited data hamper their effectiveness.

A critical concern arises regarding to the possibility of erroneous detection. The AD could misidentify certain observations as anomalies, even though they are actually normal. These errors are commonly referred to as *false positives*, which can cause serious consequences in high-stake decision making. Especially, when conducting AD under DA, there is an increased risk of misclassifying normal instances as anomalies due to potential DA errors. For instance, in the medical field, some unhealthy individuals transformed from the source domain may become closely similar to healthy individuals in the target domain. Thus, we mistakenly classify a healthy individual as unhealthy, and we could inadvertently administer drugs that may harm their well-being. Hence, there is a critical need of an inference method for controlling the false positive rate (FPR).

In AD-DA, controlling the false negative rate (FNR) is also important. In the literature of statistics, a common procedure is to initially control the FPR at a specified level α, e.g., 0.05, while concurrently seeking to minimize the FNR, i.e., maximizing the true positive rate (TPR $= 1 - $ FNR) by empirical evidences. In this paper, we follow this established practice.

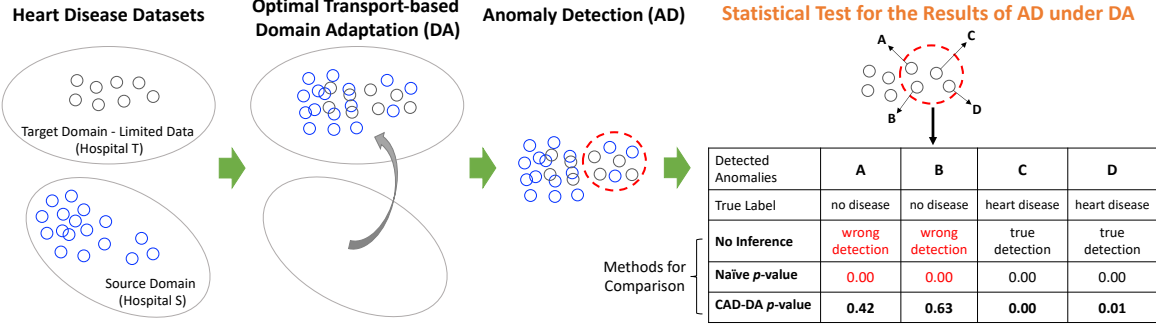*Correspondence: ichiro.takeuchi@mae.nagoya-u.ac.jp

Figure 1: Illustration of the proposed method. Conducting AD-DA without inference results wrong anomalies (**A**, **B**). The naive p-values are small even for falsely detected anomalies. The proposed CAD-DA can identify both false positive (FP) and true positive (TP) detections, i.e., large p-values for FPs and small p-values for TPs.

To our knowledge, none of the existing method can control the FPR of AD under the context of DA. The main challenge is that, without accounting for the influence of DA, the FPR can not be properly controlled. Tsukurimichi et al. (2022) proposed a method for testing the anomalies when they are detected by a class of robust regression methods (Huber et al., 1973; Andrews, 1974; Zaman et al., 2001; Rousseeuw and Leroy, 2005; Maronna et al., 2019). However, this method supposes that the data comes from the same distribution, and it is invalid in the scenario that a distribution shift occurs and DA must be applied.

Our idea is to leverage *Selective Inference (SI)* (Lee et al., 2016) for resolving this challenge. However, directly incorporating SI in our setting is still non-trivial because SI is highly *problem-specific*. Therefore, we need to carefully examine the selection strategy of the algorithm in the context of AD after DA. Moreover, if we naively leverage the ideas of existing SI methods (Lee et al., 2016; Duy and Takeuchi, 2023), the power of the test is low, i.e, the FNR is high. Therefore, we need to introduce an approach to minimize the FNR while properly controlling the FPR. We would like to note that we start this new research direction with the Optimal Transport (OT)-based DA Flamary et al. (2016), which is recently promising and popular in the OT community. The detailed discussions on future extensions to other types of DA are provided in §5.

Contributions. Our contributions are as follows:

• We mathematically formulate the problem of testing AD results under DA and introduce *CAD-DA*, a novel method to conduct the statistical test with controllable FPR. Compared to the literature on controllable AD, CAD-DA presents a unique challenge in addressing the effect of DA to ensure the validity of FPR control.

• We overcome the challenge by leveraging the SI framework to handle the influence of DA. We carefully examine the selection strategy for OT-based

Table 1: The importance of the proposed method lies in its ability to control the False Positive Rate (FPR).

	No Inference	**Naive**	**CAD-DA**
$N = 120$	FPR = 1.0	0.6	**0.04**
$N = 240$	FPR = 1.0	0.7	**0.05**

DA, whose operations can be characterized by linear/quadratic inequalities, and prove that achieving controllable AD-DA is indeed possible. Furthermore, we introduce a more strategic approach to enhance the TPR. To our knowledge, this is the first work capable of conducting valid inference within the context of DA.

• We conduct experiments on both synthetic and real-world datasets to support our theoretical results, showcasing superior performance of the CAD-DA.

Example 1. *To show the importance of the proposed method, we provide an example presented in Fig. 1. Our goal is to detect patients in Hospital T with heart disease, treated as anomalies, in a scenario where the number of patients is limited. Here, the source domain consists of patients in Hospital S, while the target domain comprises patients in Hospital T. We employ the OT-based DA approach to transfer the data from the source domain to the target domain. Subsequently, we apply an AD algorithm, i.e., mean absolute deviation. The AD algorithm erroneously identified two healthy individuals as anomalies. To address this issue, we conducted an additional inference step using the proposed p-values, allowing us to identify both true positive and false positive detections. Furthermore, we repeated the experiments N times and the FPR results are shown in Tab. 1. With the proposed method, we were able to control the FPR under $\alpha = 0.05$, which other competitors were unable to achieve.*

Related works. Although there exists an extensive body of literature on AD methods (Aggarwal, 2017), there has been a limited exploration of applying the hypothesis testing framework to assess the results of

AD. For instance, in Srivastava and von Rosen (1998) and Pan and Fang (1995), the likelihood ratio test has been discussed to determine if an individual data point is an anomaly, employing the mean-shift model. However, these classical anomaly inference methods hold validity exclusively when the target anomalies are predetermined in advance. When we employ these classical techniques on the anomalies detected by an AD algorithm, they become invalid in the sense that the FPR cannot be controlled at the desired level.

In order to control the FPR when using classical methods, the use of multiple testing correction is essential. The most popular technique is Bonferroni correction, in which the correction factor scales exponentially with the number of instances, denoted as n. Specifically, it grows to a value of 2^n. However, this correction factor can become prohibitively large unless n fairly small, which leads to overly conservative statistical inference.

In recent years, SI has emerged as a promising approach to resolve the invalidity of the traditional inference method without the need of employing conservative multiple testing correction. Specifically, instead of considering the exponentially increasing value of 2^n, we consider the correction factor of 1 by conducting the *conditional inference* conditioning on the single set of detected anomalies. This is the basic concept of the conditional SI introduced in the seminal work of Lee et al. (2016).

The seminal paper has not only laid the foundation for research on SI for feature selection (Loftus and Taylor, 2014; Fithian et al., 2015; Tibshirani et al., 2016; Yang et al., 2016; Hyun et al., 2018a; Sugiyama et al., 2021a; Fithian et al., 2014; Duy and Takeuchi, 2022a) but has also spurred the development of SI for more complex supervised learning algorithms, such as boosting (Rügamer and Greven, 2020), decision trees (Neufeld et al., 2022), kernel methods (Yamada et al., 2018), higher-order interaction models (Suzumura et al., 2017; Das et al., 2021) and deep neural networks (Duy et al., 2022; Miwa et al., 2023).

Moreover, SI is also valuable for unsupervised learning problems, such as change point detection (Umezu and Takeuchi, 2017; Hyun et al., 2018b; Duy et al., 2020; Sugiyama et al., 2021b; Jewell et al., 2022), clustering (Lee et al., 2015; Inoue et al., 2017; Gao et al., 2022; Chen and Witten, 2022), and segmentation (Tanizaki et al., 2020; Duy et al., 2022). Furthermore, SI can be applied to statistical inference on the DTW distance (Duy and Takeuchi, 2022b) and the Wasserstein distance (Duy and Takeuchi, 2023).

The studies most related to this paper are Chen and Bien (2019) and Tsukurimichi et al. (2022). The authors of (Chen and Bien, 2019) introduced a method for testing the features of a linear model after removing anomalies. Although their work did not directly address the problem of testing the anomalies, it was the inspiration for Tsukurimichi et al. (2022). The contribution of Tsukurimichi et al. (2022) centered on introducing a SI approach for testing the anomalies identified by a class of robust regression methods. However, a notable limitation of Chen and Bien (2019) and Tsukurimichi et al. (2022) is their assumption that the data comes from the same distribution. Therefore, when applied in the context of DA, their methods loses its validity, making them unsuitable for scenarios involving DA. Besides, their primary focus revolves around the context of linear regression, which differs from the setting we consider in this paper.

2 PROBLEM SETUP

Let us consider two random vectors

$$\boldsymbol{X}^s = (X_1^s, ..., X_{n_s}^s)^\top = \boldsymbol{\mu}^s + \boldsymbol{\varepsilon}^s, \quad \boldsymbol{\varepsilon}^s \sim \mathbb{N}(\mathbf{0}, \Sigma^s),$$
$$\boldsymbol{X}^t = (X_1^t, ..., X_{n_t}^t)^\top = \boldsymbol{\mu}^t + \boldsymbol{\varepsilon}^t, \quad \boldsymbol{\varepsilon}^t \sim \mathbb{N}(\mathbf{0}, \Sigma^t),$$

where n_s and n_t are the number of instances in the source and target domains, $\boldsymbol{\mu}^s$ and $\boldsymbol{\mu}^t$ are unknown signals, $\boldsymbol{\varepsilon}^s$ and $\boldsymbol{\varepsilon}^t$ are the Gaussian noise vectors with the covariance matrices Σ^s and Σ^t assumed to be known or estimable from independent data. We assume that the number of instances in the target domain is limited, i.e., n_t is much smaller than n_s. The goal is to statistically test the results of AD after DA.

Optimal Transport (OT)-based DA (Flamary et al., 2016). Let us define the cost matrix as

$$C(\boldsymbol{X}^s, \boldsymbol{X}^t) = \left[(X_i^s - X_j^t)^2 \right]_{ij} \in \mathbb{R}^{n_s \times n_t}.$$

The OT problem between the source and target domain is then defined as

$$\hat{T} = \underset{T \geq 0}{\arg\min} \ \langle T, C(\boldsymbol{X}^s, \boldsymbol{X}^t) \rangle \tag{1}$$
$$\text{s.t.} \quad T \mathbf{1}_{n_t} = \mathbf{1}_{n_s}/n_s, T^\top \mathbf{1}_{n_s} = \mathbf{1}_{n_t}/n_t,$$

where $\mathbf{1}_n \in \mathbb{R}^n$ is the vector whose elements are set to 1. After obtaining the optimal transportation matrix $\hat{T}$, source instances are transported in the target domain. The transformation $\tilde{\boldsymbol{X}}^s$ of $\boldsymbol{X}^s$ is defined as:

$$\tilde{\boldsymbol{X}}^s = n_s \hat{T} \boldsymbol{X}^t.$$

More details are provided in Sec 3.3 of Flamary et al. (2016).

Anomaly detection. After transforming the data from source domain to the target domain, we apply

an AD algorithm $\mathcal{A}$ on $\left\{ \tilde{\boldsymbol{X}}^s, \boldsymbol{X}^t \right\}$ to obtain a set $\mathcal{O}$ of indices of anomalies in the target domain:

$$\mathcal{A} : \left\{ \tilde{\boldsymbol{X}}^s, \boldsymbol{X}^t \right\} \mapsto \mathcal{O} \in [n_t].$$

In this paper, we used the Median Absolute Deviation (MAD) as an example of the AD algorithm. Our proposed CAD-DA framework is not specialized for a specific AD algorithm but can also be applied to other AD algorithms (see §3.3 for more details).

Statistical inference and decision making with a p-value. To statistically quantifying the significance of the detected anomalies, we consider the statistical test on the following null and alternative hypotheses:

$$\mathrm{H}_{0,j} : \mu_j^t = \bar{\boldsymbol{\mu}}_{-\mathcal{O}}^t \quad \text{vs.} \quad \mathrm{H}_{1,j} : \mu_j^t \neq \bar{\boldsymbol{\mu}}_{-\mathcal{O}}^t, \quad \forall j \in \mathcal{O},$$

where

$$\bar{\boldsymbol{\mu}}_{-\mathcal{O}}^t = \frac{1}{n_t - |\mathcal{O}|} \sum_{\ell \in [n_t] \backslash \mathcal{O}} \mu_\ell^t.$$

In other words, our goal is to test if each of the detected anomalies $j \in \mathcal{O}$ is truly deviated from the remaining data points after removing the set of anomalies $\mathcal{O}$.

To test the hypotheses, the test statistic is defined as:

$$T_j = X_j^t - \bar{\boldsymbol{X}}_{-\mathcal{O}}^t = \boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix}, \tag{2}$$

where $\boldsymbol{\eta}_j$ is the direction of the test statistic:

$$\boldsymbol{\eta}_j = \begin{pmatrix} \boldsymbol{0}^s \\ \boldsymbol{e}_j^t - \frac{1}{n_t - |\mathcal{O}|} \boldsymbol{e}_{-\mathcal{O}}^t \end{pmatrix}, \tag{3}$$

$\boldsymbol{0}^s \in \mathbb{R}^{n_s}$ represents a vector where all entries are set to 0, $\boldsymbol{e}_j^t \in \mathbb{R}^{n_t}$ is a vector in which the jth entry is set to 1, and 0 otherwise, $\boldsymbol{e}_{-\mathcal{O}}^t \in \mathbb{R}^{n_t}$ is a vector in which the jth entry is set to 0 if $j \in \mathcal{O}$, and 1 otherwise.

After obtaining the test statistic in (2), we need to compute a p-value. Given a significance level $\alpha \in [0, 1]$, e.g., 0.05, we reject the null hypothesis $\mathrm{H}_{0,j}$ and assert that X_j^t is an anomaly if the p-value $\leq \alpha$. Conversely, if the p-value $> \alpha$, we infer that there is insufficient evidence to conclude that X_j^t is an anomaly.

Challenge of computing a valid p-value. The traditional (naive) p-value, which does not properly consider the effect of DA and AD, is defined as:

$$p_j^{\text{naive}} = \mathbb{P}_{\mathrm{H}_{0,j}} \left(\left| \boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix} \right| \geq \left| \boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{X}_{\text{obs}}^s \\ \boldsymbol{X}_{\text{obs}}^t \end{pmatrix} \right| \right),$$

where $\boldsymbol{X}_{\text{obs}}^s$ and $\boldsymbol{X}_{\text{obs}}^t$ are the observations of the random vectors $\boldsymbol{X}^s$ and $\boldsymbol{X}^t$, respectively. If the vector

$\boldsymbol{\eta}_j$ is independent of the DA and AD algorithms, the naive p-value is valid in the sense that

$$\mathbb{P} \left(\underbrace{p_j^{\text{naive}} \leq \alpha \mid \mathrm{H}_{0,j} \text{ is true}}_{\text{a false positive}} \right) = \alpha, \quad \forall \alpha \in [0, 1], \quad (4)$$

i.e., the probability of obtaining a false positive is controlled under a certain level of guarantee. However, in our setting, the vector $\boldsymbol{\eta}_j$ actually depends on the DA and AD. The property of a valid p-value in (4) is no longer satisfied. Hence, the naive p-value is *invalid*.

3 PROPOSED CAD-DA METHOD

In this section, we introduce the technical details of computing the valid p-value in CAD-DA.

3.1 The valid p-value in CAD-DA

To calculate the valid p-value, we need to derive the sampling distribution of the test statistic in (2). To achieve the task, we utilize the concept of conditional SI, i.e., we consider the sampling distribution of the test statistic *conditional* on the AD results after DA:

$$\mathbb{P} \left(\boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix} \mid \mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{O}_{\text{obs}} \right), \tag{5}$$

where $\mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t}$ is the results of AD after DA *for any random $\boldsymbol{X}^s$ and $\boldsymbol{X}^t$*, and $\mathcal{O}_{\text{obs}} = \mathcal{O}_{\boldsymbol{X}_{\text{obs}}^s, \boldsymbol{X}_{\text{obs}}^t}$.

Next, we introduce the *selective p-value* defined as

$$p_j^{\text{sel}} = \mathbb{P}_{\mathrm{H}_{0,j}} \left(\left| \boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix} \right| \geq \left| \boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{X}_{\text{obs}}^s \\ \boldsymbol{X}_{\text{obs}}^t \end{pmatrix} \right| \mid \mathcal{E} \right), \quad (6)$$

where the conditioning event $\mathcal{E}$ is defined as

$$\mathcal{E} = \left\{ \mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{O}_{\text{obs}}, \mathcal{Q}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{Q}_{\text{obs}} \right\}. \tag{7}$$

The $\mathcal{Q}_{\boldsymbol{X}^s, \boldsymbol{X}^t}$ is the *nuisance component* defined as

$$\mathcal{Q}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \left(I_{n_s + n_t} - \boldsymbol{b} \boldsymbol{\eta}_j^\top \right) \begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix}, \tag{8}$$

where $\boldsymbol{b} = \frac{\Sigma \boldsymbol{\eta}_j}{\boldsymbol{\eta}_j^\top \Sigma \boldsymbol{\eta}_j}$ and $\Sigma = \begin{pmatrix} \Sigma^s & 0 \\ 0 & \Sigma^t \end{pmatrix}$.

Remark 1. *The nuisance component $\mathcal{Q}_{\boldsymbol{X}^s, \boldsymbol{X}^t}$ corresponds to the component $\boldsymbol{z}$ in the seminal paper of Lee et al. (2016) (see Sec. 5, Eq. (5.2), and Theorem 5.2)). We note that additionally conditioning on $\mathcal{Q}_{\boldsymbol{X}^s, \boldsymbol{X}^t}$, which is required for technical purpose, is a standard approach in the SI literature and it is used in almost all the SI-related works that we cited.*

Lemma 1. *The selective p-value proposed in (6) satisfies the property of a valid p-value:*

$$\mathbb{P}_{\mathrm{H}_{0,j}} \left(p_j^{\text{sel}} \leq \alpha \right) = \alpha, \quad \forall \alpha \in [0, 1].$$
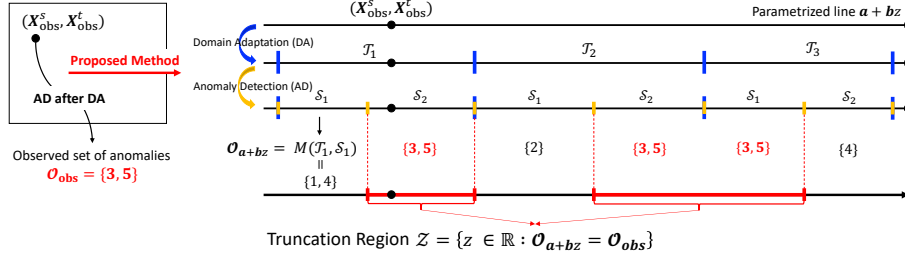
Figure 2: A schematic illustration of the proposed method. By applying AD after DA to the observed data, we obtain a set of anomalies. Then, we parametrize the data with a scalar parameter z in the dimension of the test statistic to identify the truncation region $\mathcal{Z}$ whose data have the *same* result of anomaly detection as the observed data. Finally, the valid inference is conducted conditional on $\mathcal{Z}$. We utilize the concept of "divide-and-conquer" and introduce a hierarchical line search method for efficiently characterizing the truncation region $\mathcal{Z}$.
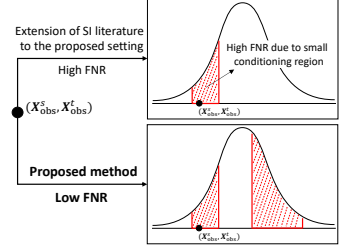
Figure 3: If we leverage the idea of existing SI literature and apply to our setting, the FNR will be high due to small conditioning region by extra-conditioning. In contrast, the FNR is minimized with the proposed method.

Proof. The proof is deferred to Appendix 6.1. □

Lemma 1 indicates that, by using the selective p-value, the FPR is theoretically controlled at the significance level α. To compute the selective p-value, we need to identify the conditioning event $\mathcal{E}$ in (7) whose characterization will be introduced in the next section.

3.2 Conditional Data Space Characterization

We define the set of $\begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix} \in \mathbb{R}^{n_s+n_t}$ that satisfies the conditions in (7) as:

$$\mathcal{D} = \left\{ \begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix} \middle| \mathcal{O}_{\boldsymbol{X}^s,\boldsymbol{X}^t} = \mathcal{O}_{\mathrm{obs}}, \mathcal{Q}_{\boldsymbol{X}^s,\boldsymbol{X}^t} = \mathcal{Q}_{\mathrm{obs}} \right\}. \quad (9)$$

In fact, the conditional data space $\mathcal{D}$ is *a line* in $\mathbb{R}^{n_s+n_t}$ as stated in the following lemma.

Lemma 2. *The set $\mathcal{D}$ in (9) can be rewritten using a scalar parameter $z \in \mathbb{R}$ as follows:*

$$\mathcal{D} = \left\{ \begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix} = \boldsymbol{a} + \boldsymbol{b}z \mid z \in \mathcal{Z} \right\}, \quad (10)$$

where vector $\boldsymbol{a} = \mathcal{Q}_{\mathrm{obs}}$, $\boldsymbol{b}$ is defined in (8), and

$$\mathcal{Z} = \left\{ z \in \mathbb{R} \mid \mathcal{O}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{O}_{\mathrm{obs}} \right\}. \quad (11)$$

Here, with a slight abuse of notation, $\mathcal{O}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{O}_{\begin{pmatrix} \boldsymbol{x}^s \\ \boldsymbol{x}^t \end{pmatrix}}$ is equivalent to $\mathcal{O}_{\boldsymbol{X}^s,\boldsymbol{X}^t}$.

Proof. The proof is deferred to Appendix 6.2. □

Remark 2. *Lemma 2 indicates that we need NOT consider the (n_s+n_t)-dimensional data space. Instead, we need only consider the* one-dimensional projected data space $\mathcal{Z}$ in (11). *The fact of restricting the data to a line has been implicitly exploited in Lee et al. (2016), and explicitly discussed in Sec. 6 of Liu et al. (2018).*

Reformulation of the selective p-value. Let us consider a random variable and its observation:

$$Z = \boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix} \in \mathbb{R} \ \ \text{and} \ \ Z_{\mathrm{obs}} = \boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{X}^s_{\mathrm{obs}} \\ \boldsymbol{X}^t_{\mathrm{obs}} \end{pmatrix} \in \mathbb{R}.$$

Then, the selective p-value in (6) can be rewritten as

$$p_j^{\mathrm{sel}} = \mathbb{P}_{\mathrm{H}_{0,j}} \left(|Z| \geq |Z_{\mathrm{obs}}| \mid Z \in \mathcal{Z} \right). \quad (12)$$

Because $Z \sim \mathbb{N}(0, \boldsymbol{\eta}_j^\top \Sigma \boldsymbol{\eta}_j)$ under the null hypothesis, $Z \mid Z \in \mathcal{Z}$ follows a *truncated* normal distribution. Once the truncation region $\mathcal{Z}$ is identified, computation of the selective p-value in (12) is straightforward. Therefore, the remaining task is to identify $\mathcal{Z}$.

3.3 Identification of Truncation Region $\mathcal{Z}$

As discussed in §3.2, to calculate the selective p-value (12), we must identify the truncation region $\mathcal{Z}$ in (11). However, there is no direct way to identify $\mathcal{Z}$. To overcome the difficulty, we utilize the concept the "*divide-and-conquer*" and introduce an approach (illustrated in Fig. 2) to efficiently identify $\mathcal{Z}$, described as follows:

● **Step 1 (extra-conditioning)**: breaking down the problem into multiple sub-problems by additionally conditioning the transportation for DA and all the steps of the AD after DA. Each sub-problem can be directly solved with a finite number of operations.

● **Step 2 (hierarchical line search)**: hierarchically combining multiple extra-conditioning steps and check the condition $\mathcal{O}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{O}_{\mathrm{obs}}$ to obtain $\mathcal{Z}$.

Specifically, let us denote by U a number of all possible transportations for DA along the parametrized line. We denote by V^u a number of all possible sets of steps performed by the AD algorithm after the $\mathcal{T}_u$ transportation, $u \in [U]$. The entire one-dimensional

---

Algorithm 1 `hierarchical_line_search`

Input: $\boldsymbol{a}, \boldsymbol{b}, z_{\min}, z_{\max}$

1: Initialization: $u = 1$, $v = 1$, $z_{u,v} = z_{\min}$, $\mathcal{V} = \emptyset$

2: **while** $z_{u,v} < z_{\max}$ **do**

3:    $\mathcal{T}_u \leftarrow$ DA on data $\boldsymbol{a} + \boldsymbol{b}z_{u,v}$

4:    Compute $[L_u, R_u] = \mathcal{Z}_u \leftarrow$ Lemma 3

5:    **while** *true* **do**

6:       $\mathcal{S}_v \leftarrow$ AD after DA on data $\boldsymbol{a} + \boldsymbol{b}z_{u,v}$

7:       Compute $\mathcal{Z}_v \leftarrow$ Lemma 4

8:       $[L_{u,v}, R_{u,v}] = \mathcal{Z}_{u,v} \leftarrow \mathcal{Z}_u \cap \mathcal{Z}_v$

9:       **if** $\mathcal{M}(\mathcal{T}_u, \mathcal{S}_v) = \mathcal{O}_{\text{obs}}$ **then**

10:          $\mathcal{V} \leftarrow \mathcal{V} \cup \{(u,v)\}$

11:       **end if**

12:       **if** $R_{u,v} = R_u$ **then**

13:          $v \leftarrow 1$, $u \leftarrow u + 1$, $z_{u,v} = R_{u,v}$, **break**

14:       **end if**

15:       $v \leftarrow v + 1$, $z_{u,v} = R_{u,v}$

16:    **end while**

17: **end while**

Output: $\mathcal{V}$

Algorithm 2 `CAD-DA`

Input: $\boldsymbol{X}_{\text{obs}}^s, \boldsymbol{X}_{\text{obs}}^t, z_{\min}, z_{\max}$

1: $\mathcal{O}_{\text{obs}} \leftarrow$ AD after DA on $\{\boldsymbol{X}_{\text{obs}}^s, \boldsymbol{X}_{\text{obs}}^t\}$

2: **for** $j \in \mathcal{O}_{\text{obs}}$ **do**

3:    Compute $\boldsymbol{\eta}_j \leftarrow$ Eq. (3), $\boldsymbol{a}$ and $\boldsymbol{b} \leftarrow$ Eq. (10)

4:    $\mathcal{V} \leftarrow$ `hierarchical_line_search` $(\boldsymbol{a}, \boldsymbol{b}, z_{\min}, z_{\max})$

5:    $p_j^{\text{sel}} \leftarrow$ Eq. (12) with $\mathcal{Z} \leftarrow$ Eq. (14)

6: **end for**

Output: $\{p_j^{\text{sel}}\}_{j \in \mathcal{O}_{\text{obs}}}$

---

space $\mathbb{R}$ can be decomposed as:

$$\mathbb{R} = \bigcup_{u \in [U]} \bigcup_{v \in [V^u]} \left\{ \underbrace{z \in \mathbb{R} \mid \mathcal{T}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{T}_u, \mathcal{S}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{S}_v}_{\text{a sub-problem of extra-conditioning}} \right\},$$

where $\mathcal{T}_{\boldsymbol{a}+\boldsymbol{b}z}$ denotes the OT-based DA on $\boldsymbol{a} + \boldsymbol{b}z$, $\mathcal{S}_{\boldsymbol{a}+\boldsymbol{b}z}$ denote a set of steps performed by the AD algorithm after DA. Our goal is to search a set

$$\mathcal{V} = \left\{ (u,v) : \mathcal{M}(\mathcal{T}_u, \mathcal{S}_v) = \mathcal{O}_{\text{obs}}, \right\}, \qquad (13)$$

for all $u \in [U], v \in [V^u]$, the function $\mathcal{M}$ is defined as:

$$\mathcal{M} : (\mathcal{T}_{\boldsymbol{a}+\boldsymbol{b}z}, \mathcal{S}_{\boldsymbol{a}+\boldsymbol{b}z}) \mapsto \mathcal{O}_{\boldsymbol{a}+\boldsymbol{b}z}.$$

Finally, the region $\mathcal{Z}$ in (11) can be obtained as follows:

$$\mathcal{Z} = \left\{ z \in \mathbb{R} \mid \mathcal{O}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{O}_{\text{obs}} \right\}$$
$$= \bigcup_{(u,v) \in \mathcal{V}} \left\{ z \in \mathbb{R} \mid \mathcal{T}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{T}_u, \mathcal{S}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{S}_v \right\}. \quad (14)$$

Extra-conditioning (step 1). For any $u \in [U]$ and $v \in [V^u]$, we define the subset of one-dimensional projected dataset on a line for the extra-conditioning as:

$$\mathcal{Z}_{u,v} = \left\{ z \in \mathbb{R} \mid \mathcal{T}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{T}_u, \mathcal{S}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{S}_v \right\}. \quad (15)$$

The extra-conditioning region can be re-written as:

$$\mathcal{Z}_{u,v} = \mathcal{Z}_u \cap \mathcal{Z}_v,$$
$$\mathcal{Z}_u = \{ z \mid \mathcal{T}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{T}_u \}, \quad \mathcal{Z}_v = \{ z \mid \mathcal{S}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{S}_v \}.$$

Lemma 3. *The set $\mathcal{Z}_u$ can be characterized by a set of quadratic inequalities w.r.t. z described as follows:*

$$\mathcal{Z}_u = \left\{ z \in \mathbb{R} \mid \boldsymbol{w} + \boldsymbol{r}z + \boldsymbol{o}z^2 \geq \boldsymbol{0} \right\},$$

where vectors $\boldsymbol{w}$, $\boldsymbol{r}$, and $\boldsymbol{o}$ are defined in Appendix 6.3.

Lemma 4. *The set $\mathcal{Z}_v$, which represents the set of operation performed by the MAD algorithm, can be characterized by a set of linear inequalities:*

$$\mathcal{Z}_v = \left\{ z \in \mathbb{R} \mid \boldsymbol{p}z \leq \boldsymbol{q} \right\}, \qquad (16)$$

where vectors $\boldsymbol{p}$ and $\boldsymbol{q}$ are provided in Appendix 6.4.

The proofs of Lemmas 3 and 4 are deferred to Appendices 6.3 and 6.4. Since the definitions of $\boldsymbol{w}$, $\boldsymbol{r}$, $\boldsymbol{o}$, $\boldsymbol{p}$, and $\boldsymbol{q}$ are complex and require extensive descriptions, we deferred them to the Appendices. Briefly speaking, they are used for ensuring that the transportation and all the steps of the AD after DA remains the same for all $z \in \mathcal{Z}_{u,v}$. Lemmas 3 and 4 indicate that $\mathcal{Z}_u$ and $\mathcal{Z}_v$ can be *analytically obtained* by solving the systems of quadratic and linear inequalities, respectively. After computing $\mathcal{Z}_u$ and $\mathcal{Z}_v$, the extra conditioning region $\mathcal{Z}_{u,v}$ in (15) is obtained by $\mathcal{Z}_{u,v} = \mathcal{Z}_u \cap \mathcal{Z}_v$. As mentioned in §2, our proposed CAD-AD can be applied to other AD algorithms whose operations can be characterized by sets of linear/quadratic inequalities (e.g., least absolute deviations, Huber regression).

Remark 3. *The selective p-value computed with the extra-conditioning region $\mathcal{Z}_{u,v}$ is still valid and this fact is well-known in the literature of conditional SI. The computation of $\mathcal{Z}_{u,v}$ can be considered as an extension of the methods presented in Lee et al. (2016) and Duy and Takeuchi (2023) into our proposed setting. However, the major drawback of this case is that the TPR is low, i.e., the FNR is high. Therefore, we introduce the line search step to remove the extra-conditioning for the purpose of minimizing the FNR. The illustration is shown in Fig. 3.*

Hierarchical line search (step 2). Our strategy is to identify $\mathcal{V}$ in (13) by repeatedly applying OT-based

DA and MAD after DA to a sequence of datasets $\boldsymbol{a} + \boldsymbol{b}z$ within sufficiently wide range of $z \in [z_{\min}, z_{\max}]^1$. For simplicity, we consider the case in which $\mathcal{Z}_u$ is an interval [2]. Since $\mathcal{Z}_v$ is also an interval, $\mathcal{Z}_{u,v}$ is an interval. We denote $\mathcal{Z}_u = [L_u, R_u]$ and $\mathcal{Z}_{u,v} = [L_{u,v}, R_{u,v}]$. The hierarchical line search procedure can be summarized in Algorithm 1. After obtaining $\mathcal{V}$ by Algorithm 1. We can compute $\mathcal{Z}$ in (14), which is subsequently used to obtain the proposed selective p-value in (12). The entire steps of the proposed CAD-DA method is summarized in Algorithm 2.

3.4 Extension to Multi-Dimension

In this section, we generalize the problem setup and the proposed method in multi-dimension. We consider two random sets $X^s \in \mathbb{R}^{n_s \times d}$ and $X^t \in \mathbb{R}^{n_t \times d}$ of d-dimensional vectors which are random samples from

$$\text{vec}(X^s) \sim \mathbb{N}\Big(\text{vec}(M^s),\ \Sigma^s_{\text{vec}(X^s)}\Big),$$
$$\text{vec}(X^t) \sim \mathbb{N}\Big(\text{vec}(M^t),\ \Sigma^t_{\text{vec}(X^t)}\Big),$$

where $\text{vec}(\cdot)$ is an operator that transforms a matrix into a vector with concatenated rows, M^s and M^t are signal matrices that are unknown, $\Sigma^s_{\text{vec}(X^s)}$ and $\Sigma^t_{\text{vec}(X^t)}$ are variance matrices that are known. After obtaining $\hat{T}$ by solving the OT problem, we obtain $\tilde{X}^s = n_s \hat{T} X^t$. Then, the AD result after DA is:

$$\mathcal{A} : \big\{\tilde{X}^s, X^t\big\} \mapsto \mathcal{O} \in [n_t].$$

For $j \in \mathcal{O}$, we consider the following hypotheses:

$$\text{H}_{0,j} : M^t_{j,\kappa} = \bar{M}^t_{-\mathcal{O},\kappa} \quad \text{vs.} \quad \text{H}_{1,j} : M^t_{j,\kappa} \neq \bar{M}^t_{-\mathcal{O},\kappa},$$

for all $\kappa \in [d]$, $\bar{M}^t_{-\mathcal{O},\kappa} = \frac{1}{n_t - |\mathcal{O}|} \sum_{\ell \in [n_t] \setminus \mathcal{O}} M^t_{\ell,\kappa}$.

The test-statistic is then can be defined as:

$$\Gamma_j = \sum_{\kappa \in [d]} |X^t_{j,\kappa} - \bar{X}^t_{-\mathcal{O},\kappa}| \qquad (17)$$

In order to compute the valid p-values, we consider the following conditional distribution of the test statistic:

$$\mathbb{P}\Big(\Gamma_j \mid \mathcal{O}_{X^s,X^t} = \mathcal{O}_{\text{obs}},\ \mathcal{I}_{X^s,X^t} = \mathcal{I}_{\text{obs}}\Big), \qquad (18)$$

where $\mathcal{I}_{X^s,X^t}$ is a set of all the signs of the subtractions in (17). By extending the techniques in §3.2 and §3.3, the conditional space in (18) can be obtained which is subsequently used to compute the p-value.

---

[1] We set $z_{\min} = -20\sigma$ and $z_{\max} = 20\sigma$, σ is the standard deviation of the distribution of the test statistic, because the probability mass outside this range is negligibly small.

[2] If $\mathcal{Z}_u$ is a union of intervals, we can select the interval containing the data point that we are currently considering.
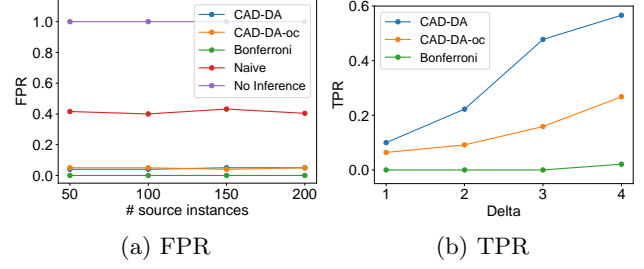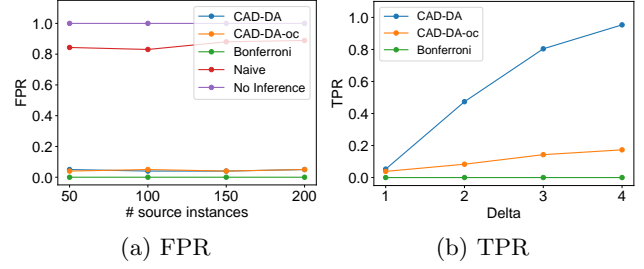

Figure 4: FPR and TPR in univariate case

Figure 5: FPR and TPR in multi-dimensional case
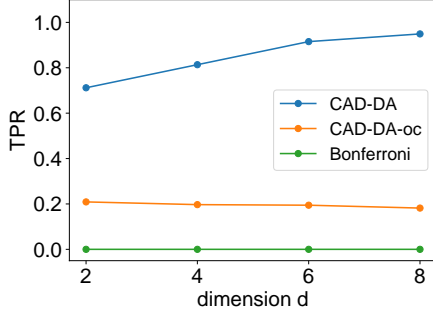
4 EXPERIMENT

In this section, we demonstrate the performance of the proposed method. We compared the performance of the following methods in terms of FPR and TPR:

- **CAD-DA**: proposed method

- **CAD-DA-oc**: proposed method with only the extra-conditioning described in §3.2 (extension the ideas in Lee et al. (2016) and Duy and Takeuchi (2023) to our proposed setting)

- **Bonferroni**: the most popular multiple hypothesis testing approach

- **Naive**: traditional statistical inference

- **No Inference**: AD after DA without inference

We note that if a method cannot control the FPR under α, it is *invalid* and we do not need to consider its TPR. A method has high TPR indicates that it has low FNR. We set $\alpha = 0.05$. We executed the code on Intel(R) Xeon(R) CPU E5-2687W v4 @ 3.00GHz.

4.1 Numerical Experiments

Univariate case. We generated $\boldsymbol{X}^s$ and $\boldsymbol{X}^t$ with $\mu_i^s = 0$, $\varepsilon_i^s \sim \mathbb{N}(0,1)$, for all $i \in [n_s]$, and $\mu_j^t = 2, \varepsilon_j^t \sim \mathbb{N}(0,1)$, for all $j \in [n_t]$. We randomly selected 5 data points in the target domain and made them to be abnormal by setting $\mu_j^t = \mu_j^t + \Delta$. Regarding the FPR experiments, we set $n_s \in \{50, 100, 150, 200\}$, $n_t = 25$, and $\Delta = 0$. In regard to the TPR experiments, we set $n_s = 150$, $n_t = 25$, and $\Delta \in \{1, 2, 3, 4\}$. Each experi-

Figure 6: TPR when changing d

(a) FPR

(b) TPR

Figure 7: FPR and TPR in the case of correlated data

(a) FPR

(b) TPR

Figure 8: FPR and TPR when changing ρ

Figure 9: Comparison with Tsukurimichi et al. (2022).

ment was repeated 120 times. The results are shown in Fig. 4. In the plot on the left, the `CAD-DA`, `CAD-DA-oc` and `Bonferroni` controlled the FPR under α whereas the `Naive` and `No Inference` *could not*. Because the `Naive` and `No Inference` failed to control the FPR, we no longer considered the TPR . In the plot on the right, we can see that the `CAD-DA` has highest TPR compared to other methods in all the cases, i.e., the `CAD-DA` has lowest FNR compared to the competitors.

Multi-dimensional case. We generated X^s and X^t with $X^s_{i,:} \sim \mathbb{N}(\mathbf{0}_d, I_d), \forall i \in [n_s]$, $X^t_{j,:} \sim \mathbb{N}(\mathbf{2}_d, I_d), \forall j \in [n_t]$, and the dimension $d = 10$. The settings for FPR and TPR experiments were the same as univariate case. The results are shown in Fig. 5. The interpretation of the results is similar and consistent with the univariate case. The important point we would like to note is that the difference in TPR between the `CAD-DA` and `CAD-DA-oc` in the multi-dimensional case is larger than that in the univariate case. This indicates that the issue of extra-conditioning is more serious when increasing d. If we naively extend the ideas in Lee et al. (2016); Duy and Takeuchi (2023) to our setting, the TPR will be low, i.e., the FNR is high. In contrast, by introducing the hierarchical linear search step to remove the extra-conditioning, the FNR is significantly decreased by the proposed `CAD-DA` method.

TPR comparison when changing d. We additionally conducted the TPR experiments by varying the dimension $d \in \{2, 4, 6, 8\}$. The results are shown in Fig. 6. Overall, both the `CAD-DA` and `CAD-DA-oc` consistently exhibited higher TPR compared to `Bonferroni` in all cases. Moreover, as the dimension d increased, the TPR disparity between `CAD-DA` and `CAD-DA-oc` became larger. This observation underscores the advantageous performance of the `CAD-DA`, particularly in higher-dimensional scenarios.

Correlated data. In this setting, we consider the case where the data is not independent. We generated X^s and X^t with $X^s_{i,:} \sim \mathbb{N}(\mathbf{0}_d, \Xi), \forall i \in [n_s]$, $X^t_{j,:} \sim \mathbb{N}(\mathbf{2}_d, \Xi), \forall j \in [n_t]$, the matrix $\Xi = \left[\rho^{|i-j|}\right]_{ij}, \forall i, j \in [d]$, $\rho = 0.5$, and $d = 10$. The settings for FPR
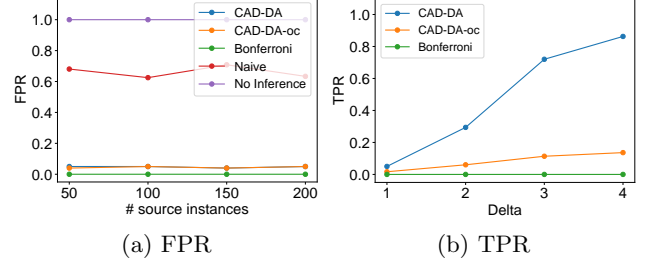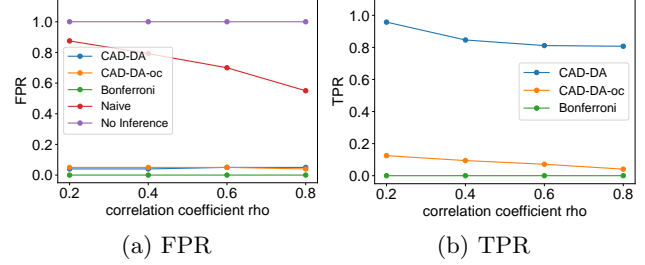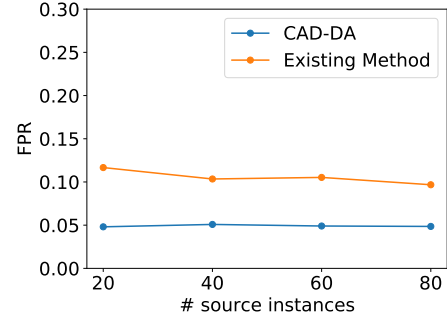
and TPR experiments were also the same as univariate case. The results are shown in Fig. 7. Additionally, we also conducted FPR and TPR experiments when changing $\rho \in \{0.2, 0.4, 0.6, 0.8\}$. We set $n_s = 150, n_t = 25$, $\Delta = 0$ for FPR experiments, and $\Delta = 4$ for the TPR experiments. The results are shown in Fig. 8. In essence, the correlated data contains redundant information. This means that the effective sample size is smaller than the actual sample size. A smaller effective sample size reduces the amount of information available for the statistical test, making it less powerful. Therefore, the TPR tends to decrease when increasing the value of ρ. However, in all the case, the proposed `CAD-DA` method consistently achieves the highest TPR while controlling the FPR.

Comparison with Tsukurimichi et al. (2022) and robustness experiments. Since the method of Tsukurimichi et al. (2022) is not applicable to our setting, we have to introduce an extended setting for

Table 2: Comparison on real-data.

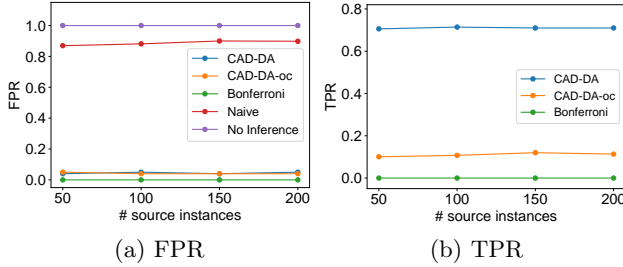|	$N = 120$		$N = 240$	
	FPR	TPR	FPR	TPR
No Inference	1.00	N/A	1.00	N/A
Naive	0.90	N/A	0.90	N/A
Bonferroni	**0.00**	0.00	**0.00**	0.00
CAD-DA-oc	**0.05**	0.12	**0.04**	0.07
CAD-DA	**0.04**	**0.71**	**0.04**	**0.73**

(a) FPR (b) TPR

Figure 10: FPR and TPR comparison on real-data.

conducting the experiments of FPR comparison with their method. The details are provided in Appendix 6.5. The results are shown in Fig. 9. The `CAD-DA` could properly control the FPR under $\alpha = 0.05$ whereas the existing method (Tsukurimichi et al., 2022) failed because it could not account for the influence of DA process. Additionally, we conducted the experiments on the robustness of the `CAD-DA` when the noise follows Laplace distribution, skew normal distribution, t_{20} distribution, and variance is estimated from the data. The details are shown in Appendix 6.6. Overall, the `CAD-DA` method still maintained good performance.

4.2   Real-data Experiments

We performed comparison on real-data. We used the *NeuroKit2* simulator (Makowski et al., 2021) to generate realistic respiration signals ($n_s = 150$) used for source dataset, and heart-beat signals ($n_t = 25$) used for target dataset, each with a dimensionality of $d = 12$. We repeated the experiment $N \in \{120, 240\}$ times. The results are shown in Tab. 2. While the `Bonferroni`, `CAD-DA-oc` and `CAD-DA` could control the FPR, the `CAD-DA` had the highest TPR in all the cases. We note that, in the case of Bonferroni correction, the TPRs were all 0.0 which indicates that this method is too conservative and all the true anomalies were missed out even though there exists, i.e., FNR = 1.0. We also conducted the FPR and TPR comparison by varying $n_s \in \{50, 100, 150, 200\}$. The corresponding results are shown in Fig. 10.

Additionally, we compared the p-values of the
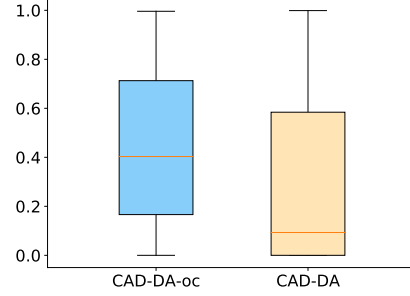
Figure 11: Boxplots of p-values

`CAD-DA-oc` and `CAD-DA` on Heart Disease dataset, which is available at the UCI Machine Learning Repository. We randomly selected $n_s = 150$ patients whose gender is male (source domain), $n_t = 25$ patients whose gender is female (target), and conducted the experiment to compute the p-value. The experiments was repeated 120 times. The boxplots of the distribution of the p-values are illustrated in Fig. 11. The p-values of the `CAD-DA` tend to be smaller than those of `CAD-DA-oc`, which indicates that the proposed `CAD-DA` method has higher power than the `CAD-DA-oc`.

5   DISCUSSION

We propose a novel setup of testing the results of AD after DA and introduce a method to compute a valid p-value for conducting the statistical test. We believe that this study stands as a significant step toward controllable machine leaning under DA. Some open questions remain. Our method currently does not support MMD-based DAs (Gong et al., 2013; Pan et al., 2010; Baktashmotlagh et al., 2013) or metric learning-based DA (Saenko et al., 2010) because the selection event of these methods are more complicated than OT-based DA. A potential solution could involve a sampling-based approach to approximate the truncation region for p-value computation. Similarly, our method is applicable to AD algorithms with analytically characterizable event, such as MAD and robust regression. Extending the proposed method to more complex AD algorithms would represent a valuable contribution.

Another important future direction could be expanding the proposed framework for testing the anomalies in the context of Deep Learning (DL)-based DA. Simply conducting AD after DL-based DA often reveals many FPs, thereby compromising the reliability of the model. Consequently, enhancing the proposed methodology to accommodate DL-based DA would represent a significant contribution to the existing literature. Nevertheless, it is crucial to acknowledge that the DL training is inherently complex, necessitating careful analysis when addressing DL-based DA.

References

C. C. Aggarwal. On abnormality detection in spuriously populated data streams. In *Proceedings of the 2005 siam international conference on data mining*, pages 80–91. SIAM, 2005.

C. C. Aggarwal. *Outlier Analysis*. Springer, 2017.

D. F. Andrews. A robust method for multiple linear regression. *Technometrics*, 16(4):523–531, 1974.

O. Avci, O. Abdeljaber, S. Kiranyaz, M. Hussein, M. Gabbouj, and D. J. Inman. A review of vibration-based damage detection in civil structures: From traditional methods to machine learning and deep learning applications. *Mechanical systems and signal processing*, 147:107077, 2021.

M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE international conference on computer vision*, pages 769–776, 2013.

S. Chen and J. Bien. Valid inference corrected for outlier removal. *Journal of Computational and Graphical Statistics*, pages 1–12, 2019.

Y. T. Chen and D. M. Witten. Selective inference for k-means clustering. *arXiv preprint arXiv:2203.15267*, 2022.

D. Das, V. N. L. Duy, H. Hanada, K. Tsuda, and I. Takeuchi. Fast and more powerful selective inference for sparse high-order interaction model. *arXiv preprint arXiv:2106.04929*, 2021.

Y. Du, S. Zhou, X. Jing, Y. Peng, H. Wu, and N. Kwok. Damage detection techniques for wind turbine blades: A review. *Mechanical Systems and Signal Processing*, 141:106445, 2020.

V. N. L. Duy and I. Takeuchi. More powerful conditional selective inference for generalized lasso by parametric programming. *The Journal of Machine Learning Research*, 23(1):13544–13580, 2022a.

V. N. L. Duy and I. Takeuchi. Exact statistical inference for time series similarity using dynamic time warping by selective inference. *arXiv preprint arXiv:2202.06593*, 2022b.

V. N. L. Duy and I. Takeuchi. Exact statistical inference for the wasserstein distance by selective inference. *Annals of the Institute of Statistical Mathematics*, 75(1):127–157, 2023.

V. N. L. Duy, H. Toda, R. Sugiyama, and I. Takeuchi. Computing valid p-value for optimal changepoint by selective inference using dynamic programming. In *Advances in Neural Information Processing Systems*, 2020.

V. N. L. Duy, S. Iwazaki, and I. Takeuchi. Quantifying statistical significance of neural network-based image segmentation by selective inference. *Advances in Neural Information Processing Systems*, 35:31627–31639, 2022.

W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

W. Fithian, J. Taylor, R. Tibshirani, and R. Tibshirani. Selective sequential model selection. *arXiv preprint arXiv:1512.02565*, 2015.

R. Flamary, N. Courty, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 1:1–40, 2016.

L. L. Gao, J. Bien, and D. Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, pages 1–11, 2022.

B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *International conference on machine learning*, pages 222–230. PMLR, 2013.

P. J. Huber et al. Robust regression: asymptotics, conjectures and monte carlo. *The annals of statistics*, 1(5):799–821, 1973.

S. Hyun, M. G'sell, and R. J. Tibshirani. Exact post-selection inference for the generalized lasso path. *Electronic Journal of Statistics*, 12(1):1053–1097, 2018a.

S. Hyun, K. Lin, M. G'Sell, and R. J. Tibshirani. Post-selection inference for changepoint detection algorithms with application to copy number variation data. *arXiv preprint arXiv:1812.03644*, 2018b.

S. Inoue, Y. Umezu, S. Tsubota, and I. Takeuchi. Post clustering inference for heterogeneous data. *IEICE*

Technical Report; IEICE Tech. Rep., 117(293):69–76, 2017.

S. Jewell, P. Fearnhead, and D. Witten. Testing for a change in mean after changepoint detection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1082–1104, 2022.

J. D. Lee, Y. Sun, and J. E. Taylor. Evaluating the statistical significance of biclusters. *Advances in neural information processing systems*, 28, 2015.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

K. Liu, J. Markovic, and R. Tibshirani. More powerful post-selection inference, with application to the lasso. *arXiv preprint arXiv:1801.09037*, 2018.

J. R. Loftus and J. E. Taylor. A significance test for forward stepwise model selection. *arXiv preprint arXiv:1405.3920*, 2014.

D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen. Neurokit2: A python toolbox for neurophysiological signal processing. *Behavior Research Methods*, Feb 2021. ISSN 1554-3528.

R. A. Maronna, R. D. Martin, V. J. Yohai, et al. *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

D. Miwa, D. V. N. Le, and I. Takeuchi. Valid p-value for deep learning-driven salient region. In *Proceedings of the 11th International Conference on Learning Representation*, 2023.

A. C. Neufeld, L. L. Gao, and D. M. Witten. Treevalues: selective inference for regression trees. *Journal of Machine Learning Research*, 23(305):1–43, 2022.

J.-X. Pan and K.-T. Fang. Multiple outlier detection in growth curve model with unstructured covariance matrix. *Annals of the Institute of Statistical Mathematics*, 47(1):137–153, 1995.

S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2):199–210, 2010.

T. Pourhabibi, K.-L. Ong, B. H. Kam, and Y. L. Boo. Fraud detection: A systematic literature review of graph-based anomaly detection approaches. *Decision Support Systems*, 133:113303, 2020.

P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*, volume 589. John wiley & sons, 2005.

D. Rügamer and S. Greven. Inference for l 2-boosting. *Statistics and computing*, 30(2):279–289, 2020.

K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 213–226. Springer, 2010.

M. S. Srivastava and D. von Rosen. Outliers in multivariate regression models. *Journal of Multivariate Analysis*, 65(2):195–208, 1998.

K. Sugiyama, V. N. Le Duy, and I. Takeuchi. More powerful and general selective inference for stepwise feature selection using homotopy method. In *International Conference on Machine Learning*, pages 9891–9901. PMLR, 2021a.

R. Sugiyama, H. Toda, V. N. L. Duy, Y. Inatsu, and I. Takeuchi. Valid and exact statistical inference for multi-dimensional multiple change-points by selective inference. *arXiv preprint arXiv:2110.08989*, 2021b.

S. Suzumura, K. Nakagawa, Y. Umezu, K. Tsuda, and I. Takeuchi. Selective inference for sparse high-order interaction models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3338–3347. JMLR. org, 2017.

K. Tanizaki, N. Hashimoto, Y. Inatsu, H. Hontani, and I. Takeuchi. Computing valid p-values for image segmentation by selective inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9553–9562, 2020.

R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

T. Tsukurimichi, Y. Inatsu, V. N. L. Duy, and I. Takeuchi. Conditional selective inference for robust regression and outlier detection using piecewise-linear homotopy continuation. *Annals of the Institute of Statistical Mathematics*, 74(6):1197–1228, 2022.

Y. Umezu and I. Takeuchi. Selective inference for change point detection in multi-dimensional sequences. *arXiv preprint arXiv:1706.00514*, 2017.

W.-K. Wong, A. Moore, G. Cooper, and M. Wagner. Rule-based anomaly pattern detection for detecting disease outbreaks. In *AAAI/IAAI*, pages 217–223, 2002.

M. Yamada, Y. Umezu, K. Fukumizu, and I. Takeuchi. Post selection inference with kernels. In *International conference on artificial intelligence and statistics*, pages 152–160. PMLR, 2018.

F. Yang, R. F. Barber, P. Jain, and J. Lafferty. Selective inference for group-sparse linear models. In *Ad-*

vances in Neural Information Processing Systems, pages 2469–2477, 2016.

A. Zaman, P. J. Rousseeuw, and M. Orhan. Econometric applications of high-breakdown robust regression techniques. *Economics Letters*, 71(1):1–8, 2001.

Checklist

1. For all models and algorithms presented, check if you include:

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:

   (a) Statements of the full set of assumptions of all theoretical results. [Yes]

   (b) Complete proofs of all theoretical results. [Yes]

   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Yes]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

6 APPENDIX

6.1 Proof of Lemma 1

We have

$$\boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix} \mid \{\mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{O}_{\mathrm{obs}}, \mathcal{Q}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{Q}_{\mathrm{obs}}\} \sim \mathrm{TN}\left(\boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{\mu}^s \\ \boldsymbol{\mu}^t \end{pmatrix}, \boldsymbol{\eta}_j^\top \Sigma \boldsymbol{\eta}_j, \mathcal{Z}\right),$$

which is a truncated normal distribution with a mean $\boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{\mu}^s \\ \boldsymbol{\mu}^t \end{pmatrix}$, variance $\boldsymbol{\eta}_j^\top \Sigma \boldsymbol{\eta}_j$ in which $\Sigma = \begin{pmatrix} \Sigma^s & 0 \\ 0 & \Sigma^t \end{pmatrix}$, and the truncation region $\mathcal{Z}$ described in Sec. 3.2. Therefore, under the null hypothesis,

$$p_j^{\mathrm{sel}} \mid \{\mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{O}_{\mathrm{obs}}, \mathcal{Q}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{Q}_{\mathrm{obs}}\} \sim \mathrm{Unif}(0, 1).$$

Thus, $\mathbb{P}_{\mathrm{H}_{0,j}}\left(p_j^{\mathrm{sel}} \leq \alpha \mid \mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{O}_{\mathrm{obs}}, \mathcal{Q}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{Q}_{\mathrm{obs}}\right) = \alpha, \forall \alpha \in [0, 1]$. Under the null hypothesis, by conditioning on the results of AD under DA, we can exactly derive the sampling distribution of the test statistic, which is a truncated normal distribution. Then, by applying *the probability integral transform*, the selective p-value follows uniform distribution. We would like to remind that, in probability theory, the probability integral transform states that if X is a continuous random variable with cumulative distribution function F_X, then the random variable $F_X(X)$ has a uniform distribution on $[0, 1]$. Consequently, the p-value, defined as $1 - F_X(X)$, also adheres to a uniform distribution on the same interval $[0, 1]$.

Next, we have

$$\mathbb{P}_{\mathrm{H}_{0,j}}\left(p_j^{\mathrm{sel}} \leq \alpha \mid \mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{O}_{\mathrm{obs}}\right)$$

$$= \int \mathbb{P}_{\mathrm{H}_{0,j}}\left(p_j^{\mathrm{sel}} \leq \alpha \mid \mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{O}_{\mathrm{obs}}, \mathcal{Q}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{Q}_{\mathrm{obs}}\right) \mathbb{P}_{\mathrm{H}_{0,j}}\left(\mathcal{Q}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{Q}_{\mathrm{obs}} \mid \mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{O}_{\mathrm{obs}}\right) d\mathcal{Q}_{\mathrm{obs}}$$

$$= \int \alpha \, \mathbb{P}_{\mathrm{H}_{0,j}}\left(\mathcal{Q}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{Q}_{\mathrm{obs}} \mid \mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{O}_{\mathrm{obs}}\right) d\mathcal{Q}_{\mathrm{obs}}$$

$$= \alpha \int \mathbb{P}_{\mathrm{H}_{0,j}}\left(\mathcal{Q}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{Q}_{\mathrm{obs}} \mid \mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{O}_{\mathrm{obs}}\right) d\mathcal{Q}_{\mathrm{obs}}$$

$$= \alpha.$$

Finally, we obtain the result in Lemma 1 as follows:

$$\mathbb{P}_{\mathrm{H}_{0,j}}\left(p_j^{\mathrm{sel}} \leq \alpha\right) = \sum_{\mathcal{O}_{\mathrm{obs}}} \mathbb{P}_{\mathrm{H}_{0,j}}\left(p_j^{\mathrm{sel}} \leq \alpha \mid \mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{O}_{\mathrm{obs}}\right) \mathbb{P}_{\mathrm{H}_{0,j}}\left(\mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{O}_{\mathrm{obs}}\right)$$

$$= \sum_{\mathcal{O}_{\mathrm{obs}}} \alpha \, \mathbb{P}_{\mathrm{H}_{0,j}}\left(\mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{O}_{\mathrm{obs}}\right)$$

$$= \alpha \sum_{\mathcal{O}_{\mathrm{obs}}} \mathbb{P}_{\mathrm{H}_{0,j}}\left(\mathcal{O}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{O}_{\mathrm{obs}}\right)$$

$$= \alpha.$$

6.2 Proof of Lemma 2

According to the second condition in (9), we have

$$\mathcal{Q}_{\boldsymbol{X}^s, \boldsymbol{X}^t} = \mathcal{Q}_{\mathrm{obs}}$$

$$\Leftrightarrow \left(I_{n_s + n_t} - \boldsymbol{b}\boldsymbol{\eta}_j^\top\right) \begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix} = \mathcal{Q}_{\mathrm{obs}}$$

$$\Leftrightarrow \begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix} = \mathcal{Q}_{\mathrm{obs}} + \boldsymbol{b}\boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix}.$$

By defining $\boldsymbol{a} = \mathcal{Q}_{\mathrm{obs}}$, $z = \boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix}$, and incorporating the first condition of (9), we obtain Lemma 2.

6.3 Proof of Lemma 3

The proof is constructed based on the results presented in Duy and Takeuchi (2023), in which the authors introduced an approach to characterize the event of OT by using the concept of *parametric linear programming*. Let us re-written the OT problem between the source and target domain in (1) as:

$$\hat{\boldsymbol{t}} = \underset{\boldsymbol{t} \in \mathbb{R}^{n_s n_t}}{\arg\min} \quad \boldsymbol{t}^\top \boldsymbol{c}(\boldsymbol{X}^s, \boldsymbol{X}^t)$$

$$\text{s.t. } H\boldsymbol{t} = \boldsymbol{h}, \ \boldsymbol{t} \geq \boldsymbol{0},$$

where $\boldsymbol{t} = \text{vec}(T)$, $\boldsymbol{c}(\boldsymbol{X}^s, \boldsymbol{X}^t) = \text{vec}(C(\boldsymbol{X}^s, \boldsymbol{X}^t)) = \left[\Omega\binom{\boldsymbol{X}^s}{\boldsymbol{X}^t}\right] \circ \left[\Omega\binom{\boldsymbol{X}^s}{\boldsymbol{X}^t}\right]$,

$$\Omega = \text{hstack}\left(I_{n_s} \otimes \mathbf{1}_{n_t}, -\mathbf{1}_{n_s} \otimes I_{n_t}\right) \in \mathbb{R}^{n_s n_t \times (n_s + n_t)},$$

$\text{vec}(\cdot)$ is an operator that transforms a matrix into a vector with concatenated rows, the operator $\circ$ is element-wise product, $\text{hstack}(\cdot, \cdot)$ is horizontal stack operation, $I_n \in \mathbb{R}^{n \times n}$ is the identity matrix, and $\mathbf{1}_m \in \mathbb{R}^m$ is a vector of ones. The matrix H is defined as $H = (H_r \ H_c)^\top \in \mathbb{R}^{(n_s + n_t) \times n_s n_t}$ in which

$$H_r = \begin{bmatrix} 1 \ldots 1 & 0 \ldots 0 & \ldots & 0 \ldots 0 \\ 0 \ldots 0 & 1 \ldots 1 & \ldots & 0 \ldots 0 \\ \ldots & & \ldots & \ldots & \ldots \\ 0 \ldots 0 & 0 \ldots 0 & \ldots & 1 \ldots 1 \end{bmatrix} \in \mathbb{R}^{n_s \times n_s n_t}$$

that performs the sum over the rows of T and

$$H_c = \begin{bmatrix} I_{n_t} & I_{n_t} & \ldots & I_{n_t} \end{bmatrix} \in \mathbb{R}^{n_t \times n_s n_t}$$

that performs the sum over the columns of T, and $\boldsymbol{h} = \left(\mathbf{1}_{n_s}/n_s \ \mathbf{1}_{n_t}/n_t\right)^\top \in \mathbb{R}^{n_s + n_t}$.

Next, we consider the OT problem with the parametrized data $\boldsymbol{a} + \boldsymbol{b}z$:

$$\underset{\boldsymbol{t} \in \mathbb{R}^{n_s n_t}}{\min} \quad \boldsymbol{t}^\top [\Omega(\boldsymbol{a} + \boldsymbol{b}z) \circ \Omega(\boldsymbol{a} + \boldsymbol{b}z)] \ \text{ s.t. } H\boldsymbol{t} = \boldsymbol{h}, \boldsymbol{t} \geq \boldsymbol{0}$$

$$\Leftrightarrow \underset{\boldsymbol{t} \in \mathbb{R}^{n_s n_t}}{\min} \quad (\tilde{\boldsymbol{w}} + \tilde{\boldsymbol{r}}z + \tilde{\boldsymbol{o}}z^2)^\top \boldsymbol{t} \ \text{ s.t. } H\boldsymbol{t} = \boldsymbol{h}, \boldsymbol{t} \geq \boldsymbol{0},$$

where

$$\tilde{\boldsymbol{w}} = (\Omega\boldsymbol{a}) \circ (\Omega\boldsymbol{a}), \quad \tilde{\boldsymbol{r}} = (\Omega\boldsymbol{a}) \circ (\Omega\boldsymbol{b}) + (\Omega\boldsymbol{b}) \circ (\Omega\boldsymbol{a}), \quad \text{and} \quad \tilde{\boldsymbol{o}} = (\Omega\boldsymbol{b}) \circ (\Omega\boldsymbol{b}).$$

By fixing $\mathcal{B}_u$ as the optimal basic index set of the linear program, the *relative cost vector* w.r.t to the set of non-basic variables $\mathcal{B}_u^c$ is defines as

$$\boldsymbol{r}_{\mathcal{B}_u^c} = \boldsymbol{w} + \boldsymbol{r}z + \boldsymbol{o}z^2,$$

where

$$\boldsymbol{w} = \left(\tilde{\boldsymbol{w}}_{\mathcal{B}_u^c}^\top - \tilde{\boldsymbol{w}}_{\mathcal{B}_u}^\top H_{:,\mathcal{B}_u}^{-1} H_{:,\mathcal{B}_u^c}\right)^\top, \quad \boldsymbol{r} = \left(\tilde{\boldsymbol{r}}_{\mathcal{B}_u^c}^\top - \tilde{\boldsymbol{r}}_{\mathcal{B}_u}^\top H_{:,\mathcal{B}_u}^{-1} H_{:,\mathcal{B}_u^c}\right)^\top, \quad \boldsymbol{o} = \left(\tilde{\boldsymbol{o}}_{\mathcal{B}_u^c}^\top - \tilde{\boldsymbol{o}}_{\mathcal{B}_u}^\top H_{:,\mathcal{B}_u}^{-1} H_{:,\mathcal{B}_u^c}\right)^\top, \quad (19)$$

$H_{:,\mathcal{B}_u}$ is a sub-matrix of S made up of all rows and columns in the set $\mathcal{B}_u$. The requirement for $\mathcal{B}_u$ to be the optimal basis index set is $\boldsymbol{r}_{\mathcal{B}_u^c} \geq \boldsymbol{0}$ (i.e., the cost in minimization problem will never decrease when the non-basic variables become positive and enter the basis). We note that the optimal basis index set $\mathcal{B}_u$ corresponds to the transportation $\mathcal{T}_u$. Therefore, the set $\mathcal{Z}_u$ can be defined as

$$\mathcal{Z}_u = \left\{z \in \mathbb{R} \mid \mathcal{T}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{T}_u\right\},$$

$$= \left\{z \in \mathbb{R} \mid \mathcal{B}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{B}_u\right\},$$

$$= \left\{z \in \mathbb{R} \mid \boldsymbol{r}_{\mathcal{B}_u^c} = \boldsymbol{w} + \boldsymbol{r}z + \boldsymbol{o}z^2 \geq \boldsymbol{0}\right\}.$$

Thus, we obtain the result in Lemma 3.

6.4 Proof of Lemma 4

For notational simplicity, let us denote the original data and the data after OT-based DA as follows:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{X}^s \\ \boldsymbol{X}^t \end{pmatrix} \in \mathbb{R}^{n_s + n_t}, \quad \tilde{\boldsymbol{Y}} = \begin{pmatrix} \hat{\boldsymbol{X}}^s \\ \boldsymbol{X}^t \end{pmatrix} = \Theta \boldsymbol{Y}, \text{ where } \Theta = \begin{pmatrix} 0_{n_s \times n_s} & n_s \hat{T} \\ 0_{n_t \times n_s} & I_{n_t} \end{pmatrix} \in \mathbb{R}^{(n_s + n_t) \times (n_s + n_t)}.$$

The procedure of applying MAD on $\tilde{\boldsymbol{Y}}$ is described as follows:

1. $\tilde{Y}_{k_1} = \text{median}(\hat{\boldsymbol{Y}})$. This step can be represented by the following sets:

$$\mathcal{S}_v^{1a} = \left\{ (k, k_1) : \tilde{Y}_k \leq \tilde{Y}_{k_1} \right\}, \quad \mathcal{S}_v^{1b} = \left\{ (k, k_1) : \tilde{Y}_k \geq \tilde{Y}_{k_1} \right\}, \quad k \in [n_s + n_t]. \tag{20}$$

2. $|\tilde{Y}_{k_2} - \tilde{Y}_{k_1}| = \text{median}\left(\left\{ |\tilde{Y}_k - \tilde{Y}_{k_1}| \right\}_{k \in [n_s + n_t]} \right)$. This step can be represented by the following sets:

$$\mathcal{S}_v^{2a} = \left\{ s_k : s_k = \text{sign}\left(\tilde{Y}_k - \tilde{Y}_{k_1} \right) \right\}, \tag{21a}$$

$$\mathcal{S}_v^{2b} = \left\{ (k, k_2) : s_k\left(\tilde{Y}_k - \tilde{Y}_{k_1} \right) \leq s_{k_2}\left(\tilde{Y}_{k_2} - \tilde{Y}_{k_1} \right) \right\}, \tag{21b}$$

$$\mathcal{S}_v^{2c} = \left\{ (k, k_2) : s_k\left(\tilde{Y}_k - \tilde{Y}_{k_1} \right) \geq s_{k_2}\left(\tilde{Y}_{k_2} - \tilde{Y}_{k_1} \right) \right\}, \tag{21c}$$

for any $k \in [n_s + n_t]$.

3. Given γ, $\tilde{Y}_k$ is considered to be an anomaly if $\tilde{Y}_k \notin \left[\tilde{Y}_{k_1} - \gamma |\tilde{Y}_{k_2} - \tilde{Y}_{k_1}|, \tilde{Y}_{k_1} + \gamma |\tilde{Y}_{k_2} - \tilde{Y}_{k_1}| \right]$. This step can be represented by the following sets:

$$\mathcal{S}_v^{3a} = \left\{ k \in [n_s + n_t] : \tilde{Y}_k < \tilde{Y}_{k_1} - \gamma s_{k_2}\left(\tilde{Y}_{k_2} - \tilde{Y}_{k_1} \right) \right\}, \tag{22a}$$

$$\mathcal{S}_v^{3b} = \left\{ k \in [n_s + n_t] : \tilde{Y}_k > \tilde{Y}_{k_1} + \gamma s_{k_2}\left(\tilde{Y}_{k_2} - \tilde{Y}_{k_1} \right) \right\}, \tag{22b}$$

$$\mathcal{S}_v^{3c} = \left\{ k \in [n_s + n_t] : \tilde{Y}_{k_1} - \gamma s_{k_2}\left(\tilde{Y}_{k_2} - \tilde{Y}_{k_1} \right) \leq \tilde{Y}_k \leq \tilde{Y}_{k_1} + \gamma s_{k_2}\left(\tilde{Y}_{k_2} - \tilde{Y}_{k_1} \right) \right\}. \tag{22c}$$

The entire MAD algorithm can be represented as $\mathcal{S}_v = \mathcal{S}_v^{1a} \cup \mathcal{S}_v^{1b} \cup \mathcal{S}_v^{2a} \cup \mathcal{S}_v^{2b} \cup \mathcal{S}_v^{2c} \cup \mathcal{S}_v^{3a} \cup \mathcal{S}_v^{3b} \cup \mathcal{S}_v^{3c}$.

For any data point $\boldsymbol{a} + \boldsymbol{b}z$, if it satisfies all the inequalities from (20) to (22), $\mathcal{S}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{S}_v$. Regarding the inequalities in $\mathcal{S}_v^{1a}$ of (20) w.r.t. $\boldsymbol{a} + \boldsymbol{b}z$, we can write

$$\boldsymbol{e}_k^\top \Theta(\boldsymbol{a} + \boldsymbol{b}z) \leq \boldsymbol{e}_{k_1}^\top \Theta(\boldsymbol{a} + \boldsymbol{b}z)$$
$$\Leftrightarrow (\boldsymbol{e}_k - \boldsymbol{e}_{k_1})^\top \Theta \boldsymbol{b}z \leq (\boldsymbol{e}_{k_1} - \boldsymbol{e}_k)^\top \Theta \boldsymbol{a},$$

for all $(k, k_1) \in \mathcal{S}_v^{1a}$, and $\boldsymbol{e}_k \in \mathbb{R}^{n_s + n_t}$ is a vector with 1 at the kth, and 0 otherwise. Then we have a system of linear inequalities $\boldsymbol{p}^{1a} z \leq \boldsymbol{q}^{1a}$ where

$$\boldsymbol{p}^{1a} = \text{vector}\left(\left\{ \boldsymbol{e}_{k,k_1}^\top \Theta \boldsymbol{b} \right\}_{(k,k_1) \in \mathcal{S}_v^{1a}} \right), \quad \boldsymbol{q}^{1a} = \text{vector}\left(\left\{ \boldsymbol{e}_{k_1,k}^\top \Theta \boldsymbol{a} \right\}_{(k,k_1) \in \mathcal{S}_v^{1a}} \right),$$

vector$(\cdot)$ is the operation that converts a set to a vector, and $\boldsymbol{e}_{k,k_1} = \boldsymbol{e}_k - \boldsymbol{e}_{k_1}$. Similarly, we obtain $\boldsymbol{p}^{1b} z \leq \boldsymbol{q}^{1b}$, where

$$\boldsymbol{p}^{1b} = \text{vector}\left(\left\{ \boldsymbol{e}_{k_1,k}^\top \Theta \boldsymbol{b} \right\}_{(k,k_1) \in \mathcal{S}_v^{1a}} \right), \quad \boldsymbol{q}^{1b} = \text{vector}\left(\left\{ \boldsymbol{e}_{k,k_1}^\top \Theta \boldsymbol{a} \right\}_{(k,k_1) \in \mathcal{S}_v^{1a}} \right).$$

In regard to (21), we have three systems of linear inequalities $\boldsymbol{p}^{2a} z \leq \boldsymbol{q}^{2a}$, $\boldsymbol{p}^{2b} z \leq \boldsymbol{q}^{2b}$, and $\boldsymbol{p}^{2c} z \leq \boldsymbol{q}^{2c}$, where

$$\boldsymbol{p}^{2a} = \text{vector}\left(\left\{ s_k \boldsymbol{e}_{k_1,k}^\top \Theta \boldsymbol{b} \right\}_{s_k \in \mathcal{S}_v^{2a}} \right), \quad \boldsymbol{q}^{2a} = \text{vector}\left(\left\{ s_k \boldsymbol{e}_{k,k_1}^\top \Theta \boldsymbol{a} \right\}_{s_k \in \mathcal{S}_v^{2a}} \right),$$

$$\boldsymbol{p}^{2b} = \text{vector}\left(\left\{ (s_k \boldsymbol{e}_{k,k_1} - s_{k_2} \boldsymbol{e}_{k_2,k_1})^\top \Theta \boldsymbol{b} \right\}_{(k,k_2) \in \mathcal{S}_v^{2b}} \right), \quad \boldsymbol{q}^{2b} = \text{vector}\left(\left\{ (s_{k_2} \boldsymbol{e}_{k_2,k_1} - s_k \boldsymbol{e}_{k,k_1})^\top \Theta \boldsymbol{a} \right\}_{(k,k_2) \in \mathcal{S}_v^{2b}} \right),$$

$$\boldsymbol{p}^{2c} = \text{vector}\left(\left\{ (s_{k_2} \boldsymbol{e}_{k_2,k_1} - s_k \boldsymbol{e}_{k,k_1})^\top \Theta \boldsymbol{b} \right\}_{(k,k_2) \in \mathcal{S}_v^{2c}} \right), \quad \boldsymbol{q}^{2c} = \text{vector}\left(\left\{ (s_k \boldsymbol{e}_{k,k_1} - s_{k_2} \boldsymbol{e}_{k_2,k_1})^\top \Theta \boldsymbol{a} \right\}_{(k,k_2) \in \mathcal{S}_v^{2c}} \right).$$

With regard to (22), we have four systems of linear inequalities $\boldsymbol{p}^{3a}z \le \boldsymbol{q}^{3a}$, $\boldsymbol{p}^{3b}z \le \boldsymbol{q}^{3b}$, $\boldsymbol{p}^{3c,1}z \le \boldsymbol{q}^{3c,1}$, and $\boldsymbol{p}^{3c,2}z \le \boldsymbol{q}^{3c,2}$ (the (22c) corresponds to two systems of linear inequalities), where

$$\boldsymbol{p}^{3a} = \text{vector}\left(\left\{\left(\gamma s_{k_2}\boldsymbol{e}_{k_2,k_1} - \boldsymbol{e}_{k_1,k}\right)^\top \Theta \boldsymbol{b}\right\}_{k \in \mathcal{S}_v^{3a}}\right), \; \boldsymbol{q}^{3a} = \text{vector}\left(\left\{\left(\boldsymbol{e}_{k_1,k} - \gamma s_{k_2}\boldsymbol{e}_{k_2,k_1}\right)^\top \Theta \boldsymbol{a}\right\}_{k \in \mathcal{S}_v^{3a}}\right),$$

$$\boldsymbol{p}^{3b} = \text{vector}\left(\left\{\left(\gamma s_{k_2}\boldsymbol{e}_{k_2,k_1} - \boldsymbol{e}_{k,k_1}\right)^\top \Theta \boldsymbol{b}\right\}_{k \in \mathcal{S}_v^{3b}}\right), \; \boldsymbol{q}^{3b} = \text{vector}\left(\left\{\left(\boldsymbol{e}_{k,k_1} - \gamma s_{k_2}\boldsymbol{e}_{k_2,k_1}\right)^\top \Theta \boldsymbol{a}\right\}_{k \in \mathcal{S}_v^{3b}}\right),$$

$$\boldsymbol{p}^{3c,1} = \text{vector}\left(\left\{\left(\boldsymbol{e}_{k_1,k} - \gamma s_{k_2}\boldsymbol{e}_{k_2,k_1}\right)^\top \Theta \boldsymbol{b}\right\}_{k \in \mathcal{S}_v^{3c}}\right), \; \boldsymbol{q}^{3c,1} = \text{vector}\left(\left\{\left(\gamma s_{k_2}\boldsymbol{e}_{k_2,k_1} - \boldsymbol{e}_{k_1,k}\right)^\top \Theta \boldsymbol{a}\right\}_{k \in \mathcal{S}_v^{3c}}\right),$$

$$\boldsymbol{p}^{3c,2} = \text{vector}\left(\left\{\left(\boldsymbol{e}_{k,k_1} - \gamma s_{k_2}\boldsymbol{e}_{k_2,k_1}\right)^\top \Theta \boldsymbol{b}\right\}_{k \in \mathcal{S}_v^{3c}}\right), \; \boldsymbol{q}^{3c,2} = \text{vector}\left(\left\{\left(\gamma s_{k_2}\boldsymbol{e}_{k_2,k_1} - \boldsymbol{e}_{k,k_1}\right)^\top \Theta \boldsymbol{a}\right\}_{k \in \mathcal{S}_v^{3c}}\right).$$

Finally, by defining

$$\boldsymbol{p} = \text{vstack}\left(\boldsymbol{p}^{1a}, \boldsymbol{p}^{1b}, \boldsymbol{p}^{2a}, \boldsymbol{p}^{2b}, \boldsymbol{p}^{2c}, \boldsymbol{p}^{3a}, \boldsymbol{p}^{3b}, \boldsymbol{p}^{3c,1}, \boldsymbol{p}^{3c,2}\right), \tag{23}$$

$$\text{and} \quad \boldsymbol{q} = \text{vstack}\left(\boldsymbol{q}^{1a}, \boldsymbol{q}^{1b}, \boldsymbol{q}^{2a}, \boldsymbol{q}^{2b}, \boldsymbol{q}^{2c}, \boldsymbol{q}^{3a}, \boldsymbol{q}^{3b}, \boldsymbol{q}^{3c,1}, \boldsymbol{q}^{3c,2}\right), \tag{24}$$

we obtain the result in Lemma 4.

6.5 Comparison with Tsukurimichi et al. (2022)

Since the method of Tsukurimichi et al. (2022) is not applicable to our proposed setting, we have to introduce an extended setting for conducting the experiments of FPR comparison with their method. The method of Tsukurimichi et al. (2022) primarily focus on a regression problem. Let us consider $(X^s, \boldsymbol{Y}^s)$ and $(X^t, \boldsymbol{Y}^t)$, where $X^s \in \mathbb{R}^{n_s \times p}$ and $X^t \in \mathbb{R}^{n_t \times p}$ are given feature matrices assumed to be non-random,

$$\mathbb{R}^{n_s} \ni \boldsymbol{Y}^s \sim \mathbb{N}(\boldsymbol{\mu}^s, \Sigma^s) \quad \text{and} \quad \mathbb{R}^{n_t} \ni \boldsymbol{Y}^t \sim \mathbb{N}(\boldsymbol{\mu}^t, \Sigma^t).$$

The cost matrix is defined as

$$C(\boldsymbol{Y}^s, \boldsymbol{Y}^t) = \left[(Y_i^s - Y_j^t)^2\right]_{ij} \in \mathbb{R}^{n_s \times n_t}.$$

After obtaining $\hat{T}$ by solving the OT problem with $C(\boldsymbol{Y}^s, \boldsymbol{Y}^t)$, we transform the data from the source domain to the target domain and conduct a robust regression in the target domain. The problem of hypothesis testing is the same as Tsukurimichi et al. (2022). However, when computing the p-value for conducting the test, the method of Tsukurimichi et al. (2022) does not take into account the influence of DA. Therefore, their method is invalid and could not control the FPR. In contrast, with the proposed method, we can successfully control the FPR by handling the effect of DA process. In the experiment for FPR comparison, we set $n_s \in \{20, 40, 60, 80\}$, $n_t = 15$, $p = 5$ and used LAD (least absolute deviations) as the robust regression method. The results are shown in Fig. 9. The proposed `CAD-DA` could properly control the FPR under $\alpha = 0.05$ whereas the existing method (Tsukurimichi et al., 2022) failed because it could not account for the influence of DA process.

6.6 Robustness experimentes

We conducted the following experiments:

- Non-normal data: we considered the noise following Laplace distribution, skew normal distribution (skewness coefficient 10) and t_{20} distribution. We set $n_s \in \{100, 150, 200, 250\}$, $n_t = 45$, and $d = 1$. Each experiment was repeated 120 times. We tested the FPR for both $\alpha = 0.05$ and $\alpha = 0.1$. The FPR results are shown in Figs. 12a, 12b and 12c. We confirmed that the `CAD-DA` still maintained good performance on FPR control.

- Estimated variance: the variances of the noises were estimated from the data by using empirical variance. Our proposed `CAD-DA` method could properly control the FPR (Fig. 12d).
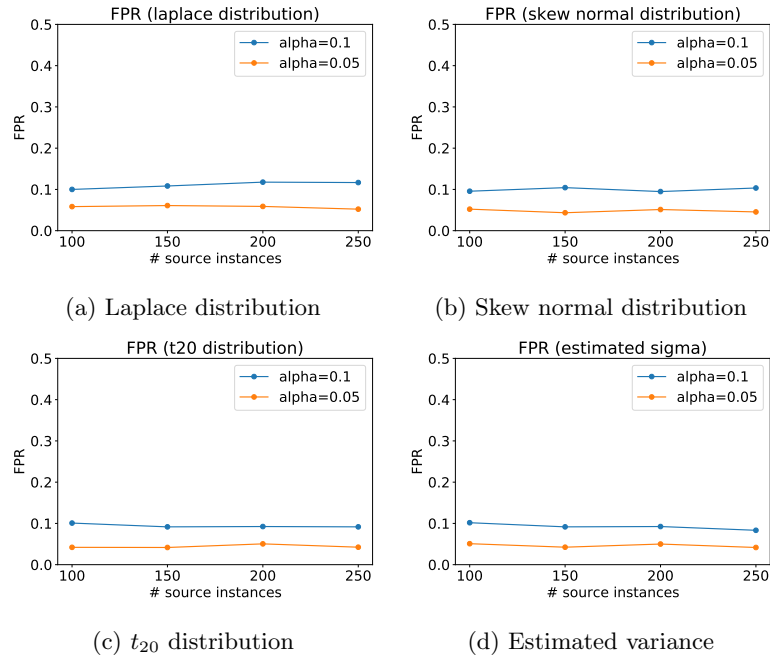
(a) Laplace distribution

(b) Skew normal distribution

(c) t_{20} distribution

(d) Estimated variance

Figure 12: False positive rate of the proposed `CAD-DA` method when data is non-normal or variance is unknown.