# Statistical Inference for Feature Selection after Optimal Transport-based Domain Adaptation

**Nguyen Thang Loi**[1,2], **Duong Tan Loc**[1,2], **Vo Nguyen Le Duy**[1,2,3,*]

[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
[3]RIKEN

## Abstract

Feature Selection (FS) under domain adaptation (DA) is a critical task in machine learning, especially when dealing with limited target data. However, existing methods lack the capability to guarantee the reliability of FS under DA. In this paper, we introduce a novel statistical method to statistically test FS reliability under DA, named SFS-DA (statistical FS-DA). The key strength of SFS-DA lies in its ability to control the false positive rate (FPR) below a pre-specified level $\alpha$ (e.g., 0.05) while maximizing the true positive rate. Compared to the literature on statistical FS, SFS-DA presents a unique challenge in addressing the effect of DA to ensure the validity of the inference on FS results. We overcome this challenge by leveraging the Selective Inference (SI) framework. Specifically, by carefully examining the FS process under DA whose operations can be characterized by linear and quadratic inequalities, we prove that achieving FPR control in SFS-DA is indeed possible. Furthermore, we enhance the true detection rate by introducing a more strategic approach. Experiments conducted on both synthetic and real-world datasets robustly support our theoretical results, showcasing the SFS-DA's superior performance.

## 1 INTRODUCTION

Feature selection (FS) is an important task in machine learning (ML), aimed at identifying the most relevant features from a dataset while discarding redundant or irrelevant ones. By reducing the dimensionality of the data, FS enhances model interpretability, reduces overfitting, and improves computational efficiency. It plays a critical role in high-dimensional data settings, where the number of features often exceeds the number of observations. Common FS techniques such as Lasso (Tibshirani, 1996) and stepwise feature selection plays a critical role in several applications and has been widely applied in many areas such as economics and finance (Tian et al., 2015; Coad and Srhoj, 2020), bioinformatics (Wu et al., 2009; Ma and Huang, 2008), and chemoinformatics (Lo et al., 2018).

In many applications, limited data availability can impair the performance of FS models. Domain adaptation (DA) offers a solution by allowing models to transfer data points from a source domain with abundant labeled data to a target domain with limited labeled data. This approach capitalizes the similarities between the two domains and utilizes techniques such as optimal transport to align their distributions, thereby improving the efficacy of FS in the target domain where limited data hinder effectiveness and boosting model performance in practical applications.

When conducting FS under DA, there is a critical risk of mistakenly selecting irrelevant features as relevant. These erroneous FS results are commonly referred to as *false positives*. In high-stakes applications, such as medical diagnostics, these false positives can cause serious consequences. For example, selecting irrelevant features may result in incorrectly identifying a patient as high-risk for breast cancer due to unrelated genetic markers. Such a misidentification could lead to unnecessary procedures, such as biopsies or preventive surgeries, causing emotional distress, financial burdens, and potential physical harm to the patient. This highlights the importance of developing a statistical method that controls the false positive rate (FPR).

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. 

*Corresponding author. Email: duyvnl@uit.edu.vn

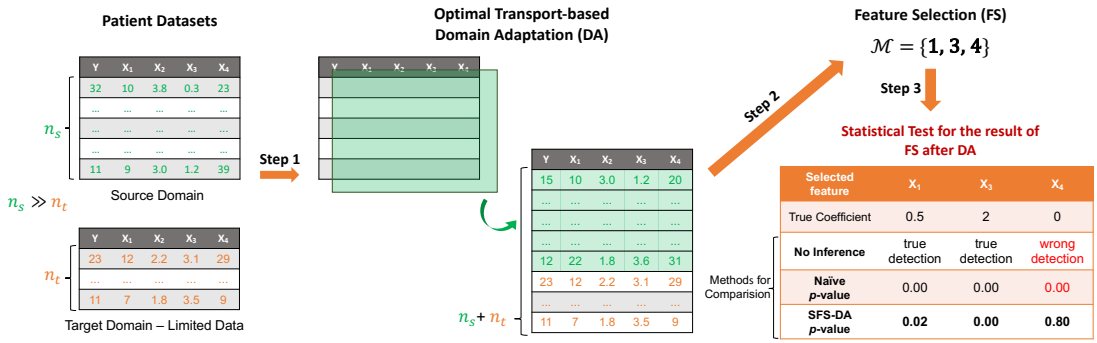



Figure 1: Illustration of the proposed SFS-DA method. Performing FS-DA without statistical inference leads to the selection of irrelevant feature ($X_4$). The naive $p$-value is small even for a falsely detected feature. With the proposed SFS-DA method, we successfully identify both false positives (FPs) and true positives (TPs), yielding large $p$-values for FPs and small $p$-values for TPs.

We also emphasize the critical need to manage the false negative rate (FNR). In statistical literature, a common strategy involves initially controlling the false positive rate (FPR) at a pre-determined level, such as $\alpha = 0.05$, while concurrently aiming to minimize the FNR, which is equivalent to maximizing the true positive rate (TPR = 1 - FNR) through empirical evidence. Following this established methodology, this paper adopts a similar approach. We propose a method to theoretically control the probability of misclassifying an irrelevant feature as relevant while simultaneously minimizing the probability of misclassifying a relevant feature as irrelevant.

To the best of our knowledge, *none* of the existing methods can control the FPR of FS under DA. Several methods have been proposed in the literature for FPR control in FS techniques, including Lasso (Berk et al., 2013; Lee et al., 2016; Duy and Takeuchi, 2022) and stepwise feature selection (Tibshirani et al., 2016; Sugiyama et al., 2021). However, these methods assume that the data originates from the same distribution, which becomes *invalid* in scenarios where a distribution shift occurs and DA must be employed. Duy et al. (2024) is the first work capable of conducting the inference under DA. However, their method primarily focuses on the unsupervised anomaly detection problem, which completely differs from the setup of FS under DA we consider in this paper. Consequently, their approach cannot be applied to our setting.

Conducting valid inference for controlling the FPR in FS under DA is challenging because the features selected are dependent on the application of FS-DA to the data. This violates the assumption of traditional inference methods, which require that the selected features be fixed in advance. To overcome the challenge, our idea is to leverage the concept of *Selective Inference (SI)* (Lee et al., 2016). However, directly applying SI in our setting is non-trivial, as SI is inherently problem- and model-specific. This necessitates the development of a new method tailored to the specific setting and the structure of the ML model. Consequently, we need to thoroughly examine the algorithm's selection strategy in the context of FS under DA.

In this paper, we focus on Optimal Transport (OT)-based DA (Flamary et al., 2016), which has gained popularity in the OT community, as well as the Lasso, a well-established method for FS. Additionally, we extend our method to the elastic net (Zou and Hastie, 2005). The detailed discussions on future extensions to other types of FS and DA are provided in §5.

**Contributions.** Our contributions are as follows:

- We introduce and mathematically formulate the problem setup of testing FS results in the context of DA within the hypothesis testing framework. We presents a unique challenge in addressing the impact of FS under DA to ensure the validity of FPR control.
- We propose a novel statistical method, named *SFS-DA* (statistical FS-DA), to conduct the introduced hypothesis test. To our knowledge, this is the first method capable of properly the FPR in FS under DA by providing valid $p$-values for the selected features. Additionally, we introduce a more strategic approach to maximize the TPR, i.e., reducing the FNR.
- We conduct extensive experiments on both synthetic and real-world datasets to thoroughly validate our theoretical findings, demonstrating the superior performance of the proposed SFS-DA method. For reproducibility, our implementation is available at:

`https://github.com/NT-Loi/SFS_DA.git`

**Example 1.** *To demonstrate the importance of the proposed SFS-DA method, we present an example in Fig. 1. Our objective is to perform FS to identify the relevant features that influence blood glucose level*

Table 1: The key strength of the proposed SFS-DA lies in its ability to control the False Positive Rate (FPR).

| | **No Inference** | **Naive** | **SFS-DA** |
|---|---|---|---|
| $N = 120$ | FPR = 1.0 | 0.15 | **0.05** |
| $N = 240$ | FPR = 1.0 | 0.12 | **0.04** |

*in the target domain, e.g., a hospital with a limited number of patients. We employ the OT-based DA approach to transfer the data from the source domain, where we have a substantial patient dataset, to the target domain. Subsequently, we apply a FS algorithm, i.e., the Lasso. The FS after DA erroneously identified an irrelevant feature as relevant. To resolve this issue, we introduced an additional inference step using the SFS-DA p-values, enabling us to identify both true positive and false positive detections. Additionally, we repeated the experiments N times, with the FPR results presented in Tab. 1. Using the proposed method, we successfully controlled the FPR at $\alpha = 0.05$, which other competing methods were unable to achieve.*

**Related works.** Traditional statistical inference in feature selection often faces challenges regarding the validity of $p$-values. A common issue arises from the reliance on naive $p$-values, which are computed under the assumption that the selected features are fixed in advance. However, when features are selected using the FS-DA method, this assumption is violated, which makes the naive $p$-values invalid. Data splitting (DS) offers a solution by dividing the data into two parts: one for selection and the other for inference. This ensures that the feature selection phase is independent of the testing phase, making the $p$-values computed via DS valid. However, DS reduces the amount of data available for both phases, potentially weakening the statistical power. Additionally, it is not always possible to split the data, e.g., when the data is correlated.

In recent years, SI has been actively studied for conducting inference on the features of linear models that are selected by FS methods. SI was first introduced for the Lasso (Lee et al., 2016). The basic idea of SI is to conduct the inference conditional on the FS process. This approach mitigates the bias of the FS step, allowing for the computation of valid $p$-values. The seminal paper laid the foundation for subsequent research on SI for FS (Loftus and Taylor, 2014; Fithian et al., 2014; Tibshirani et al., 2016; Yang et al., 2016; Suzumura et al., 2017; Hyun et al., 2018; Sugiyama et al., 2021; Das et al., 2022; Duy and Takeuchi, 2022). However, these methods assume the data is drawn from the same distribution. Therefore, they lose the validity in the context of DA, where distribution shifts occur, making them inappropriate for such scenarios.

A closely related work, and the main motivation for this study, is Duy et al. (2024), where the authors propose a framework for computing valid $p$-values for anomalies detected by an anomaly detection method within an OT-based DA setting. However, their focus is on unsupervised learning and anomaly detection task, which completely differs from the problem setup of supervised FS under DA that we consider in this paper. As a result, their method is not directly applicable to our setting.

## 2 PROBLEM SETUP

To formulate the problem, we consider a regression setup with two random response vectors defined by

$$\boldsymbol{Y}^s = (Y_1^s, \ldots, Y_{n_s}^s)^\top \sim \mathbb{N}(\boldsymbol{\mu}^s, \Sigma^s),$$
$$\boldsymbol{Y}^t = (Y_1^t, \ldots, Y_{n_t}^t)^\top \sim \mathbb{N}(\boldsymbol{\mu}^t, \Sigma^t),$$

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 $\Sigma^s$ and $\Sigma^t$ assumed to be known or estimable from independent data. We denote the feature matrices in the source and target domains, which are non-random, by $X^s \in \mathbb{R}^{n_s \times p}$ and $X^t \in \mathbb{R}^{n_t \times p}$, respectively, where $p$ is the number of features. 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 Lasso results after DA.

### 2.1 Optimal Transport (OT)-based DA

We leverage the OT-based DA proposed by Flamary et al. (2016) and apply it to our supervised setting. Let us define the source and target data as:

$$D^s = \begin{pmatrix} X^s & \boldsymbol{Y}^s \end{pmatrix} \text{ and } D^t = \begin{pmatrix} X^t & \boldsymbol{Y}^t \end{pmatrix}, \tag{1}$$

$D^s \in \mathbb{R}^{n_s \times (p+1)}$, $D^t \in \mathbb{R}^{n_t \times (p+1)}$. Then, we define the the cost matrix as:

$$C(D^s, D^t) = \left[ \left\| D_i^s - D_j^t \right\|_2^2 \right]_{ij} \in \mathbb{R}^{n_s \times n_t},$$

for any $i \in [n_s] = \{1, 2, ..., n_s\}$ and $j \in [n_t]$. We note that $D_i^s$ and $D_j^t$ are the $i^{\text{th}}$ and $j^{\text{th}}$ rows of $D^s$ and $D^t$, respectively, which are the $(p+1)$-dimensional vectors.

**Optimal transport**. The OT problem for DA between the source and target domains is defined as:

$$\hat{T} = \underset{T \in \mathbb{R}^{n_s \times n_t},\ T \geq 0}{\arg\min} \left\langle T, C(D^s, D^t) \right\rangle \tag{2}$$
$$\text{s.t.} \quad T\mathbf{1}_{n_t} = \mathbf{1}_{n_s}/n_s,\ T^\top \mathbf{1}_{n_s} = \mathbf{1}_{n_t}/n_t,$$

where $\langle \cdot, \cdot \rangle$ is the Frobenius inner product, $\mathbf{1}_n \in \mathbb{R}^n$ is the vector whose elements are set to 1. Once the optimal transportation matrix $\hat{T}$ is obtained, the source instances are transported into the target domain.

**Transformed data after DA.** The transformation $\tilde{D}^s$ of $D^s$ is defined as:

$$\tilde{D}^s = n_s \hat{T} D^t \in \mathbb{R}^{n_s \times (p+1)}. \tag{3}$$

More details are provided in Sec 3.3 of Flamary et al. (2016). Let us decompose $\tilde{D}^s$ into $\tilde{D}^s = \left(\tilde{X}^s \ \tilde{\boldsymbol{Y}}^s\right)$, the matrix $\tilde{X}^s$ and vector $\tilde{\boldsymbol{Y}}^s$ can be defined as:

$$\tilde{X}^s = n_s \hat{T} X^t \quad \text{and} \quad \tilde{\boldsymbol{Y}}^s = n_s \hat{T} \boldsymbol{Y}^t, \tag{4}$$

according to (3) and the definition of $D^t$ in (1). Here, $\tilde{X}^s$ and $\tilde{\boldsymbol{Y}}^s$ represent the transformations of $X^s$ and $\boldsymbol{Y}^s$ to the target domain, respectively.

### 2.2 Feature Selection by Lasso after DA

After transforming the data from the source domain to the target domain, we apply the Lasso to the combined dataset of the transformed and target data:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{2} \left\| \tilde{\boldsymbol{Y}} - \tilde{X}\boldsymbol{\beta} \right\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1, \tag{5}$$

where $\lambda \geq 0$ is a regularization parameter,

$$\tilde{X} = \begin{pmatrix} \tilde{X}^s \\ X^t \end{pmatrix} \in \mathbb{R}^{(n_s+n_t)\times p}, \ \tilde{\boldsymbol{Y}} = \begin{pmatrix} \tilde{\boldsymbol{Y}}^s \\ \boldsymbol{Y}^t \end{pmatrix} \in \mathbb{R}^{n_s+n_t}$$

Since the Lasso produces sparse solutions, the set of selected features is defined as:

$$\mathcal{M} = \left\{ j : \hat{\beta}_j \neq 0 \right\}. \tag{6}$$

While our primary focus in the main paper is on the Lasso, we also present an extension to the elastic net (Zou and Hastie, 2005). Detailed information is provided in §3.4. Future extensions to other types of FS methods are discussed in §5.

### 2.3 Statistical Inference and Decision Making

Our goal is to assess if the selected features in (6) are truly relevant or just selected by chance. To conduct the inference on the $j^{\text{th}}$ selected feature, we consider the statistical test on the following hypotheses:

$$\mathrm{H}_{0,j} : \beta_j = 0 \quad \text{vs.} \quad \mathrm{H}_{1,j} : \beta_j \neq 0,$$

where $\beta_j = \left[ \left( {X^t_{\mathcal{M}}}^\top X^t_{\mathcal{M}} \right)^{-1} {X^t_{\mathcal{M}}}^\top \boldsymbol{\mu}^t \right]_j$ and $X^t_{\mathcal{M}}$ is the sub-matrix of $X^t$ made up of columns in the set $\mathcal{M}$. To test these hypotheses, a natural choice of the test statistic is the least square estimate defined as:

$$\tau_j = \left[ \left( {X^t_{\mathcal{M}}}^\top X^t_{\mathcal{M}} \right)^{-1} {X^t_{\mathcal{M}}}^\top \boldsymbol{Y}^t \right]_j = \boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{Y}^s \\ \boldsymbol{Y}^t \end{pmatrix}, \tag{7}$$

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

$$\boldsymbol{\eta}_j = \begin{pmatrix} \mathbf{0}^s \\ X^t_{\mathcal{M}} \left( {X^t_{\mathcal{M}}}^\top X^t_{\mathcal{M}} \right)^{-1} \boldsymbol{e}_j \end{pmatrix}, \tag{8}$$

in which $\mathbf{0}^s \in \mathbb{R}^{n_s}$ represents a vector where all entries are set to 0, $\boldsymbol{e}_j \in \mathbb{R}^{|\mathcal{M}|}$ is a vector in which the $j^{\text{th}}$ entry is set to 1, and 0 otherwise.

**Compute $p$-value and decision making.** After obtaining the test statistic in (7), we proceed 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 the $j^{\text{th}}$ feature is relevant if the $p$-value $\leq \alpha$. Conversely, if the $p$-value $> \alpha$, there is not enough evidence to conclude that the $j^{\text{th}}$ feature is relevant.

**Challenge of computing a valid $p$-value.** The conventional (naive) $p$-value is defined as:

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

where $\boldsymbol{Y}^s_{\text{obs}}$ and $\boldsymbol{Y}^t_{\text{obs}}$ are the observations of the random vectors $\boldsymbol{Y}^s$ and $\boldsymbol{Y}^t$, respectively. If the vector $\boldsymbol{\eta}_j$ is independent of the FS and DA algorithms, the naive $p$-value is valid in the sense that

$$\mathbb{P}\Big( \underbrace{p_j^{\text{naive}} \leq \alpha \mid \mathrm{H}_{0,j} \text{ is true}}_{\text{a false positive}} \Big) = \alpha, \quad \forall \alpha \in [0, 1], \tag{9}$$

i.e., the probability of obtaining a false positive is controlled under a certain level of guarantee. However, in our setting, the vector $\boldsymbol{\eta}_j$ is influenced by the FS and DA, i.e., it is defined based on the set of selected features after performing FS under DA. As a result, the property of a valid $p$-value in (9) is no longer satisfied. Consequently, the naive $p$-value is *invalid* because it does not account for the effect of FS and DA.

## 3 PROPOSED SFS-DA METHOD

In this section, we present the details of the proposed SFS-DA method for computing the valid $p$-value.

### 3.1 The valid $p$-value in SFS-DA

To compute the valid $p$-value, we first need to determine the distribution of the test statistic defined in (7). We achieve this by leveraging the concept of SI, specifically by examining the distribution of the test statistic *conditional* on the FS results after DA:

$$\mathbb{P}\left( \boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{Y}^s \\ \boldsymbol{Y}^t \end{pmatrix} \ \middle| \ \mathcal{M}_{\boldsymbol{Y}^s, \boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}} \right), \tag{10}$$

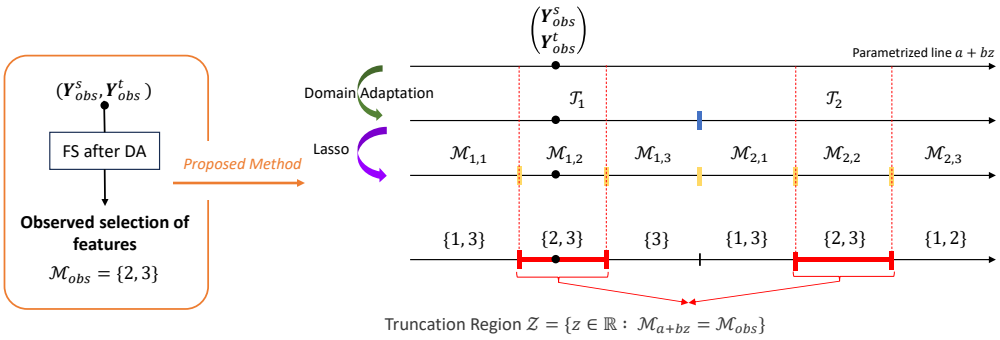



Figure 2: After performing DA, we apply FS to identify the relevant features. Next, we parametrize the data using a scalar parameter $z$ in the dimension of the test statistic to define the truncation region $\mathcal{Z}$, whose the data have the *same* FS results as the observed data. Finally, we conduct the inference by conditioning on $\mathcal{Z}$. To enhance the efficiency, we utilize a divide-and-conquer strategy to effectively identify the region $\mathcal{Z}$.

where $\mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t}$ is the set of selected features of Lasso FS after DA *for any random* vectors $\boldsymbol{Y}^s$ and $\boldsymbol{Y}^t$, and $\mathcal{M}_{\text{obs}} = \mathcal{M}_{\boldsymbol{Y}^s_{\text{obs}},\boldsymbol{Y}^t_{\text{obs}}}$ is the observed selected features.

Based on the distribution in (10), we introduce the selective $p$-value which is defined as:

$$p_j^{\text{sel}} = \mathbb{P}_{\text{H}_{0,\text{j}}}\left( \left| \boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{Y}^s \\ \boldsymbol{Y}^t \end{pmatrix} \right| \geq \left| \boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{Y}^s_{\text{obs}} \\ \boldsymbol{Y}^t_{\text{obs}} \end{pmatrix} \right| \;\middle|\; \mathcal{E} \right), \quad (11)$$

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

$$\mathcal{E} = \left\{ \mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}},\ \mathcal{Q}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{Q}_{\text{obs}} \right\}. \quad (12)$$

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

$$\mathcal{Q}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \left( I_{n_s+n_t} - \boldsymbol{b}\boldsymbol{\eta}_j^\top \right) \begin{pmatrix} \boldsymbol{Y}^s \\ \boldsymbol{Y}^t \end{pmatrix}, \quad (13)$$

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 $\mathcal{Q}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t}$ corresponds to the component $\boldsymbol{z}$ in the seminal paper of Lee et al. (2016) (see Sec. 5, Eq. (5.2)). The additional conditioning on $\mathcal{Q}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t}$ is required for technical reasons, specifically to facilitate tractable inference. This is the standard approach in SI literature and it is used in almost all the SI-related works that we cite.*

**Lemma 1.** *The selective p-value proposed in (11) satisfies the property of a valid p-value:*

$$\mathbb{P}_{\text{H}_{0,\text{j}}}\left( p_j^{\text{sel}} \leq \alpha \right) = \alpha, \quad \forall \alpha \in [0,1].$$

*Proof.* The proof is deferred to Appendix 6.1. □

Lemma 1 indicates that, by using the proposed selective $p$-value, the FPR is theoretically controlled for any level $\alpha \in [0,1]$. Once $\mathcal{E}$ in (12) is identified, the selective $p$-value can be computed. We will present the characterization of $\mathcal{E}$ in the next section.

### 3.2 Characterization of Conditioning Event $\mathcal{E}$

Let us define the set of $\begin{pmatrix} \boldsymbol{Y}^s \\ \boldsymbol{Y}^t \end{pmatrix} \in \mathbb{R}^{n_s+n_t}$ that satisfies the conditions in $\mathcal{E}$ defined in (12) as:

$$\mathcal{D} = \left\{ \begin{pmatrix} \boldsymbol{Y}^s \\ \boldsymbol{Y}^t \end{pmatrix} \in \mathbb{R}^{n_s+n_t} \;\middle|\; \begin{matrix} \mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}}, \\ \mathcal{Q}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{Q}_{\text{obs}} \end{matrix} \right\}. \quad (14)$$

In the following lemma, we show that the conditional data space $\mathcal{D}$ is, in fact, restricted to a *line* in $\mathbb{R}^n$.

**Lemma 2.** *Let us define $\boldsymbol{a} = \mathcal{Q}_{\text{obs}}$, the set $\mathcal{D}$ in (14) can be rewritten using a scalar parameter $z \in \mathbb{R}$ as:*

$$\mathcal{D} = \left\{ \begin{pmatrix} \boldsymbol{Y}^s \\ \boldsymbol{Y}^t \end{pmatrix} = \boldsymbol{Y}(z) = \boldsymbol{a} + \boldsymbol{b}z \mid z \in \mathcal{Z} \right\}, \quad (15)$$

*where $\boldsymbol{b}$ is defined in (13) and $\mathcal{Z}$ is defined as:*

$$\mathcal{Z} = \left\{ z \in \mathbb{R} \mid \mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{M}_{\text{obs}} \right\}. \quad (16)$$

*Here, $\mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{M}_{\begin{pmatrix} \boldsymbol{Y}^s \\ \boldsymbol{Y}^t \end{pmatrix}}$ is equivalent to $\mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t}$.*

The proof is deferred to Appendix 6.2. The fact that the conditional space can be restricted to a line was implicitly exploited in Lee et al. (2016) and discussed in Sec. 6 of Liu et al. (2018). Lemma 2 shows that it is not necessary to consider the $n$-dimensional data space. Instead, we only need to focus on the *one-dimensional projected* data space $\mathcal{Z}$ in (16).

**Reformulation of the selective $p$-value computation with $\mathcal{Z}$.** Let us denote a random variable $Z \in \mathbb{R}$ and its observation $Z_{\text{obs}} \in \mathbb{R}$ as follows:

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

The selective $p$-value in (11) can be rewritten as

$$p_j^{\text{sel}} = \mathbb{P}_{\text{H}_{0,\text{j}}}\left( |Z| \geq |Z_{\text{obs}}| \mid Z \in \mathcal{Z} \right). \quad (17)$$

Once the truncation region $\mathcal{Z}$ is identified, computation of the selective $p$-value in (17) is straightforward. Therefore, the remaining task is to identify $\mathcal{Z}$.

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

To identify $\mathcal{Z}$, the naive approach is to apply Lasso FS under DA on $\binom{\boldsymbol{Y}^s}{\boldsymbol{Y}^t} = \boldsymbol{a} + \boldsymbol{b}z$ for *infinitely many* values of $z \in \mathbb{R}$ to obtain the set of features $\mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z}$ and check if it is the same as the observed $\mathcal{M}_{\text{obs}}$ or not, which is *computationally intractable*. To resolve the difficulty, we introduce an efficient approach (illustrated in Fig. 2), inspired by Duy and Takeuchi (2022) and Duy et al. (2024), to identify $\mathcal{Z}$ in finite operations as follows:

- We divide the problem into multiple sub-problems, conditioning not only on the set of selected features but also on the DA transportation and the signs of the coefficients of the selected feature.
- We show that the sub-problem is efficiently solvable.
- We combine multiple sub-problems to obtain $\mathcal{Z}$.

**Divide-and-conquer strategy.** Let us denote by $U$ a total number of possible transportations for DA along the parametrized line. We define $V_u$ as a number of all possible sets of features can be obtained by Lasso FS after the $\mathcal{T}_u$ transportation, $u \in [U]$. The entire one-dimensional space $\mathbb{R}$ can be decomposed as:

$$\mathbb{R} = \bigcup_{u\in[U]} \bigcup_{v\in[V_u]} \underbrace{\left\{ z \in \mathbb{R} \;\middle|\; \begin{array}{l} \mathcal{T}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{T}_u, \\ \mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{M}_v, \\ \mathcal{S}_{\mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z}} = \mathcal{S}_{\mathcal{M}_v} \end{array} \right\}}_{\text{a sub-problem of additional conditioning}},$$

where $\mathcal{T}_{\boldsymbol{a}+\boldsymbol{b}z}$ denotes the OT-based DA on $\boldsymbol{a} + \boldsymbol{b}z$, $\mathcal{S}_{\mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z}}$ denotes a set of signs of the coefficients for the selected features in $\mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z}$. For $u \in [U], v \in [V_u]$, we aim to identify a set:

$$\mathcal{R} = \Big\{(u, v) : \mathcal{M}_v = \mathcal{M}_{\text{obs}}\Big\}. \tag{18}$$

The region $\mathcal{Z}$ in (16) then can be identified as follows:

$$\begin{aligned} \mathcal{Z} &= \Big\{ z \in \mathbb{R} \mid \mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{M}_{\text{obs}} \Big\} \\ &= \bigcup_{(u,v)\in\mathcal{R}} \left\{ z \in \mathbb{R} \;\middle|\; \begin{array}{l} \mathcal{T}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{T}_u, \\ \mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{M}_v, \\ \mathcal{S}_{\mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z}} = \mathcal{S}_{\mathcal{M}_v} \end{array} \right\}. \end{aligned} \tag{19}$$

**Solving of each sub-problem.** For any $u \in [U]$ and $v \in [V_u]$, we define the subset of one-dimensional projected dataset on a line for the sub-problem as:

$$\mathcal{Z}_{u,v} = \left\{ z \;\middle|\; \begin{array}{l} \mathcal{T}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{T}_u, \\ \mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{M}_v, \mathcal{S}_{\mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z}} = \mathcal{S}_{\mathcal{M}_v} \end{array} \right\}. \tag{20}$$

The sub-problem region $\mathcal{Z}_{u,v}$ can be re-written as:

$$\mathcal{Z}_{u,v} = \mathcal{Z}_u \cap \mathcal{Z}_v, \text{where } \mathcal{Z}_u = \Big\{ z \in \mathbb{R} \mid \mathcal{T}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{T}_u \Big\},$$
$$\mathcal{Z}_v = \Big\{ z \in \mathbb{R} \mid \mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{M}_v, \mathcal{S}_{\mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z}} = \mathcal{S}_{\mathcal{M}_v} \Big\}.$$

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

$$\mathcal{Z}_u = \Big\{ z \in \mathbb{R} \mid \boldsymbol{p} + \boldsymbol{q}z + \boldsymbol{r}z^2 \geq \boldsymbol{0} \Big\},$$

*where vectors $\boldsymbol{p}$, $\boldsymbol{q}$, and $\boldsymbol{r}$ are defined in Appendix 6.3.*

The proof is deferred to Appendix 6.3. The purpose of Lemma 3 is to ensure that the transportation $\mathcal{T}_u$ remains the same for all $z \in \mathcal{Z}_u$.

**Lemma 4.** *Let us define the Lasso optimization problem after the $\mathcal{T}_u$ transportation as:*

$$\hat{\boldsymbol{\beta}}(z) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \frac{1}{2} \big\| \tilde{\boldsymbol{Y}}_u(z) - \tilde{X}_u \boldsymbol{\beta} \big\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1,$$

*where $\tilde{\boldsymbol{Y}}_u(z) = \Omega_u \boldsymbol{Y}(z)$ and $\tilde{X}_u = \Omega_u X$. Here,*

$$\Omega_u = \begin{pmatrix} 0_{n_s \times n_s} & n_s \mathcal{T}_u \\ 0_{n_t \times n_s} & I_{n_t} \end{pmatrix} \in \mathbb{R}^{(n_s+n_t)\times(n_s+n_t)},$$

*where $0_{n\times m} \in \mathbb{R}^{n\times m}$ is the zero matrix, $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix, and $X = (X^s \; X^t)^\top$. The set $\mathcal{Z}_v$ can be identified as follows:*

$$\mathcal{Z}_v = \left\{ z \in \mathbb{R} \;\middle|\; \begin{array}{l} \hat{\boldsymbol{\beta}}_j(z) \neq 0, \; \forall j \in \mathcal{M}_v, \\ \hat{\boldsymbol{\beta}}_j(z) = 0, \; \forall j \notin \mathcal{M}_v, \\ \text{sign}\Big(\hat{\boldsymbol{\beta}}_{\mathcal{M}_v}(z)\Big) = \mathcal{S}_{\mathcal{M}_v} \end{array} \right\},$$

*which can be efficiently computed by solving a set of linear inequalities w.r.t $z$, derived from the Karush–Kuhn–Tucker (KKT) conditions of the Lasso.*

The proof is deferred to Appendix 6.4. Lemma 4 guarantees that, for any $z \in \mathcal{Z}_v$, the selected features and the signs of their coefficients remain the same when conducting FS after the $\mathcal{T}_u$ transportation.

In Lemmas 3 and 4, we demonstrate that $\mathcal{Z}_u$ and $\mathcal{Z}_v$ can be *analytically obtained* by solving the systems of quadratic/linear inequalities, respectively. Once $\mathcal{Z}_u$ and $\mathcal{Z}_v$ are computed, the sub-problem region $\mathcal{Z}_{u,v}$ in (20) is obtained by $\mathcal{Z}_{u,v} = \mathcal{Z}_u \cap \mathcal{Z}_v$.

**Computation of truncation region $\mathcal{Z}$ by combining multiple sub-problems and algorithm.** To identify $\mathcal{R}$ in (18), the OT-based DA and Lasso FS after DA are repeatedly applied to a series of datasets $\boldsymbol{a} + \boldsymbol{b}z$, over a sufficiently wide range off $z \in [z_{\min}, z_{\max}]$[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 $\mathcal{Z}_u = [\ell_u, r_u]$ and

[1] We set $z_{\min} = -20\sigma$ and $z_{\max} = 20\sigma$, $\sigma$ 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.

**Algorithm 1** `SFS-DA`

**Input:** $X^s, \boldsymbol{Y}^s_{\text{obs}}, X^t, \boldsymbol{Y}^t_{\text{obs}}, z_{\min}, z_{\max}$

$\mathcal{M}_{\text{obs}} \leftarrow$ FS after DA on $(X^s, \boldsymbol{Y}^s_{\text{obs}})$ and $(X^t, \boldsymbol{Y}^t_{\text{obs}})$
**for** $j \in \mathcal{M}_{\text{obs}}$ **do**
  Compute $\boldsymbol{\eta}_j \leftarrow$ Eq. (8), $\boldsymbol{a}$ and $\boldsymbol{b} \leftarrow$ Eq. (15)
  $X = (X^s \; X^t)^\top$
  $\mathcal{R} \leftarrow$ `divide_and_conquer` (X, $\boldsymbol{a}, \boldsymbol{b}, z_{\min}, z_{\max}$)
  Identify $\mathcal{Z} \leftarrow$ Eq. (19) with $\mathcal{R}$
  Compute $p^{\text{sel}}_j \leftarrow$ Eq. (17) with $\mathcal{Z}$
**end for**

**Output:** $\{p^{\text{sel}}_i\}_{i \in \mathcal{M}^{\text{obs}}}$

$\mathcal{Z}_{u,v} = [\ell_{u,v}, r_{u,v}]$. The divide-and-conquer procedure can be summarized in Algorithm 2. After obtaining $\mathcal{R}$ by Algorithm 2. We can compute $\mathcal{Z}$ in (19), which is subsequently used to obtain the proposed selective $p$-value in (17). The entire steps of the proposed SFS-DA method is summarized in Algorithm 1.

**Remark 2.** *The selective p-value computed using the truncation region $\mathcal{Z}_{u,v}$ for each sub-problem in Eq. (20) is still valid. This can be viewed as an extension of the seminal SI framework introduced by Lee et al. (2016) to our setting. However, a major drawback of this method is its notably low statistical power, primarily due to over-conditioning. Specifically, the inference is conducted not only by conditioning on the set of selected features but also on several redundant conditions, such as the DA transportation and the signs of the coefficients of the selected features. Excessive redundant conditioning leads to less information for the inference, thereby lowering the statistical power of the test. Therefore, in this paper, we introduce an approach to remove redundant conditioning, thereby improving statistical power.*

### 3.4 Extension to Elastic Net

In certain cases, the Lasso solutions can be unstable. Adding an $\ell_2$ penalty to the objective function of Lasso yields the elastic net (Zou and Hastie, 2005), which helps stabilize the FS results. Therefore, we extend our proposed method to elastic net case. The sub-problem of FS after DA in the elastic net case is similar to that in the Lasso case, i.e.,

$$\mathcal{Z}_{u,v} = \mathcal{Z}_u \cap \mathcal{Z}^{\text{enet}}_v \text{ with } \mathcal{Z}_u = \Big\{ z \in \mathbb{R} \mid \mathcal{T}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{T}_u \Big\},$$
$$\mathcal{Z}^{\text{enet}}_v = \Big\{ z \in \mathbb{R} \mid \mathcal{M}^{\text{enet}}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{M}^{\text{enet}}_v, \mathcal{S}^{\text{enet}}_{\mathcal{M}_{\boldsymbol{a}+\boldsymbol{b}z}} = \mathcal{S}^{\text{enet}}_{\mathcal{M}_v} \Big\}.$$

Here, $\mathcal{Z}_u$ is the same as in the Lasso case, and $\mathcal{Z}^{\text{enet}}_v$ is the region corresponding to the elastic net case, whose characterization is detailed in the following lemma.

**Algorithm 2** `divide_and_conquer`

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

Initialization: $u = 1$, $v = 1$, $z_{u,v} = z_{\min}$, $\mathcal{R} = \emptyset$
**while** $z_{u,v} < z_{\max}$ **do**
  $\mathcal{T}_u \leftarrow$ DA on $\boldsymbol{a} + \boldsymbol{b} z_{u,v}$
  Compute $[\ell_u, r_u] = \mathcal{Z}_u \leftarrow$ Lemma 3
  $r_{u,v} = \ell_u$
  **while** $r_{u,v} < r_u$ **do**
    $\tilde{X}_u, \tilde{\boldsymbol{Y}}_u(z_{u,v}) \leftarrow$ Lemma 4
    $\mathcal{M}_v$ and $\mathcal{S}_{\mathcal{M}_v} \leftarrow$ FS after DA on $\big(\tilde{X}_u, \tilde{\boldsymbol{Y}}_u(z_{u,v})\big)$
    $\mathcal{Z}_v \leftarrow$ Lemma 4
    $[\ell_{u,v}, r_{u,v}] = \mathcal{Z}_{u,v} \leftarrow \mathcal{Z}_u \cap \mathcal{Z}_v$
    $\mathcal{R} \leftarrow \mathcal{R} \cup \{(u,v)\}$ **if** $\mathcal{M}_v = \mathcal{M}_{\text{obs}}$
    $v \leftarrow v + 1$, $z_{u,v} = r_{u,v}$
  **end while**
  $v \leftarrow 1$, $u \leftarrow u + 1$, $z_{u,v} = r_{u,v}$
**end while**

**Output:** $\mathcal{R}$

**Lemma 5.** *Let us define the elastic net optimization problem after the $\mathcal{T}_u$ transportation as follows:*

$$\hat{\boldsymbol{\beta}}^{\text{enet}}(z) = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \frac{1}{2} \big\| \tilde{\boldsymbol{Y}}_u(z) - \tilde{X}_u \boldsymbol{\beta} \big\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 + \frac{\gamma}{2} \|\boldsymbol{\beta}\|_2^2,$$

*where $\lambda$ and $\gamma$ are the regularization parameters, $\tilde{\boldsymbol{Y}}_u(z)$ and $\tilde{X}_u$ are defined in Lemma 4. Then, the set $\mathcal{Z}^{\text{enet}}_v$ can be identified as follows:*

$$\mathcal{Z}^{\text{enet}}_v = \left\{ z \in \mathbb{R} \;\middle|\; \begin{array}{l} \hat{\boldsymbol{\beta}}^{\text{enet}}_j(z) \neq 0, \; \forall j \in \mathcal{M}^{\text{enet}}_v, \\ \hat{\boldsymbol{\beta}}^{\text{enet}}_j(z) = 0, \; \forall j \notin \mathcal{M}^{\text{enet}}_v, \\ \operatorname{sign}\Big(\hat{\boldsymbol{\beta}}^{\text{enet}}_{\mathcal{M}_v}(z)\Big) = \mathcal{S}_{\mathcal{M}^{\text{enet}}_v} \end{array} \right\},$$

*which can be efficiently computed by solving a set of linear inequalities w.r.t $z$.*

The proof of Lemma 5 is deferred to Appendix 6.5.

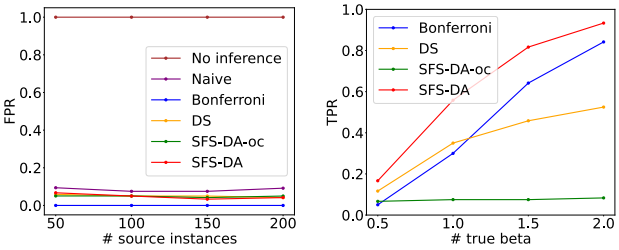
## 4 EXPERIMENT

We demonstrate the performance of the proposed SFS-DA. Here, we present the main results. Several additional experiments can be found in Appendix 6.6.

### 4.1 Experimental Setup

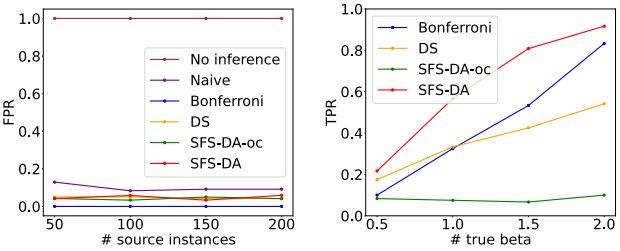
**Methods for comparison.** We compared the performance of the following methods:

- `SFS-DA`: proposed method.
- `SFA-DA-oc`: proposed method, which considers only one sub-problem, i.e., over-conditioning, described in §3.3 (extension of Lee et al. (2016) to our setting).




(a) FPR (b) TPR

Figure 3: FPR and TPR in the case of Lasso




(a) FPR (b) TPR

Figure 4: FPR and TPR in the case of elastic net

- `DS:` data splitting.
- `Bonferroni:` the most popular multiple testing.
- `Naive:` traditional statistical inference.
- `No inference:` FS after DA without inference.

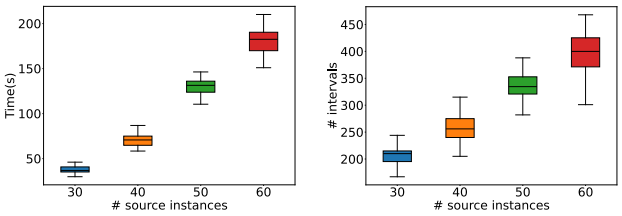
We note that if a method fails to control the FPR at $\alpha$, it is *invalid*, and its TPR becomes irrelevant. A method with a high TPR implies a low FNR.

**Synthetic data generation.** We generated $\boldsymbol{Y}^s$ with $\boldsymbol{Y}_i^s = X_i^{s\top}\boldsymbol{\beta}^s + \varepsilon$, $X_i^s \sim \mathbb{N}(\mathbf{0}, I_p), \forall i \in [n_s]$, and $\varepsilon \sim \mathbb{N}(0,1)$. Similarly, $\boldsymbol{Y}^t$ is generated with $\boldsymbol{Y}_i^t = X_i^{t\top}\boldsymbol{\beta}^t + \varepsilon$ in which $X_i^t \sim \mathbb{N}(\mathbf{0}, I_p)$. We set $p = 5$, $\lambda = 10$, $\gamma = 1$ (elastic net), and $\alpha = 0.05$. For the FPR experiments, all elements of $\boldsymbol{\beta}^t$ were set to 0 and $n_s \in \{50, 100, 150, 200\}$. For the TPR experiments, all elements of $\boldsymbol{\beta}^t$ were set to 0.5 and $n_s = 100$. We set $n_t = 10$, indicating that the target data is limited. In all experiments, elements of $\boldsymbol{\beta}^s$ are set to 2. Note that we only conduct the inference on the target data. Therefore, the values of $\boldsymbol{\beta}^s$ do not affect the inference. Each experiment was repeated 120 times.

### 4.2 Numerical results

**The results of FPRs and TPRs.** The results of FPR and TPR in two cases of Lasso and elastic net are shown in Figs. 3 and 4. Each point in these two figures represents $\mathbb{P}(p-\text{value} \leq \alpha)$ and is computed by counting the number of $p$-values satisfying $p - \text{value} \leq \alpha$ across 120 runs, then dividing by 120. In the plots on the left, the `SFS-DA`, `SFS-DA-oc`, `Bonferroni`, `DS` controlled the FPR whereas the `Naive` and `No Inference` *could not*. Because the `Naive` and `No Inference` failed to control the FPR, we no longer considered their TPRs. In the plots on the right, the `SFS-DA` has the highest TPR compared to other methods in all the cases, i.e., the `SFS-DA` has the lowest FNR.

**Computational time.** In Fig. 5, we show the boxplots of the time for computing each $p$-value as well as actual number of intervals of $z$ that we encountered on the line when constructing the truncation region $\mathcal{Z}$ w.r.t. $n_s$. The data generation parameters used in Fig. 5 are configured as follows. We set $n_s \in \{30, 40, 50, 60\}$, $n_t = 10$, $p = 10$ and $\lambda = 10$. Each experiment was repeated 10 times. The computation time for calculating the $p$-value of a selected feature was measured in each experiment. The plots demonstrate that the complexity of the `SFS-DA` increases linearly w.r.t. $n_s$.

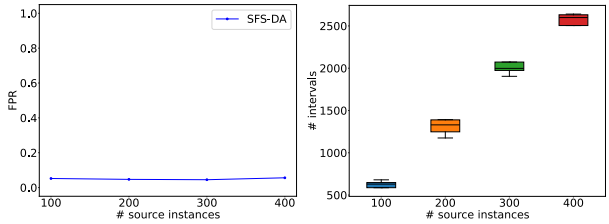



(a) Computational time (b) Encountered intervals

Figure 5: Computational cost of the proposed SFS-DA

**Complexity.** The complexity of Algorithm 1 is $\mathcal{O}(|\mathcal{M}_{\text{obs}}| \times \rho)$ where $|\mathcal{M}_{\text{obs}}|$ is the number of selected features and $\rho$ is the number of sub-problems encountered along the line defined in Eq. (15), which is used to compute the $p$-value for a single selected feature. We note that, in the worst-case, the value of $\rho$ still grows exponentially. However, it is known that the actual computational cost differs significantly from the worst case, as evidenced by numerous references in the parametric programming literature (Efron et al., 2004; Hastie et al., 2004; Mairal and Yu, 2012). A well-known example is the Lasso regularization path, which also has the worst-case computational cost on the exponential order of the number of features, but the actual cost is known to be nearly linear order. Similarly, in the proposed SFS-DA method, it is also evident from our experimental results that the number of sub-problems (i.e., intervals) encounted along the line is almost linearly increasing in practice (Fig. 5).

**Scalability.** We evaluate the scalability of our approach by examining its performance across varying sample sizes of the source data and in high-dimensional settings. The results are shown in Fig. 6 and Tab. 2. They demonstrate that the proposed SFS-DA method effectively controls the FPR while maintaining reasonable computational times.




(a) FPR (b) Encountered intervals

Figure 6: Scalability of the proposed method

Table 2: Performance in high-dimensional settings.

| Num. Predictors ($p$) | FPR | Avg. Time ($s$) |
|---|---|---|
| 1000 | 0.042 | 239.8 |
| 1500 | 0.056 | 264.4 |
| 2000 | 0.059 | 285.6 |
| 2500 | 0.047 | 295.8 |

Table 3: Computational cost on riboflavin dataset.

| Avg. Encountered Intervals | Avg. Time ($s$) |
|---|---|
| 140 | 70.14 |

### 4.3 Results on Real-World Datasets

We performed comparison on five real-world datasets. In this section, we present the experimental results for three datasets: the Diabetes dataset (Efron et al., 2004), the Heart Failure dataset, and the Seoul Bike dataset, all available in the UCI Machine Learning Repository. The results for the remaining two datasets can be found in Appendix 6.6. For each dataset, we present the distribution of $p$-value for each feature. For each dataset, we randomly selected instances from source and target domain, with $n_s = 100$ and $n_t = 20$. We used Lasso for FS. The results are shown in Figs. 7, 8, 9. The $p$-values of `Bonferroni` are equal to one in almost all cases, indicating that this method is *conservative*. While the $p$-value of `DS` is smaller than that of `SFS-DA` in a few cases (S5 in Diabetes dataset and Temperature in Seoul Bike dataset), in all remaining cases, the $p$-value of the proposed `SFS-DA` tends to be smaller than those of the competitors, demonstrating that `SFS-DA` exhibits the highest statistical power.

We also demonstrated the computational performance of our method on a high-dimensional real-world dataset, specifically the riboflavin production dataset, which contains 4,088 features (Bühlmann et al., 2014). In this experiment, the source domain consists of data from riboflavinGrouped, while the target domain consists of data from riboflavin. We set $p = 4088$ and randomly selected instances from the source and target domains, with $n_s = 100$ and $n_t = 10$, respectively. The results, presented in Tab. 3, demonstrate a reasonable computational cost of the proposed method.

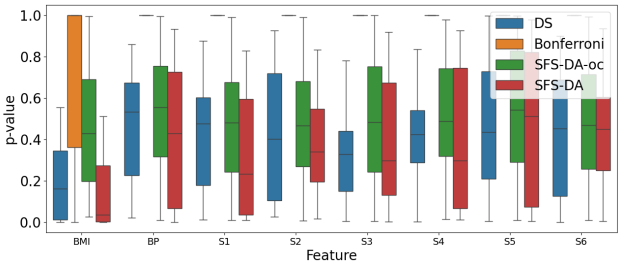



Figure 7: Diabetes dataset. The source domain consists of "people over 50 years old", while the target domain consists of "people under 50 years old".

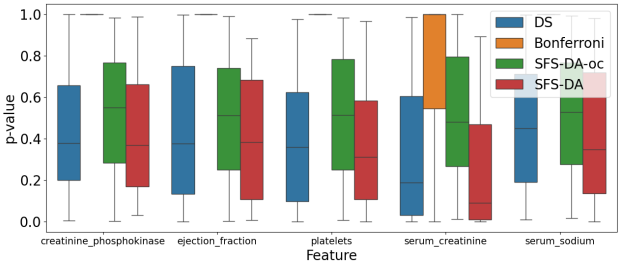



Figure 8: Heart Failure dataset. The source and target domain settings match the Diabetes dataset.

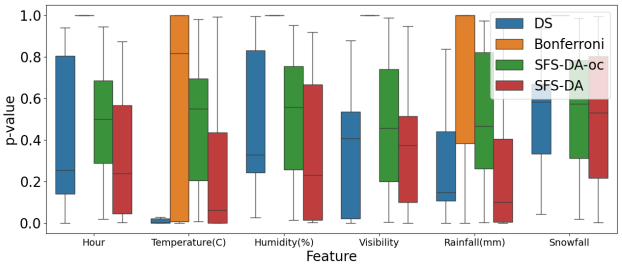



Figure 9: Seoul Bike dataset. The source is "regular-day bike renters", and the target is "holiday renters".

## 5 DISCUSSION

We introduce a novel testing setup for feature selection (FS) after domain adaptation (DA), including a method for calculating valid p-values using a divide-and-conquer approach within the selective inference (SI) framework. While currently limited to OT-based DA (due to simpler selection events), future work could extend this to more complex DA methods like MMD-based or metric learning-based approaches, potentially using sampling-based approximations. Similarly, our approach is also applicable to other FS algorithms with linearly/quadratically constrained selection events (e.g., step-wise selection) within the context of FS after DA. Another important future direction is expanding the proposed framework to include adversarial training or transfer learning settings, as FS after these methods often leads to false positives, reducing model reliability, enhancing the proposed method to address these settings would make a significant contribution to the existing literature.

**Acknowledgements**

We thank the anonymous reviewers and area chair for their comments. This work was supported by the Domestic Postdoctoral Fellowship Program of Vingroup Innovation Foundation (VINIF), code VINIF.2024.STS.36.

**References**

R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

P. Bühlmann, M. Kalisch, and L. Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1:255–278, 01 2014. doi: 10.1146/annurev-statistics-022513-115545.

A. Coad and S. Srhoj. Catching gazelles with a lasso: Big data techniques for the prediction of high-growth firms. *Small Business Economics*, 55(3):541–565, 2020.

D. Das, V. N. Le Duy, H. Hanada, K. Tsuda, and I. Takeuchi. Fast and more powerful selective inference for sparse high-order interaction model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9999–10007, 2022.

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, 2022.

V. N. L. Duy, H.-T. Lin, and I. Takeuchi. Cadda: Controllable anomaly detection after domain adaptation by statistical inference. In *International Conference on Artificial Intelligence and Statistics*, pages 1828–1836. PMLR, 2024.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004. doi: 10.1214/009053604000000067. URL `https://doi.org/10.1214/009053604000000067`.

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

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

T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5 (Oct):1391–1415, 2004.

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, 2018.

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.

Y.-C. Lo, S. E. Rensi, W. Torng, and R. B. Altman. Machine learning in chemoinformatics and drug discovery. *Drug discovery today*, 23(8):1538–1546, 2018.

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

S. Ma and J. Huang. Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics*, 9(5):392–403, 2008.

J. Mairal and B. Yu. Complexity analysis of the lasso regularization path. *arXiv preprint arXiv:1205.0079*, 2012.

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, 2021.

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.

S. Tian, Y. Yu, and H. Guo. Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, 52:89–100, 2015.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

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. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

F. Yang, R. F. Barber, P. Jain, and J. Lafferty. Selective inference for group-sparse linear models. In *Advances in Neural Information Processing Systems*, pages 2469–2477, 2016.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.

## 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. [No]

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). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

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. [No]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [No]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [No]

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

$$\boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{Y}^s \\ \boldsymbol{Y}^t \end{pmatrix} \Big| \left\{ \mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}},\ \mathcal{Q}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{Q}_{\text{obs}} \right\} \sim TN\left( \boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{\mu}^s \\ \boldsymbol{\mu}^t \end{pmatrix}, \boldsymbol{\eta}_j^\top \Sigma \boldsymbol{\eta}_j, \mathcal{Z} \right),$$

which is a truncated normal distribution with mean $\boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{\mu}^s \\ \boldsymbol{\mu}^t \end{pmatrix}$, variance $\eta_j^\top \Sigma \eta_j$, in which $\Sigma = \begin{pmatrix} \Sigma^s & 0 \\ 0 & \Sigma^t \end{pmatrix}$, and the truncation region $\mathcal{Z}$ described in §3.3. Therefore, under null hypothesis,

$$p_j^{\text{sel}} \ \Big|\ \left\{ \mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}},\ \mathcal{Q}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{Q}_{\text{obs}} \right\} \sim \text{Unif}(0,1)$$

Thus, $\mathbb{P}_{\text{H}_{0,\text{j}}} \left( p_j^{\text{sel}} \ \Big|\ \mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}},\ \mathcal{Q}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{Q}_{\text{obs}} \right) = \alpha, \forall \alpha \in [0,1]$.

Next, we have

$$\begin{aligned}
&\mathbb{P}_{\text{H}_{0,\text{j}}} \left( p_j^{\text{sel}} \ \Big|\ \mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}} \right) \\
&= \int \mathbb{P}_{H_{0,j}} \left( p_j^{\text{sel}} \le \alpha \,\middle|\, \mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}},\ \mathcal{Q}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{Q}_{\text{obs}} \right) \mathbb{P}_{H_{0,j}} \left( \mathcal{Q}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{Q}_{\text{obs}} \,\middle|\, \mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}} \right) d\mathcal{Q}_{\text{obs}} \\
&= \int \alpha\, \mathbb{P}_{H_{0,j}} \left( \mathcal{Q}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{Q}_{\text{obs}} \,\middle|\, \mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}} \right) d\mathcal{Q}_{\text{obs}} \\
&= \alpha \int \mathbb{P}_{H_{0,j}} \left( \mathcal{Q}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{Q}_{\text{obs}} \,\middle|\, \mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}} \right) d\mathcal{Q}_{\text{obs}} \\
&= \alpha.
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}_{\text{H}_{0,\text{j}}} \left( p_j^{\text{sel}} \le \alpha \right) &= \sum_{\mathcal{M}_{\text{obs}}} \mathbb{P}_{H_{0,j}} \left( p_j^{\text{sel}} \le \alpha \,\middle|\, \mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}} \right) \mathbb{P}_{H_{0,j}} \left( \mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}} \right) \\
&= \sum_{\mathcal{M}_{\text{obs}}} \alpha\, \mathbb{P}_{H_{0,j}} \left( \mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}} \right) \\
&= \alpha \sum_{\mathcal{M}_{\text{obs}}} \mathbb{P}_{H_{0,j}} \left( \mathcal{M}_{\boldsymbol{Y}^s,\boldsymbol{Y}^t} = \mathcal{M}_{\text{obs}} \right) \\
&= \alpha.
\end{aligned}$$

## 6.2 Proof of Lemma 2

Base on the second condition in (14), we have

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

By defining $\boldsymbol{a} = \mathcal{Q}_{obs}, z = \boldsymbol{\eta}_j^\top \begin{pmatrix} \boldsymbol{Y}^s \\ \boldsymbol{Y}^t \end{pmatrix}$, and incorporating the second condition of (14), we obtain Lemma 2.

### 6.3 Proof of Lemma 3

The proof is constructed based on the results presented in Duy et al. (2024), 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 (2) as:

$$\hat{\boldsymbol{t}} = \underset{\boldsymbol{t}\in\mathbb{R}^{n_s n_t}}{\arg\min}\ \boldsymbol{t}^\top \boldsymbol{c}\left(D^s, D^t\right)$$
$$\text{s.t.} \quad H\boldsymbol{t} = \boldsymbol{h}, \boldsymbol{t} \geq 0,$$

where $\boldsymbol{t} = \text{vec}(T), \boldsymbol{c}\left(D^s, D^t\right) = \text{vec}\left(C\left(D^s, D^t\right)\right) = \boldsymbol{c}' + \left[\Theta\begin{pmatrix}\boldsymbol{Y}^s\\ \boldsymbol{Y}^t\end{pmatrix}\right] \circ \left[\Theta\begin{pmatrix}\boldsymbol{Y}^s\\ \boldsymbol{Y}^t\end{pmatrix}\right]$,

$$\boldsymbol{c}' = \text{vec}\left(\left[\left\|X_i^s - X_j^t\right\|_2^2\right]_{ij}\right) \in \mathbb{R}^{n_s n_t},$$

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

the cost vector $\boldsymbol{c}'$ once computed from $X^s$ and $X^t$ remains fixed, $\text{vec}(\cdot)$ 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, the operator $\otimes$ is Kronecker product, $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix, and $\mathbf{1}_m \in \mathbb{R}^m$ is a vector of ones. The matrix $H$ is defined as $H = \begin{pmatrix}H_r & H_c\end{pmatrix}^\top \in \mathbb{R}^{(n_s+n_t)\times n_s n_t}$ in which

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

that performs the sum over the rows of $T$ and

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

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

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

$$\begin{aligned} &\min_{\boldsymbol{t}\in\mathbb{R}^{n_s n_t}} \boldsymbol{t}^T\left[(\boldsymbol{c}' + \Theta(\boldsymbol{a} + \boldsymbol{b}z)) \circ (\boldsymbol{c}' + \Theta(\boldsymbol{a} + \boldsymbol{b}z))\right] \quad \text{s.t.} \quad H\boldsymbol{t} = h, \quad \boldsymbol{t} \geq 0, \\ \Leftrightarrow &\min_{\boldsymbol{t}\in\mathbb{R}^{n_s n_t}} (\tilde{\boldsymbol{p}} + \tilde{\boldsymbol{q}}z + \tilde{\boldsymbol{r}}z^2)^\top \boldsymbol{t} \quad \text{s.t.} \quad H\boldsymbol{t} = h, \quad \boldsymbol{t} \geq 0. \end{aligned}$$

where

$$\tilde{\boldsymbol{p}} = (\boldsymbol{c}' + \Theta\boldsymbol{a}) \circ (\boldsymbol{c}' + \Theta\boldsymbol{a}), \quad \tilde{\boldsymbol{q}} = (\Theta\boldsymbol{a}) \circ (\Theta\boldsymbol{b}) + (\Theta\boldsymbol{b}) \circ (\Theta\boldsymbol{a}), \quad \text{and} \quad \tilde{\boldsymbol{r}} = (\Theta\boldsymbol{b}) \circ (\Theta\boldsymbol{b}).$$

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-basis variables $\mathcal{B}_u^c$ is defined as

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

where

$$\boldsymbol{p} = (\tilde{\boldsymbol{p}}_{\mathcal{B}_u^c}^\top - \tilde{\boldsymbol{p}}_{\mathcal{B}_u}^\top H_{:,\mathcal{B}_u}^{-1} H_{:,\mathcal{B}_u^c})^\top, \quad \boldsymbol{q} = (\tilde{\boldsymbol{q}}_{\mathcal{B}_u^c}^\top - \tilde{\boldsymbol{q}}_{\mathcal{B}_u}^\top H_{:,\mathcal{B}_u}^{-1} H_{:,\mathcal{B}_u^c})^\top, \quad \boldsymbol{r} = (\tilde{\boldsymbol{r}}_{\mathcal{B}_u^c}^\top - \tilde{\boldsymbol{r}}_{\mathcal{B}_u}^\top H_{:,\mathcal{B}_u}^{-1} H_{:,\mathcal{B}_u^c})^\top, \tag{21}$$

$H_{:,\mathcal{B}_u}^{-1}$ is a sub-matrix of $H$ 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 $\boldsymbol{r}_{\mathcal{B}_u^c} \geq \mathbf{0}$ (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 $\mathcal{Z}_u$ is defined as

$$\begin{aligned} \mathcal{Z}_u &= \{z \in \mathbb{R} \mid \mathcal{T}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{T}_u\}, \\ &= \{z \in \mathbb{R} \mid \mathcal{B}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{B}_u\}, \\ &= \{z \in \mathbb{R} \mid \boldsymbol{r}_{\mathcal{B}_u^c} = \boldsymbol{p} + \boldsymbol{q}z + \boldsymbol{r}z^2 \geq 0\}. \end{aligned}$$

Thus, we obtain the result in Lemma 3.

### 6.4 Proof of Lemma 4

The identification of $\mathcal{Z}_v$ is constructed based on the results presented in Lee et al. (2016), in which the authors characterized conditioning event of Lasso by deriving from the KKT conditions. Let us define the KKT conditions of the Lasso after the $\mathcal{T}_u$ transportation as following:

$$\begin{aligned} \tilde{X}_u^\top \left(\tilde{X}_u \hat{\boldsymbol{\beta}}(z) - \tilde{\boldsymbol{Y}}_u(z)\right) + \lambda \mathcal{S} = 0, \\ \mathcal{S}_j = \text{sign}(\hat{\boldsymbol{\beta}}_j(z)), \quad \text{if } \hat{\boldsymbol{\beta}}_j(z) \neq 0, \\ \mathcal{S}_j \in (-1, 1), \qquad \text{if } \hat{\boldsymbol{\beta}}_j(z) = 0. \end{aligned} \tag{22}$$

The two first conditions of the set:

$$\mathcal{Z}_v = \left\{ z \in \mathbb{R} \;\middle|\; \begin{array}{l} \hat{\boldsymbol{\beta}}_j(z) \neq 0, \ \forall j \in \mathcal{M}_v, \\ \hat{\boldsymbol{\beta}}_j(z) = 0, \ \forall j \notin \mathcal{M}_v, \\ \text{sign}\Big(\hat{\boldsymbol{\beta}}_{\mathcal{M}_v}(z)\Big) = \mathcal{S}_{\mathcal{M}_v} \end{array} \right\}$$

lead to the set $\mathcal{M}_v$ being the result of the Lasso after DA. Then, by partitioning Eq. (22) according to the active set $\mathcal{M}_v$, adopting the convention that $\mathcal{M}_{v_c}$ means "$\mathcal{M}_v$'s complement", the KKT conditions in (22) can be rewritten as following:

$$\begin{aligned} \tilde{X}_{u_{\mathcal{M}_v}}^\top \left(\tilde{X}_{u_{\mathcal{M}_v}} \hat{\boldsymbol{\beta}}_{\mathcal{M}_v}(z) - \tilde{\boldsymbol{Y}}_u(z)\right) + \lambda \mathcal{S}_{\mathcal{M}_v} = 0, \\ \tilde{X}_{u_{\mathcal{M}_{v_c}}}^\top \left(\tilde{X}_{u_{\mathcal{M}_v}} \hat{\boldsymbol{\beta}}_{\mathcal{M}_v}(z) - \tilde{\boldsymbol{Y}}_u(z)\right) + \lambda \mathcal{S}_{\mathcal{M}_{v_c}} = 0, \\ \text{sign}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_v}(z)) = \mathcal{S}_{\mathcal{M}_v}, \\ \left\|\mathcal{S}_{\mathcal{M}_{v_c}}\right\|_\infty < \mathbf{1}. \end{aligned} \tag{23}$$

By solving the first two equations (23) for $\hat{\boldsymbol{\beta}}_{\mathcal{M}_v}(z)$ and $\mathcal{S}_{\mathcal{M}_{v_c}}$, we obtain the equivalent set of conditions:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathcal{M}_v}(z) &= (\tilde{X}_{u_{\mathcal{M}_v}}^\top \tilde{X}_{u_{\mathcal{M}_v}})^{-1} (\tilde{X}_{u_{\mathcal{M}_v}}^\top \tilde{\boldsymbol{Y}}_u(z) - \lambda \mathcal{S}_{\mathcal{M}_v}), \\ \mathcal{S}_{\mathcal{M}_{v_c}} &= \tilde{X}_{u_{\mathcal{M}_{v_c}}}^\top (\tilde{X}_{u_{\mathcal{M}_v}}^\top)^+ \mathcal{S}_{\mathcal{M}_v} + \frac{1}{\lambda} \tilde{X}_{u_{\mathcal{M}_{v_c}}}^\top (I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v}} (\tilde{X}_{u_{\mathcal{M}_v}})^+) \tilde{\boldsymbol{Y}}_u(z), \\ \text{sign}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_v}(z)) &= \mathcal{S}_{\mathcal{M}_v}, \\ \left\|\mathcal{S}_{\mathcal{M}_{v_c}}\right\|_\infty &< \mathbf{1}, \end{aligned}$$

where $(X)^+ = (X^\top X)^{-1} X^\top$, $\left(X^\top\right)^+ = X(X^\top X)^{-1}$. Then, the set $\mathcal{Z}_v$ can be rewritten as:

$$\mathcal{Z}_v = \left\{ z \in \mathbb{R} \;\middle|\; \begin{array}{r} (\tilde{X}_{u_{\mathcal{M}_v}}^\top \tilde{X}_{u_{\mathcal{M}_v}})^{-1} (\tilde{X}_{u_{\mathcal{M}_v}}^\top \tilde{\boldsymbol{Y}}_u(z) - \lambda \mathcal{S}_{\mathcal{M}_v}) = \hat{\boldsymbol{\beta}}_{\mathcal{M}_v}(z), \\ \tilde{X}_{u_{\mathcal{M}_{v_c}}}^\top (\tilde{X}_{u_{\mathcal{M}_v}}^\top)^+ \mathcal{S}_{\mathcal{M}_v} + \frac{1}{\lambda} \tilde{X}_{u_{\mathcal{M}_{v_c}}}^\top (I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v}} (\tilde{X}_{u_{\mathcal{M}_v}})^+) \tilde{\boldsymbol{Y}}_u(z) = \mathcal{S}_{\mathcal{M}_{v_c}}, \\ \text{sign}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_v}(z)) = \mathcal{S}_{\mathcal{M}_v}, \\ \left\|\mathcal{S}_{\mathcal{M}_{v_c}}\right\|_\infty < \mathbf{1}. \end{array} \right\}$$

The two last conditions of $\mathcal{Z}_v$ then can be rewritten as:

$$\begin{aligned} &\left\{\text{sign}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_v}(z)) = \mathcal{S}_{\mathcal{M}_v}\right\} \\ &= \left\{\mathcal{S}_{\mathcal{M}_v} \circ \hat{\boldsymbol{\beta}}_{\mathcal{M}_v}(z) > \mathbf{0}\right\}, \\ &= \left\{\mathcal{S}_{\mathcal{M}_v} \circ (\tilde{X}_{u_{\mathcal{M}_v}}^\top \tilde{X}_{u_{\mathcal{M}_v}})^{-1} (\tilde{X}_{u_{\mathcal{M}_v}}^\top \tilde{\boldsymbol{Y}}_u(z) - \lambda \mathcal{S}_{\mathcal{M}_v}) > \mathbf{0}\right\}, \\ &= \left\{\mathcal{S}_{\mathcal{M}_v} \circ \left(\tilde{X}_{u_{\mathcal{M}_v}}\right)^+ \tilde{\boldsymbol{Y}}_u(z) > \lambda \mathcal{S}_{\mathcal{M}_v} \circ \left(\left(\tilde{X}_{u_{\mathcal{M}_v}}^\top \tilde{X}_{u_{\mathcal{M}_v}}\right)^{-1} \mathcal{S}_{\mathcal{M}_v}\right)\right\}, \\ &= \left\{\mathcal{S}_{\mathcal{M}_v} \circ \left(\tilde{X}_{u_{\mathcal{M}_v}}\right)^+ \Omega_u \boldsymbol{Y}(z) > \lambda \mathcal{S}_{\mathcal{M}_v} \circ \left(\left(\tilde{X}_{u_{\mathcal{M}_v}}^\top \tilde{X}_{u_{\mathcal{M}_v}}\right)^{-1} \mathcal{S}_{\mathcal{M}_v}\right)\right\}, \\ &= \left\{\mathcal{S}_{\mathcal{M}_v} \circ \left(\tilde{X}_{u_{\mathcal{M}_v}}\right)^+ \Omega_u (\boldsymbol{a} + \boldsymbol{b}z) > \lambda \mathcal{S}_{\mathcal{M}_v} \circ \left(\left(\tilde{X}_{u_{\mathcal{M}_v}}^\top \tilde{X}_{u_{\mathcal{M}_v}}\right)^{-1} \mathcal{S}_{\mathcal{M}_v}\right)\right\}, \\ &= \{\boldsymbol{\psi}_0 z \leq \boldsymbol{\phi}_0\}, \end{aligned}$$

$$
\begin{aligned}
&\left\{\left\|\mathcal{S}_{\mathcal{M}_{v_c}}\right\|_\infty < \mathbf{1}\right\}\\
&= \left\{-\mathbf{1} < \mathcal{S}_{\mathcal{M}_{v_c}} < \mathbf{1}\right\},\\
&= \left\{-\mathbf{1} < \tilde{X}^\top_{u_{\mathcal{M}_{v_c}}}(\tilde{X}^\top_{u_{\mathcal{M}_v}})^+\mathcal{S}_{\mathcal{M}_v} + \frac{1}{\lambda}\tilde{X}^\top_{u_{\mathcal{M}_{v_c}}}(I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v}}(\tilde{X}_{u_{\mathcal{M}_v}})^+)\tilde{\boldsymbol{Y}}_u(z) < \mathbf{1}\right\},\\
&= \left\{\begin{array}{l} \frac{1}{\lambda}\tilde{X}^\top_{u_{\mathcal{M}_{v_c}}}(I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v}}(\tilde{X}_{u_{\mathcal{M}_v}})^+)\Omega_u\boldsymbol{Y}(z) < \mathbf{1} - \tilde{X}^\top_{u_{\mathcal{M}_{v_c}}}(\tilde{X}^\top_{u_{\mathcal{M}_v}})^+\mathcal{S}_{\mathcal{M}_v}\\ -\frac{1}{\lambda}\tilde{X}^\top_{u_{\mathcal{M}_{v_c}}}(I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v}}(\tilde{X}_{u_{\mathcal{M}_v}})^+)\Omega_u\boldsymbol{Y}(z) < \mathbf{1} + \tilde{X}^\top_{u_{\mathcal{M}_{v_c}}}(\tilde{X}^\top_{u_{\mathcal{M}_v}})^+\mathcal{S}_{\mathcal{M}_v}\end{array}\right\},\\
&= \left\{\begin{array}{l} \frac{1}{\lambda}\tilde{X}^\top_{u_{\mathcal{M}_{v_c}}}(I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v}}(\tilde{X}_{u_{\mathcal{M}_v}})^+)\Omega_u(\boldsymbol{a}+\boldsymbol{b}z) < \mathbf{1} - \tilde{X}^\top_{u_{\mathcal{M}_{v_c}}}(\tilde{X}^\top_{u_{\mathcal{M}_v}})^+\mathcal{S}_{\mathcal{M}_v}\\ -\frac{1}{\lambda}\tilde{X}^\top_{u_{\mathcal{M}_{v_c}}}(I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v}}(\tilde{X}_{u_{\mathcal{M}_v}})^+)\Omega_u(\boldsymbol{a}+\boldsymbol{b}z) < \mathbf{1} + \tilde{X}^\top_{u_{\mathcal{M}_{v_c}}}(\tilde{X}^\top_{u_{\mathcal{M}_v}})^+\mathcal{S}_{\mathcal{M}_v}\end{array}\right\},\\
&= \left\{\begin{pmatrix}\boldsymbol{\psi}_{10}\\ \boldsymbol{\psi}_{11}\end{pmatrix} z \le \begin{pmatrix}\boldsymbol{\phi}_{10}\\ \boldsymbol{\phi}_{11}\end{pmatrix}\right\},\\
&= \{\boldsymbol{\psi}_1 z \le \boldsymbol{\phi}_1\}.
\end{aligned}
$$

Finally, the set $\mathcal{Z}_v$ can be defined as:

$$\mathcal{Z}_v = \{z \in \mathbb{R} \mid \boldsymbol{\psi} z \le \boldsymbol{\phi}\},$$

where $\boldsymbol{\psi} = (\boldsymbol{\psi}_0 \quad \boldsymbol{\psi}_1)^\top$, $\boldsymbol{\phi} = (\boldsymbol{\phi}_0 \quad \boldsymbol{\phi}_1)^\top$.
Thus, the set $\mathcal{Z}_v$ can be identified by solving a set of linear inequalities w.r.t $z$.

### 6.5 Proof of Lemma 5

The identification of $\mathcal{Z}_v^{\text{enet}}$ is constructed similar to that in Lemma 4. Let us define the KKT conditions of the elastic net after the $\mathcal{T}_u$ transportation as following:

$$
\begin{aligned}
(\tilde{X}_u^\top\tilde{X}_u + \gamma I_p)\hat{\boldsymbol{\beta}}^{\text{enet}}(z) - \tilde{X}_u^\top\tilde{\boldsymbol{Y}}_u(z) + \lambda\mathcal{S} &= 0,\\
\mathcal{S}_j = \text{sign}(\hat{\boldsymbol{\beta}}_j^{\text{enet}}(z)), \quad &\text{if } \hat{\boldsymbol{\beta}}_j^{\text{enet}}(z) \neq 0,\\
\mathcal{S}_j \in (-1, 1), \quad &\text{if } \hat{\boldsymbol{\beta}}_j^{\text{enet}}(z) = 0.
\end{aligned}
\tag{24}
$$

By following the same approach as in Lemma 4, the KKT conditions of elastic net in (24) can be partitioned according to $\mathcal{M}_v^{\text{enet}}$ as below:

$$
\begin{aligned}
(\tilde{X}^\top_{u_{\mathcal{M}_v^{\text{enet}}}}\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}} + \gamma I)\hat{\boldsymbol{\beta}}^{\text{enet}}(z) - \tilde{X}^\top_{u_{\mathcal{M}_v^{\text{enet}}}}\tilde{\boldsymbol{Y}}_u(z) + \lambda\mathcal{S} &= 0,\\
\tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}}\left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\hat{\boldsymbol{\beta}}_{\mathcal{M}_v^{\text{enet}}}(z) - \tilde{\boldsymbol{Y}}_u(z)\right) + \lambda\mathcal{S}_{\mathcal{M}_{v_c}^{\text{enet}}} &= 0,\\
\text{sign}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_v}(z)) &= \mathcal{S}_{\mathcal{M}_v},\\
\left\|\mathcal{S}_{\mathcal{M}_{v_c}}\right\|_\infty &< \mathbf{1}.
\end{aligned}
\tag{25}
$$

By solving the first two equations (25) for $\hat{\boldsymbol{\beta}}_{\mathcal{M}_v^{\text{enet}}}(z)$ and $\mathcal{S}_{\mathcal{M}_{v_c}^{\text{enet}}}$, we obtain the equivalent set of conditions:

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_{\mathcal{M}_v^{\text{enet}}}(z) &= (\tilde{X}^\top_{u_{\mathcal{M}_v^{\text{enet}}}}\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}} + \gamma I)^{-1}(\tilde{X}^\top_{u_{\mathcal{M}_v^{\text{enet}}}}\tilde{\boldsymbol{Y}}_u(z) - \lambda\mathcal{S}_{\mathcal{M}_v^{\text{enet}}}),\\
\mathcal{S}_{\mathcal{M}_{v_c}^{\text{enet}}} &= \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}}(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}})^{**}\mathcal{S}_{\mathcal{M}_v^{\text{enet}}} + \frac{1}{\lambda}\tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}}(I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}})^*)\tilde{\boldsymbol{Y}}_u(z),\\
\text{sign}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_v^{\text{enet}}}(z)) &= \mathcal{S}_{\mathcal{M}_v^{\text{enet}}},\\
\left\|\mathcal{S}_{\mathcal{M}_{v_c}^{\text{enet}}}\right\|_\infty &< \mathbf{1},
\end{aligned}
$$

where $(X)^* = (X^\top X + \gamma I)^{-1}X^\top$, $(X)^{**} = X(X^\top X + \gamma I)^{-1}$. The set $\mathcal{Z}_v^{\text{enet}}$ can be rewritten as:

$$
\mathcal{Z}_v^{\text{enet}} = \left\{ z \in \mathbb{R} \;\middle|\; \begin{array}{r} (\tilde{X}^\top_{u_{\mathcal{M}_v^{\text{enet}}}}\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}} + \gamma I)^{-1}(\tilde{X}^\top_{u_{\mathcal{M}_v^{\text{enet}}}}\tilde{\boldsymbol{Y}}_u(z) - \lambda\mathcal{S}_{\mathcal{M}_v^{\text{enet}}}) = \hat{\boldsymbol{\beta}}_{\mathcal{M}_v^{\text{enet}}}(z),\\ \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}}(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}})^{**}\mathcal{S}_{\mathcal{M}_v^{\text{enet}}} + \frac{1}{\lambda}\tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}}(I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}})^*)\tilde{\boldsymbol{Y}}_u(z) = \mathcal{S}_{\mathcal{M}_{v_c}^{\text{enet}}},\\ \text{sign}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_v^{\text{enet}}}(z)) = \mathcal{S}_{\mathcal{M}_v^{\text{enet}}},\\ \left\|\mathcal{S}_{\mathcal{M}_{v_c}^{\text{enet}}}\right\|_\infty < \mathbf{1}. \end{array} \right\}
$$

The last two conditions of $\mathcal{Z}_v^{\text{enet}}$ can be characterized by a set of inequalities as in Lemma 4 with similar approach:

$$
\begin{aligned}
&\left\{\text{sign}(\hat{\boldsymbol{\beta}}_{\mathcal{M}_v^{\text{enet}}}(z)) = \mathcal{S}_{\mathcal{M}_v^{\text{enet}}}\right\} \\
&= \left\{\mathcal{S}_{\mathcal{M}_v^{\text{enet}}} \circ \hat{\boldsymbol{\beta}}_{\mathcal{M}_v^{\text{enet}}}(z) > \mathbf{0}\right\}, \\
&= \left\{\mathcal{S}_{\mathcal{M}_v^{\text{enet}}} \circ (\tilde{X}^\top_{u_{\mathcal{M}_v^{\text{enet}}}} \tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}} + \gamma I)^{-1} (\tilde{X}^\top_{u_{\mathcal{M}_v^{\text{enet}}}} \tilde{\boldsymbol{Y}}_u(z) - \lambda \mathcal{S}_{\mathcal{M}_v^{\text{enet}}}) > \mathbf{0}\right\}, \\
&= \left\{\mathcal{S}_{\mathcal{M}_v^{\text{enet}}} \circ \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^* \tilde{\boldsymbol{Y}}_u(z) > \lambda \mathcal{S}_{\mathcal{M}_v^{\text{enet}}} \circ \left(\left(\tilde{X}^\top_{u_{\mathcal{M}_v^{\text{enet}}}} \tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}} + \gamma I\right)^{-1} \mathcal{S}_{\mathcal{M}_v^{\text{enet}}}\right)\right\}, \\
&= \left\{\mathcal{S}_{\mathcal{M}_v^{\text{enet}}} \circ \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^* \Omega_u \boldsymbol{Y}(z) > \lambda \mathcal{S}_{\mathcal{M}_v^{\text{enet}}} \circ \left(\left(\tilde{X}^\top_{u_{\mathcal{M}_v^{\text{enet}}}} \tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}} + \gamma I\right)^{-1} \mathcal{S}_{\mathcal{M}_v^{\text{enet}}}\right)\right\}, \\
&= \left\{\mathcal{S}_{\mathcal{M}_v^{\text{enet}}} \circ \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^* \Omega_u (\boldsymbol{a} + \boldsymbol{b}z) > \lambda \mathcal{S}_{\mathcal{M}_v^{\text{enet}}} \circ \left(\left(\tilde{X}^\top_{u_{\mathcal{M}_v^{\text{enet}}}} \tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}} + \gamma I\right)^{-1} \mathcal{S}_{\mathcal{M}_v^{\text{enet}}}\right)\right\}, \\
&= \left\{\boldsymbol{\psi}_0^{\text{enet}} z \le \boldsymbol{\phi}_0^{\text{enet}}\right\}, \\
&\left\{\left\|\mathcal{S}_{\mathcal{M}_{v_c}^{\text{enet}}}\right\|_\infty < \mathbf{1}\right\} \\
&= \left\{-\mathbf{1} < \mathcal{S}^{\text{enet}}_{\mathcal{M}_{v_c}} < \mathbf{1}\right\}, \\
&= \left\{-\mathbf{1} < \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}} (\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}})^{**} \mathcal{S}_{\mathcal{M}_v^{\text{enet}}} + \frac{1}{\lambda} \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}} (I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}} (\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}})^*) \tilde{\boldsymbol{Y}}_u(z) < \mathbf{1}\right\}, \\
&= \left\{\begin{array}{l} \frac{1}{\lambda} \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}} \left(I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}} \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^*\right) \tilde{\boldsymbol{Y}}_u(z) < \mathbf{1} - \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}} \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^{**} \mathcal{S}_{\mathcal{M}_v^{\text{enet}}} \\ -\frac{1}{\lambda} \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}} \left(I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}} \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^*\right) \tilde{\boldsymbol{Y}}_u(z) < \mathbf{1} + \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}} \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^{**} \mathcal{S}_{\mathcal{M}_v^{\text{enet}}} \end{array}\right\}, \\
&= \left\{\begin{array}{l} \frac{1}{\lambda} \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}} \left(I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}} \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^*\right) \Omega_u \boldsymbol{Y}(z) < \mathbf{1} - \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}} \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^{**} \mathcal{S}_{\mathcal{M}_v^{\text{enet}}} \\ -\frac{1}{\lambda} \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}} \left(I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}} \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^*\right) \Omega_u \boldsymbol{Y}(z) < \mathbf{1} + \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}} \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^{**} \mathcal{S}_{\mathcal{M}_v^{\text{enet}}} \end{array}\right\}, \\
&= \left\{\begin{array}{l} \frac{1}{\lambda} \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}} \left(I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}} \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^*\right) \Omega_u (\boldsymbol{a} + \boldsymbol{b}z) < \mathbf{1} - \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}} \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^{**} \mathcal{S}_{\mathcal{M}_v^{\text{enet}}} \\ -\frac{1}{\lambda} \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}} \left(I_{n_s+n_t} - \tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}} \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^*\right) \Omega_u (\boldsymbol{a} + \boldsymbol{b}z) < \mathbf{1} + \tilde{X}^\top_{u_{\mathcal{M}_{v_c}^{\text{enet}}}} \left(\tilde{X}_{u_{\mathcal{M}_v^{\text{enet}}}}\right)^{**} \mathcal{S}_{\mathcal{M}_v^{\text{enet}}} \end{array}\right\}, \\
&= \left\{\begin{pmatrix} \boldsymbol{\psi}_{10}^{\text{enet}} \\ \boldsymbol{\psi}_{11}^{\text{enet}} \end{pmatrix} z \le \begin{pmatrix} \boldsymbol{\phi}_{10}^{\text{enet}} \\ \boldsymbol{\phi}_{11}^{\text{enet}} \end{pmatrix}\right\}, \\
&= \left\{\boldsymbol{\psi}_1^{\text{enet}} z \le \boldsymbol{\phi}_1^{\text{enet}}\right\}.
\end{aligned}
$$

Finally, the set $\mathcal{Z}_v^{\text{enet}}$ can also be identified by solving a set of linear inequalities w.r.t $z$:

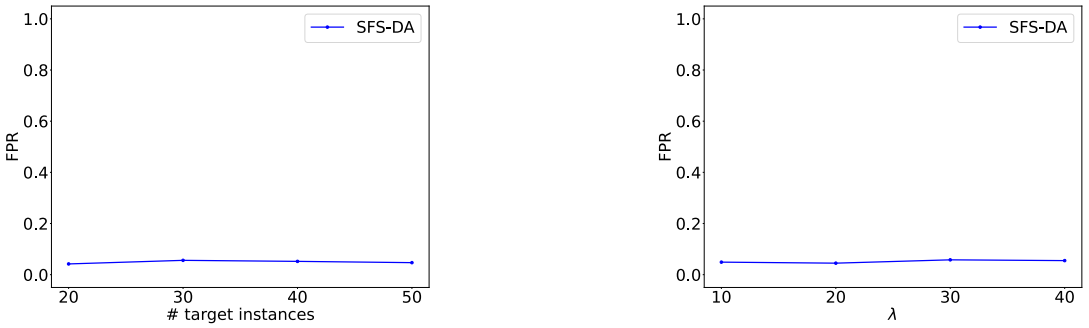
$$\mathcal{Z}_v^{\text{enet}} = \left\{z \in \mathbb{R} \mid \boldsymbol{\psi}^{\text{enet}} z \le \boldsymbol{\phi}^{\text{enet}}\right\},$$

where $\boldsymbol{\psi}^{\text{enet}} = (\boldsymbol{\psi}_0^{\text{enet}} \quad \boldsymbol{\psi}_1^{\text{enet}})^\top$, $\boldsymbol{\phi}^{\text{enet}} = (\boldsymbol{\phi}_0^{\text{enet}} \quad \boldsymbol{\phi}_1^{\text{enet}})^\top$.

### 6.6 Additional Experiments

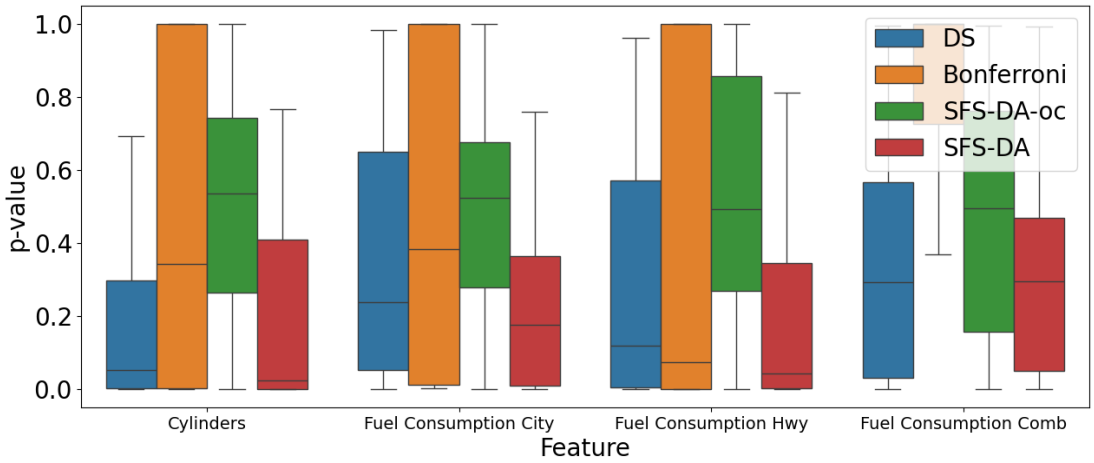
**FPR results with varying $n_t$ and the hyper-parameter $\lambda$.** We conducted additional experiments to demonstrate SFS-DA's ability to control the FPR across varying target sample sizes and different values of the hyper-parameter $\lambda$. For the experiments with varying target sample size, we set $n_t = \{20, 30, 40, 50\}$ and $n_s = 150$. For the experiments with different values of the hyper-parameter $\lambda$, we set $\lambda = \{10, 20, 30, 40\}, n_s = 100$, and $n_t = 10$. The results are shown in Fig. 10. They demonstrate that the proposed SFS-DA maintains its FPR control at $\alpha = 0.05$.

**Real-world datasets.** As noted in §4, we also we conducted a comparison of different methods on two datasets: the CO2 Emissions Canada dataset and the Walmart dataset, both available on Kaggle. We also present the percentage of $p$-value less than or equal $\alpha$ on each dataset, which can be used for providing insights into the statistical power of each method.

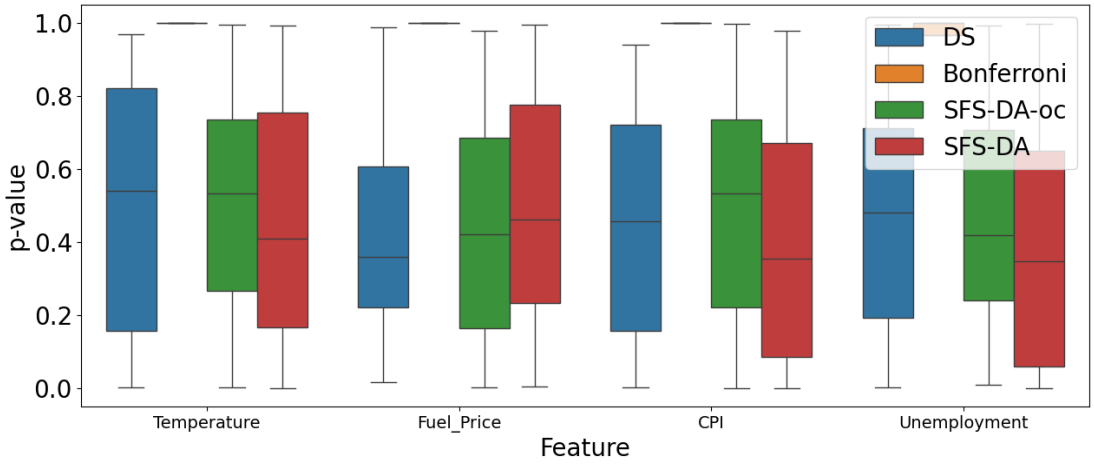



(a) Varying $n_t$ (b) Varying $\lambda$

Figure 10: FPR results with varying $n_t$ and the hyper-parameter $\lambda$.

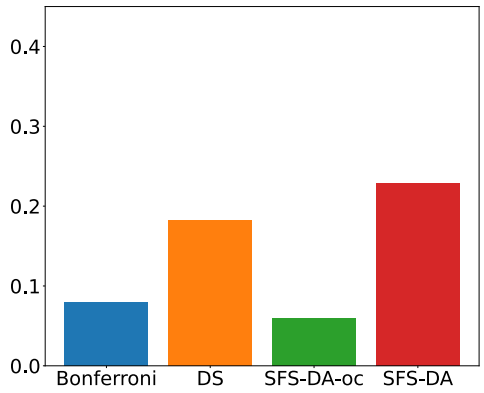



(a) CO2 Emissions Canada dataset. The source domain comprises "vehicles using gasoline fuel", while the target domain includes "vehicles that use other types of fuel".

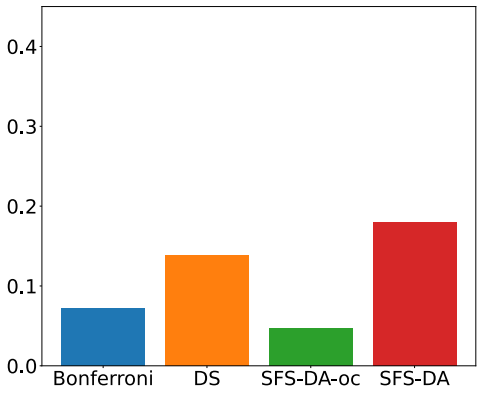



(b) Walmart dataset. The source domain is "people who go shopping at Walmart on regular days", while the target domain is "people who go shopping at Walmart on holidays".

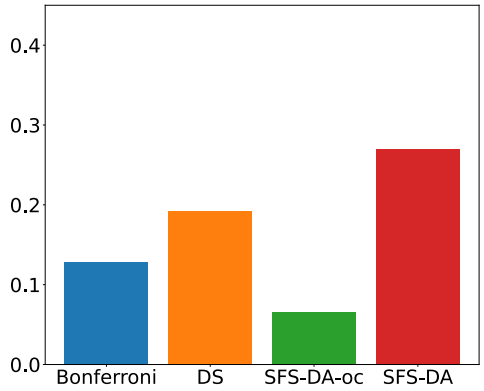
Figure 11: Distributions of $p$-values for each feature in the CO2 Emissions Canada and Walmart datasets. While the median $p$-value for `DS` is smaller than that of the proposed `SFS-DA` in one case (Fuel Price in the Walmart dataset), in all other cases, the median $p$-value for `SFS-DA` is smaller than those of the other methods, indicating that `SFS-DA` demonstrates the superior statistical power.



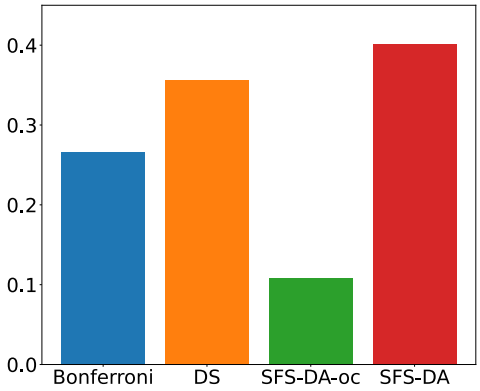

(a) Diabetes



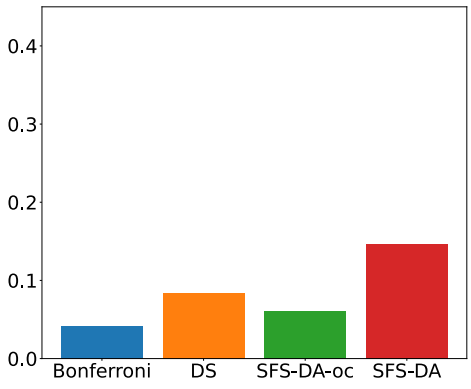

(b) Heart Failure




(c) Seoul Bike




(d) CO2 Emissions Canada




(e) Walmart

Figure 12: $\mathbb{P}(p\text{-value} \leq \alpha)$ on each dataset, where $\alpha = 0.05$. In this setting, we randomly select one feature from the set of selected features to conduct inference. In all cases, the percentage of significant $p$-values from the proposed method is higher than that of the competing methods. This suggests that our proposed `SFS-DA` method exhibits higher statistical power compared to the alternatives.