Robust Causal Transfer Learning

Anonymous Author
Anonymous Institution

Abstract

One of the most important problems in transfer learning is the task of domain adaptation, where the goal is to apply an algorithm trained in one or more source domains to a different (but related) target domain. In many cases, the differences between the source and target domains can be modeled and characterized by a set of variables that are observed under different contexts (e.g. experimental settings) particular to each domain. These context variables can often be linked to real-world environmental variables, such as the organism in a genetics study or the environment in a data-collecting research institute. We propose `RCTL`, an algorithm that identifies invariant causal features across source and target domains based on Markov blankets. `RCTL` is scalable and robust, thanks to local causal discovery that increases the power of computational independence tests and makes the task of domain adaptation computationally tractable. We show the scalability and robustness of `RCTL` using synthetic and a real data set in different domain adaptation scenarios.

1 Introduction

Standard supervised learning usually assume that both training and test data are drawn from the same distribution. However, this is a strong assumption, and often violated in practice if the training and test data are sampled under different conditions, due to either differing data collection processes or differing applications (Csurka, 2017; Li et al., 2019). Domain adaptation approaches aim to learn domain invariant features to mitigate the problem of data shift and enhance the quality of predictions (Redko et al., 2019; Stojanov et al., 2019a; Park et al., 2020).

Many of domain adaptation techniques in the literature consider the *covariate shift* situation (Chen et al., 2016; von Kügelgen, Mey, and Loog, 2019; Stojanov et al., 2019b; Li, Lam, and Prusty, 2020; Kisamori, Kanagawa, and Yamazaki, 2020), where the marginal distribution of the features differs across the source and target domains, while the conditional distribution of the target given the features does not change. From the causal inference point of view, it has been noted in (Schoelkopf et al., 2012) that covariate shift corresponds to causal learning i.e., predicting effects from causes. Taking into account the causal structure of a system of interest and finding causally invariant features in both source and target domains, enable us to safely transfer the predictions of the target variable based only on causally invariant features to the target domain (Magliacane et al., 2018; Rojas-Carulla et al., 2018a; Subbaswamy and Saria, 2018; Subbaswamy, Schulam, and Saria, 2019). To illustrate the importance and effectiveness of this approach consider the following example.

Example 1 (Prediction of Diabetes at Early Stage)**.** *According to Diabetes Australia, failure in early detection of TYPE II could cost the Australian healthcare system more than 700 million dollars each year (Australia, 2020). Total annual cost impact of diabetes in Australia estimated at 14.6 billion dollars (Australia, 2020). In 2017, the total expenditure of diagnosed diabetes in the United States alone was 327 billion USD (Association, 2020). Therefore, early diagnosis and initiation of appropriate therapeutic treatment may play a pivotal role in: (1) helping patients to manage the disease early and potentially prevent or delay the serious disease complications that can decrease quality of life and (2) reducing huge economic impact of diabetes on the healthcare systems and national economies. To predict diabetes using data mining and machine learning techniques, we need its symptoms along with clinical data. The common symptoms and possible causes of diabetes Type II are weakness, obesity, delayed healing, visual blurring, partial paresis,*

Preliminary work. Under review by AISTATS 2021. Do not distribute.

muscle stiffness, alopecia, among others (Association, 2020). The pancreas is the organ that produces insulin, and it plays a major role in regulating blood glucose levels. Although we do not know what causes diabetes Type II (Australia, 2020), one may argue that the occurrence, rate, or frequency of Delayed Healing, Blurred Vision, Partial Paresis, and Weakness increases with age, and Delayed Healing and Partial Paresis may cause some complications that result in pancreas malfunction.

Prediction of Diabetes and Domain Adaptation Problem. The causal graph of this scenario can be represented as the directed acyclic graph (DAG) in Figure 1 (a). To provide an instance of a domain adaptation problem, we divided the diabetes dataset (Islam et al., 2020) into two sub-populations: (1) source domain with patients of the age less than 50 (we call them young patients) and (2) target domain with patients of the age greater than or equals to 50 (we call them old patients). As shown in Figure 1 (b) and (c), intervention in age leads to shift distributions across domains. In our experiments, feature selection methods that do not take into account the causal structure select all four variables *Delayed Healing*, *Blurred Vision*, *Partial Paresis*, and *Weakness* (highly related features to diabetes in this scenario) that lead to worse predictions in the target domain (MSE = 0.29 and SSE = 65.322) than the time that we only consider *Delayed Healing* and *Partial Paresis* for predictions in the target domain (MSE = 0.221 and SSE = 49.8927). The reason is that conditioning on the variables *Blurred Vision* and *Weakness* makes the paths between age and diabetes open. In other words, in this case, the set of features does not generalize to the target domain and lead to bad predictions. However, conditioning only on *Delayed Healing* and *Partial Paresis* blocks the paths between age and diabetes. In other words, *Delayed Healing* and *Partial Paresis* are causally invariant features that can be used for the prediction of diabetes even in the presence of distribution shifts due to the age intervention.

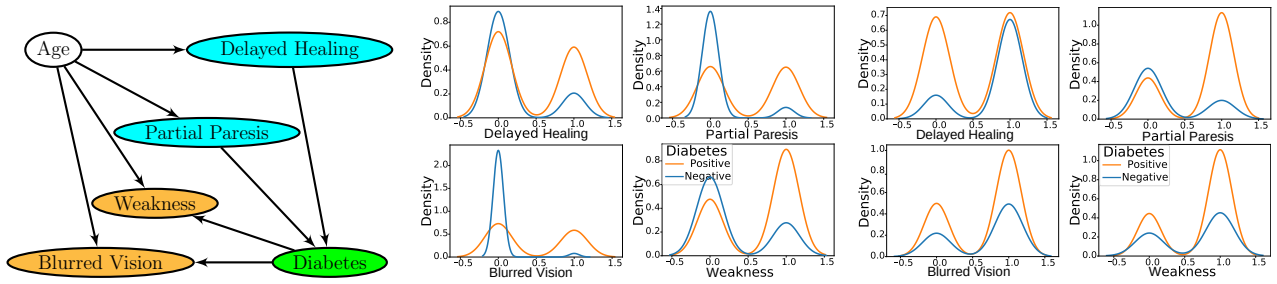
Our methodology in this paper is premised upon the recent work (Magliacane et al., 2018) that proposes a method for solving the problem of domain adaptation that exploits causal inference and does not even rely on prior knowledge of the causal graph, we refer to this approach as **E**xhaustive **S**ubset **S**earch (`ESS`). `ESS` simply uses a brute-force search algorithm to find a subset of variables (i.e., causally invariant features) that separates (more precisely, d-separates) the target variable from those variables (known as *context variables*) that are responsible for data shift across the source and target domains. Under certain assumptions, called *causal domain adaption assumptions* (see section 4.2 for details), `ESS` correctly detects causally invariant features. However, `ESS` can handle about only seven variables on a laptop computer due to the brute-force algorithm used in `ESS` (Magliacane et al., 2018). Moreover, `ESS` has been designed base on conditional independence (CI) tests, and an increased cardinality of the conditioning set reduces the robustness of most independence tests (Spirtes, Glymour, and Scheines, 2001). To overcome the problems of `ESS` regarding *scalability* and *robustness*, we propose an algorithm based on Markov blanket discovery (see section 3 for the definition) that (1) takes advantage of local computations and makes our proposed algorithm scalable, and (2) reduces the search space for finding causally invariant features drastically due to the cardinality of the Markov blankets which are fairly small (≤ 10) in (almost all) causal models. As a result the CI tests are more robust than `ESS` when the number of variables in the systems goes beyond dozen of variables. Our main theoretical and empirical contributions are as follows:

- We address the main limitations of causal inference techniques (i.e., scalability and robustness) in the literature for solving the problem of domain adaptation in the presence of covariate shift and we propose a new algorithm, called **R**obust **C**ausal **T**ransfer **L**earning (`RCTL`), to solve the problem of domain adaptation in the presence of covariate shift, and we prove its correctness (section 4).
- We prove that the standard Markov blanket discovery algorithms such as Grow-Shrink (GSMB), IAMB algorithm and its variants are still correct under the faithfulness assumption where causal sufficiency is not assumed (section 4).
- Experimentally, we demonstrate on synthesized and real-world data that our proposed algorithm improves performance over numerous state-of-the-art algorithms in the presence of covariate shift (section 6).

2 Related Work

The contributions of this paper regarding a robust causal inference technique for domain adaptation task intersect with various works in the literature as follows.

Common Domain Adaptation Scenarios. Here, we provide a brief overview of the main domain adaptation scenarios that can be found in the literature. There are three main domain adaptation problems: (1) **Covariate shift**, which is one of the most studied forms of data shift, occurs if the marginal distributions of *context variables* change across source and target domains while the posterior (conditional)



(a) Causal Graph for Diabetes Dataset

(b) Diabetes Distribution in the Source (Young Patients)

(c) Diabetes Distribution in the Target (Old Patients)

Figure 1: Prediction of Diabetes at Early Stage: In this scenario, age intervention leads to shift of distributions between source (Young Patients < 50) and target (Old Patients ≥ 50) domain (see also Example 1). In all cases, the orange curves indicate the distribution of tested positive patients and the blue curves indicate the distribution of tested negative patients. A standard feature selection method that does not take into account the causal structure, but would use *Weakness* and *Blurred Vision* to predict *Diabetes* (because *Weakness* and *Blurred Vision* are in the Markov blanket of *Diabetes*, and hence they are not only highly relevant to *Diabetes* but also good predictors of *Diabetes* in the source domain), would obtain biased predictions in the target domain (MSE = 0.29 and SSE = 65.322). Using *Delayed Healing* and *Partial Paresis* instead yields less accurate predictions in the source domain, but much more accurate ones in the target domain (MSE = 0.221 and SSE = 49.8927).

distributions are the same between source and target domains (Shimodaira, 2000; Sugiyama et al., 2008; Gretton et al., 2009). (2) **Target shift** occurs if the marginal distributions of the *target variable* changes across source and target domains while the posterior distributions are the same between source and target domains (Storkey, 2009; Zhang et al., 2013; Lipton, Wang, and Smola, 2018). (3) **Concept shift** occurs if marginal distributions between source and target domains remain unchanged while the posteriors change across the domains (Moreno-Torres et al., 2012; Zhang, Gong, and Scholkopf, 2015; Gong et al., 2016). We only focus on the covariate shift in this paper. For a more comprehensive discussion about domain adaptation problems and methods see (Kouw and Loog, 2019).

Causal Inference and Domain Adaptation Methods. Here, we briefly provide an overview of causal inference methods that address the problem of covariate shift: (1) **Transportability** in causal inference formalized in (Pearl and Bareinboim, 2011; Bareinboim and Pearl, 2012, 2014) by introducing a formal representation called *selection diagram* for expressing knowledge about differences and commonalities between the source and target domains. Using this representation along with the *do-calculus* (Pearl, 2009), enable us to derive a procedure for deciding whether effects in the target domain can be inferred from experiments conducted in the source domain(s). (2) **Graph surgery** is a *proactive* approach proposed in (Subbaswamy and Saria, 2018; Subbaswamy, Schulam, and Saria, 2019) that uses the graphical knowledge of the causal mechanisms to proactively remove variables generated by unstable mechanisms from the joint factorization to yield a distribution that is *invariant* to the differences across domains. (3) **Graph pruning** methods (Magliacane et al., 2018; Rojas-Carulla et al., 2018a) formalized as a feature selection problem in which the goal is to find the *optimal subset* that the conditional distribution of the target variable given this subset of predictors is invariant across domains. Both transportability and graph surgery methods assume that the structure of the causal mechanism as an *acyclic directed mixed graph* (ADMG) is the same in the source and target domains and also the causal model is considered to be known in advance. However, learning causal models from data is a challenging task, especially in the presence of *unmeasured confounders* (Glymour, Zhang, and Spirtes, 2019). On the other hand, graph pruning methods do not rely on prior knowledge of the causal graph but they do not allow for interventions on the target variable and concentrate on strong perturbations. Another main limitation of these approaches is that they, currently, *do not scale beyond dozens of variables* (Kouw and Loog, 2019). In this paper we propose a graph pruning approach that takes advantage of local causal discovery to overcome the problem of *scalability*. For this purpose, our proposed method uses Markov blanket recovery algorithms to reduce the search space for finding causally invariant features. This in turn results in a significant reduction in the number of conditional independence tests, and therefore of the overall computational complexity of the proposed algorithm. This also increases the *reliability* of independence tests for small data sets, especially in high-dimensional settings, due to the ex-

istence of sample efficient Markov blanket recovery algorithms (Aliferis et al., 2010).

Markov Blanket Recovery Here, we briefly provide an overview of methods that address the problem of Markov blanket discovery: (1) Markov Blanket Recovery for *Bayesian Networks with Causal Sufficiency Assumption*: In (Margaritis and Thrun, 1999), the authors presented the first provably correct algorithm, called Grow-Shrink Markov Blanket (GSMB), that discovers the Markov blanket of a variable from a *faithful* data under the *causal sufficiency assumption*. Variants of GSMB were proposed to improve *speed* and *reliability* such as the Incremental Association Markov Blanket (IAMB) and its variants (Tsamardinos et al., 2003), Fast-IAMB (Yaramakala and Margaritis, 2005), and IAMB with false discovery rate control (IAMB-FDR) (Peña, 2008). Since in discrete data the sample size required for high-confidence statistical tests of conditional independence in GSMB and IAMB algorithms grows exponentially in the size of the Markov blanket, several sample-efficient algorithms e.g., HITON-MB (Aliferis et al., 2010) and Max–Min Markov Blanket (MMMB) (Tsamardinos et al., 2006) were proposed to overcome the data inefficiency of GSMB and IAMB algorithms. One can find alternative computational methods for Markov blanket discovery that were developed in the past two decades in (Peña, 2007; Liu and Liu, 2016; Ling et al., 2019), among others. (2) Markov Blanket Recovery for *Chain Graphs with Causal Sufficiency Assumption*: In (Javidian, P. Jamshidi, and Valtorta, 2020), the authors extended the concept of Markov blankets to chain graphs (Lauritzen, 1996), which is different from Markov blankets defined in DAGs. (3) Markov Blanket Recovery for *Causal Graphs without Causal Sufficiency Assumption*: Gao and Ji Gao and Ji (2016) proposed the latent Markov blanket learning with constrained structure EM algorithm (LMB-CSEM) to discover the Markov blanket of a given target variable in the presence of unmeasured confounders. However, LMB-CSEM was proposed to find the Markov blankets in a DAGs and provides no theoretical guarantees for finding all possible unmeasured confounders in the Markov blanket of the target variable. Recently, Yu et. al. Yu et al. (2018) proposed a new algorithm, called M3B, to discover Markov blanket of a given target variable in the presence of unmeasured confounders. In section 4 we prove that the GSMB, IAMB algorithm and its variants are still correct under the faithfulness assumption where causal sufficiency is not assumed.

3 Basic Definitions and Concepts

Assume that $G = (V, E)$ is a directed graph, where $V = \{v_1, v_2, v_3 \cdots v_n\}$ is the set of nodes (variables), $n \geq 1$, and E is the set of directed or bidirected edges. We say v_i is a *parent* of v_j and v_j is a *child* of v_i if $v_i \rightarrow v_j$ is an edge in G. We denote the set of parents and children of a variable v by $\mathbf{pa}(v)$ and $\mathbf{ch}(v)$, respectively. Any bidirected edge $v_i \leftrightarrow v_j$ means that there exists a node $v_k \notin V$ as a hidden confounder such that $v_i \leftarrow v_k \rightarrow v_j$. If G is acyclic, then G is an *Acyclic Directed Mixed Graph* (ADMG). Formally, $\mathbf{ne}(T) = \mathbf{pa}(T) \cup \mathbf{ch}(T) \cup \{v \in V | v \leftrightarrow T$ is a bidirected edge in $G\}$ refers to the *neighbours* of T. We define *spouses* of v as $\mathbf{sp}(v) = \{u \in V | \exists w \in V$ s.t. $u \rightarrow w \leftarrow v$ in $G\}$. We define *Markov blanket* of node T as $\mathbf{Mb}(T) = \mathbf{pa}(T) \cup \mathbf{ch}(T) \cup \mathbf{sp}(T)$ when G is a directed acyclic graph (DAG) and the set of children, parents, and spouses of T, and vertices connected with T or children of T by a bidirected path (i.e., only with edges $\leftrightarrow$) and their respective parents is the Markov blanket of T when G is an ADMG (Statnikov, Lemeir, and Aliferis, 2013).

A vertex α is said to be an *ancestor* of a vertex β if either there is a directed path $\alpha \rightarrow \cdots \rightarrow \beta$ from α to β, or $\alpha = \beta$. We apply this definition to sets: $\mathbf{an}(X) = \{\alpha | \alpha$ is an ancestor of β for some $\beta \in X\}$.

Definition 1. *A nonendpoint vertex ζ on a path is a* collider *on the path if the edges preceding and succeeding ζ on the path have an arrowhead at ζ, that is, $\rightarrow \zeta \leftarrow$, or $\leftrightarrow \zeta \leftrightarrow$, or $\leftrightarrow \zeta \leftarrow$, or $\rightarrow \zeta \leftrightarrow$. A nonendpoint vertex ζ on a path which is not a collider is a noncollider on the path. A path between vertices α and β in an ADMG G is said to be* m*-connecting given a set Z (possibly empty), with $\alpha, \beta \notin Z$, if:*

(i) every noncollider on the path is not in Z, and

(ii) every collider on the path is in $\boldsymbol{an}_G(Z)$.

If there is no path m-connecting α and β given Z, then α and β are said to be m-separated *given Z. Sets X and Y are* m*-separated given Z, if for every pair α, β, with $\alpha \in X$ and $\beta \in Y$, α and β are* m*-separated given Z (X, Y, and Z are disjoint sets; X, Y are nonempty). This criterion is referred to as a* global Markov property*. We denote the independence model resulting from applying the* m*-separation criterion to G, by $\Im_m(G)$. This is an extension of Pearl's d-separation criterion to mixed graphs in that in a DAG D, a path is d-connecting if and only if it is m-connecting.*

Definition 2. *Let G_A denote the induced subgraph of G on the vertex set A, formed by removing from G all vertices that are not in A, and all edges that do not have both endpoints in A. Two vertices x and y in an*

ADMG G are said to be collider connected if there is a path from x to y in G on which every non-endpoint vertex is a collider; such a path is called a collider path. (Note that a single edge trivially forms a collider path, so if x and y are adjacent in an ADMG then they are collider connected.) The augmented graph *derived from G, denoted $(G)^a$, is an undirected graph with the same vertex set as G such that $c-d$ in $(G)^a \Leftrightarrow$ c and d are collider connected in G.*

Definition 3. *Disjoint sets $X, Y \neq \emptyset$, and Z (Z may be empty) are said to be* m^*-separated *if X and Y are separated by Z in $(G_{\boldsymbol{an}(X\cup Y\cup Z)})^a$. Otherwise X and Y are said to be m^*-connected given Z. The resulting independence model is denoted by $\Im_{m^*}(G)$.*

Richardson in (Richardson, 2003, Theorem 1) shows that for an ADMG G, $\Im_m(G) = \Im_{m^*}(G)$.

The *Markov condition* is said to hold for $G = (V, E)$ and a probability distribution $P(V)$ if $\langle G, P\rangle$ satisfies the following implication: $\forall X, Y \in V, \forall Z \subseteq V \setminus \{X, Y\} : (X \perp\!\!\!\perp_d Y|Z \Longrightarrow X \perp\!\!\!\perp_p Y|Z)$. The *faithfulness condition* states that the only conditional independencies to hold are those specified by the Markov condition, formally: $\forall X, Y \in V, \forall Z \subseteq V \setminus \{X, Y\} : (X \not\perp\!\!\!\perp_d Y|Z \Longrightarrow X \not\perp\!\!\!\perp_p Y|Z)$.

4 Methodology

In this section, first, we prove that the domain adaptation task under the **C**ausal **D**omain **A**daptation (CDA) assumptions can be done effectively via searching inside the set of *neighbors* or *Markov blanket* of the target variable T rather than an inefficient brute-force algorithm `ESS` (Magliacane et al., 2018). Then, we present an *efficient* and *scalable* algorithm, called the `RCTL` Algorithm, that shows how *neighbors* or *Markov blanket* discovery of the target variable T can be used to facilitate the task of domain adaptation. To apply `RCTL` in practice, we need some algorithms for Markov blanket discovery in the presence of unmeasured confounders. For this purpose, we prove that GSMB and IAMB and its variants are still sound under the faithfulness assumption, even when the causal sufficiency assumption does not hold. The proof of theorems can be found in the supplementary material document.

4.1 Problem: Causal Domain Adaptation

Here we formally state the causal domain adaptation task that we address in this work:

For predicting T from a subset of features $S \subseteq V \setminus \{C, T\}$, where T is the target variable, $C = \{c_s, c_t\}$ is the context variable with the value c_s in the source domain(s) and the value c_t in the target domain, and S is a set that d-separates T from C, we define the *transfer bias* as $\hat{T}^t_S - \hat{T}^s_S$, where $\hat{T}^t_S := \mathbb{E}(T|S, C = c_t)$ and $\hat{T}^s_S := \mathbb{E}(T|S, C = c_s)$. Also, we define the *incomplete information bias* as $\hat{T}^t_{V\setminus\{C,T\}} - \hat{T}^t_S$. The *total bias* when using $\hat{T}^s_S$ to predict T is the sum of the transfer bias and the incomplete information bias:

$$\underbrace{\hat{T}^t_{V\setminus\{C,T\}} - \hat{T}^s_S}_{\text{total bias}} = \underbrace{\hat{T}^t_S - \hat{T}^s_S}_{\text{transfer bias}} + \underbrace{\hat{T}^t_{V\setminus\{C,T\}} - \hat{T}^t_S}_{\text{incomplete information bias}}$$

For more details see (Magliacane et al., 2018). Our task here is to find a separating set S that *minimizes* the total bias.

4.2 Assumptions

In this paper we consider the same assumptions as in (Magliacane et al., 2018). Here we briefly introduce them in Assumption 1 and 2:

Assumption 1 (**J**oint **C**ausal **I**nference (JCI) assumptions)**.** *We consider that causal graph $G = (V, E)$ is an ADMG with the variable set V. From now on, we will distinguish system variables $X_{j,j\in J}$ describing the system of interest, and context variables $C_{i,i\in I}$ describing the context in which the system has been observed:*

(a) A context variable is never caused by a system variable i.e., $(\forall j \in J, i \in I : X_j \to C_i \notin G)$,

(b) System variables are not confounded by context variables $(\forall j \in J, i \in I : X_j \leftrightarrow C_i \notin G)$, and

(c) All pairs of context variables are confounded i.e., $(\forall i, k \in I : C_i \leftrightarrow C_k \in G$, and $\forall i, k \in I : C_i \to C_k \notin G)$.

The first assumption is called *exogenity* and captures what we mean by "context". The second and third assumptions are not as important as the exogenity and can be relaxed, depending on the application (Magliacane et al., 2018). To address the causal domain adaptation task, we need the following assumptions as well:

Assumption 2 (**C**ausal **D**omain **A**daptation (CDA) assumptions)**.** *We consider causal graph G that satisfies Assumption 1. We say G satisfies CDA assumptions if*

(a) The probability distribution of V follows Markov condition and faithful assumption w.r.t. G,

(b) For a target variable T that $T \perp\!\!\!\perp A|S$ in the source domain we have the same conditional independency in the target domain, where $T \not\subseteq A \cup S$,

(c) For any context variable C, $C \to T \notin G$.

4.3 Theoretical Results

The following theorem enables us to localize the task of finding domain invariant features under the CDA assumptions.

Theorem 1. *Assume that causal domain adaptation assumptions hold for the context $C_{i,i\in I}$ and system variables $X_{j,j\in J}$ regarding given data for a single or multiple source domains. To find the best separating set(s) of features that d-separate(s) $C_{i,i\in I}$ from the target variable T in the causal graph G and minimize(s) the total bias, it is enough to restrict our search to the set of neighbors $\boldsymbol{ne}(T)$ or Markov blankets $\boldsymbol{Mb}(T)$ of the target variable T.*

Theorem 1 enables us to develop a new *efficient* and *scalable* algorithm, called **R**obust **C**ausal **T**ransfer **L**earning (`RCTL`), that exploits locality for learning *invariant causal features* across environments and can be used for the task of domain adaptation.

Since CDA assumptions does not require *causal sufficiency assumption*, we need sound and scalable algorithms for Markov blanket discovery in the presence of unmeasured confounders. For this purpose, we first need to provide a graphical characterization of Markov blankets in ADMGs.

Let $G = (V, E, P)$ be an ADMG model. Then, V is a set of random variables, (V, E) is an ADMG, and P is a joint probability distribution over V. Let $T \in V$. Then the *Markov blanket* $\mathbf{Mb}(T)$ is the set of all variables that there is a *collider path* between them and T. We now show that the Markov blanket of the target variable T in an ADMG probabilistically shields T from the rest of the variables. Under the faithfulness assumption, the Markov blanket is the smallest set with this property.

Theorem 2. *Let $G = (V, E, P)$ be an ADMG model. Then, $T \perp\!\!\!\perp_p V \setminus \{T, \boldsymbol{Mb}(T)\} | \boldsymbol{Mb}(T)$.*

The following theorem safely enables us to use standard Markov blanket recovery algorithms for domain adaptation task without causal sufficiency assumption:

Theorem 3. *Given the Markov assumption and the faithfulness assumption, a causal system represented by an ADMG, and i.i.d. sampling, in the large sample limit, the Markov blanket recovery algorithms GSMB Margaritis (2003), IAMB Tsamardinos et al. (2003), Fast-IAMB Yaramakala and Margaritis (2005), Interleaved Incremental Association (IIAMB) Tsamardinos et al. (2003), and IAMB-FDR Peña (2008) correctly identify all Markov blankets for each variable. (Note that Causal Sufficiency is not assumed.)*

4.4 `RCTL`: Robust Causal Transfer Learning

Our `RCTL` algorithm addresses the task of domain adaptation by finding a separating set $S \subset V$, where V is the set of context variables $C_{i,i\in I}$ and system variables $X_{j,j\in J}$ such that for the target variable T we have $C_i \perp\!\!\!\perp T | S$, for every $i \in I$ in the source domain. Since Assumption 2(b) implies that this conditional independence holds across domains, if such a separating set S can be found, S is considered as a set of causally *invariant* features for T across environments. Our `RCTL` algorithm, formally described in Algorithm 1, consists of two main steps. First, we find the Markov blanket of the target variable T, i.e., $\mathbf{Mb}(T)$ (line 5 in Algorithm 1). Using the property of Markov blankets, i.e., $T \perp\!\!\!\perp V \setminus \{T, \mathbf{Mb}(T)\} | \mathbf{Mb}(T)$, if there is no context variable in the Markov blanket of the target variable T, then $T \perp\!\!\!\perp C_{i,i\in I} | \mathbf{Mb}(T)$. This means that strongly relevant features, i.e., the Markov blanket of the target variable T, provides a minimal feature set required for prediction of the target variable T with maximum predictivity (Aliferis et al., 2010). However, if there exist a context variable C in $\mathbf{Mb}(T)$, we find the neighbours of the target variable T, i.e., $\mathbf{ne}(T)$ (line 10 in Algorithm 1). This step can be divided into two cases. In the first case, there exist a context variable C in the neighbors of the target variable T, i.e., $C \in \mathbf{ne}(T)$. This means that the data set does not satisfy the CDA assumptions because it violates the Assumption 1(a) or Assumption 2(c). In this case, the algorithm throws a failure because there is no subset of feature that makes T conditionally independent of the context variables. In other words, there is no subset of feature that are causally invariant across domains w.r.t. the target variable T. In the second case, there is no context variable C in the neighbors of the target variable T, i.e., $C \notin \mathbf{ne}(T)$ (line 13-23 in Algorithm 1). In this case, we consider all possible subsets of the neighbors $\mathbf{ne}(T)$ to find those subsets S that satisfy the separating condition $T \perp\!\!\!\perp C_{i,i\in I} | S$. For this purpose, the algorithm filter out those subsets for which p_{value} is below the significance level α, i.e., T is *not* conditionally independent of the context variables given those sets. At the end of line 18, we have a list S that contains all possible separating sets. If S is empty, we abstain from making prediction as the domain adaptation task is not successful. Otherwise, we sort the subsets in S based on the obtained p_{value}s (line 23). Then, the algorithm returns those subsets sub_i in S with the greatest p_{value} as the best possible separating sets, because according to Theorem 1, these sets d-separates T from the context variable(s) and minimize(s) the total bias.

Algorithm 1: RCTL: A Robust Transfer Learning Algorithm for Causal Domain Adaptation

Input: A Dataset with variable set V is the set of context variables $C_{i,i\in I}$ and system variables $X_{j,j\in J}$, target variable T, significance level α;
Output: A separating set S;
Let Set S be an *empty list*;
// Step 1: Find $\mathbf{Mb}(T)$, Markov blanket of T.
Set $S_1 = \mathbf{Mb}(T)$;
if $S_1 \cap C_{i,i\in I} = \emptyset$ **then**
  return S_1;
else
  // Step 2: Find $\mathbf{ne}(T)$, neighbours of T.
  Set $S_2 = \mathbf{ne}(T)$;
  if $S_2 \cap C_{i,i\in I} = \emptyset$ **then**
    Set Subs = **Subset**(S_2) ; /* All possible combinations of subsets for given set S_2 */
    for *each* $sub_i \in Subs$ **do**
      $p_{val} = p_{value}(C_{i,i\in I} \perp\!\!\!\perp T | sub_i)$;
      if $p_{val} > \alpha$ **then**
        // This means T is conditionally independent of the context variables given the set sub_i.
        Add sub_i to S;
      end
    end
    if $S = \emptyset$ **then**
      return null;
    end
    Sort(S) ; /* Sorts the list in descending order S based on p_{value} */
    return *subsets from* S *with highest* p_{value};
  else
    return null;
  end
end

5 Experimental Evaluation

We empirically validated the robustness and scalability of RCTL on various synthetic and real world datasets. We contrast RCTL against various other feature selection approaches including Conditional Mutual Information Maximization (CMIM) (Fleuret, 2004), Mutual Information Maximization (MIM) (Lewis, 1992), Adaboost feature selection (Adast) (Freund and Schapire, 1999), Greedy Subset Search (GSS) (Rojas-Carulla et al., 2018b), C4.5 using Decision Trees (C4.5 + DTR) (Quinlan, 1986) and Exhaustive Subset Search (ESS) (Magliacane et al., 2018). These algorithms were coupled individually with machine learning based regressors like, Random Forest Regressor (RFR) (Breiman, 2001) , kNN Regressor (kNNR) (Yao and Ruzzo, 2006) and Support Vector Regressor (SVR) (Drucker et al., 1997) for generating predictions on their respective selected feature sets. The data generation processes, conditional independence tests and Markov blanket Algorithms were implemented in *R* by extending the *bnlearn* (Scutari, 2017) and *pclag* (Kalisch et al., 2012) packages. We used Python 3.6 with *scikit-learn* (Pedregosa et al., 2011) library for implementing the above mentioned feature selection and machine learning algorithms to compare our approach against. The library is also used for prediction over subset of features selected by RCTL using *RandomForestRegressor* function. The generated predictions from each of the algorithms are compared to the actual results for finding the error and all experiments have been reported using various comparison metrics.

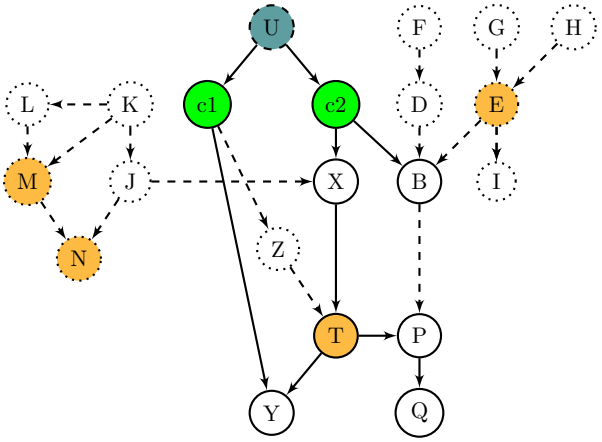


Figure 2: Ground Truth Synthetic Graph: In this figure, the grey node U represents the unknown confounder connecting the context variables filled in green C_1 & C_2. The nodes in orange like E, T, represent possible target variables. The central structure represented by solid lines is common to all graphs. The dashed lines and nodes are used to introduce variability in the structure across source and target in the synthetic experiments as explained in the setting.

5.1 Experimental Settings

In this section, we mention the different criteria used for data generation, parameter-tuning and validation. Based on changes in parameters and user dependent variables, we divide this section into three key parts. In each sub-section we describe the combinations of hyperparameters used, why we used them and how they were implemented in practice.

5.1.1 Synthetic Data Generation

Let us consider an ADMG G as shown in Figure 2. The basic synthetic dataset based on G consists of 19 randomly generated variables, with 16 system variables, 2 context variables (C_1 & C_2) and 1 unknown confounder (U) (to be removed while using data). A dataset is generated such that it is faithful w.r.t. G. The data was generated in Gaussian as well as Discrete distributions. For Gaussian setting, a model graph as shown in Figure 2 was generated using *model2network* function from the *graphviz* package. In Figure 2 the dashed lines and circles annotate the supplementary additions to the central structure which is represented by solid lines. The obtained DAG is passed to the *cus-*

tom.fit function from *bnlearn* which defines the mean, variance and relationship (edges values) between the nodes. This process ensures that the generated dataset will be faithful w.r.t G. Once the graph is set and all variables have been properly defined, we use the *rbn* function from *bnlearn* to generate required number of samples. The dataset now obtained will have the desired number of samples and is faithful w.r.t G, this dataset (not the graph structure) is used further for our experiments. For the Discrete setting, the same process is followed, but instead of mean, variance and edge values, we use a matrix of random probabilities. For each node, the size of the matrix would be the number of discrete values for the node, plus the number of discrete values for each of its connecting nodes multiplied by the number of modes connected to it. We generated multiple datasets using different configurations following the above mentioned process, to simulate the behaviour across different environment changes. These configurations were carefully generated to exploit characteristics of real-world scenarios. These changes in configurations were random, but followed a pattern such that they can mimic almost all possible environments, to ensure adaptability of approach across domains. The perturbations in configurations can be broadly classified into:

(1) **Changes to the sample size:** We generated distributions with high disparity in the "amount" of data. The generated datasets contained 50, 1000 and 10000 rows of data. This would help us to see the difference in accuracy and robustness of the feature selection algorithms under extreme data-sensitive conditions (i.e. low sample size, moderate size, big data).

(2) **Changes to the network size:** We change the size of the graph, by increasing or dropping the number of nodes to monitor change in performance, ensure scalability of approach to changes in the number of environment variable. For this, we consider 3 different network sizes 20, 12 and 8 nodes. As shown in Figure 2, the central structure made of solid lines consists of 8 nodes which additional dotted nodes were added to increase size of graph. However, to allow a level of experimental flexibility we do not rigidly conform to limiting our experiments to the above mentioned number of nodes. During our experimentation we tried multiple combinations of edges and nodes and have added results based on them to this extended version of the paper. To remove ambiguity on the ground truth structures used to obtain the results in Figure **??**, in this extended version we also provide the ground truth graphs for the synthetic dataset alongside the results obtained on it using various approaches. We also experiment by making changes to and on the number of context variables. This was done by using variables that are either partly, or completely unaffected by them (For example, in Figure 2, we can use the whole or parts of graph with nodes M, N, E as target variables).

(3) **Changes to the complexity:** The difference in complexity of the dataset can be further sub-divided into *structural* change i.e. changing the edge connections or relations among the nodes (like in Figure 2, we can add Z to induce direct affect of C_1 on target, or connect B and P, to affect the Markov blanket of target), and *domain-distribution* change i.e. changes between the source and target domain. Since we consider two context variables, *domain-distribution* change can be brought about by changes in either one or both. These controlled changes we made across source and target would enable us to evaluate the robustness and check for domain adaptability of features in various approaches we compared. *Domain-distribution* change is the only setting which is different in source and target as it helps us to measure domain independence, the rest of the above mentioned setting changes are base settings i.e. for a given experiment they do not change between the source and target. The changes to the domain can be classified further into three sub-settings: smooth (very little change to context variable), mild (small change in context variables) and severe (drastic change in context variables between source and target). The change in context variables between source and target were practically brought about by changing the mean and variance for Gaussian distributions and changing the probabilities associated with the variables in discrete distributions. We have experimented on all possible scenarios using complex and random combinations of the above mentioned settings to generate all possible scenarios and summarized our results in Section 6. In practice, for example, for a central structure consist of 8 nodes as shown in Figure 2 having 10000 samples with severe changes in context, we train by generating one dataset on the base setting. Then make severe changes to the domain-distribution (i.e. by drastically changing the values of the context variables) and then generate a new dataset which is used for testing. To ensure consistency in our results across all our experiments, for a given complex setting we generate 8 test datasets, by making slight or almost unnoticeable changes to any one of the system variables at random. The slight change is made so that the machine learning algorithms used for prediction over the selected feature set do not report the same error each time, instead an error range which serves as a better depiction of the robustness (error range will be small for more robust approaches as they learn invariant features across environments).

5.1.2 Real-world Dataset

We used the diabetes dataset (Islam et al., 2020) that contains reports of common diabetic symptoms of 520 persons. This includes data about symptoms that may cause or are potentially caused by diabetes. The dataset has been created from a direct questionnaire to people who have either recently been diagnosed as diabetic, or who are still non-diabetic but having show few or more symptoms of diabetes. The diabetes dataset contains a total of 17 features, with 320 cases of diabetes Type II positive and 200 of diabetes negative patients. The data has been collected from the patients of the Sylhet Diabetes Hospital, Sylhet, Bangladesh. In the dataset we consider two context variables Age and Gender. We perform 3 sets of data shifts between training and testing domains using these context variables: (1) *Gender Shift*: We train on the Male and test on Female. (2) *Age Shift*: We split the data using an arbitrary threshold (we choose 50 as threshold for our experimentation as it gives a fair distribution of samples between training and testing data), train on young (less than 50) and test on the older (greater than 50). (3) *Double context shift*: We do a dataset shift for both context variables, in our case, we train on young male patients and test on old female patients. In practice, while experimenting on a context shift setting we train by sampling 100 samples from the source domain keeping the samples in target domain constant. This allows us to generates conditions similar to having multiple source domains and single target domain. Such conditions will help further test the robustness of the feature selection algorithm.

5.1.3 Implementation Parameters

`RCTL` requires the use of Markov-blanket and neighbourhood for feature selection. We use multiple Markov blanket discovery-algorithms to evaluate the effect the algorithm-choice can have on final prediction. For this purpose, we select, GS, IAMB, interIAMB, fastIAMB, , fdrIAMB, MMMB, SI.HINTON.MB. These algorithms were implemented in *R* using the *bnlearn* package. The variable *subs* mentioned on line 12 of Algorithm 1 contains all possible combination of subsets from the feature set S_2 . To find all possible subsets of a given feature set we use Python 3.5, one or more of these subsets will act as the separating set. To find the separating set, we sort the subsets based on p_{value} of conditional independence tests and choosing the subset/s with highest score. We use a significance level α of 0.05 as threshold for the conditional independence tests. The tests for different configurations are based on the type of data being used. For Gaussian data, we used *gaussCItest* from the *pclag* package, which uses Fisher's z-transformation of the partial correlation, for testing correlation for sets of normally distributed random variables. For discrete data, we use *mi* (mutual information - an information-theoretic distance measure. It is somewhat proportional to the log-likelihood ratio) test from the *bnlearn* package.

5.2 Evaluation Metrics

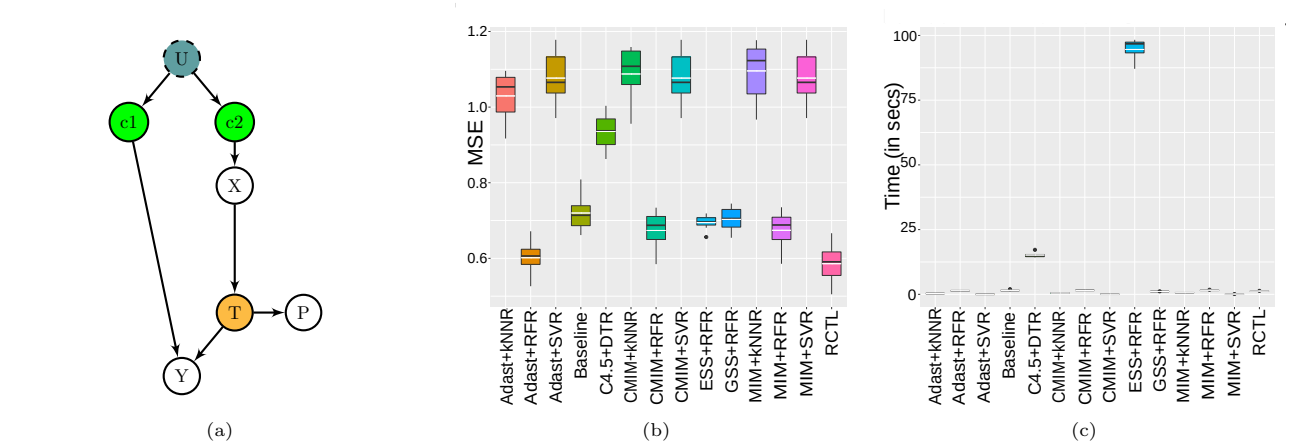
For evaluation of `RCTL`'s feature selection method we used a standard Machine leanring technique (Random Forest Regression). This was built on top of `RCTL` for evaluating the selected *domain-invariant* separating set by training on source and prediction on target domain. We used the RandomForestRegressor function from *scikit-learn* package, using the out-of-bag (OOB) scoring approach on the default parameters, for this process. Without any feature selection, directly using the available feature set, and feeding it to the Random Forest Regressor is considered as *Baseline* approach for this paper. We also tested `RCTL` on other regression techniques like Lasso, Polynomial, Decision Tree and Support Vector Regression, in order to mitigate doubts about approach specific biases. In all the ML techniques, we found the same error more or less, but we chose Random Forest Regressor as it not only performed well, but also, gave consistent results in across all settings. Our evaluation is done on two commonly used comparison metrics for evaluating model performance: (a) Mean-Square-Error (MSE), (b) Sum-of-Square-Errors (SSE). In principle, low values of MSE and SSE indicate better performance. We also report time taken in seconds as a metric for our approach which shows the improvement in scalability over other approaches.

6 Experimental Results

Our experimental results over various combinations of environment settings as discussed in Section 5.1, have been shown in Figure **??**, the remaining experimentation has been added to the supplementary material.

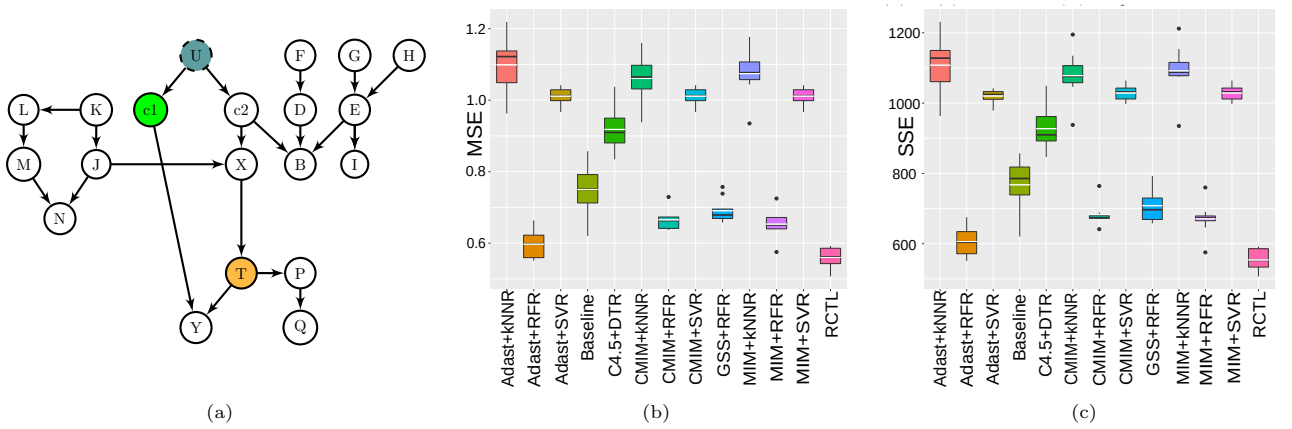
6.1 Synthetic Dataset

`RCTL` outperforms (in some cases a comparable performance) other approaches in all environment settings we consider on the synthetic dataset. In Figure **??**, we report the most complicated scenarios involving severe domain shifts, extreme sample sizes and multiple ground truth models, the rest of the experimentation can be found in the supplementary material. As shown in our experimental results, `RCTL` is, overall, as good as or even better than the state-of the-art algorithms.



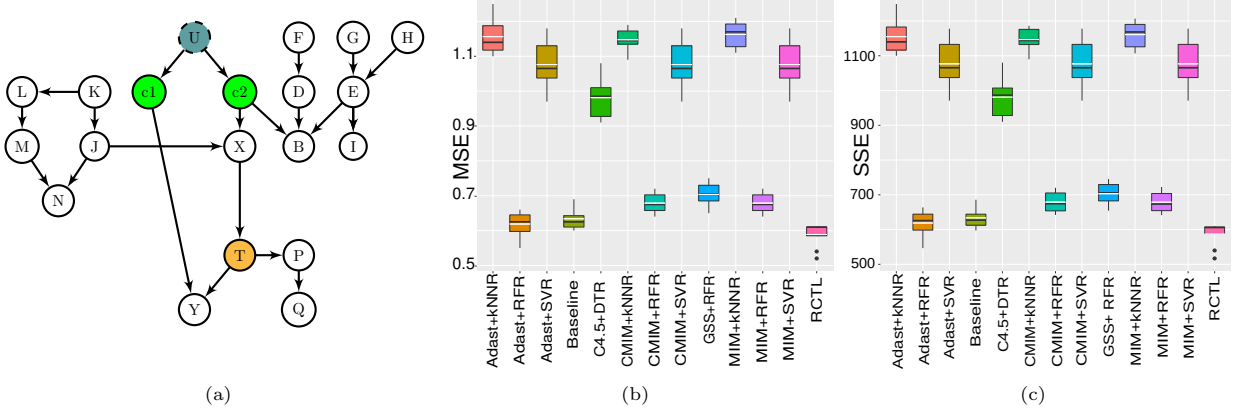
(a) (b) (c)

Figure 3: Results on (b), (c) are based on dataset generated from the ground truth graph (a). Data shift between the source and the target occurs by change in distribution of C_1& C_2 across domains for prediction on the target variable T, where sample size = 1000 for a discrete distribution.



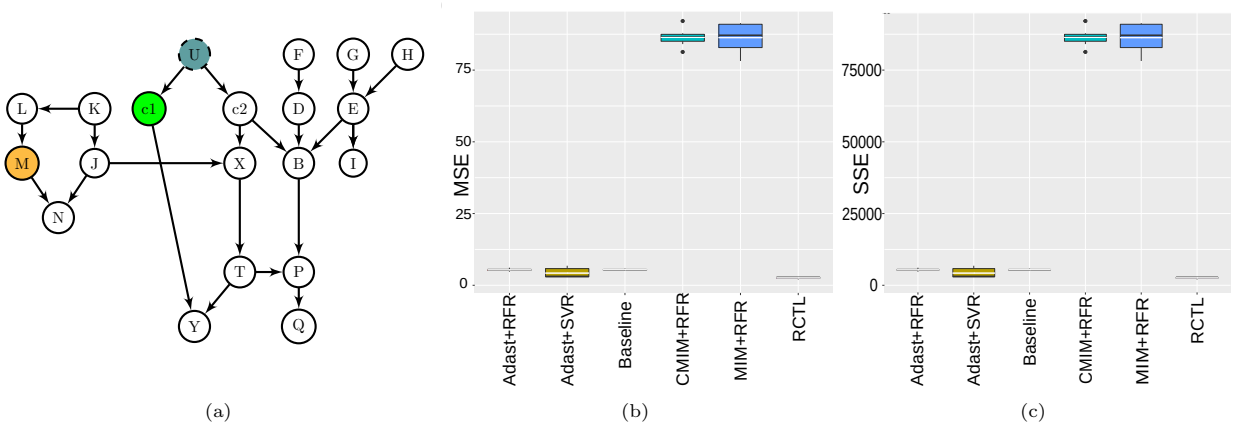
(a) (b) (c)

Figure 4: Results on (b), (c) are based on dataset generated from the ground truth graph (a). Data shift between the source and the target occurs by change in distribution of C_1 across domains for prediction on the target variable T, where sample size = 1000 for a discrete distribution.



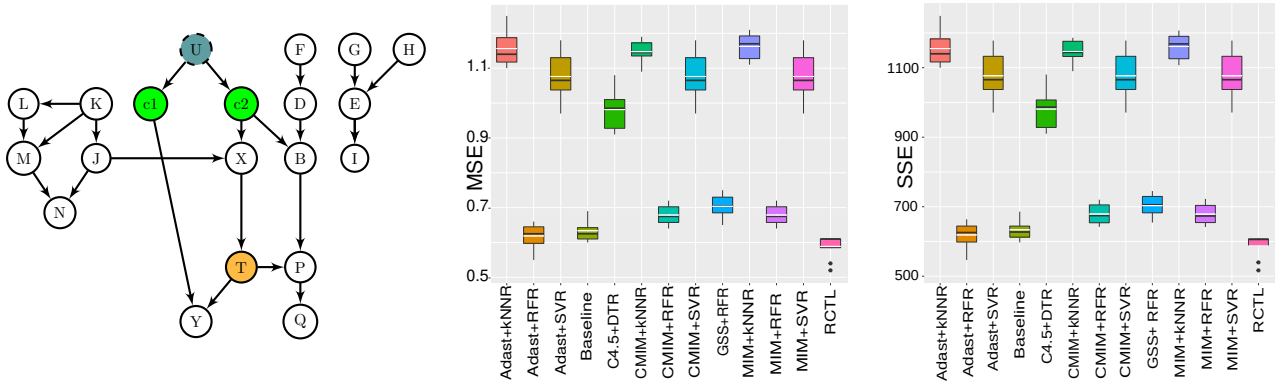
(a) (b) (c)

Figure 5: Results on (b),(c) are based on dataset generated from the ground truth graph (a). Data shift between the source and the target occurs by change in distribution of C_1& C_2 across domains for prediction on the target variable T, where sample size = 1000 for a discrete distribution.



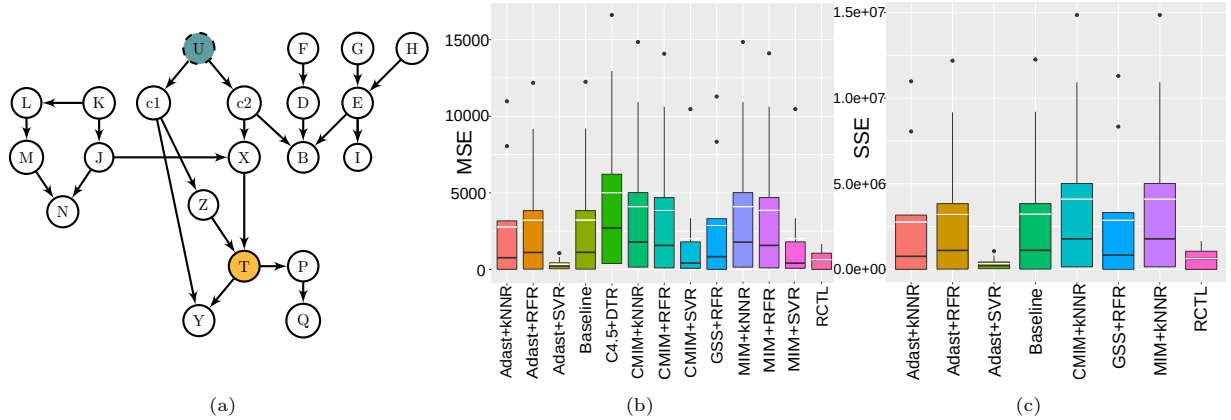
(a) (b) (c)

Figure 6: Results on (b), (c) are based on dataset generated from the ground truth graph (a). Data shift between the source and the target occurs by change in distribution of C_1 across domains for prediction on the target variable M, where sample size = 1000 for a discrete distribution.



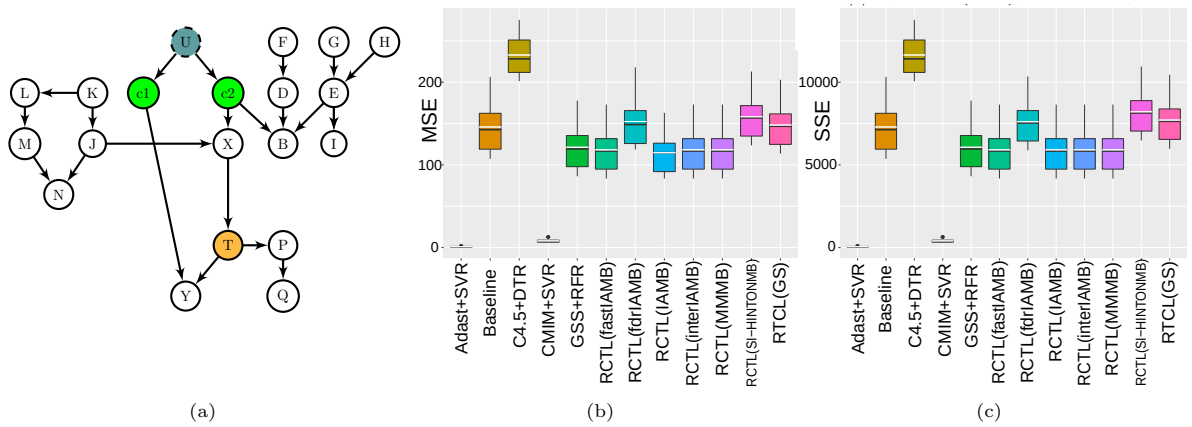
(a) (b) (c)

Figure 7: Results on (b), (c) are based on dataset generated from the ground truth graph (a). Data shift between the source and the target occurs by change in distribution of C_1& C_2 across domains for prediction on the target variable T, where sample size = 1000 for a discrete distribution.



(a) (b) (c)

Figure 8: Results on (b), (c) are based on dataset generated from the ground truth graph (a). No Data shift between the source and the target occurs across domains, prediction on the target variable T, where sample size = 1000 for a Gaussian distribution.



(a) (b) (c)

Figure 9: Results on (b), (c) are based on dataset generated from the ground truth graph (a). Data shift between the source and the target occurs by change in distribution of C_1& C_2 across domains for prediction on the target variable T, where sample size = 50 for a Gaussian distribution.

Based on our findings, we summarize some significant highlights.

Key highlights of the experimentation: (1) *Scalability comparison between* `ESS` *(Magliacane et al., 2018) and* `RCTL`: `ESS`, the closest work similar to `RCTL`, employs conditional independence test based feature selection for domain adaptation. However, it uses exhaustive feature search over the entire feature set, which increases the time required for subset generation exponentially and makes it virtually impossible for the use on datasets with roughly more than 10 variables. We used `ESS` for computation on a 12 variable synthetic dataset for 72 hours, it crashed without generating any results. `RCTL` drastically reduces time required for subset search as it searches on only those variables that are in the locality of the target variable. Since Markov blanket algorithms make use of local computation, by computing the Markov blanket of the multiple variables in parallel we can reduce inference time even further. This makes `RCTL` scalable to high dimensional datasets and have potential applications in Big Data. (2) *Robustness to Conditional Independence Tests*: In cases where conditional independence tests have to be estimated from data, mistakes occur in keeping or removing members from the estimated separating sets. The main source of erroneous CI tests are large condition sets given a limited sample size and the curse-of-dimensionality (Cheng, Bell, and Liu, 1997). In such cases, the resulting changes in the CI tests can lead to different separating sets. However, `RCTL` that is based on Markov blanket discovery only uses a small fraction of the dataset included in the vicinity of the target, increasing the reliance on conditional independence tests, making the algorithm robust. Our experimentation has also shown that for most environment settings, where we know the ground truth causal graph, `RCTL` is able to find the correct causally invariant features. (3) *Robustness on multiple environment settings*: We found that many feature selection algorithms like, Adaboost, CMIM despite of having lower rates of MSE, SSE on Gaussian data, show a high variation in error for discrete settings. A possible explanation for this might be that, some of these algorithms do not generalize well to all types of data distributions. For example, Adaboost has a completely separate implementation in Python for discrete settings, however, one can apply `RCTL` to both types of distributions without making any modification to the underlying algorithm. This is because Markov blanket algorithms detect the type of data distributions early on, and have separate tests for each (like *mi*-test for discrete setting), hence the feature set learned is robust and unbiased to the type of data distribution being used. This makes `RCTL` robust and usable on Gaussian, Discrete as well as hybrid data distributions across environments (refer to Figure 7, 8). (4) *Quantitative analysis of results on Gaussian settings*: Although we report only the most interesting scenarios in Figures 3-9, we notice that, for Gaussian settings (see Figure 6 (b), (c)), the error rates of many of these feature selection algorithms seem lower than `RCTL`. For example, in Figure 7 (c) it seems that Adaboost has a slightly lower MSE than `RCTL` for the specified setting. For all such scenarios, we perform t-tests and F-tests to confirm our observation from the results. In these tests, we found that for all such cases the difference in error rates is almost insignificant. This shows that `RCTL` gives comparable if not lower MSE, SSE to other feature selection algorithms (see supplementary material). (5) *Increase in Error on small sample size*: We do see a slight increase in the error on small sample size setting for `RCTL` (see Figure 9 (b), (c)). Further in-

vestigation reveals that, since, we do not utilize the underlying ground truth graph structure, the lack of data is straining the Markov blanket Algorithms, causing them to sometimes incorrectly learn the Markov blanket and neighbourhood of the target variable. For example, for the setting in Figure 9 (b) using Figure 9 (a) as the ground truth, we found that the Markov blanket of the target variable T using IAMB is P, X, C_1, G. In this case the Markov blanket is wrong, but the conditioning set learned was satisfactory (P, X) and causally invariant. But still, we found that the predictions are not accurate. This is also because, since the sample size is already so small, and we are further reducing the number of variables being used for prediction (down to 2 in this case) it significantly reduces size of training data. The reduction in the dataset size would make us miss out on relevant information which causes *incomplete information bias* as explained in section 4.1. This would strain the machine learning algorithm being used for prediction, as it cannot learn the trend for such a small dataset. An increase in the sample size though, not only improves the accuracy of the predictions by Markov blanket algorithms but also, reduces the error on machine learning algorithms used for prediction over selected feature set as the amount of information increases. (6) *Choice of Markov Blanket Algorithms*: The Markov blanket algorithm being used also holds key significance and their practical uses may show results different from the theoretical perception. For example, GS does not consider the ordering and the strength of the association of candidate variable and the target variable T. On the other hand, IAMB first orders variables based on the strength of their association with the target variable T and then check their membership in the $\mathbf{Mb}(T)$. Since choosing different p-values, sample size of the data, and the maximum size of the conditioning sets have a big effect of the quality of learned $\mathbf{Mb}(T)$, it is necessary to choose a robust and effective Markov blanket approach depending on given setting. We found in our algorithms that fdrIAMB to be the best in small sample size settings, which is consistent with previously reported results (Peña, 2008). We show in Figure 9 (b), (c), the effect of different Markov blanket approaches on the error rates.

6.2 Real-world data: Diabetes Dataset

Since we do not have the ground truth causal graph for the real world diabetes dataset (Islam et al., 2020), we assume that it follows Assumption 1 & 2. We consider *Age* as context variable, because previous studies (Kirkman et al., 2012) have shown that older people are at higher risk for the development of Type II diabetes. Firstly, although this is a dataset with only 16 variables, we could not compute the feature set

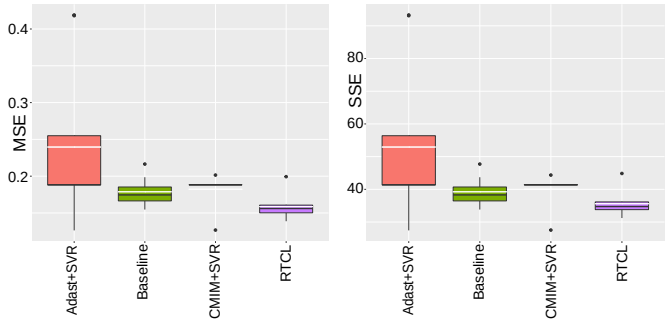


Figure 10: Error rate comparison of multiple feature selection approaches on a simulation of the Diabetes dataset with *Age Shift* (discussed in Section 5.1.2).

for `ESS` (Magliacane et al., 2018). Similar to synthetic dataset, we ran it on the computer for 72 hours after which it was crashed, confirming that it is not a scalable method. The results in Figure 10 show that `RCTL` uses fewer number of features for prediction and it provides the lowest error compared with other approaches. Also, we noticed that the total sample size for both training and testing scenarios had 200 entries for each. This signifies that even for a moderate number of samples, the Markov blanket algorithms can learn the vicinity of the target well enough for giving low error rates in the target environment.

Table 1: Results on Real World Dataset (Diabetes)

Methodology	SSE	MSE	Time (in s)
Adaboost+SVR	38.99	0.173	**0.156**
CMIM+RFR	42.59	0.189	0.352
Baseline	45.34	0.225	0.331
RCTL	**37.42**	**0.166**	0.593

7 Conclusion

In this paper we address the problems regarding *scalability* and *robustness* of the causal domain adaptation approach (`ESS`) proposed in (Magliacane et al., 2018). To overcome these difficulties, we proposed an algorithm based on Markov blanket discovery that resolves scalability and robustness of `ESS` due to the locality nature of Markov blankets. One interesting direction for future work is relaxing restrictive assumptions such as faithfulness, and designing new scalable and robust causal domain adaptation algorithms based on weaker assumptions.

References

Aliferis, C. F.; Statnikov, A.; Tsamardinos, I.; Mani, S.; and Koutsoukos, X. D. 2010. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of Machine Learning Research* 11(7):171–234.

Association, A. D. 2020. Statistics about diabetes. `https://www.diabetes.org/`.

Australia, D. 2020. Diabetes in australia. `https://www.diabetesaustralia.com.au`.

Bareinboim, E., and Pearl, J. 2012. Transportability of causal effects: Completeness results. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, 698–704. Toronto, Ontario, Canada: AAAI Press.

Bareinboim, E., and Pearl, J. 2014. Transportability from multiple environments with limited experiments: Completeness results. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc. 280–288.

Breiman, L. 2001. Random forests. *Machine learning* 45(1):5–32.

Chen, X.; Monfort, M.; Liu, A.; and Ziebart, B. D. 2016. Robust covariate shift regression. volume 51 of *Proceedings of Machine Learning Research*, 1270–1279. Cadiz, Spain: PMLR.

Cheng, J.; Bell, D. A.; and Liu, W. 1997. Learning belief networks from data: An information theory based approach. In *Proceedings of the 6th CIKM*, 325–331.

Csurka, G. 2017. *Domain adaptation in computer vision applications*, volume 8. Springer.

Drucker, H.; Burges, C. J.; Kaufman, L.; Smola, A. J.; and Vapnik, V. 1997. Support vector regression machines. In *Advances in neural information processing systems*, 155–161.

Fleuret, F. 2004. Fast binary feature selection with conditional mutual information. *Journal of Machine learning research* 5(Nov):1531–1555.

Freund, Y., and Schapire, R. 1999. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence* 14(771-780):1612.

Gao, T., and Ji, Q. 2016. Constrained local latent variable discovery. In *Proceedings of IJCAI'16*, 1490–1496.

Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of causal discovery methods based on graphical models. *Frontiers in Genetics* 10:524.

Gong, M.; Zhang, K.; Liu, T.; Tao, D.; Glymour, C.; and Schölkopf, B. 2016. Domain adaptation with conditional transferable components. In *International conference on machine learning*, 2839–2848.

Gretton, A.; Smola, A.; Huang, J.; Schmittfull, M.; Borgwardt, K.; and Schölkopf, B. 2009. *Covariate shift and local learning by distribution matching*. Cambridge, MA, USA: MIT Press. 131–160.

Islam, M. F.; Ferdousi, R.; Rahman, S.; and Bushra, H. Y. 2020. Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis*. Springer. 113–125.

Javidian, M. A.; P. Jamshidi; and Valtorta, M. 2020. Learning LWF chain graphs: A Markov blanket discovery approach. In *Proceedings of the Uncertainty in Artificial Intelligence (UAI'20)*.

Kalisch, M.; Mächler, M.; Colombo, D.; Maathuis, M.; and Bühlmann, P. 2012. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software, Articles* 47(11):1–26.

Kirkman, M. S.; Briscoe, V. J.; Clark, N.; Florez, H.; Haas, L. B.; Halter, J. B.; Huang, E. S.; Korytkowski, M. T.; Munshi, M. N.; Odegard, P. S.; et al. 2012. Diabetes in older adults. *Diabetes care* 35(12):2650–2664.

Kisamori, K.; Kanagawa, M.; and Yamazaki, K. 2020. Simulator calibration under covariate shift with kernels. volume 108 of *Proceedings of Machine Learning Research*, 1244–1253. Online: PMLR.

Kouw, W. M., and Loog, M. 2019. A review of domain adaptation without target labels. *IEEE transactions on pattern analysis and machine intelligence*.

Lauritzen, S. 1996. *Graphical Models*. Oxford Science Publications.

Lewis, D. D. 1992. Feature selection and feature extraction for text categorization. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Li, Y.; Murias, M.; Major, S.; Dawson, G.; and Carlson, D. 2019. On target shift in adversarial domain adaptation. volume 89 of *Proceedings of Machine Learning Research*, 616–625. PMLR.

Li, F.; Lam, H.; and Prusty, S. 2020. Robust importance weighting for covariate shift. volume 108 of *Proceedings of Machine Learning Research*, 352–362. Online: PMLR.

Ling, Z.; Yu, K.; Wang, H.; Liu, L.; Ding, W.; and Wu, X. 2019. Bamb: A balanced Markov blanket discovery approach to feature selection. *ACM Trans. Intell. Syst. Technol.* 10(5).

Lipton, Z. C.; Wang, Y.; and Smola, A. J. 2018. Detecting and correcting for label shift with black box predictors. In Dy, J. G., and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, 3128–3136. PMLR.

Liu, X., and Liu, X. 2016. Swamping and masking in Markov boundary discovery. *Machine Learning* 104(1):25–54.

Magliacane, S.; van Ommen, T.; Claassen, T.; Bongers, S.; Versteeg, P.; and Mooij, J. M. 2018. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 10869–10879. Red Hook, NY, USA: Curran Associates Inc.

Margaritis, D., and Thrun, S. 1999. Bayesian network induction via local neighborhoods. In *Proceedings of the NIPS'99*, 505–511.

Margaritis, D. 2003. *Learning Bayesian Network Model Structure from Data*. Ph.D. Dissertation, Carnegie-Mellon University.

Moreno-Torres, J. G.; Raeder, T.; Alaiz-RodríGuez, R.; Chawla, N. V.; and Herrera, F. 2012. A unifying view on dataset shift in classification. *Pattern recognition* 45(1):521–530.

Park, S.; Bastani, O.; Weimer, J.; and Lee, I. 2020. Calibrated prediction with covariate shift via unsupervised domain adaptation. volume 108 of *Proceedings of Machine Learning Research*, 3219–3229. Online: PMLR.

Pearl, J., and Bareinboim, E. 2011. Transportability of causal and statistical relations: A formal approach. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 247–254.

Pearl, J. 2009. *Causality. Models, reasoning, and inference*. Cambridge University Press.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research* 12:2825–2830.

Peña, J. M. 2007. Towards scalable and data efficient learning of Markov boundaries. *International Journal of Approximate Reasoning* 45(2):211 – 232.

Peña, J. M. 2008. Learning Gaussian graphical models of gene networks with false discovery rate control. In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, 165–176.

Quinlan, J. R. 1986. Induction of decision trees. *Machine learning* 1(1):81–106.

Redko, I.; Courty, N.; Flamary, R.; and Tuia, D. 2019. Optimal transport for multi-source domain adaptation under target shift. volume 89 of *Proceedings of Machine Learning Research*, 849–858. PMLR.

Richardson, T. 2003. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics* 30(1):145–157.

Rojas-Carulla, M.; Schölkopf, B.; Turner, R.; and Peters, J. 2018a. Invariant models for causal transfer learning. *The Journal of Machine Learning Research* 19(1):1309–1342.

Rojas-Carulla, M.; Schölkopf, B.; Turner, R.; and Peters, J. 2018b. Invariant models for causal transfer learning.

Sadeghi, K. 2017. Faithfulness of probability distributions and graphs. *The Journal of Machine Learning Research* 18(1):5429–5457.

Schoelkopf, B.; Janzing, D.; Peters, J.; Sgouritsa, E.; Zhang, K.; and Mooij, J. 2012. On causal and anticausal learning. In Langford, J., and Pineau, J., eds., *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, ICML '12, 1255–1262. New York, NY, USA: Omnipress.

Scutari, M. 2017. Bayesian network constraint-based structure learning algorithms: Parallel and optimized implementations in the bnlearn R package. *Journal of Statistical Software, Articles* 77(2):1–20.

Shimodaira, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference* 90(2):227 – 244.

Spirtes, P.; Glymour, C.; and Scheines, R. 2001. *Causation, Prediction and Search, second ed.* Cambridge, MA.: MIT Press.

Statnikov, A.; Lemeir, J.; and Aliferis, C. F. 2013. Algorithms for discovery of multiple Markov boundaries. *The Journal of Machine Learning Research* 14(1):499–566.

Stojanov, P.; Gong, M.; Carbonell, J.; and Zhang, K. 2019a. Data-driven approach to multiple-source domain adaptation. volume 89 of *Proceedings of Machine Learning Research*, 3487–3496. PMLR.

Stojanov, P.; Gong, M.; Carbonell, J.; and Zhang, K. 2019b. Low-dimensional density ratio estimation for covariate shift correction. volume 89 of *Proceedings of Machine Learning Research*, 3449–3458. PMLR.

Storkey, A. J. 2009. *When Training and Test Sets Are Different: Characterizing Learning Transfer*. MIT Press. 3–28.

Subbaswamy, A., and Saria, S. 2018. Counterfactual normalization: Proactively addressing dataset shift using causal mechanisms. In Globerson, A., and Silva, R., eds., *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, 947–957. AUAI Press.

Subbaswamy, A.; Schulam, P.; and Saria, S. 2019. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 3118–3127. PMLR.

Sugiyama, M.; Suzuki, T.; Nakajima, S.; Kashima, H.; von Bünau, P.; and Kawanabe, M. 2008. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics* 60(4):699–746.

Tsamardinos, I.; Aliferis, C.; Statnikov, A.; and Statnikov, E. 2003. Algorithms for large scale Markov blanket discovery. In *In The 16th International FLAIRS Conference, St*, 376–380. AAAI Press.

Tsamardinos, I.; ; Brown, L. E.; and Aliferis, C. F. 2006. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning* 65(1).

von Kügelgen, J.; Mey, A.; and Loog, M. 2019. Semi-generative modelling: Covariate-shift adaptation with cause and effect features. volume 89 of *Proceedings of Machine Learning Research*, 1361–1369. PMLR.

Yao, Z., and Ruzzo, W. L. 2006. A regression-based k nearest neighbor algorithm for gene function prediction from heterogeneous data. In *BMC bioinformatics*, volume 7, 1–11. BioMed Central.

Yaramakala, S., and Margaritis, D. 2005. Speculative Markov blanket discovery for optimal feature selection. In *Proceedings of the ICDM'05*.

Yu, K.; Liu, L.; Li, J.; and Chen, H. 2018. Mining Markov blankets without causal sufficiency. *IEEE Transactions on Neural Networks and Learning Systems* 29(12):6333–6347.

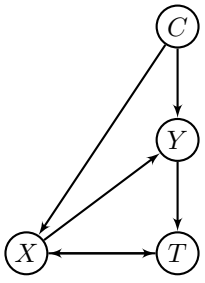
Zhang, K.; Schölkopf, B.; Muandet, K.; and Wang, Z. 2013. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, 819–827.

Zhang, K.; Gong, M.; and Scholkopf, B. 2015. Multi-source domain adaptation: A causal view. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, 3150–3157. AAAI Press.

Appendix A. Proofs of Theoretical Results

Proof of Theorem 1. Let G be a causal graph with variable set V consisting of system variables $X_{j,j\in J}$ and context variables $C_{i,i\in I}$. Assume that $C \in C_{i,i\in I}$, T is the target variable, $\mathbf{Mb}(T)$ is the Markov blanket of T, and $\mathbf{ne}(T)$ is the set of neighbors of T. We have two cases:

- $C \notin \mathbf{Mb}(T)$: In this case, $C \perp\!\!\!\perp T | \mathbf{Mb}(T)$ due to the definition of Markov blanket.
- $C \in \mathbf{Mb}(T)$. In this case we show that considering the CDA assumptions, every path between C and T is blocked by a subset of $\mathbf{ne}(T)$ or there is no subset of variables that d-separates C from T. We have the following subcases:
  (a) The path between C and T is of the form $C \cdots X \to T$. $X \in \mathbf{pa}(T)$ and this path is blocked by X.
  (b) The path between C and T is of the form $C \cdots X \leftarrow T$. Due to the CDA assumptions $C \neq T$. Using the JCI assumptions 1 (a) and (b), $X \not\to C$ and $X \not\leftrightarrow T$. So, there is a collider between X and T, say Y, that $Y \notin \mathbf{pa}(T)$ due to the acyclicity of the causal graph. This means that this path is blocked at the collider Y.
  (c) The path between C and T is of the form $C \cdots X \leftrightarrow T$ and $Y \leftarrow X \leftrightarrow T$ is part of the path such that $Y \in \mathbf{pa}(T)$. In this case, $X \in \mathbf{ne}(T)$ blocks the path.
  (d) The path between C and T is of the form $C \cdots Y \to X \leftrightarrow T$ or $C \cdots Y \leftrightarrow X \leftrightarrow T$. In this case there is a collider at X that blocks the path.
  (e) It is quite possible that there exist a situation that we have paths of the form (c) and (d) simultaneously. In this case, there is no subset of variables that d-separates C from T. The following simple example illustrates such situations:



Note that in all cases the Markov condition and faithful assumptions guarantee the correctness of independence relationships. As we shown in all cases under the CDA assumptions, to find a separating set of features that d-separates C from the target variable T in the causal graph G, it is enough to restrict our search to the set of neighbors $\mathbf{ne}(T)$ or Markov blankets $\mathbf{Mb}(T)$ of the target variable T. In the case that $C_{i,i\in I}$ has more than one element, similar argument can be used to prove the theorem.

Using any subset S for prediction that satisfies the d-separating set property, implies zero transfer bias. So, the best predictions are then obtained by selecting a separating subset that also minimizes the source domains risk (i.e., minimizes the incomplete information bias). □

Correctness of Algorithm 1. The correctness of Algorithm 1 follows from Theorem 1. □

Proof of Theorem 2. It is enough to show that for any $A \in V \setminus \{T, \mathbf{Mb}(T)\}$, $T \perp\!\!\!\perp_d A | \mathbf{Mb}(T)$. For this purpose, we prove that any path between A and T in G is blocked by $\mathbf{Mb}(T)$. In the following cases ($A \multimap B$, where means $A - B$ or $A \to B$ and $A \ast\!\!\ast B$ means $A - B$, $A \to B$, or $A \leftarrow B$) we have:

1. The path ρ between A and T is of the form $A \ast\!\!\ast \cdots \ast\!\!\ast B \to T$. Clearly, $B \in \mathbf{Mb}(T)$ blocks the path ρ.
2. The path ρ between A and T is of the form $A \ast\!\!\ast \cdots \ast\!\!\ast C \leftarrow B \leftarrow T$. Clearly, $B \in \mathbf{Mb}(T)$ blocks the path ρ.
3. The path ρ between A and T is of the form $A \ast\!\!\ast \cdots \ast\!\!\ast C \to B \leftarrow T$. $C \in \mathbf{Mb}(T)$ blocks the path ρ.
4. The path ρ between A and T is of the form $A \ast\!\!\ast \cdots \ast\!\!\ast C \to D \leftrightarrow \cdots \leftrightarrow B \leftarrow T$, where $\omega = C \to D \leftrightarrow \cdots \leftrightarrow B \leftarrow T$ is the largest collider path between T and a node on the path ρ. Since all nodes of ω are in the Markov blanket of T, $\forall X \in \omega, X \neq A$. So, $C \in \mathbf{Mb}(T)$ blocks the path ρ.
5. The path ρ between A and T is of the form $A \ast\!\!\ast \cdots \ast\!\!\ast C \to D \leftrightarrow \cdots \leftrightarrow B \leftrightarrow T$, where $\omega = C \to D \leftrightarrow \cdots \leftrightarrow B \leftarrow T$ is the largest collider path between T and a node on the path ρ. Since all nodes of ω are in the Markov blanket of T, $\forall X \in \omega, X \neq A$. So, $C \in \mathbf{Mb}(T)$ blocks the path ρ.

From the global Markov property it follows that every m-separation relation in G implies conditional independence in every joint probability distribution P that satisfies the global Markov property for G. Thus, we have $T \perp\!\!\!\perp_p V \setminus \{T, \mathbf{Mb}(T)\} | \mathbf{Mb}(T)$. □

Now, we are ready to prove Theorem 3.

Sketch of proof of Theorem 3. If a variable belongs to $\mathbf{Mb}(T)$, then it will be admitted in the first step (Growing phase) at some point, since it will be dependent on T given the candidate set of $\mathbf{Mb}(T)$. This holds because of the causal faithfulness and because the set $\mathbf{Mb}(T)$ is the minimal set with that property. If a variable X is not a member of $\mathbf{Mb}(T)$, then conditioned on $\mathbf{Mb}(T) \setminus \{X\}$, it will be independent of T and thus will be removed from the candidate set of $\mathbf{Mb}(T)$ in the second phase (Shrinking phase) because the causal Markov condition entails that independencies in the distribution are represented in the graph. Since the causal faithfulness condition entails dependencies in the distribution from the graph, we never remove any variable X from the candidate set of $\mathbf{Mb}(T)$ if $X \in \mathbf{Mb}(T)$. Using this argument inductively we will end up with the $\mathbf{Mb}(T)$. □

In order to show the details of the proof of the Theorem 3, we prove only the correctness of the Grow-Shrink Markov blanket (GSMB) algorithm without causal sufficiency assumption in details as following (for the other algorithms listed in Theorem 3, a similar argument can be used):

Proof of Correctness of the GSMB Algorithm. By "correctness" we mean that GSMB is able to produce the true Markov blanket of any variable in the ground truth ADMG under Markov condition and the faithfulness assumption if all conditional independence tests done during its course are assumed to be correct.

Algorithm 2: The GS Markov Blanket Algorithm (Margaritis, 2003).

Input: a set V of nodes, a target variable T, and a probability distribution p faithful to an unknown ADMG G.
Output: The Markov blanket of T i.e., $\mathbf{Mb}(T)$.

Set $S = \emptyset$;
`/* Grow Phase: */`
while $\exists X \in V \setminus \{T\}$ *such that* $X \not\perp\!\!\!\perp_p T | S$ **do**
  $S \leftarrow S \cup \{X\}$;
`/* Shrink Phase: */`
while $\exists X \in S$ *such that* $X \perp\!\!\!\perp_p T | S \setminus \{X\}$ **do**
  $S \leftarrow S \setminus \{X\}$;
return ($\mathbf{Mb}(T) \leftarrow S$);

We first prove that there does not exist any variable $X \in \mathbf{Mb}(T)$ at the end of the growing phase that is not in S. The proof is by induction (a semi-induction approach on finite subset of natural numbers) on the *length* of the collider path(s) between X and T. We define the length of a collider path between X and T as the number of edges between them. Let s be the length of a largest collider path between X and T.

- (*Base case*) For the base of induction, consider the set of adjacents of T i.e., $\mathbf{adj}(T) = \{X \in V | X \rightarrow T, X \leftarrow T, \text{ or } X \leftrightarrow T\}$. In this case, $\nexists S \subseteq V \setminus \{T, X\}$ such that $X \perp\!\!\!\perp_m T | S$ in G. The faithfulness assumption implies that $\nexists S \subseteq V \setminus \{T, X\}$ such that $X \perp\!\!\!\perp_p T | S$. So, at the end of the grow phase, all of adjacents of T are in the candidate set for the Markov blanket of T.

- (*Induction hypothesis*) For all $1 \leq n < s$, if there is a collider path between X and T of length n then $X \in \mathbf{Mb}(T)$ at the end of the growing phase.

- (*Induction step*) To prove the inductive step, we assume the induction hypothesis for $n = s - 1$ and then use this assumption to prove that the statement holds for s. Assume that $\rho = (v_0 = T) \ast\!\!\rightarrow v_1 \leftrightarrow \cdots \leftrightarrow v_{s-1} \leftarrow\!\!\ast v_s$ is a collider path of length s. From the induction hypothesis we know that $v_i \in \mathbf{Mb}(T), \forall 1 \leq i < s$ at the end of the grow phase. Using Definition 1 implies that ρ is m-connected to T. This means that at some point of the grow phase v_s falls into the set S. The faithfulness assumption implies that $\exists S \subseteq V \setminus \{T, v_s\}$ such that $v_s \not\perp\!\!\!\perp_p T | S$. So, at the end of the grow phase, all of X's that there is a collider path between X and T fall into the candidate set for the Markov blanket of T.

For the correctness of the shrinking phase, we have to prove two things: (1) we never remove any variable X from S if $X \in \mathbf{Mb}(T)$, and (2) if $X \notin \mathbf{Mb}(T)$, X is removed in the shrink phase.

Now we prove case (1) by contradiction. Assume that $X \in \mathbf{Mb}(T)$, $X \in S$ at the end of the grow phase, and we remove X from S in the shrink phase. This means $X \perp\!\!\!\perp_p T | S \setminus \{X\}$. Using the faithful assumption implies that $X \perp\!\!\!\perp_m T | S \setminus \{X\}$ i.e., $S \setminus \{X\}$ m-separates X from T in G, which is a contradiction because there is a collider path between X and T and $\mathbf{Mb}(T) \setminus \{X\} \subseteq S$. In other word, the collider path between X and T is m-connected by $S \setminus \{X\}$.

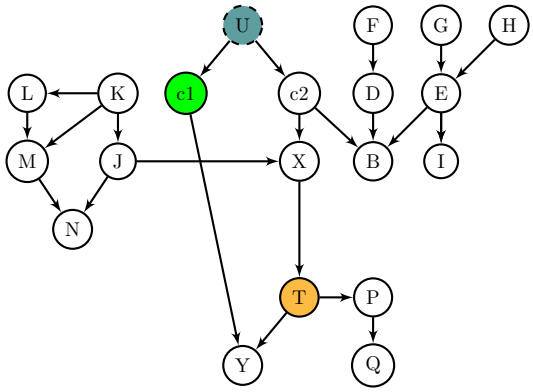
To prove the case (2), assume that $X \notin \mathbf{Mb}(T)$, $\mathbf{Y} = S \setminus \{\mathbf{Mb}(T)\}$, and $X \in \mathbf{Y}$. Due to the Markov blanket property, $\mathbf{Mb}(T)$ m-separates $\mathbf{Y}$ from T in G. Using the Markov condition implies that $\mathbf{Y} \perp\!\!\!\perp_p T | \mathbf{Mb}(T)$.

Since the probability distribution p satisfies the faithfulness assumption, it satisfies the *weak union* condition (Sadeghi, 2017). We recall the weak union property here: $A \perp\!\!\!\perp_p BD|C \Rightarrow (A \perp\!\!\!\perp_p B|DC$ and $A \perp\!\!\!\perp_p D|BC)$. Using the weak union property for $\mathbf{Y} \perp\!\!\!\perp_p T|\mathbf{Mb}(T)$ implies that $X \perp\!\!\!\perp_p T|(\mathbf{Mb}(T) \cup (\mathbf{Y} \setminus \{X\}))$, which is the same as $X \perp\!\!\!\perp_p T|(S \setminus \{X\})$. This means, $X \notin \mathbf{Mb}(T)$ will be removed at the end of the shrink phase. □

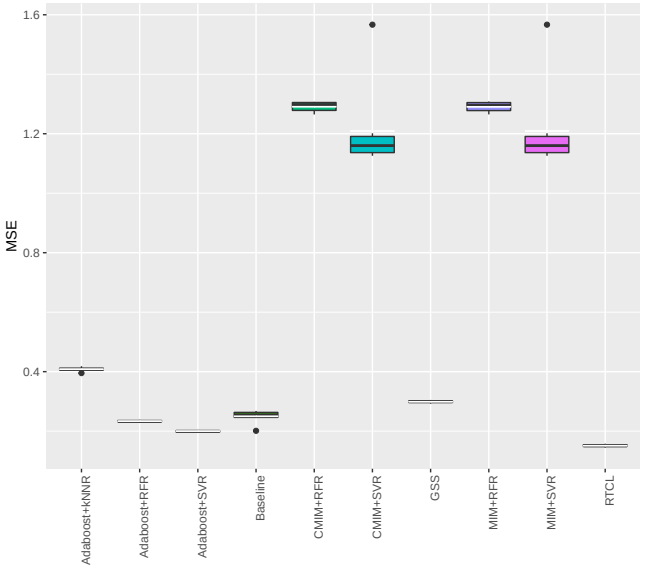
Appendix B. More Experimental Results

In this section we include the remaining results that we got during experimentation, along with subsequent ground truth graph for representing the various settings we experiment on synthetic data. In the later part of this section we also include t-test tables as mentioned in Section 6.

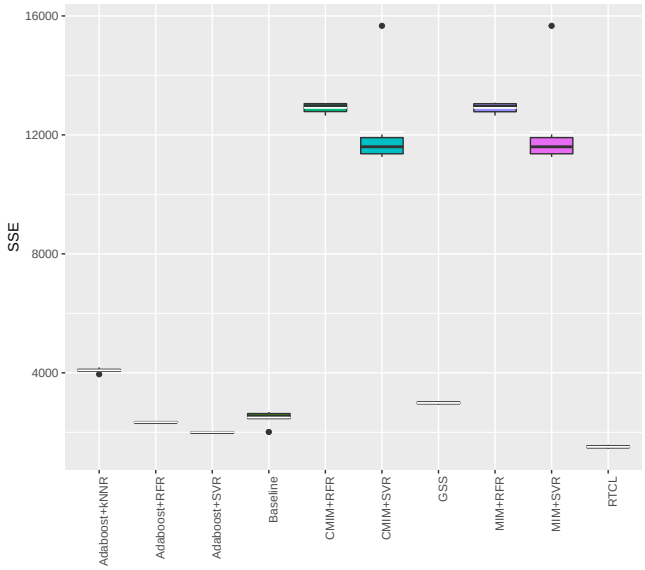
Results on Synthetic dataset



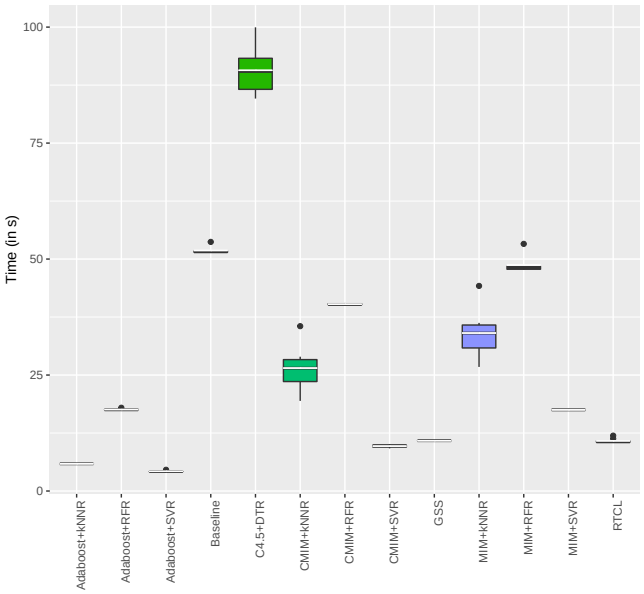
(a)



(b)

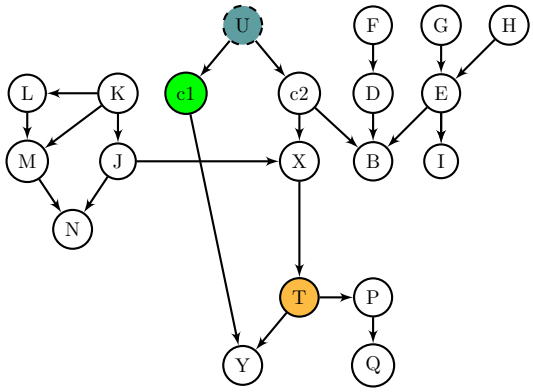


(c)

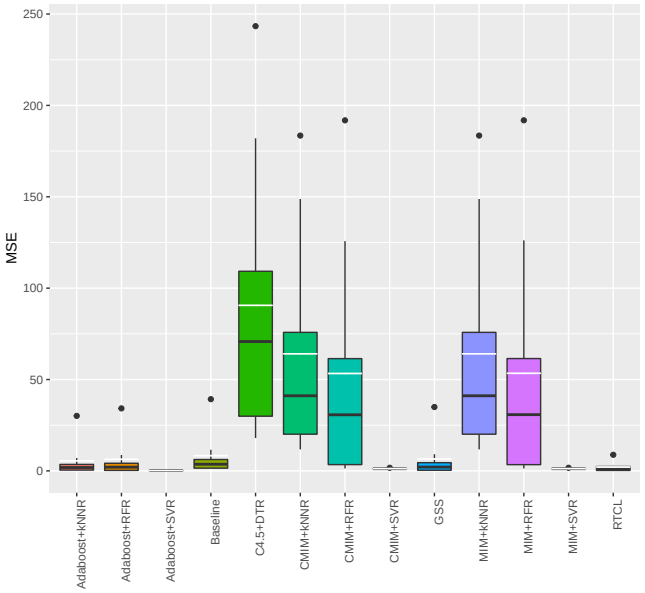


(d)

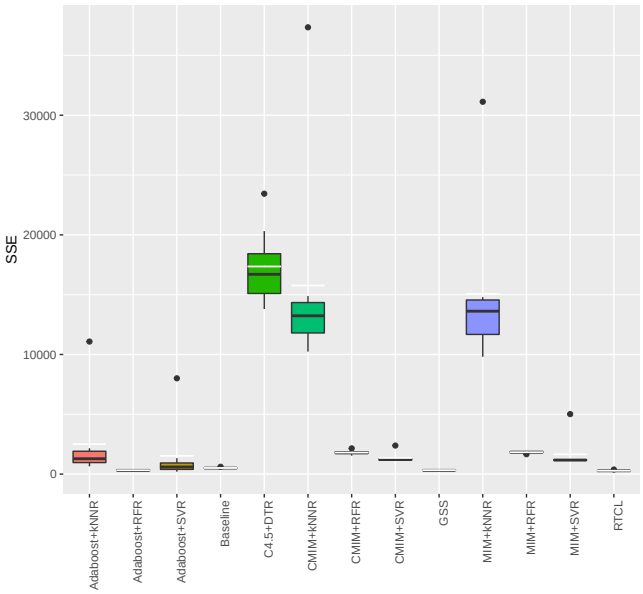
Figure 11: Results on (b), (c), (d) are generated over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the probability distribution of the context variable C_1, where sample size = 10000 for a Gaussian distribution.



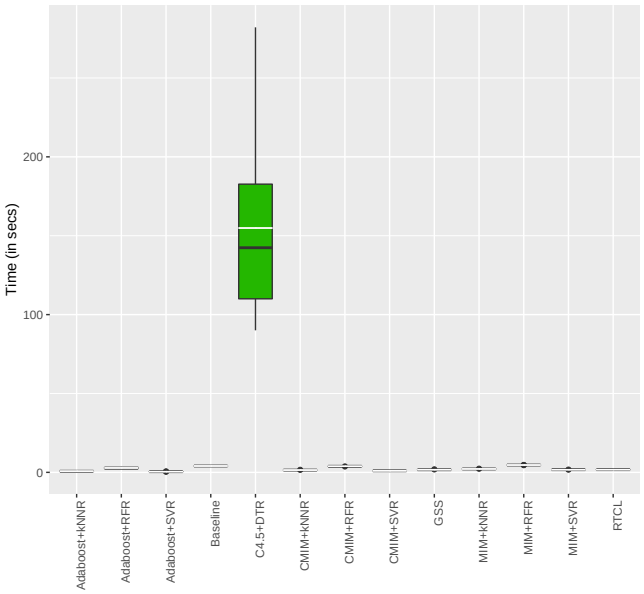
(a)



(b)

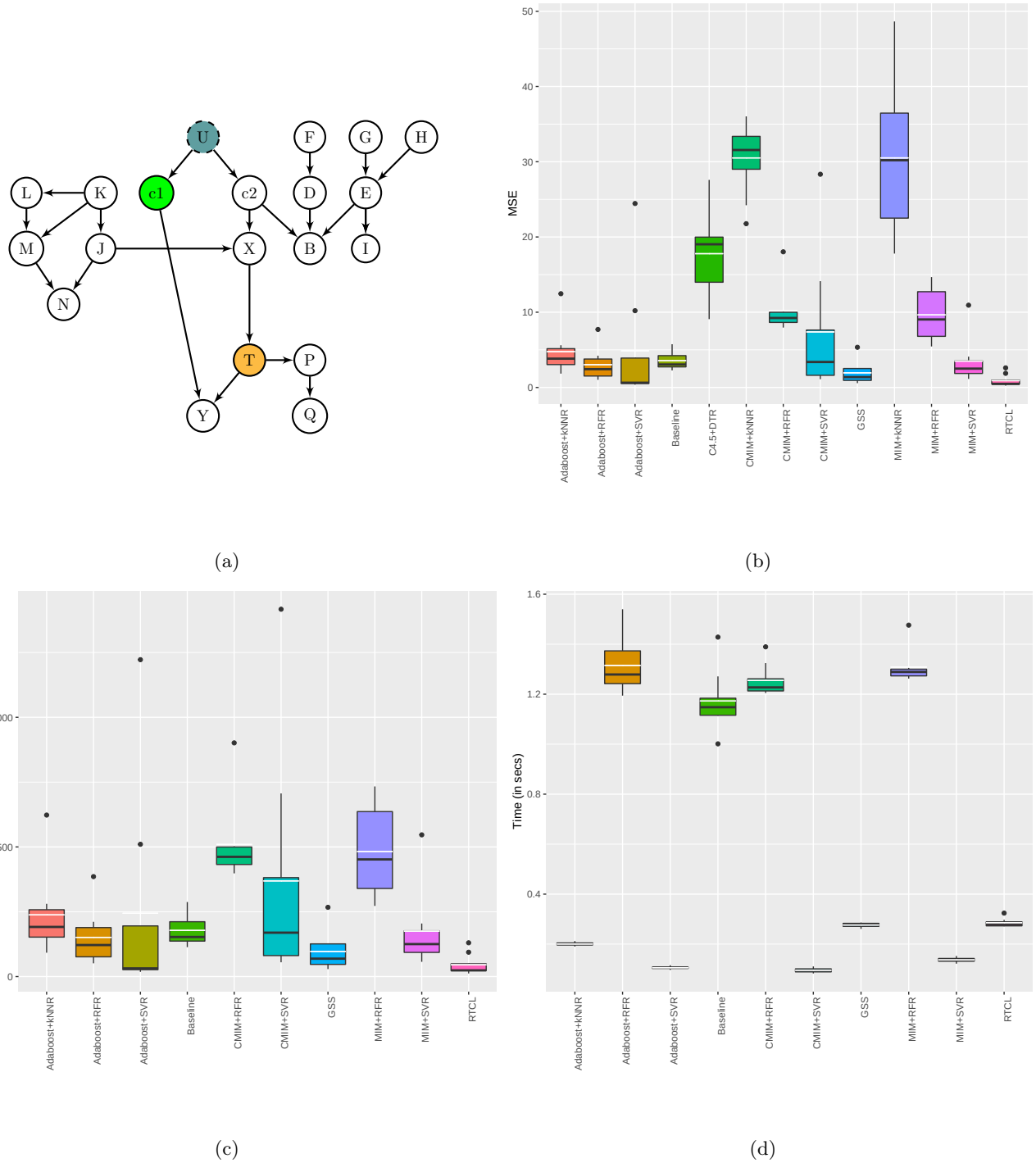


(c)



(d)

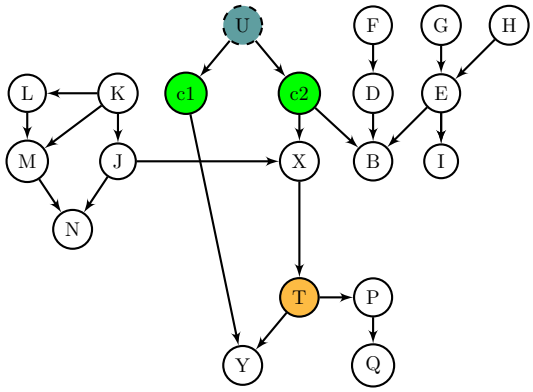
Figure 12: Results on (b), (c), (d) are generated over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the probability distribution of the context variable C_1, where sample size = 1000 for a Gaussian distribution.



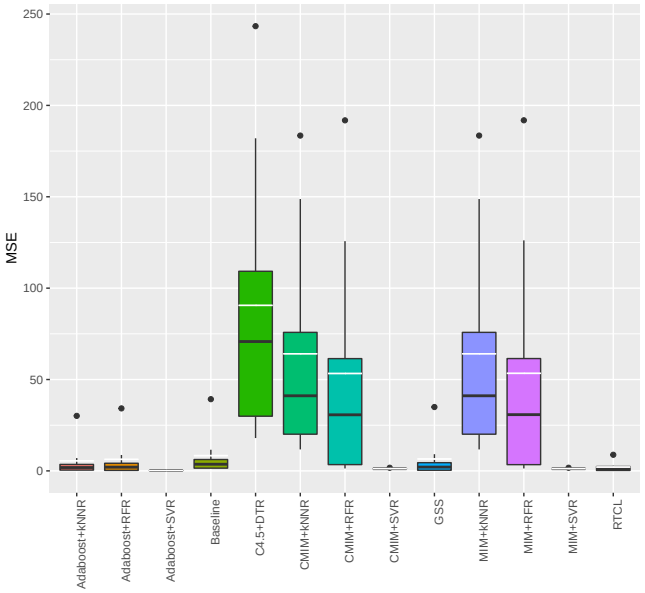
(a) (b)

(c) (d)

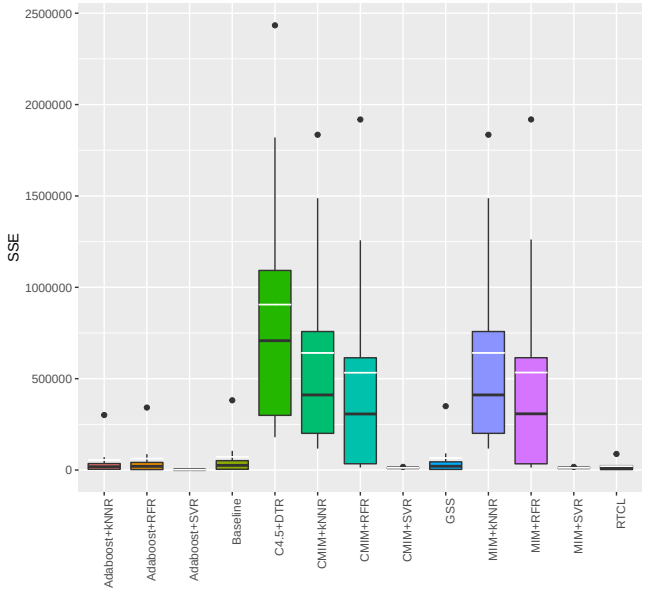
Figure 13: Results on (b), (c), (d) are generated over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the probability distribution of the context variable C_1, where sample size = 50 for a Gaussian distribution.



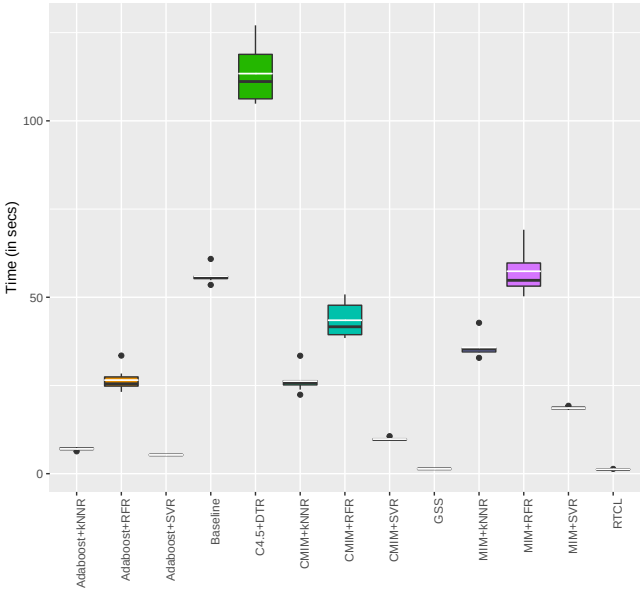
(a)



(b)

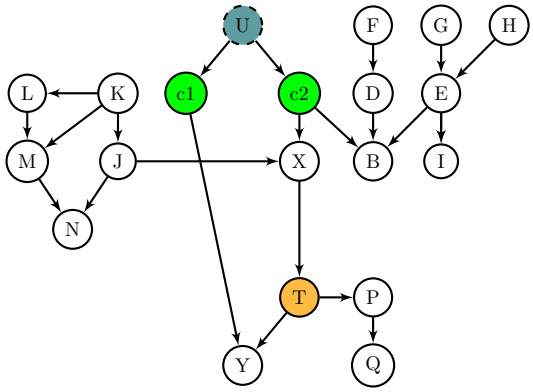


(c)

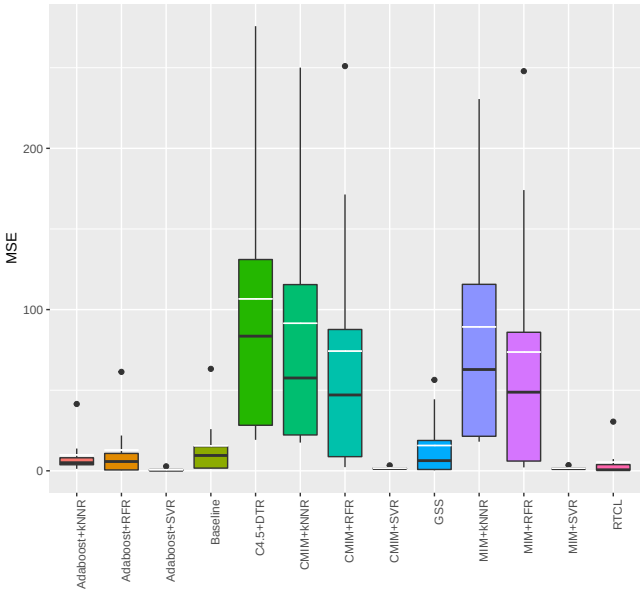


(d)

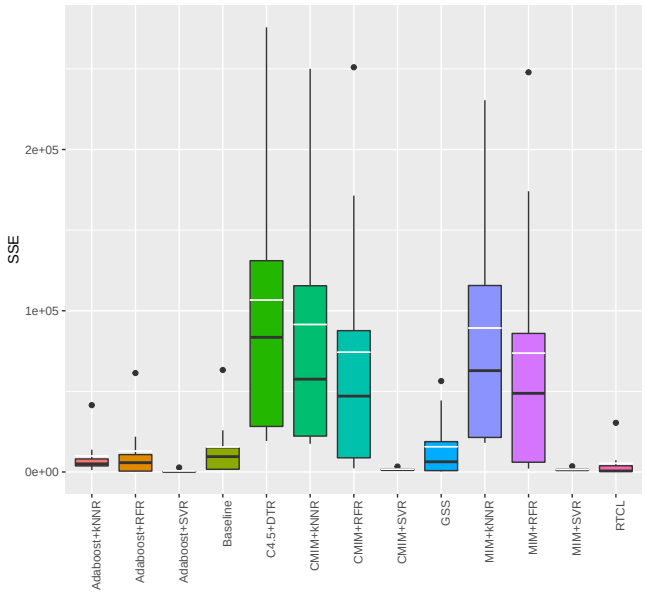
Figure 14: Results on (b), (c), (d) are generated over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the probability distribution of the context variable C_1& C_2, where sample size = 10000 for a Gaussian distribution.



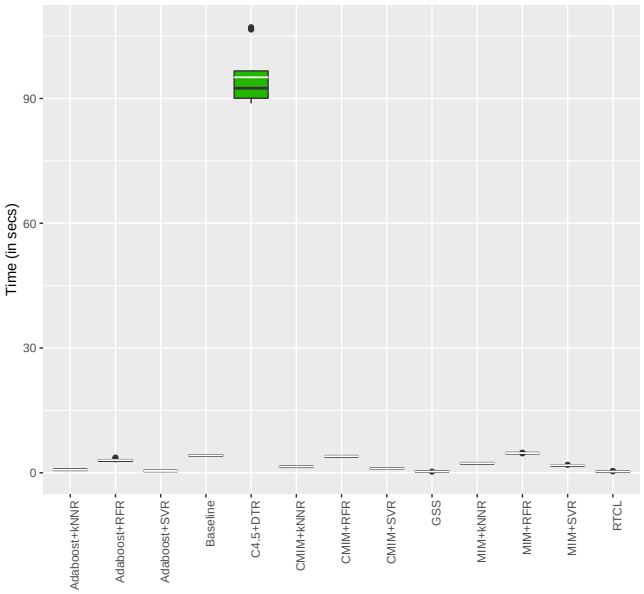
(a)



(b)

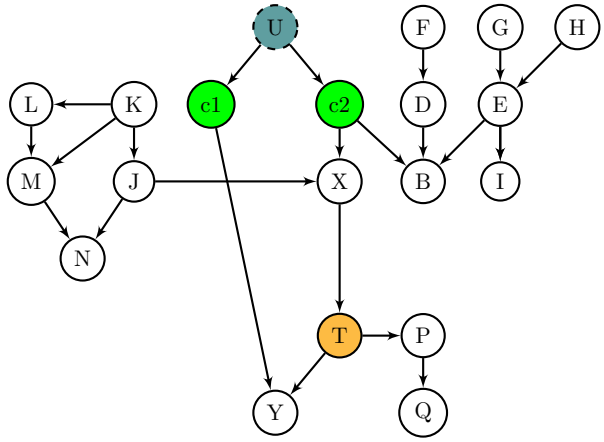


(c)

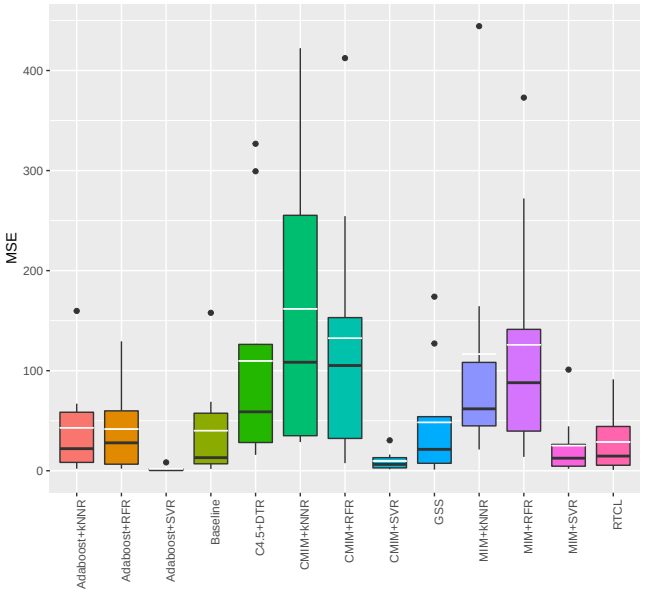


(d)

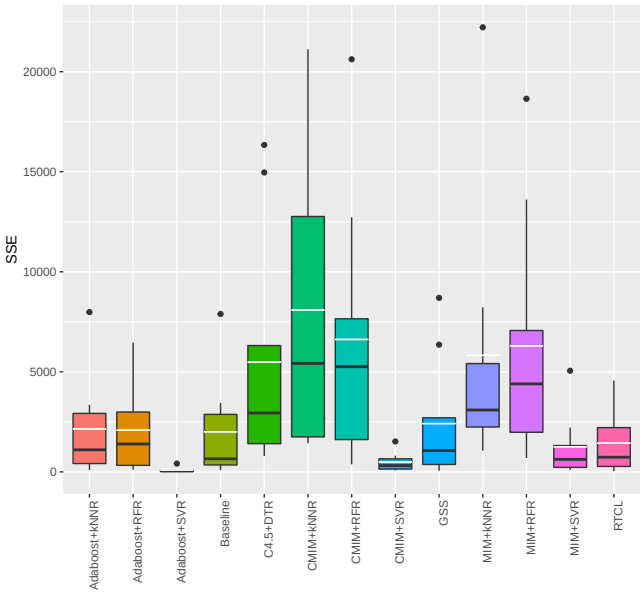
Figure 15: Results on (b), (c), (d) are generated over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the probability distribution of the context variable C_1 & C_2, where sample size = 1000 for a Gaussian distribution.



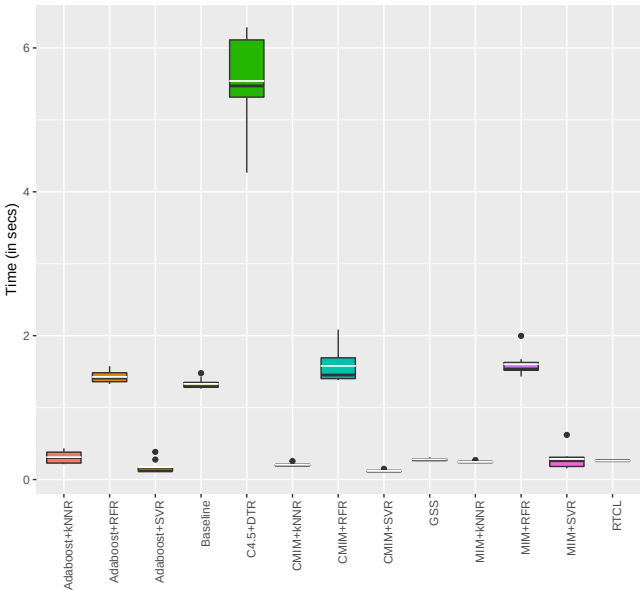
(a)



(b)

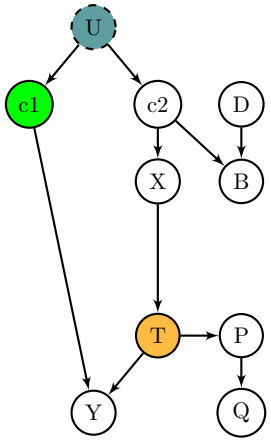


(c)

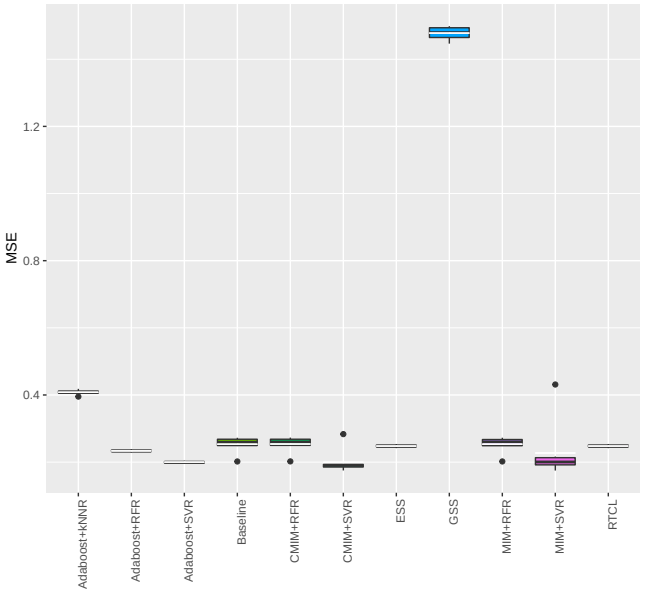


(d)

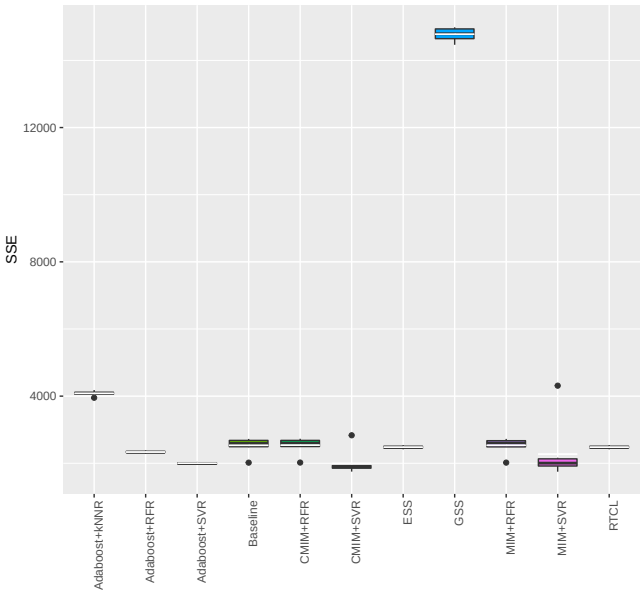
Figure 16: Results on (b), (c), (d) are generated over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the distribution of the context variable C_1 & C_2, where sample size = 50 for a Gaussian distribution.



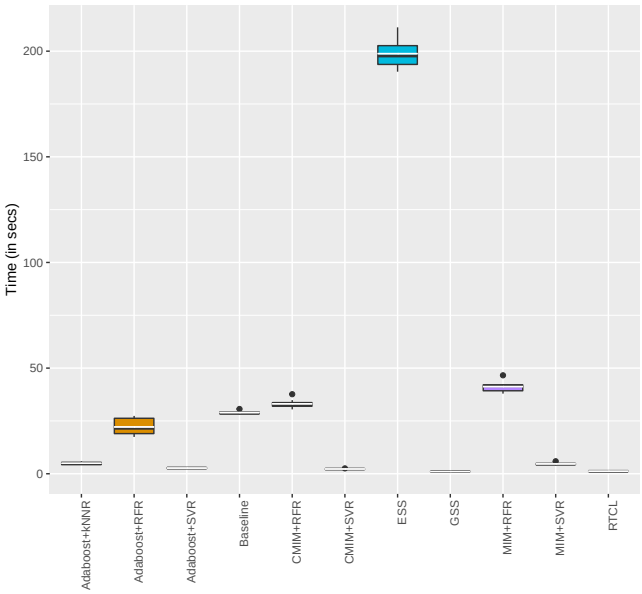
(a)



(b)

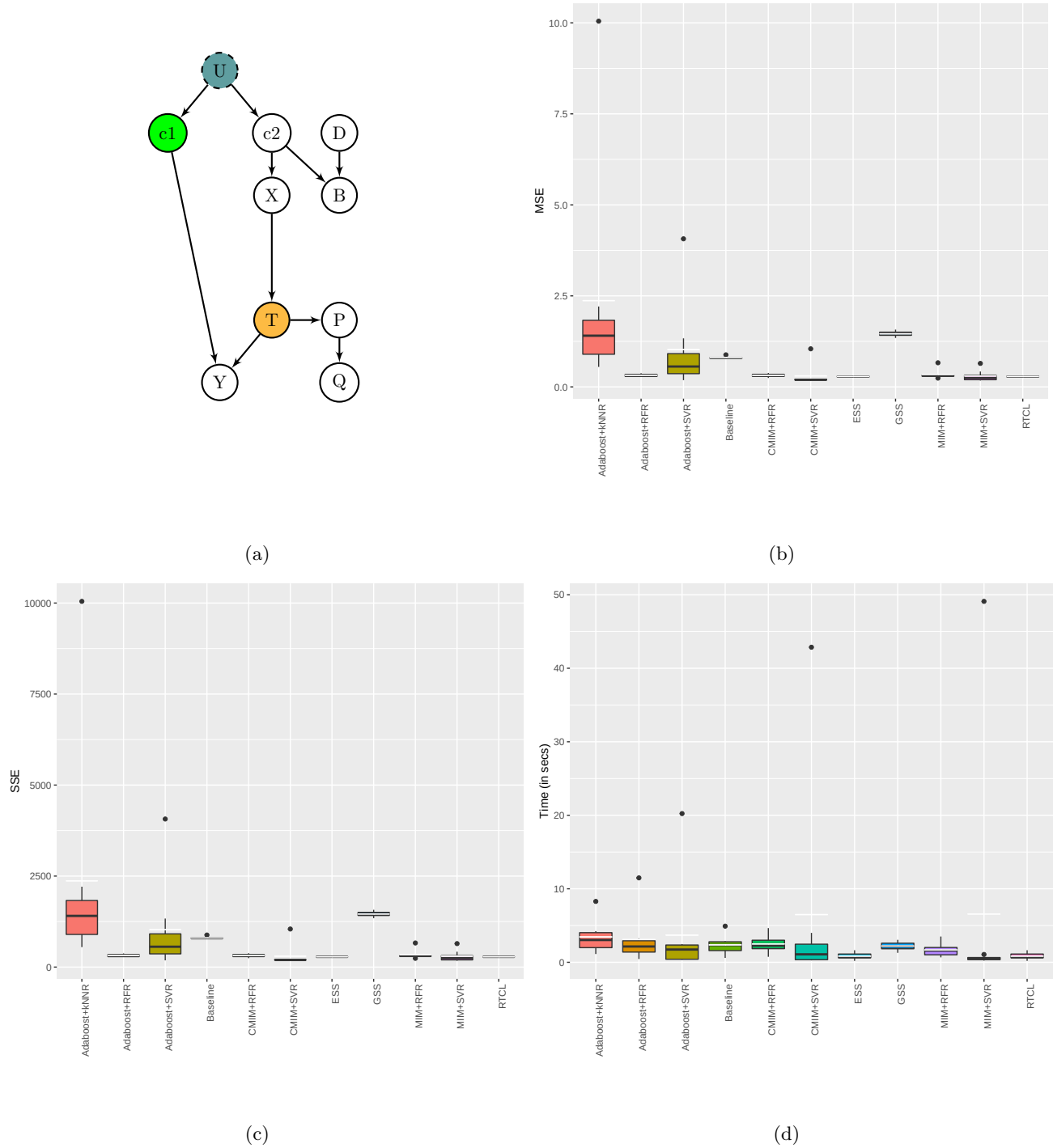


(c)



(d)

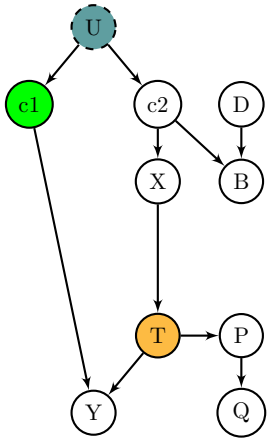
Figure 17: Results on (b), (c), (d) are generated over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the distribution of the context variable C_1, where sample size = 10000 for a Gaussian distribution.



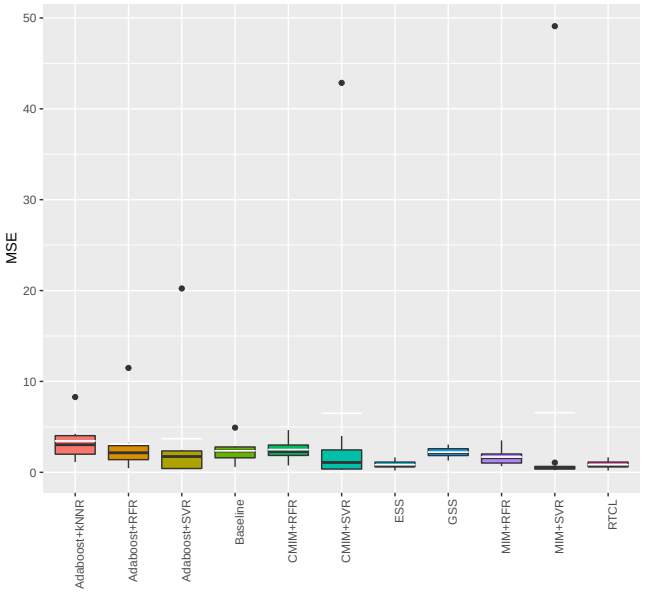
(a) (b)

(c) (d)

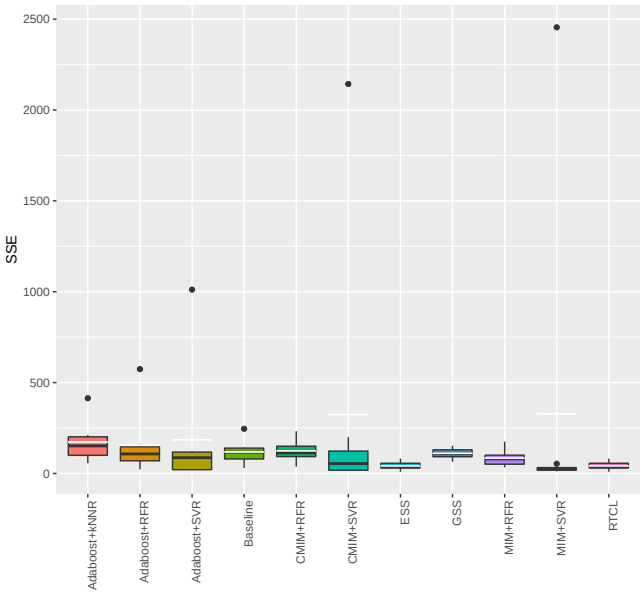
Figure 18: Results on (b), (c), (d) are generated over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the distribution of the context variable C_1, where sample size = 1000 for a Gaussian distribution.



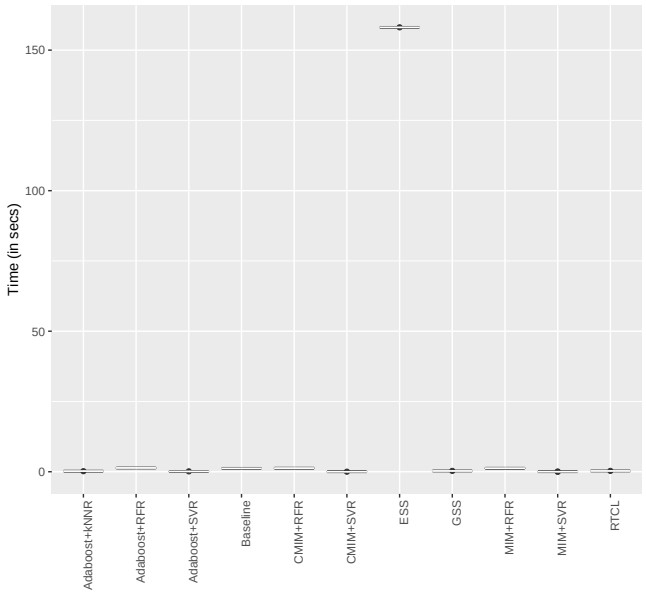
(a)



(b)

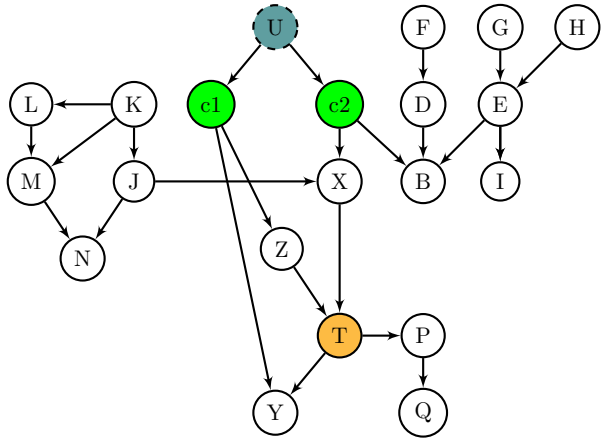


(c)

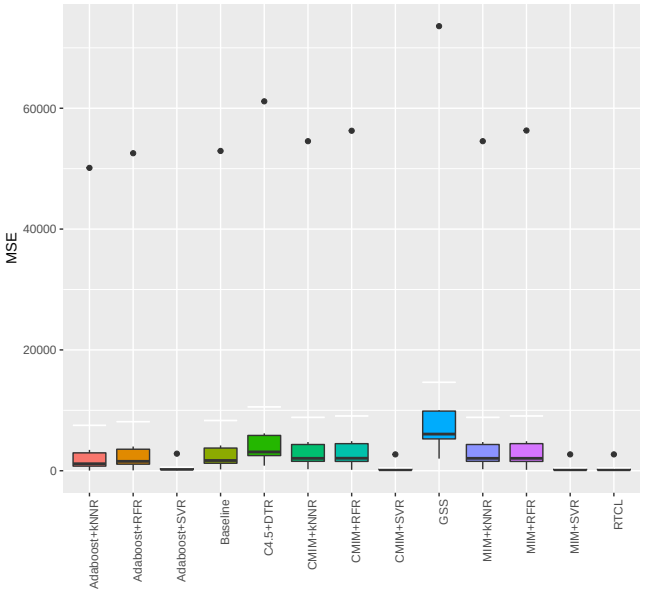


(d)

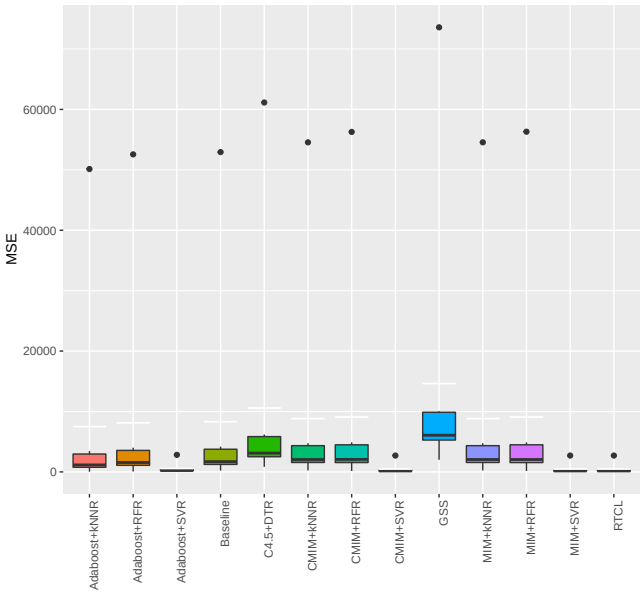
Figure 19: Results on (b), (c), (d) are generated over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the distribution of the context variable C_1, where sample size = 50 for a Gaussian distribution.



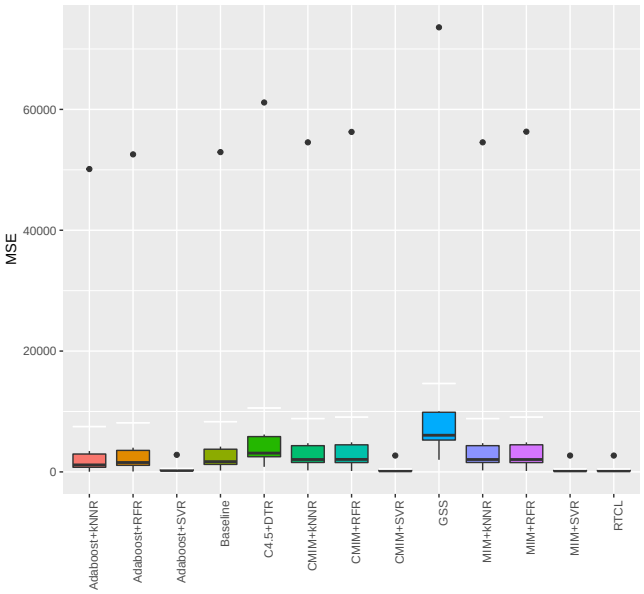
(a)



(b)



(c)



(d)

Figure 20: Results on (b), (c), (d) are generated over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the distribution of the context variable C_1 , where sample size = 1000 for a Gaussian distribution.

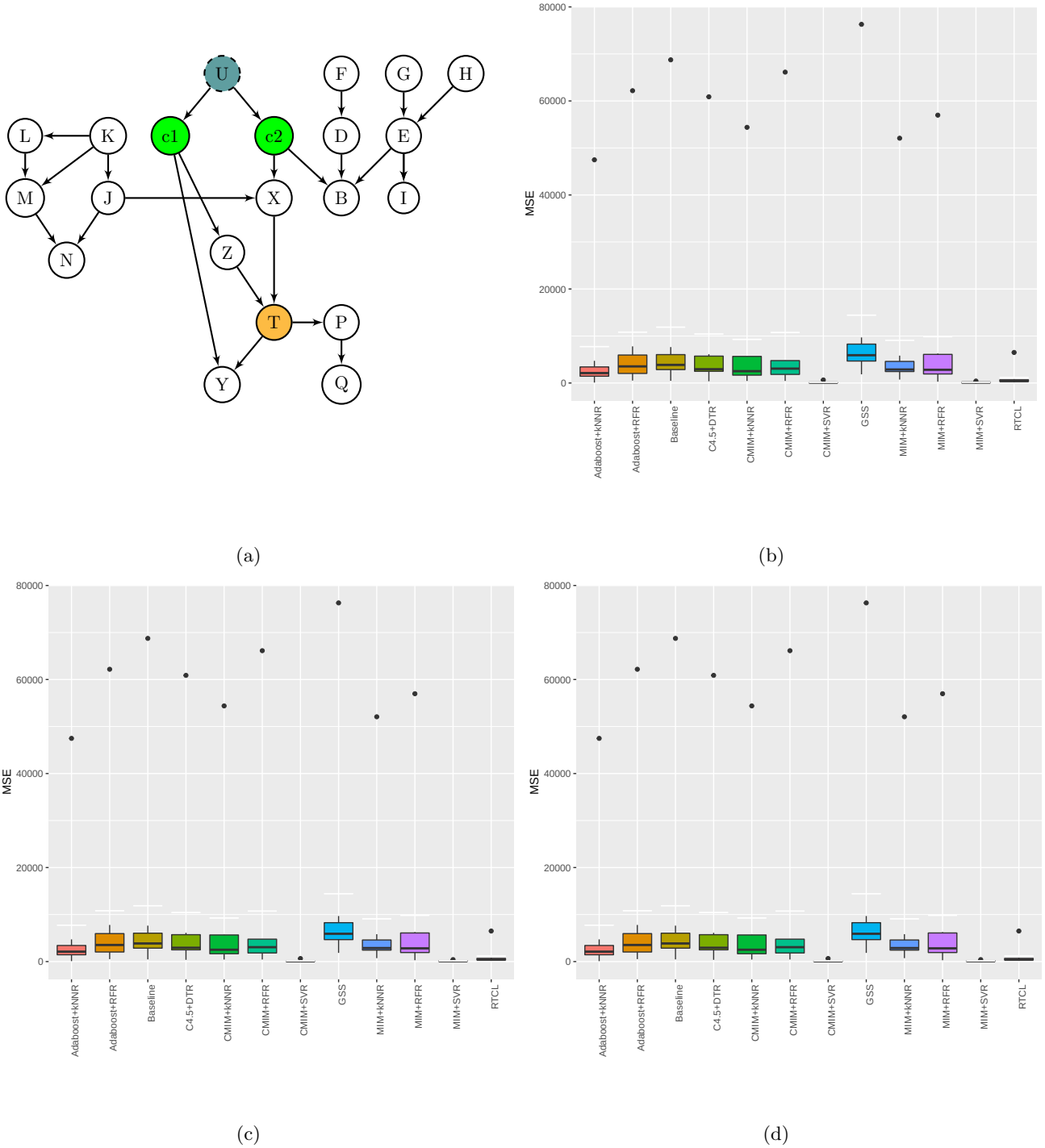
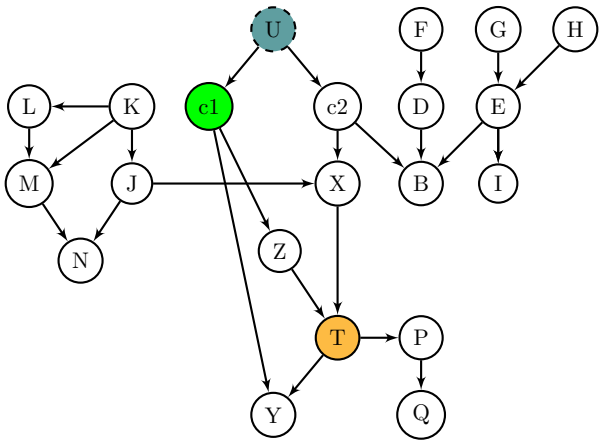


Figure 21: Results on (b), (c), (d) are generated over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the distribution of the context variable C_1 , where sample size = 50 for a Gaussian distribution.

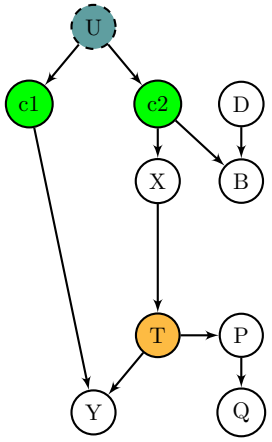


(a)

Methodology	p-value
Baseline	**0.1694**
GSS	**0.19617**
CMIM+SVR	**0.3187**
CMIM+kNNR	**0.1258**
CMIM+RFR	**0.1376**
MIM+SVR	**0.318705**
MIM+kNNR	**0.1258**
MIM+RFR	**0.13759**
Adaboost+SVR	**0.279**
Adaboost+kNNR	**0.20346**
Adaboost+RFR	**0.169**
C4.5+ DTR	**0.0889**

(b)

Figure 22: Results in the table are p-values generated by performing T-test on SSE over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the distribution of the context variable C_1, over Gaussian distribution with sample size = 1000. The bold results indicate that the average performance of `RCTL` is not significantly different from the average performance of other approaches for the p-value 0.05.

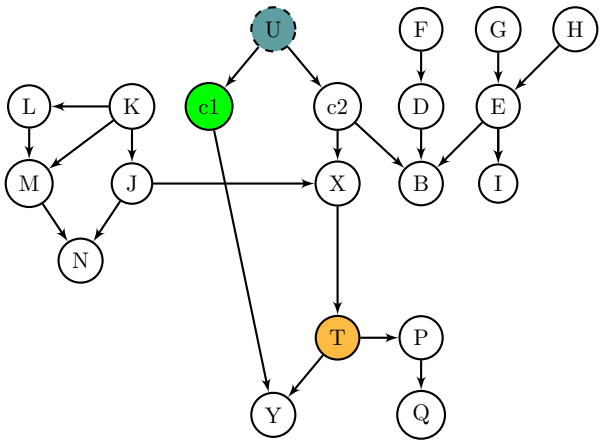


(a)

Methodology	p-value
Baseline	**0.1535**
GSS	**0.1681**
CMIM+SVR	**0.2712**
CMIM+kNNR	**0.1224**
CMIM+RFR	**0.1885**
MIM+SVR	**0.65673**
MIM+kNNR	0.04307
MIM+RFR	**0.16345**
Adaboost+SVR	**0.2660**
Adaboost+kNNR	**0.0526**
Adaboost+RFR	**0.1185**
C4.5+ DTR	0.0295
ESS	0.0222

(b)

Figure 23: Results in the table are p-values generated by performing T-test on MSE over the ground truth graph (a) for the target variable T. Datashift between source and target occurs by change in the distribution of the context variable C_1 & C_2 , over Gaussian distribution with sample size = 1000. The bold results indicate that the average performance of `RCTL` is not significantly different from the average performance of other approaches for the p-value 0.05.

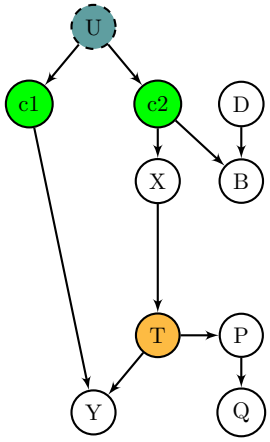


(a)

Methodology	p-value
Baseline	**0.01747**
GSS	**6.14934e-08**
CMIM+SVR	**2.9747e-08**
CMIM+kNNR	**0.00445**
CMIM+RFR	**1.7186e-17**
MIM+SVR	**5.415e-07**
MIM+kNNR	**0.0057**
MIM+RFR	**8.48536e-17**
Adaboost+SVR	**0.0001**
Adaboost+kNNR	**1.0681e-07**
Adaboost+RFR	0.16304
C4.5+ DTR	**4.2182e-08**

(b)

Figure 24: Results in the table are p-values generated by performing T-test on MSE over the ground truth graph (a) for the target variable T. Datashift between source and target occurs by change in the distribution of the context variable C_1 , over Gaussian distribution with sample size = 1000. The bold results indicate that the average performance of `RCTL` is significantly different from the average performance of other approaches for the p-value 0.05.

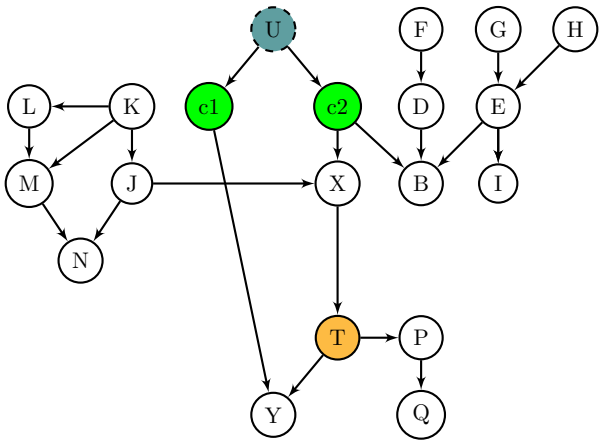


(a)

Methodology	p-value
Baseline	**2.5901e-09**
GSS	**9.1341e-06**
CMIM+SVR	0.7350
CMIM+kNNR	0.1154
CMIM+RFR	**0.0265**
MIM+SVR	0.94294
MIM+kNNR	0.1687
MIM+RFR	**0.0190**
Adaboost+SVR	0.87851
Adaboost+kNNR	**9.9208e-09**
Adaboost+RFR	0.1205
C4.5+ DTR	**2.6341e-05**
ESS	0.6341

(b)

Figure 25: Results in the table are p-values generated by performing T-test on MSE over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the distribution of the context variable C_1, over Gaussian data with sample size = 1000. The bold results indicate that the average performance of `RCTL` is significantly better than the average performance of other approaches for the p-value 0.05.

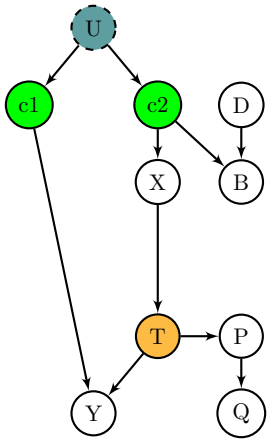


(a)

Methodology	p-value
Baseline	**0.0225**
GSS	**1.2961e-05**
CMIM+SVR	**1.212e-10**
CMIM+kNNR	**1.6860e-14**
CMIM+RFR	**0.0001**
MIM+SVR	**1.212e-10**
MIM+kNNR	**3.344e-14**
MIM+RFR	**0.0001**
Adaboost+SVR	**1.212e-10**
Adaboost+kNNR	**5.148e-13**
Adaboost+RFR	0.1290
C4.5+ DTR	**2.562e-10**

(b)

Figure 26: Results in the table are p-values generated by performing T-test on SSE over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the distribution of the context variables C_1& C_2, over Discrete data with sample size = 1000. The bold results indicate that the average performance of `RCTL` is significantly better than the average performance of other approaches for the p-value 0.05.

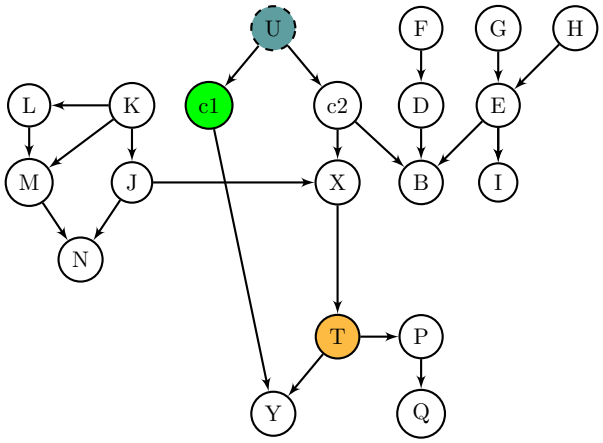


(a)

Methodology	p-value
Baseline	**0.00013**
GSS	**0.0001**
CMIM+SVR	**4.785e-10**
CMIM+kNNR	**5.4748e-10**
CMIM+RFR	**0.0059**
MIM+SVR	**4.7853e-10**
MIM+kNNR	**4.7034e-10**
MIM+RFR	**0.0056**
Adaboost+SVR	**4.785e-10**
Adaboost+kNNR	**7.088e-10**
Adaboost+RFR	0.5531
C4.5+ DTR	**3.136e-09**
ESS	**0.0006**

(b)

Figure 27: Results in the table are p-values generated by performing T-test on SSE over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the distribution of the context variables C_1& C_2, over Discrete data. The bold results indicate that the average performance of `RCTL` is significantly better than the average performance of other approaches for the p-value 0.05.

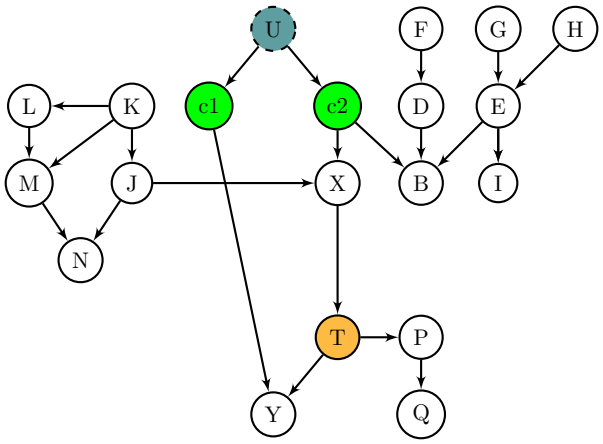


(a)

Methodology	p-value
Baseline	**0.0001**
GSS	**1.585e-06**
CMIM+SVR	**7.227e-15**
CMIM+kNNR	**4.352e-09**
CMIM+RFR	**5.212e-06**
MIM+SVR	**7.227e-15**
MIM+kNNR	**7.327e-09**
MIM+RFR	**0.0001**
Adaboost+SVR	**7.229e-15**
Adaboost+kNNR	**3.760e-08**
Adaboost+RFR	0.054
C4.5+ DTR	**8.380e-10**

(b)

Figure 28: Results in the table are p-values generated by performing T-test on MSE over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the distribution of the context variable C_1, over Discrete data. The bold results indicate that the average performance of `RCTL` is significantly better than the average performance of other approaches for the p-value 0.05.

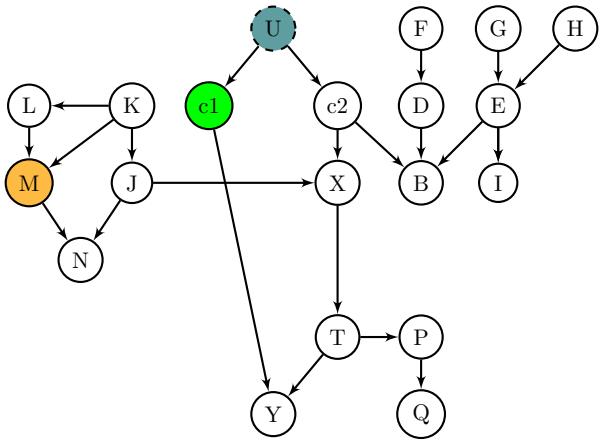


(a)

Methodology	p-value
Baseline	**0.1651**
GSS	**0.1687**
CMIM+SVR	3.5196e-07
CMIM+kNNR	0.0277
CMIM+RFR	**0.066**
MIM+SVR	3.519e-07
MIM+kNNR	0.0277
MIM+RFR	**0.0664**
Adaboost+SVR	**0.3315**
Adaboost+kNNR	**0.1858**
Adaboost+RFR	**0.1865**
C4.5+ DTR	0.0228

(b)

Figure 29: Results in the table are p-values generated by performing T-test on MSE over the ground truth graph (a) for the target variable T. Data shift between source and target occurs by change in the distribution of the context variable C_1& C_2, over Discrete data with sample size = 10000. The bold results indicate that the average performance of `RCTL` is comparable to the average performance of other approaches for the p-value 0.05.



(a)

Methodology	p-value
Baseline	**6.7513e-12**
GSS	**0.0009**
CMIM+SVR	**0.0056**
CMIM+kNNR	**1.03614e-06**
CMIM+RFR	**1.2254e-11**
MIM+SVR	**1.0319e-05**
MIM+kNNR	**1.1642e-07**
MIM+RFR	**4.537e-10**
Adaboost+SVR	**0.0305**
Adaboost+kNNR	**0.0138**
Adaboost+RFR	**3.993e-10**
C4.5+ DTR	**2.0194e-07**

(b)

Figure 30: Results in the table are p-values generated by performing T-test on MSE over the ground truth graph (a) for the target variable M. Data shift between source and target occurs by change in the distribution of the context variable C_1, over Discrete data with sample size = 1000. The bold results indicate that the average performance of `RCTL` is significantly better than the average performance of other approaches for the p-value 0.05.