Knowledge Transfer for Learning Markov Equivalence Classes

Verónica Rodríguez-López VERORL@INAOEP.MX

Luis Enrique Sucar ESUCAR@INAOEP.MX

Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, México

Abstract

There are domains, such as in biology, medicine, and neuroscience, where the causal relations vary across members of a population, and where it may be difficult to collect data for some specific members. For these domains, it is convenient to develop algorithms that, from small sample sizes, can discover the specific causal relations of a subject. Learning these subject-specific models with the existing causal discovery algorithms could be difficult. Most of them were designed to find the common causal relations of a population in the large sample limit. Although transfer learning techniques have shown to be useful for improving predictive associative models learned with limited data sets, their application in the field of causal discovery has not been sufficiently explored. In this paper, we propose a knowledge transfer algorithm for discovering Markov equivalence classes for subject-specific causal models. We explore transferring weighted instances of auxiliary data sets, according to their relevance, for improving models learned with limited sample sizes. Experimental results on data sets generated from simulated and benchmark causal Bayesian networks show that our method outperforms in adjacency and arrowhead recovery the base and a similar knowledge transfer discovery methods.

Keywords: Bayesian networks, Causal discovery, Transfer learning, Subject-specific model.

1. Introduction

Probabilistic graphical models (PGMs) are useful tools for encoding causal relations between variables of closed systems and provide information to make predictions under manipulations. In their learning, either from observations, through interventions or both, discovering their causal structure is an important aspect. From observational data, it is possible to discover Markov Equivalence Classes (MECs) that represent the structure of a set of equivalent causal PGMs with the same joint probability distribution (Chickering, 2002).

Most algorithms have been designed to learn population-wide causal PGMs including the common causal relations of a population (Glymour et al., 2019; Malinsky and Danks, 2018; Mooij et al., 2019; Spirtes and Zhang, 2016; Tillman and Eberhardt, 2014). In some domains, there could be variations in causal relations across the members of a population. For example, in neuroscience, because of differences in the degree of disease affectation and the recovery process, it has been observed that causal relations between brain regions might vary across patients (Grefkes and Fink, 2014). Findings in genetics also have revealed that there are somatic genome alterations causing expression changes in specific tumors (Cooper et al., 2018). For these domains, it is convenient to develop algorithms to learn subject-specific causal models that help to capture the specific causal relations of a particular subject at some stage of interest, such as disease or recovery stage.

Learning subject-specific MECs from a limited sample size that include the specific causal structure of a particular member of a population could be challenging. Many existing algorithms find MECs that include the common causal relations of a population in the large sample limit (Zhang

et al., 2018; Glymour et al., 2019). However, because of the physical condition of the subjects, the difficulty or cost to carry out experiments, it can be complicated collecting enough data for some subjects. Transfer learning has shown to be useful for improving models learned with limited data sets, allowing the use of auxiliary data that comes from different models with different probability distributions (Pan and Yang, 2010).

In this paper, we propose a knowledge transfer algorithm for learning subject-specific MECs from limited data sets. It is an extension of the score-based algorithm Greedy Equivalence Search (GES) (Chickering, 2002) that transfers locally weighted instances of the auxiliary data sets to obtain a subject-specific MEC. We provide a strategy for transferring weighted auxiliary instances considering the probability distributions' differences between the target and the auxiliary sources. We evaluate and compare our algorithm with GES, PC (Spirtes et al., 2000), and a similar transfer learning algorithm (Jia et al., 2018) using data sets generated from simulated and benchmark Bayesian networks. Experimental results show that our algorithm recovers MECs with a higher number of correct adjacencies and v-structures than the ones recovered by the compared methods.

The paper is organized as follows. Related work to our proposal is described in Section 2. Section 3 provides an introduction to the main relevant concepts and includes a description of the Greedy Equivalence Search algorithm. The proposed transfer algorithm is presented in Section 4. The experimental evaluation and results are described in Section 5. Finally, the conclusions of this paper are presented in Section 6.

2. Related work

Several works have explored knowledge transfer for learning PGMs. However, most of these studies have relied on the learning of associative PGMs (Luis et al., 2010; Niculescu-Mizil and Caruana, 2007; Oyen and Lane, 2013). Limited work (Jia et al., 2018) has been done on knowledge transfer for learning causal PGMs from observational data. Although other algorithms have been proposed for learning MECs from multiple data sets, their aim is different of that for knowledge transfer algorithms. These algorithms aim to discover MECs that include the common causal relations in all data sets, assuming that all data sets include a representative number of samples (Claassen and Heskes, 2010; Ramsey et al., 2010; Tillman and Spirtes, 2011). On the other hand, even though some studies have explored the learning of subject-specific causal models (Cooper et al., 2018; Jabbari et al., 2018), which encoding the specific causal relations of a subject of population, they assume that there are enough data for the learning and use only data of the target subject.

The most related work to our proposal is the knowledge transfer algorithm of Jia et al. (2018). It is a modification of the PC algorithm that assumes all auxiliary data sets have the same relevance for learning a target MEC, ignoring their differences in probability distributions. Moreover, like other PC-based algorithms, it requires large sample sizes for the conditional independence tests (Glymour et al., 2019). Score-based algorithms have shown to be more accurate for learning MECs with small samples than constraint-based algorithms such as PC (Malinsky and Danks, 2018).

3. Preliminaries

3.1 Basic definitions

In this section, we present some basic definitions related to graphs and probabilistic graphical models. Throughout the paper, we will use capital letters (e.g., X, Y, Z) to denote variables and their

values with lower letters (e.g., x, y, z). We denote a set (of variables, parameters, or samples) by bold capital letters (e.g $\mathbf{X}, \mathbf{Pa}, \mathbf{Z}$). The assignment of value of each variable in a given set of variables, are denoted with bold lower letters (e.g., $\mathbf{x}, \mathbf{pa}, \mathbf{z}$).

The undirected graph resulting from ignoring the direction of edges in a directed acyclic graph (DAG) is the **skeleton** of the DAG. A **v-structure** in a DAG is an ordered triple of nodes (X, Y, Z), such that, the edges $X \rightarrow Y$ and $Y \leftarrow Z$ are in the DAG, and there is no edge between the nodes X, Z (Chickering, 2002). Two DAGs are equivalent if and only if they have the same skeletons and the same v-structures (Verma and Pearl, 1991).

A set of equivalent DAGs forms a **Markov equivalence class** (MEC) and can be described by a partial directed acyclic graph (PDAG) $\mathcal{G}$, called **completed PDAG** (CPDAG), where there is a directed edge for each edge contained in a v-structure, and a undirected edge for every other edge (Chickering, 2002).

A **Bayesian network** (BN) for a set of variables $\mathbf{X} = \{X_1, X_2, ..., X_p\}$ is a pair $(\mathcal{G}, \mathbf{\Theta})$, where $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a DAG with nodes that coincide with the variables in $\mathbf{X}$ that defines the structure of a BN, and θ is the set of parameters $\mathbf{\Theta} = \{\theta_1, \theta_2, ..., \theta_p\}$ defining the conditional probability $\theta_i = P(x_i|pa(x_i))$, of each node X_i given its parents $\mathbf{Pa}(X_i)$ in $\mathcal{G}$ (Sucar, 2015). A **causal Bayesian network** (causal BN) is a BN in which directed edges represent direct causes (Spirtes et al., 2000).

3.2 Greedy Equivalence Search

Greedy Equivalence Search (GES) (Chickering, 2002) is a score-based algorithm for learning the structure of Bayesian networks. In the GES algorithm, the problem of learning the structure of Bayesian networks is stated as follows (Alonso-Barba et al., 2013):

Definition 1 *Given a set $\mathbf{D} = \{d_1, ..., d_I\}$ containing I instances, where each d_i represent an assignment of value to each variable in $\mathbf{X} = \{X_1, X_2, X_3, ..., X_p\}$, the structure of a BN contained in a MEC represented by the CPDAG $\mathcal{G}^* = (\mathbf{X}, \mathbf{E})$ is found by maximizing a function such that:*

$$\mathcal{G}^* = \arg\max_{\mathcal{G} \in \mathcal{G}_C} S(\mathcal{G}, \mathbf{D}) \tag{1}$$

where $S(\mathcal{G}, \mathbf{D})$ is a scoring function that measures the goodness of fit of $\mathbf{D}$ with a candidate MEC $\mathcal{G}$, and $\mathcal{G}_C$ is the set of all candidate MECs defined over $\mathbf{X}$.

GES heuristically searches the structure, under causal sufficient and faithfulness conditions, in the space of Markov equivalence classes in two stages. In each step of the algorithm, every candidate MEC is evaluated, and it is selected the MEC with the highest score that improves the score function. In the first stage, starting with an empty graph, GES adds edges to candidate MECs until a local maximum is reached. Removing edges of the MEC found in the first stage, is performed in the second stage. In each stage of the algorithm, candidate MECs are generated, adding or deleting all possible single edges that yield valid CPDAGs (Alonso-Barba et al., 2013). The algorithm stops when a local maximum is reached and returns the CPDAG that represents the found MEC.

Decomposable and score-equivalent functions are used for evaluating candidate MECs. A scoring function $S(\cdot)$ is **decomposable** if it can be expressed as the product of local functions $S(X_i, \mathbf{Pa}(X_i), \mathbf{D})$, that only depend of a node $X_i \in \mathbf{X}$ and its parents $\mathbf{Pa}(X_i)$. If for any pair of equivalent DAGs $\mathcal{G}$ and $\mathcal{G}'$, a scoring functions $S(\cdot)$ assigns the same score, $S(\mathcal{G}) = S(\mathcal{G}')$, it is **score equivalent**. Decomposable functions allow to evaluate locally candidate MECs, in each

subgraph composed by a node X_i with its parents $\mathbf{Pa}(X_i)$, where the parents of each node are obtained from a DAG included in the candidate MEC. Since the scoring function is score equivalent, any DAG contained in the candidate MEC could be used for evaluating that MEC.

The Bayesian Dirichlet equivalent and Uniform (BDeU) is a decomposable and score equivalent function that evaluate MECs defined over discrete variables with complete data sets $\mathbf{D}$ (without missing values). It is defined as follows (Heckerman et al., 1995):

$$BDeU(\mathcal{G}, \mathbf{D}) = \prod_{i=1}^{p} \{S(X_i, \mathbf{Pa}(X_i), \mathbf{D})\} \tag{2}$$

$$S(X_i, \mathbf{Pa}(X_i), \mathbf{D}) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + C_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + C_{ijk})}{\Gamma(\alpha_{ijk})} \tag{3}$$

Where p is the number of nodes in $\mathcal{G}$, q_i is the number of values of $\mathbf{Pa}(X_i)$, r_i is the number of values of X_i, C_{ijk} is the number of cases in which $X_i = k$ and its parents $\mathbf{pa}(X_i = k) = j$, $C_{ij} = \sum_k C_{ijk}$, and $\alpha_{ijk} = \frac{1}{r_i q_i}$ is a Dirichlet prior parameter with $\alpha_{ij} = \sum_k \alpha_{ijk}$.

4. Knowledge Transfer Learning with Weighted Instances

This section describes an algorithm for learning subject-specific Markov equivalence classes called Knowledge Transfer Learning with Weighted instances GES (KTL-WeGES). It is an extension of the GES algorithm, that using instances of auxiliary data sets tries to improve the identification of Markov equivalence classes learned with limited sample sizes. This problem of knowledge transfer for learning MECs is formally stated as follows:

Definition 2 *Giving a target data set $\mathbf{D_T}$ and a set of M auxiliary data sets $\{\mathbf{D_s}\}, s = 1, .., M$, how to improve a target MEC $\mathcal{G}_T = (\mathbf{X}, \mathbf{E})$ learned using only $\mathbf{D_T}$, transferring instances from the auxiliary data sets.*

Where improve means that the obtained target MEC will have a higher number of correct adjacencies and v-structures than those obtained by an algorithm that only uses all the available data for the target subject. In this definition, we assume that $\mathbf{D_T}$ and $\{\mathbf{D_s}\}$ contain instances of $\mathbf{X}$ that were sampled from different probability distributions, with size $|\mathbf{D_T}| = N_T$, and each $|\mathbf{D_s}| = N_S$; where $N_T << N_S$, and N_T is small such that the instances in $\mathbf{D_T}$ are insufficient to get a good estimation for a target MEC.

KTL-WeGES extends GES for using auxiliary data sets in the evaluation of candidate MECs. Particularly, KTL-WeGES uses knowledge transfer for evaluating locally candidate MECs. In each step of the algorithm, the local score is calculated from a combination of instances of the target data set with weighted instances of auxiliary data sets. This procedure is described in Algorithm 1. KTL-WeGES considers the local evaluation of candidate MECs in two important steps. The first step is the estimation of the weight of each auxiliary data set, and the second is scoring the local structure of a candidate MEC using the weighted instances. We propose using the relevance of each auxiliary data set for defining its weight. Specifically, the weight of each auxiliary data set s is a factor W_s expressing how relevant is the auxiliary dataset for finding the local structure of a target MEC, with greater relevance for values nearly to one. This weight is estimated from the

differences in the conditional probability distribution $P(X_i|\mathbf{Pa}(X_i))$, between the target and the auxiliary sources as follows,

$$W_s = 2^{-|D_{KLD}(P_T(X_i|\mathbf{Pa}_T(X_i)),P_s(X_i|\mathbf{Pa}_T(X_i)))|} \tag{4}$$

where D_{KLD} is the Kullback-Leibler divergence (Campos Ibáñez, 2006) that estimate the difference between P_T and P_s,

$$D_{KLD}(P_T(\cdot), P_s(\cdot))) = \frac{1}{r_i q_i} \sum_{x_i, \mathbf{pa}_T(x_i)} P_T(x_i|\mathbf{pa}_T(x_i)) log\left(\frac{P_T(x_i|\mathbf{pa}_T(x_i))}{P_s(x_i|\mathbf{pa}_T(x_i))}\right) \tag{5}$$

and $\frac{1}{r_i q_i}$ is a normalization factor, over the number of possible configurations for X_i and $\mathbf{Pa}(X_i)$, which avoids increases in D_{KLD} when the number of parents for X_i is increased.

In the second step of the local scoring of candidate MECs, KTL-WeGES uses the following modification of the local BDeU score (defined in Equation 3) for evaluating the goodness of fit of the combination of the auxiliary and target instances with a candidate local structure:

$$S_{KTL}(X_i, \mathbf{Pa}(X_i), \mathbf{D_T}, \{\mathbf{D_s}\}, \{W_s\}) = \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + C'_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + C'_{ijk})}{\Gamma(\alpha_{ijk})} \tag{6}$$

where C'_{ijk} counts the combination of auxiliary and target instances,

$$C'_{ijk} = K\left((C_T)_{ijk} + \sum_s W_s (C_s)_{ijk}\right) \tag{7}$$

with $(C_T)_{ijk}$ and $(C_s)_{ijk}$ represent the number of cases, in $\mathbf{D_T}$ and $\mathbf{D_s}$ respectively, in which $X_i = k$ and its parents $\mathbf{pa}_T(X_i = k) = j$; W_s encodes the relevance of the auxiliary data set s; and K is a factor that normalizes the total number of cases to be $N = N_T + MN_S$. This normalization avoids giving a higher score to configurations with greater weight.

It is important to note that the computational complexity in the estimation of the local score from the target and auxiliary data sets depends on the total sample size N, the number of possible values of X_i and $\mathbf{Pa}(X_i)$, and $|\mathbf{Pa}(X_i)|$, the number of parents of X_i. Considering that X_i and its parents could take at most d values, the computational complexity of this estimation is $O(N(1+|\mathbf{Pa}(X_i|)+ d^{1+|\mathbf{Pa}(X_i)|})$ (Scutari et al., 2019). This expression indicates that KTL-WeGES is limited to work with small MECs which have few nodes (less than twenty).

5. Experimental Results

In this section, we evaluate and compare the performance of the KTL-WeGES with GES, PC and KTL-PC (Knowledge Transfer Learning PC), the knowledge transfer algorithm proposed by Jia et al. (2018). We evaluated the models obtained by GES and PC using only the target data, and these results are compared with those obtained by KTL-WeGES and KTL-PC using the auxiliary and target data sets. In the PC and KTL-PC algorithms, following the proposal of Jia et al. (2018), we use conditional mutual information as the conditional independence test.

Algorithm 1: KTL_WEGES

Algorithm *KTL_WeGES*

Input: the target data set $\mathbf{D_T}$, the set of auxiliary data sets $\{\mathbf{D_s}\}$

Output: CPDAG $\mathcal{G}_{MAX}$

$\mathcal{G}_{MAX} \leftarrow \emptyset$

$bestEdgeModif \leftarrow \emptyset$

$\delta_{MAX} \leftarrow -\infty$

$scoreNew \leftarrow \sum_{X_i \text{ in } \mathcal{G}_{MAX}} localScoreKTL(X_i, \emptyset, \mathbf{D_T}, \{\mathbf{D_s}\})$

$(\mathcal{G}_{MAX}, scoreMax) \leftarrow searchG(dir = adding, \mathbf{D_T}, \{\mathbf{D_s}\}, \mathcal{G}_{MAX}, scoreNew)$

$(\mathcal{G}_{MAX}, scoreMax) \leftarrow searchG(dir = deleting, \mathbf{D_T}, \{\mathbf{D_s}\}, \mathcal{G}_{MAX}, scoreMax)$

return $\mathcal{G}_{MAX}$

Function `searchG`$(dir, \mathbf{D_T}, \{\mathbf{D_s}\}, \mathcal{G}_{MAX}, scoreNew)$

$edgeModification \leftarrow dir$

repeat

$scoreMax \leftarrow scoreNew$

foreach *possible edge modification in* G_{MAX} **do**

$\mathbf{NewPa}(X_i) \leftarrow$ *modify the parents of* X_i *with the edge modification*

$\delta_S \leftarrow localScoreKTL(X_i, \mathbf{NewPa}(X_i), \mathbf{D_T}, \{\mathbf{D_s}\})$

$\delta_S \leftarrow \delta_S - localScoreKTL(X_i, \mathbf{Pa}(X_i), \mathbf{D_T}, \{\mathbf{D_s}\})$

if $\delta_S > \delta_{MAX}$ **then**

$\delta_{MAX} \leftarrow \delta_S$

save edge modification in $bestEdgeModif$

end

end

$scoreNew \leftarrow \delta_{MAX} + scoreMax$

if $scoreNew > scoreMax$ **then**

Update $\mathcal{G}_{MAX}$ *with* $bestEdgeModif$

end

until $(scoreNew <= scoreMax)$

return $\mathcal{G}_{MAX}, scoreMax$

Function `localScoreKTL`$(X, \mathbf{Pa}(X), \mathbf{D_T}, \{\mathbf{D_s}\})$

foreach $\mathbf{D_s}$ **do**

$W_s \leftarrow localRelevance(X, \mathbf{Pa}(X), \mathbf{D_T}, \{\mathbf{D_s}\})$ Equation (4)

end

$score_X \leftarrow S_{KTL}(X, \mathbf{Pa}(X), \mathbf{D_T}, \{\mathbf{D_s}\}, \{W_s\})$

return $score_X$

5.1 Generation of synthetic data sets

Target and auxiliary data sets are generated from ground truth causal Bayesian networks in the following form (Luis et al., 2010). Target data set is sampled from the ground truth causal BN, and auxiliary data sets, from auxiliary causal BNs. Auxiliary causal BNs are generated modifying in certain percent ($pMod$) the edges of the ground truth models, adding $pMod$ edges, followed by deleting edges in the same $pMod$ percent. Increasing $pMod$, we generate auxiliary BNs less related

to the ground truth model. Parameters of each auxiliary BN are estimated using a data set sampled from the ground truth BN. Each data set (target or auxiliary) is sampled from its corresponding BN with forward sampling, in which the values of each variable X_i are sampled in ancestral order (parents before their children), in such form that its values x_i are drawn from $P(x_i|\mathbf{pa}(x_i))$.

5.2 Evaluation metrics

We evaluated our algorithm in its ability for finding the skeleton and v-structures of the ground truth causal BNs. The CPDAGs estimated by our algorithm were compared with those of the ground truth causal BNs. Normalized Hamming distance (NSHD), F-measure, adjacency precision (AP) and recall (AR), arrowhead precision (AHP) and recall (AHR), were used as evaluation metrics. Normalized structural Hamming distance is the minimum number of edge insertions, deletions, and changes needed to transform a model into another. Precision (P) is the ratio $TP/(TP + FP)$, the ratio $TP/(TP + FN)$ is the recall (R), and $2PR/(P + R)$ is the F-measure. For adjacency precision and recall, TP is the number of adjacencies that are in common in the estimated model and ground truth model without considering the edge orientation; FP is the number of adjacencies that are present in the estimated model but not in the ground truth model; and FN is the number of adjacencies that are present in the ground truth model but not in the estimated model (Tillman and Spirtes, 2011). In arrowhead precision and recall, TP represents the number of common edges in the estimated and the ground truth models that share the same orientation; false orientation (FP or FN) is when an oriented edge $X \rightarrow Y$ is present in one model, but in the other one there is $X \leftarrow Y$, $X - Y$, or no edge between X and Y (Jabbari et al., 2018).

5.3 Experiments with Synthetic Causal Bayesian Networks

In this section, we present the experiments performed with synthetic data sets generated from simulated binary discrete causal Bayesian networks. A simulated causal BN is randomly generated with p nodes and with at most $k = p/2$ parents, following the procedure described in (Ide and Cozman, 2002). Parameters of each variable are sampled from uniform distributions.

For our experiments, we randomly simulated six binary causal BNs with $\{5, 6, 7, 8, 9, 10\}$ nodes as ground truth causal BNs. Three auxiliary causal BNs were generated from each simulated causal BN, two of them modifying its edges in 10%, and in 40% the other one. Parameters of auxiliary causal BNs were estimated from a data set with $100(2^{k+2} * (p - 1))$ samples, with $k = p/2$. From each auxiliary causal BN, an auxiliary data set with $N_S = 100(2^{k+1} * (p - 1))$ samples was obtained. Twenty target data sets, from each synthetic causal BN, with different size were generated, starting in the number of nodes p in the causal BN, and increasing by the same factor of p, that is $N_T = \{p, 2p, ..., 20p\}$. These steps were repeated ten times, yielding sixty random causal BNs in total as ground truth models (with twenty target data sets each one) for our experiments.

The CPDAGs obtained by the GES, PC, KTL-We and KTL-PC algorithms were compared with the CPDAGs of ground truth models, and evaluated using the metrics described in Section 5.2. The averages for each metric over all random causal BNs and over all target data sets of each random causal BN, for each algorithm, are summarized in Table 1. From these results, we note first that comparing the base algorithms, PC and GES, GES shows a better performance in adjacency and arrowhead recovery. It seems to indicate that GES has a better performance with small sample sizes, although the results in recall show that in these conditions, GES tends to recover fewer edges. Comparing KTL-PC against its base algorithm, PC, it performs better in adjacency precision, but it

Method	AP ↑	AR ↑	AHP ↑	AHR ↑	NSHD ↓
PC	$0.67(0.05)$	$0.33(0.12)$	$0.37(0.11)$	$0.16(0.05)$	$0.96(0.08)$
KTL-PC	$0.81(0.03)$	$0.21(0.05)$	$0.18(0.08)$	$0.07(0.02)$	$0.76(0.09)$
GES	$0.83(0.03)$	$0.34(0.03)$	$0.36(0.16)$	$0.12(0.02)$	$0.74(0.09)$
KTL-WeGES	$\mathbf{0.93(0.03)}$	$\mathbf{0.94(0.04)}$	$\mathbf{0.88(0.12)}$	$\mathbf{0.75(0.10)}$	$\mathbf{0.35(0.13)}$

Table 1: Averages in adjacency precision (AP) and recall (AR), arrowhead precision (AHP) and recall (AHR); and normalized structural Hamming distance (NSHD) for synthetic causal Bayesian networks. In parenthesis are shown the corresponding standard deviation, and in bold, the best performances. With ↑ are marked, metrics that are better with high values near one, and with ↓, those ones that are better with low values near zero.

has poorer performance in the other evaluation metrics. The results suggest that KTL-PC recovers fewer edges of the ground truth models, hence this knowledge transfer method does not seem to improve the MECs recovering. On the other hand, our proposal KTL-WeGES outperforms all methods. We found, using the Wilcoxon paired signed-rank test, that these observed differences between the methods are statistically significant with a level of significance of 5%. KTL-WeGES, compared to GES, improves adjacency and arrowhead recovery, showing better performance in adjacency precision and recall. In general, the results indicate that KTL-WeGES tends to recover CPDAGs with more true oriented edges than GES. Although, it also tends to include false edges in the CPDAGs that affects its performance in arrowhead recovery.

The performance of KTL-WeGES when varying the number of nodes in the target MEC and the sample size is depicted in Figures 1 and 2, respectively. In Figure 1, the x-axis represents the number of nodes, and the y-axis, the averages of the evaluation metrics over all synthetic causal BNs with p nodes. These plots show good performance in the F-measure, superior and almost similar in all cases for adjacency, slightly decreasing and with higher standard deviation for the arrowhead. The low F-measure, in adjacency and arrowhead, when $p = 6$ indicates that in this case, KTL-WeGES is adding more false edges, increasing the number of false v-structures. The NSHD plot indicates an increase in the difference between the estimated CPDAGs and those corresponding to ground truth causal BNs as the number of nodes increases. Because, in this case, the performance of KTL-WeGES decreases in arrowhead recovery. The x-axis in Figure 2 represents the index of the corresponding sample size N_T in the test set $\{p, 2p, ..., 20p\}$ for target data. The y-axis is the average of the evaluation metric over all synthetic causal BNs in the test set. It can be observed in these plots that all metrics for KTL-WeGES are almost constant across the sample size, increasing the standard deviation when the sample size increases in arrowhead F-measure. These plots suggest that the sample size of the target data set does not appear to influence the performance of KTL-WeGES, because the method gives more importance to auxiliary data sets with a significantly larger sample size than the target data sets.

5.4 Experiments with Benchmark Causal Bayesian Networks

In this section we present the experiments performed with the benchmark discrete causal Bayesian networks: COMA, ASIA and SACHS; available in the Bayesian Network Repository (Scutari, 2012). COMA with five nodes and five edges, ASIA with eight nodes and eight edges, and SACHS with eleven nodes and seventeen edges.

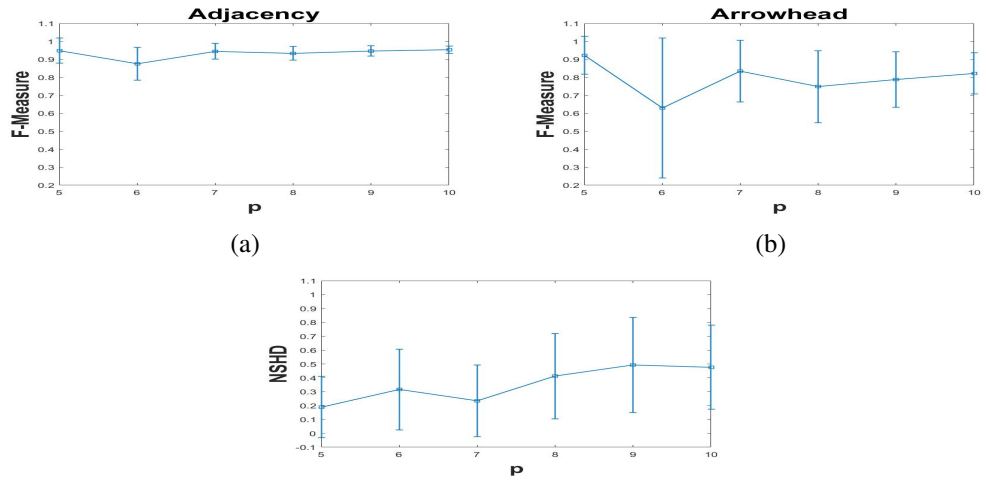

Figure 1: Plot of the averages in (a) F-measure for adjacencies, (b) F-measure for arrowhead, and (c) normalized structural Hamming distance (NSHD) across the number of nodes p. The bars represent the standard deviation of each metric.

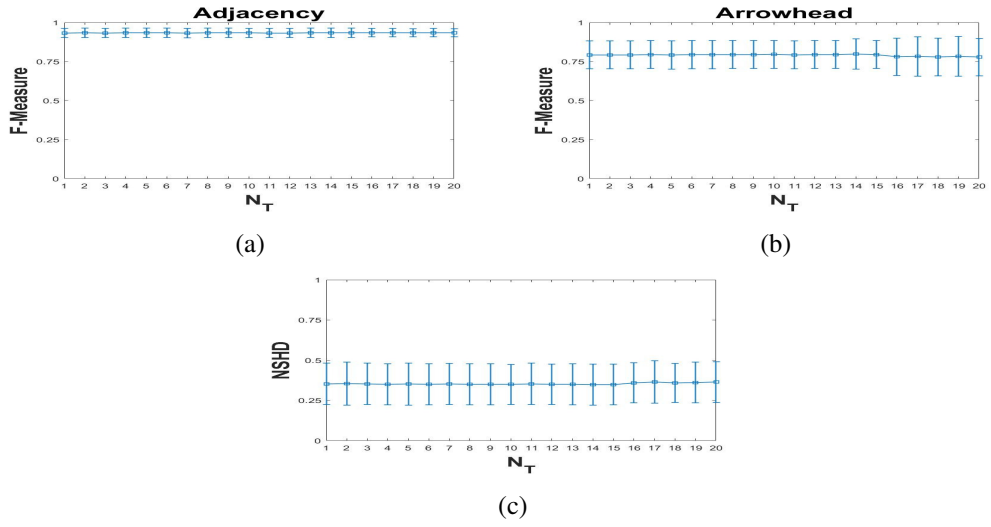

Figure 2: Plot of the averages in (a) F-measure for adjacencies, (b) F-measure for arrowhead, and (c) normalized structural Hamming distance (NSHD) across the target sample size N_T. The bars represent the standard deviation of each metric.

Four auxiliary causal BNs were generated from each benchmark causal BN, modifying them in $\{10\%, 20\%, 30\%, 40\%\}$. Following the same criteria in the generation of synthetic causal BNs, parameters of auxiliary causal BNs were estimated from a data set with $N = 100(\prod_{i=1,\ldots,k+2} r_i)$ samples. This sample size was estimated from the first $k+2$ variables of the causal BN, with the highest number of values r_i. Each auxiliary data set with size $N_S = 100(\prod_{i=1,\ldots,k+1} r_i)$ was sampled from each auxiliary causal BN. From each benchmark causal BN, twenty target data sets

$N_T = \{p, 2p, ..., 20p\}$ were generated. These steps were repeated ten times, yielding in total, for each benchmark causal BN, 200 target data sets for our experiments.

The averages over all target data sets of the evaluations metrics for each benchmark causal BN are summarized in Table 2. The results indicate a similar performance for all methods than those obtained with synthetic causal BNs. Poor performance is observed for PC and KTL-PC, and KTL-PC does not improve the models obtained by PC. Our proposal KTL-WeGES outperforms all methods showing superior performance than with synthetic causal BNs. Considering that there is less variability in the complexity of the structure of the ground truth models, KTL-WeGES shows better performance with benchmark causal BNs than with synthetic ones. The results for SACHS indicate that KTL-WeGES is discovering CPDAGs with false edges, which increases the differences with the ground truth models.

BN	**Method**	**AP** $\uparrow$	**AR** $\uparrow$	**AHP** $\uparrow$	**AHR** $\uparrow$	**NSHD** $\downarrow$
Coma	PC	$0.75(0.11)$	$0.64(0.22)$	$0.00(0.00)$	$0.00(0.00)$	$0.90(0.17)$
	KTL-PC	$0.82(0.23)$	$0.28(0.12)$	$0.00(0.00)$	$0.00(0.00)$	$0.63(0.15)$
	GES	$0.78(0.21)$	$0.37(0.18)$	$0.05(0.22)$	$0.03(0.11)$	$0.61(0.13)$
	KTL-WeGES	$\mathbf{1.00(0.00)}$	$\mathbf{1.00(0.00)}$	$\mathbf{1.00(0.00)}$	$\mathbf{1.00(0.00)}$	$\mathbf{0.00(0.00)}$
Asia	PC	$0.87(0.15)$	$0.39(0.19)$	$0.21(0.36)$	$0.07(0.11)$	$1.21(0.23)$
	KTL-PC	$0.76(0.12)$	$0.37(0.01)$	$0.08(0.20$	$0.04(0.10)$	$0.79(0.08)$
	GES	$0.75(0.11)$	$0.54(0.15)$	$0.43(0.31)$	$0.41(0.33)$	$0.82(0.34)$
	KTL-WeGES	$\mathbf{1.00(0.00)}$	$\mathbf{1.00(0.00)}$	$\mathbf{1.00(0.00)}$	$\mathbf{1.00(0.00)}$	$\mathbf{0.00(0.00)}$
Sachs	PC	$0.47(0.14)$	$0.09(0.03)$	-	-	$0.55(0.06)$
	KTL-PC	$0.76(0.05)$	$0.29(0.05)$	-	-	$0.45(0.03)$
	GES	$0.87(0.05)$	$0.47(0.11)$	-	-	$0.45(0.11)$
	KTL-WeGES	$\mathbf{0.94(0.00)}$	$\mathbf{1.00(0.00)}$	-	-	$\mathbf{0.18(0.00)}$

Table 2: Averages in adjacency precision (AP) and recall (AR), arrowhead precision (AHP) and recall (AHR); and normalized structural Hamming distance (NSHD) for benchmark causal Bayesian networks. In parenthesis are shown the corresponding standard deviation, and in bold the best performances. AHP and AHR are not reported for SACHS because it does not have v-structures. With $\uparrow$ are marked metrics that are better with high values near one, and with $\downarrow$ those ones that are better with low values near zero.

6. Conclusions

In this paper, we propose a knowledge transfer algorithm for learning subject-specific MECs with a limited sample size. Our proposal, an extension of the GES algorithm, considers transferring weighted instances of the auxiliary data sets to alleviate the lack of enough data on the target and to find the local structure of the target MECs. To estimate the weights, we introduced a strategy based on the local difference between the probability distributions of the target and the auxiliary sources.

Our experimental results, over synthetic and benchmarks causal BNs, suggest that our strategy of leveraging weighted instances of the auxiliary data sets seems to work for recovering MECs with higher quality than those discovered by the baseline methods, and an alternative knowledge transfer approach. Our proposal outperforms in adjacency and arrowhead recovery all the compared me-

thods: GES, PC, and KTL-PC (Jia et al., 2018), in all tested BNs and regardless of the sample size of target data sets.

This paper shows the results obtained from our first experiments using auxiliary data sets of the same size. However, our proposal, with minimal changes, could work with auxiliary data sets of different sizes. In the future, we plan to extend our proposal for discovering MECs of high dimensionality and with continuous nodes. We also contemplate including strategies for improving the arrowhead recovery.

Acknowledgments

This work was partially supported by CONACYT under project No. CB2017-2018 43346. The first author acknowledges scholarship support from CONACYT as a PhD student.

References

J. I. Alonso-Barba, J. A. Gámez, J. M. Puerta, et al. Scaling up the greedy equivalence search algorithm by constraining the search space of equivalence classes. *International Journal of Approximate Reasoning*, 54(4):429–451, 2013.

L. M. Campos Ibáñez. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. *Journal of Machine Learning Research*, 7(Oct): 2149–2187, 2006.

D. M. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

T. Claassen and T. Heskes. Causal discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems*, pages 415–423, 2010.

G. Cooper, C. Cai, and X. Lu. Tumor-specific causal inference (TCI): A Bayesian method for identifying causative genome alterations within individual tumors. *bioRxiv*, page 225631, 2018.

C. Glymour, K. Zhang, and P. Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:1–15, 2019.

C. Grefkes and G. R. Fink. Connectivity-based approaches in stroke and recovery of function. *The Lancet Neurology*, 13(2):206–216, 2014.

D. Heckerman, D. Geiger, and D. M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

J. S. Ide and F. G. Cozman. Random generation of Bayesian networks. In *Brazilian Symposium on Artificial Intelligence*, pages 366–376. Springer, 2002.

F. Jabbari, S. Visweswaran, and G. F. Cooper. Instance-specific Bayesian network structure learning. In *International Conference on Probabilistic Graphical Models*, pages 169–180, 2018.

H. Jia, Z. Wu, J. Chen, B. Chen, and S. Yao. Causal discovery with Bayesian networks inductive transfer. In *International Conference on Knowledge Science, Engineering and Management*, pages 351–361. Springer, 2018.

R. Luis, L. E. Sucar, and E. F. Morales. Inductive transfer for learning Bayesian networks. *Machine Learning*, 79(1-2):227–255, 2010.

D. Malinsky and D. Danks. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470, 2018.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *arXiv preprint arXiv:1611.10351*, 2019.

A. Niculescu-Mizil and R. Caruana. Inductive transfer for Bayesian network structure learning. In *Artificial Intelligence and Statistics*, pages 339–346, 2007.

D. Oyen and T. Lane. Bayesian discovery of multiple Bayesian networks via transfer learning. In *2013 IEEE 13th International Conference on Data Mining*, pages 577–586. IEEE, 2013.

S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

J. D. Ramsey, S. J. Hanson, C. Hanson, Y. O. Halchenko, R. A. Poldrack, and C. Glymour. Six problems for causal inference from fMRI. *Neuroimage*, 49(2):1545–1558, 2010.

M. Scutari. Bayesian network repository, 2012.

M. Scutari, C. Vitolo, and A. Tucker. Learning Bayesian networks from big data with greedy search: computational complexity and efficient implementation. *Statistics and Computing*, pages 1–14, 2019.

P. Spirtes and K. Zhang. Causal discovery and inference: Concepts and recent methodological advances. In *Applied Informatics*, volume 3, pages 1–28. SpringerOpen, 2016.

P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper, and T. Richardson. *Causation, prediction, and search*. MIT press, 2000.

L. E. Sucar. *Probabilistic Graphical Models*. Advances in Computer Vision and Pattern Recognition; Springer: London, UK, 2015.

R. Tillman and P. Spirtes. Learning equivalence classes of acyclic models with latent and selection variables from multiple datasets with overlapping variables. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 3–15, 2011.

R. E. Tillman and F. Eberhardt. Learning causal structure from multiple datasets with similar variable sets. *Behaviormetrika*, 41(1):41–64, 2014.

T. Verma and J. Pearl. *Equivalence and synthesis of causal models*. UCLA, Computer Science Department, 1991.

K. Zhang, B. Schölkopf, P. Spirtes, and C. Glymour. Learning causality and causality-related learning: some recent progress. *National Science Review*, 5(1):26–29, 2018.