Learning from Mixture of Experimental Data: A Constraint–Based Approach

Vincenzo Lagani[1], Ioannis Tsamardinos[1,2], and Sofia Triantafillou[1,2,⋆]

[1] BioInformatics Laboratory - FORTH-ICS
Vassilika Vouton 100, Heraklion - Greece
[2] Computer Science Department, University of Crete
Knossou Ave., Heraklion - Greece

Abstract. We propose a novel approach for learning graphical models when data coming from different experimental conditions are available. We argue that classical constraint–based algorithms can be easily applied to mixture of experimental data given an appropriate conditional independence test. We show that, when perfect statistical inference are assumed, a sound conditional independence test for mixtures of experimental data can consist in evaluating the null hypothesis of conditional independence separately for each experimental condition. We successively indicate how this test can be modified in order to take in account statistical errors. Finally, we provide "Proof-of-Concept" results for demonstrating the validity of our claims.

Keywords: Graphical Models, Mixture of Experimental data, Conditional independence test, Constraint Based learning.

1 Introduction

Graphical models are mathematical tools that have become widely known in the last decades. Structural Equation Models (SEM), Hidden Markov Models (HMM), Bayesian Networks (just to name the most common examples) are currently employed for addressing a wide range of real world applications, e.g. text recognition, information retrieval, gene regulatory network reconstruction. Despite years of research, when it comes to learning graphical models from data, there are still several open–to–debate issues; a particularly challenging problem is dealing with experimental interventions that alter the distribution of the data.

Experimental interventions are commonly employed in any area of scientific research. Patient randomization during clinical trials, as well as gene knock–outs in gene expression studies are prominent examples of experimental manipulations, that are usually essential for confirming scientific hypotheses. The same system must often been analyzed under different experimental conditions, to better investigate its operation. Unfortunately, standard graphical–model learning

[⋆] This research was partially supported by the EU funded project REACTION, Grant Agreement no. 248590.

I. Maglogiannis, V. Plagianakos, and I. Vlahavas (Eds.): SETN 2012, LNAI 7297, pp. 124–131, 2012.

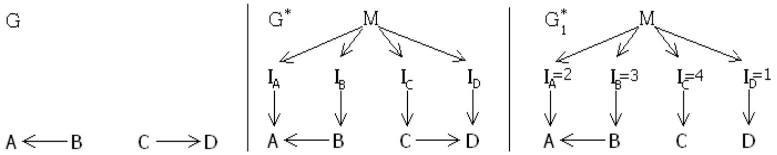

Fig. 1. A simple DAG (left), the augmented graph $\mathcal{G}^*$ (center) and the DAG corresponding to the experiment $E_1 = \{I_A = 2, I_B = 3, I_C = 4, I_D = 1\}$ (right). $I_D = 1$ is a hard intervention, causing the deletion of the edges pointing to D.

algorithms can not be directly employed on mixtures of experimental data; the naïve solution of pooling all data together can lead to spurious (in)dependencies [1], that negatively affect the learning performances. This means that a huge amount of scientific data can not be analyzed with current graphical–model learning methods.

We argue that *any constraint–based algorithm, in principle, can be extended for data coming from different experimental conditions.* Constraint–based methods are a class of algorithms for learning graphical models that have recently proved to be particularly effective[2,3]. This type of algorithms are built on the basis of conditional independence tests; depending on the embedded test, the same algorithm can deal with different types of data [4]. In this work, we employ tests specifically devised for mixtures of experimental data within algorithms originally conceived for data from a single experimental condition. This approach significantly differs from previous works, which are mainly based on a) computationally expensive Search–and–Score paradigms [5], or b) constraint–based algorithms that are specific for a particular experimentation protocol and can not address the general case [6].

To the best of our knowledge, no conditional independence test for experimental data has been proposed yet. We suggest a conceptually simple test: the null hypothesis "variables X and Y are independent given the conditioning set Z" is separately tested for each experiment, and the null hypothesis is rejected if X and Y are found associated at least once. In the next sections we demonstrate that this simple procedure is sound when perfect statistical inferences are assumed. Moreover, we propose an alternative, practical procedure for taking into account both type II and type I statistical errors. Our experimental results indicate that adopting our conditional independence tests leads to better results than naïvely pooling data from different experimental conditions.

2 Notation and Problem Statement

Simply put, a sound conditional independence test for mixtures of experimental data should be able to detect the conditional (in)dependencies characterizing the passive obeservational case on the basis of datasets sampled under various experimental conditions. Section 3 introduces a simple procedure, namely the

Basic approach, that fulfils this objective. A short introduction to the basic theories used to develop this method is presented in this section.

We assume that the data-generating procedure can be described by a Bayesian Network. Let $\mathcal{G} = \{\mathbf{N}, \mathbf{A}\}$ be a Directed Acyclic Graph (DAG), where $\mathbf{A}$ is a set of oriented edges and $\mathbf{N} = \{N_1, \ldots, N_n\}$ a set of nodes, each one representing a variable. Under the Markov Condition and the Faithfulness Assumption, $\mathcal{G}$ encodes the set of (in)dependencies of the joint probability distribution (JPD) of the set of variables $\mathbf{N}$ according to a graphical criterion, called *d*-separation [7]. We consider two types of interventions: hard (surgical) interventions, in which the manipulated variables' values are set solely by the experimental procedure, and therefore all incoming edges to the manipulated variable are removed from the graph; and soft interventions, in which the skeleton of the graph remains intact, and the parameters of the distribution of the manipulated variables are altered by the experimental procedure [8].

When we pool together data sampled under different conditions, we obtain a new JPD which encodes certain (in)dependencies. We define $\mathcal{G}^* = \{\mathbf{N}, \mathbf{A}^*, \mathbf{I}\,\mathcal{M}\}$ as the DAG representing these independencies. Node $\mathcal{M}$ corresponds to the "manipulating" variable $\mathcal{M} \in \{1, 2, \ldots\ m\}$ representing the scientist performing the m different experiments. Nodes $\mathbf{I}$ correspond to the interventional variables $\mathbf{I} = \{I_1, \ldots, I_n\}$ representing the manipulations that the scientist performs on each variable. Variable I_i can take q_i integer values $D(I_i) = \{I_i^1, I_i^2, \ldots, I_i^{q_i}\}$, each corresponding to a different manipulation of the distribution of N_i. $\mathbf{A}^* = \{\mathbf{A} \cup \{\mathcal{M} \rightarrow I_i\}_{i=1}^n \cup \{I_i \rightarrow N_i\}_{i=1}^n\}$(see Fig. 1).

Let $\{\mathbf{E}_k\}_{k=1}^m$ be a set of m different experiments. An experiment $\mathbf{E}_k$ corresponds to the k-th value of $\mathcal{M}$, and is fully identified by the fixed values that the interventional variables take during its execution. Moreover, each $\mathbf{E_k}$ is related to a DAG $\mathcal{G}_k^*$, obtained from $\mathcal{G}^*$ after removing the edges pointing to variables targeted by hard interventions. D_k is the dataset sampled/produced during the experiment $\mathbf{E}_k$, while P_k is the JPD of the variables in D_k. Similarly, the distribution over the pooled dataset $D_T = \bigcup_{k=1}^m D_k$ is indicated as P_T.

We use $(\neg)ind_k(N_i, N_j|\mathbf{C})$ to denote that "N_i and N_j are (not) independent in P_k given $\mathbf{C}$", where $N_i, N_j \in \mathbf{N}$ and $\mathbf{C} \subseteq \mathbf{N} \setminus \{N_i, N_j\}$. By convention, a conditional independence that holds in the observational case is indicated as $ind(N_i, N_j|\mathbf{C})$. Finally, a generic statistical procedure that evaluates the null hypothesis "N_i and N_j are independent given $\mathbf{C}$ in the data distribution P_k" is indicated as $TestInd(N_i, N_j|\mathbf{C}; P_k)$.

3 Assuming Perfect Statistical Inferences: Basic Approach

Let $TestInd_{Oracle}(N_i, N_j|\mathbf{C}; P_k)$ be an *oracle*, i.e. a conditional independence test that makes no statistical errors. We now define the conditional independence test for a mixture of experimental data $TestInd_{Basic}(N_i, N_j|\mathbf{C}; P_1 \ldots P_m)$:

Basic approach. *Let $D_1, \ldots, D_m$ be m datasets sampled under different experimental conditions, and $P_1, \ldots, P_m$ the respective joint probability distributions. $TestInd_{Basic}(N_i, N_j|\mathbf{C}; P_1 \ldots P_m)$ rejects the null (independence)* *hypothesis iff* $TestInd_{Oracle}(N_i, N_j|\mathbf{C}; P_k)$ *rejects the null hypothesis for at least one* $P_k, k = 1, \ldots, m$.

It is easy to demonstrate that $TestInd_{Basic}$ detects the *preserved* conditional dependencies entailed in $\mathcal{G}$:

Proposition 1. *Given the set of experiments* $\mathbf{E}_1, \ldots, \mathbf{E}_m$*, the conditional dependency* $\neg ind(N_i, N_j|\mathbf{C})$ *is* preserved *if at least one* $\neg ind_k(N_i, N_j|\mathbf{C}), k = 1, \ldots, m$ *holds.*

If the dependency is preserved, then $TestInd_{Oracle}$ rejects the null hypothesis at least once, and thus $TestInd_{Basic}$ also rejects its null hypothesis. Conversely, if $ind(N_i, N_j|\mathbf{C})$ holds, then $TestInd_{Basic}$ will accept the null hypothesis of independence, because (a) we assume that the oracle does not perform Type I statistical errors and (b) our settings ensure that *no spurious association can be created within an experiment by experimental manipulations*, as described in the following theorem.

Theorem 1. *No dependency* $\neg ind_k(N_i, N_j|\mathbf{C})$ *can hold if* $\neg ind(N_i, N_j|\mathbf{C})$ *does not hold in the observational case.*

Proof. W assume faithfulness and Markov condition for both $\mathcal{G}$ and $\mathcal{G}^*_k$, thus a spurious association in D_k can be created iff $\mathcal{G}^*_k$ encodes an artificial d-connecting path not present in $\mathcal{G}$. Such an artificial d-connecting path should either (a) be encoded in the part of the $\mathcal{G}^*_k$ structure that is in common with $\mathcal{G}$ or (b) pass through the interventional nodes and $\mathcal{M}$. Both cases are not possible, because (a) soft interventions do not change $\mathcal{G}$ structure, and deletion of arcs due to surgical interventions can only destroy d-connecting paths; (b) the values of the interventional variables are held constant during each experiment, and this implies that $ind_k(N_i, N_j|\mathbf{C}) \equiv ind_k(N_i, N_j|\mathbf{C}, I_1 = I^k_1, \ldots, I_n = I^k_n)$, where $\{I_1 = I^k_1, \ldots, I_n = I^k_n\}$ are the values assumed by the interventional variables in the k-th experiment. Conditioning on all interventional variables blocks all d-connecting paths that pass through the interventional nodes, thus excluding any spurious association. □

4 Considering Statistical Errors: Merging Approach

The *Basic* approach depends on a number of assumptions that affect its practical applicability. In a more realistic setting, type II statistical errors are possible, i.e. the statistical power may be low. A possible solution for increasing the statistical power may consist in merging different datasets; however, when data from different experimental conditions are pooled together, Theorem 1 does not hold anymore, and *spurious associations not encoded in* $\mathcal{G}$ *may be created.*

We now define a sufficient condition that ensures the absence of spurious associations in mixtures of data from different experiments. Let $D_\mathbf{u}, \mathbf{u} \subseteq \{1, \ldots, m\}$ be the pooled dataset from a subset of the experiments $\mathbf{E}_1, \ldots, \mathbf{E}_m$. $P_\mathbf{u}$ is the joint probability distribution of $D_\mathbf{u}$. We furthermore define the set of experimental modifications $\mathbf{S}_\mathbf{u}$ as the set of the interventional variables whose values change, even once, across the experiments pooled in $D_\mathbf{u}$.

Theorem 2. *The probability distribution $P_{\mathbf{u}}$ entails no spurious association iff $|\mathbf{S}_{\mathbf{u}}| \leq 1$.*

Proof. When $|\mathbf{S}_{\mathbf{u}}| \geq 2$, at least two interventional variables, namely I_r and I_s, are not conditioned upon anymore, and thus at least one *d*-connecting path, namely $N_r \leftarrow I_r \leftarrow \mathcal{M} \rightarrow I_s \rightarrow N_s$, is present.

When $|\mathbf{S}_{\mathbf{u}}| = 1$, all interventional variables are constant, except one, namely I_r. However, no *d*-connecting path between two nodes $N_i, N_j \in \mathcal{G}$ can pass through I_r, because such a path would be blocked by the other interventional variables that are all conditioned upon.

When $|\mathbf{S}_{\mathbf{u}}| = 0$ the data were produced under the same experimental conditions, i.e., results of Theorem 1 still hold. □

The results of Theorem 2 allow us to define the following conditional independence test for mixture of experimental data, namely the *S–Merging* approach:

S–Merging approach. *Given m experiments $\mathbf{E}_1, \ldots, \mathbf{E}_m$ and their respective JPDs $P_1, \ldots, P_m$, the test $TestInd_{S-Merging}(N_i, N_j|\mathbf{C}; P_1, \ldots, P_m)$ rejects the null hypothesis of independence iff $TestInd_{Oracle}(N_i, N_j|\mathbf{C}; P_{\mathbf{u}})$ rejects the null hypothesis of independence for any $P_{\mathbf{u}}$ with $|\mathbf{S}_{\mathbf{u}}| \leq 1$, $\mathbf{u} \subseteq \{1, \ldots, m\}$.*

The *S–Merging* approach tests the existence of a dependency in any single dataset *and* in any pooled dataset $D_{\mathbf{u}}$ where $|\mathbf{S}_{\mathbf{u}}| \leq 1$ (i.e., the tests performed by the *Basic* approach are always a subset of the tests performed by the *S–Merging* approach). Merging different distributions does not ensure an increment of statistical power; under this respect, the *S–Merging* approach is clearly heuristic: it *tries* to maximize the available statistical power by evaluating any subset of datasets that can be pooled together without creating spurious associations.

Finally, both the *Basic* and *S–Merging* approaches internally perform multiple statistical inferences, increasing the probability of type I statistical errors. We employ a Family Wise Error Rate (FWER) correction procedure, namely the Holm-Bonferroni method [9]. More sophisticated procedures could be adopted for correcting for multiple tests.

5 Experiments

We considered the following experimental scenario for evaluating our approaches: the observational case E_1, two experiments E_2 and E_3 with 5 randomly chosen manipulated variables each, and two experiments E_4 and E_5 such that $|\mathbf{S}_{\{2,4\}}| = 1$ and $|\mathbf{S}_{\{3,5\}}| = 1$ (i.e. the experiments within each couple differ between each other only for a single intervention). We only considered hard (surgical) interventions. We employed three prototypical Bayesian Networks, namely ALARM, INSURANCE and HAILFINDER, for generating synthetic discrete data. For each network we simulated 6 mixtures of experimental data, by varying the single–experiment sample size among $\{50, 100, 300, 500, 700, 1000\}$.

On each mixture of experimental data we applied the well known, constraint–based PC algorithm [7,10], with in turn the $TestInd_{Basic}$ and $TestInd_{S-Merging}$

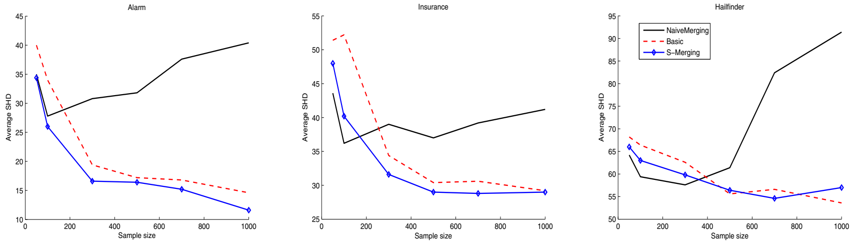

Fig. 2. Results of the experiments on synthetic data. Each graph refers to a different network (from left to right: ALARM, INSURANCE and HAILFINDER). The x axis reports the sample size for each experiment, while the y axis the SHD values averaged over 5 repetitions. Each line represents a different approach, respectively: *NaïveMerging* (black solid line), *Basic* (red dashed line) and *S–Mergin* (blue line with diamonds).

tests. The G^2 conditional independence test was employed internally, and the *fan–in* parameter of the PC algorithm was set to 3. Furthermore, we applied the PC algorithm equipped with the G^2 test on all data pooled together; we call this simple procedure the *NaïveMerging* approach. We employed the *NaïveMerging* approach in order to demonstrate that indiscriminately pooling data from different experimental conditions leads to systematic errors. The whole procedure was repeated 5 times, with significance threshold always set to 0.05.

The Structural Hamming Distance (SHD) was employed for comparing the Partially Directed Acyclic Graph (PDAG) provided by the PC method with the Complete Partially Directed Acyclic Graphs (CPDAG) corresponding to the structures of the three networks [11]. The SHD metric has an intuitive interpretation: it indicates the number of arcs that must be added, deleted, reversed or oriented in order to transform a partial directed graph into another one.

The results of our analysis are summarized in Fig. 2. $TestInd_{S-Merging}$ usually allows a better reconstruction of the true CPDAG than $TestInd_{Basic}$; this result indicates that the additional tests performed by the $S - Merging$ approach are effective in order to retrieve the true dependencies, i.e., merging data coming from different experimental conditions can lead to an increment of statistical power (given that the condition $|\mathbf{S_u}| \leq 1$ stated in Theorem 2 is respected). Moreover, both the $S - Merging$ and *Basic* approach show better performance when the sample size increments. Conversely, the performance of the *NaïveMerging* approach decreases with the increment of the sample size. This trend was expected: the spurious associations created by pooling all data together become stronger as more samples are available. Thus, the PC algorithm retrieves an increasing number of false associations, and these errors are "propagated" through the network.

6 Discussion

This work constitutes a first step towards the creation of a new class of graphical–model learning algorithms for mixtures of experimental data. Our intuition is conceptually simple: constraint–based methods, in principle, can be applied on experimental data, by simply coupling a suitable conditional independence test.

Following our intuition we provided the first conditional independence tests for mixtures of experimental data, $TestInd_{Basic}$ and $TestInd_{S-Merging}$. Our tests deal with datasets sampled under different experimental conditions, and attempt to retrieve the conditional (in)dependencies entailed in the observational data distribution. While $TestInd_{Basic}$ relies on the assumption of perfect statistical inferences, $TestInd_{S-Merging}$ is devised in order to avoid type II statistical errors by maximally exploiting all the available statistical power.

Furthermore, we provided a sufficient condition (Theorem 2) for avoiding the creation of spurious associations when data from different experiments are pooled together. Even though the rule $|\mathbf{S_u}| \leq 1$ can seem quite strict, this condition has interesting potential applications. For example, it demonstrates that a medical study where patients are randomized between two groups can be safely merged with a successive, follow up observational study carried on the same patients, for increasing the statistical power of the analysis.

Finally, the experimental results obtained with the PC algorithm seem to confirm the validity of our methods. Both the *Basic* and *S–Merging* approaches outperform the simplistic solution of pooling all data together; as the sample size increases, the spurious associations become stronger, and the difference among the approaches becomes more evident. Moreover, the *S–Merging* approach demonstrated to be usually more powerful than the *Basic* one, as expected.

The class of constraint–based algorithms is particularly large, and different algorithms show different interesting features, e.g. the possibility of learning rich causal models like Maximal Ancestral Graphs (MAGs, [12]), or the possibility of learning only part of the structure [11]. Our further researches will keep exploring the possibility of extending constraint–based methods for mixtures of experimental data.

7 Related Work

A possible approach for learning graphical models from different experiments consists in learning a first skeleton of the graph from observational data and then to exploit external interventions for orienting edges [13]. These methods consider each dataset in isolation, and thus underutilize the available information, and cannot be employed in absence observational data. Search-and-Score methods in conjunction with modified score functions [5,8] have also been employed for learning from mixtures of experimental data. The main drawback of these algorithms is that Search-and-Score procedures are usually highly computationally demanding. Other algorithms assume that interactions among variables can be represented with a specific type of function (e.g., noisy OR functions among binary variables [14]), but they are applicable only when their respective, strict assumptions hold. Constraint–based algorithms were also proposed for learning from multiple experiments. An algorithm for learning causal models in systems where only one variable is manipulated at a time (per dataset) is proposed in [6]. This algorithm is not general as it can only address cases that follow a specific experimental process. A method for inferring causal relations from (in)dependence models derived from different experiments was proposed in [15]. However, this approach can not be applied in presence of hard interventions.

Finally, to the best of our knowledge only one work identifies a sufficient condition for pooling together data from different experiments [1]. However, checking this condition requires the knowledge of the underlying causal structure, that is almost always unknown.

References

1. Eberhardt, F.: Sufficient condition for pooling data from different distributions. In: ERROR (2006)
2. Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. J. Mach. Learn. Res. 11, 171–234 (2010)
3. Aliferis, C.F., Statnikov, A., Tsamardinos, I., Mani, S., Koutsoukos, X.D.: Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part II: Analysis and Extensions. J. Mach. Learn. Res. 11, 235–284 (2010)
4. Lagani, V., Tsamardinos, I.: Structure-based variable selection for survival data. Bioinformatics 26(15), 1887–1894 (2010)
5. Cooper, G.F., Yoo, C.: Causal Discovery from a Mixture of Experimental and Observational Data. In: UAI (1999)
6. Tian, J., Pearl, J.: Causal discovery from changes. In: UAI (2001)
7. Pearl, J.: Causality: Models, Reasoning, and Inference. Cambridge University Press (March 2000)
8. Eaton, D., Murphy, K.: Exact bayesian structure learning from uncertain interventions. In: AISTAT (2007)
9. Holm, S.: A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics 6(2), 65–70 (1979)
10. Murphy, K.P.: The Bayes Net Toolbox for MATLAB
11. Tsamardinos, I., Brown, L., Constantin, A.: The max-min hill-climbing Bayesian network structure learning algorithm. Machine Learning 65(1), 31–78 (2006)
12. Richardson, T., Spirtes, P.: Ancestral Graph Markov Models. The Annals of Statistics 30(4), 962–1030 (2002)
13. He, Y.-B., Geng, Z.: Active Learning of Causal Networks with Intervention Experiments and Optimal Designs. Journal of Machine Learning Research 9, 2523–2547 (2008)
14. Hyttinen, A., Hoyer, P.O., Eberhardt, F.: Noisy-OR Models with Latent Confounding. In: UAI (2011)
15. Claassen, T., Heskes, T.: Learning causal network structure from multiple (in)dependence models. In: PGM (2010)