Journal of Machine Learning Research 13 (2012) 1097-1157 Submitted 3/11; Revised 11/11; Published 4/12

Towards Integrative Causal Analysis of Heterogeneous Data Sets and Studies

Ioannis Tsamardinos[*] TSAMARD@ICS.FORTH.GR
Sofia Triantafillou[*] STRIANT@ICS.FORTH.GR
Vincenzo Lagani VLAGANI@ICS.FORTH.GR
Institute of Computer Science
Foundation for Research and Technology - Hellas (FORTH)
N. Plastira 100 Vassilika Vouton
GR-700 13 Heraklion, Crete, Greece

Editor: Chris Meek

Abstract

We present methods able to predict the presence and strength of conditional and unconditional dependencies (correlations) between two variables Y and Z *never jointly measured* on the same samples, based on multiple data sets measuring a set of common variables. The algorithms are specializations of prior work on learning causal structures from overlapping variable sets. This problem has also been addressed in the field of *statistical matching*. The proposed methods are applied to a wide range of domains and are shown to accurately predict the presence of thousands of dependencies. Compared against prototypical statistical matching algorithms and within the scope of our experiments, the proposed algorithms make predictions that are better correlated with the sample estimates of the unknown parameters on test data ; this is particularly the case when the number of commonly measured variables is low.

The enabling idea behind the methods is to induce one or all *causal* models that are simultaneously consistent with (fit) all available data sets and prior knowledge and reason with them. This allows constraints stemming from causal assumptions (e.g., Causal Markov Condition, Faithfulness) to propagate. Several methods have been developed based on this idea, for which we propose the unifying name Integrative Causal Analysis (INCA). A contrived example is presented demonstrating the theoretical potential to develop more general methods for co-analyzing heterogeneous data sets. The computational experiments with the novel methods provide evidence that causally-inspired assumptions such as Faithfulness often hold to a good degree of approximation in many real systems and could be exploited for statistical inference. Code, scripts, and data are available at www.mensxmachina.org.

Keywords: integrative causal analysis, causal discovery, Bayesian networks, maximal ancestral graphs, structural equation models, causality, statistical matching, data fusion

1. Introduction

In several domains it is often the case that several data sets (studies) may be available related to a specific analysis question. Meta-analysis methods attempt to collect, evaluate and combine the results of several studies regarding a single hypothesis. However, studies may be heterogeneous in

[*] Also in Department of Computer Science, University of Crete.

several aspects, and thus not amenable to standard meta-analysis techniques. For example, different studies may be measuring different sets of variables or under different experimental conditions.

One approach to allow the co-analysis of heterogeneous data sets in the context of prior knowledge is to try to induce one or all *causal* models that are simultaneously consistent with all available data sets and pieces of knowledge. Subsequently, one can reason with this set of consistent models. We have named this approach *Integrative Causal Analysis* (INCA).

The use of *causal* models may allow additional inferences than what is possible with non-causal models. This is because the former employ additional assumptions connecting the concept of causality with observable and estimable quantities such as conditional independencies and dependencies. These assumptions further constrain the space of consistent models and may lead to new inferences. Two of the most common causal assumptions in the literature are the Causal Markov Condition and the Faithfulness Condition (Spirtes et al., 2001); intuitively, these conditions assume that the observed dependencies and independencies in the data are due to the causal structure of the observed system and not due to accidental properties of the distribution parameters (Spirtes et al., 2001). Another interpretation of these conditions is that the set of independencies is stable to small perturbations of the joint distribution (Pearl, 2000) of the data.

The idea of inducing causal models from several data sets has already appeared in several prior works. Methods for inducing causal models from samples measured under different experimental conditions are described in Cooper and Yoo (1999), Tian and Pearl (2001), Claassen and Heskes (2010), Eberhardt (2008); Eberhardt et al. (2010) and Hyttinen et al. (2011, 2010). Other methods deal with the co-analysis of data sets defined over different variable sets (Tillman et al., 2008; Triantafillou et al., 2010; Tillman and Spirtes, 2011). In Tillman (2009) and Tsamardinos and Borboudakis (2010) approaches that induce causal models from data sets defined over semantically similar variables (e.g., a dichotomous variable for Smoking in one data set and a continuous variable for Cigarettes-Per-Day in a second) are explored. Methods for inducing causal models in the context of prior knowledge also exist (Angelopoulos and Cussens, 2008; Borboudakis et al., 2011; Meek, 1995; Werhli and Husmeier, 2007; O'Donnell et al., 2006). INCA as a unifying common theme was first presented in Tsamardinos and Triantafillou (2009) where a mathematical formulation is given of the co-analysis of data sets that are heterogeneous in several of the above aspects. In Section 3, we present a contrived example demonstrating the theoretical potential to develop such general methods.

In this paper, we focus on the problem of analyzing data sets defined over different variable sets, as proof-of-concept of the main idea. We develop methods that could be seen as special cases of general algorithms that have appeared for this problem (Tillman et al., 2008; Triantafillou et al., 2010; Tillman and Spirtes, 2011). The methods are able to predict the presence and strength of conditional and unconditional dependencies (correlations) between two variables Y and Z *never jointly measured* on the same samples, based on multiple data sets measuring a set of common variables.

To evaluate the methods we simulate the above situation in a way that it becomes testable: a single data set is partitioned to three data sets that do not share samples. A different set of variables is excluded from each of the first two data sets, while the third is hold out for testing. Based on the first two data sets the algorithms predict certain pairs of the excluded variables should be dependent. These are then tested in the third test set containing all variables.

The proposed algorithms make numerous predictions that range in the thousands for large data sets; the predictions are highly accurate, significantly more accurate than predictions made at random. The methods also successfully predict certain conditional dependencies between pairs of variables Y, Z never measured together in a study. In addition, when linear causal relations and Gaussian error terms are assumed, the algorithms successfully predict the strength of the linear correlation between Y and Z. The latter observation is an example where the INCA approach can give rise to algorithms that provide quantitative inferences (strength of dependence), and are not limited to qualitative inferences (e.g., presence of dependencies).

Inferring the correlation between Y and Z in the above setting has also been addressed by *statistical matching* algorithms (D'Orazio et al., 2006), often found under the name of data fusion in Europe. Statistical matching algorithms make predictions based on parametric distributional assumptions, instead of causally-inspired assumptions. We have implemented two prototypical statistical matching algorithms and performed a comparative evaluation. Within the scope of our experiments, the proposed algorithms make predictions that are better correlated with the sample estimates of the unknown parameters on test data; this is particularly the case when the number of commonly measured variables is low. In addition, the proposed algorithms make predictions in cases where some statistical matching procedures fail to do so and vice versa, and thus, the two approaches can be considered complementary in this respect.

There are several philosophical and practical implications of the above results. First, the results provide ample statistical evidence that some of the typical assumptions employed in causal modeling hold abundantly (at least to a good level of approximation) in a wide range of domains and lead to accurate inferences. *To obtain the results the causal semantics are not employed per se*, that is, we do not predict the effects of experiments and manipulations. In other words, one could view the assumptions made by the causal models as constraints or priors on probability distributions encountered in Nature without any reference to causal semantics.

Second, the results point to the utility and potential impact of the approach: co-analysis provides novel inferences as a norm, not only in contrived toy problems or rare situations. Future INCA-based algorithms that are able to handle all sorts of heterogeneous data sets that vary in terms of experimental conditions, study design and sampling methodology (e.g., case-control vs. i.i.d. sampling, cross-sectional vs. temporal measurements) could potentially one day enable the automated large-scale integrative analysis of a large part of available data and knowledge to construct causal models.

The rest of this document is organized as follows: Section 2 briefly presents background on causal modeling with Maximal Ancestral Graphs. Section 3 discusses the scope and vision of the INCA approach. Section 4 presents the example scenario employed in all evaluations. Section 5 formalizes the problem of co-analysis of data sets measuring different quantities. Sections 6 and 7 present the algorithms and their comparative evaluation for predicting unconditional and conditional dependencies respectively, between variables not jointly measured. Section 8 extends the theory to devise an algorithm that can also predict the strength of the dependence. Section 9 presents the statistical matching theory and comparative evaluation. The paper concludes with Section 10 and 11 discussing the related work and the paper in general.

2. Modeling Causality with Maximal Ancestral Graphs

Maximal Ancestral Graphs (MAGs) is a type of graphical model that represents causal relations among a set of measured (observed) variables **O** as well as probabilistic properties, such as conditional independencies (independence model). *The probabilistic properties of MAGs can be de-*

veloped without any reference to their causal semantics; nevertheless, we also briefly discuss their causal interpretation.

MAGs can be viewed as a generalization of Causal Bayesian Networks. The causal semantics of an edge $A \rightarrow B$ imply that A is probabilistically causing B, that is, an (appropriate) manipulation of A results in a change of the distribution of B. Edges $A \leftrightarrow B$ imply that A and B are associated but neither A causes B nor vice-versa. Under certain conditions, the independencies implied by the model are given by a graphical criterion called *m*-separation, defined below. A desired property of MAGs is that they are closed under marginalization: the marginal of a MAG is a MAG. MAGs can also represent the presence of selection bias, but this is out of the scope of the present paper. We present the key theory of MAGs, introduced in Richardson and Spirtes (2002).

A path in a graph $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ is a sequence of distinct vertices $\langle V_0, V_1, \ldots, V_n \rangle$ all of them in $\mathbf{O}$ s.t for $0 \leq i < n$, V_i and V_{i+1} are adjacent in $\mathcal{G}$. A path from V_0 to V_n is *directed* if for $0 \leq i < n$, V_i is a parent V_{i+1}. X is called an *ancestor* of Y and Y a *descendent* of X if $X = Y$ or there is a directed path from X to Y in $\mathcal{G}$. $\mathbf{An}_{\mathcal{G}}(X)$ is used to denote the set of ancestors of node X in $\mathcal{G}$. A *directed cycle* in $\mathcal{G}$ occurs when $X \rightarrow Y \in \mathbf{E}$ and $Y \in \mathbf{An}_{\mathcal{G}}(X)$. An *almost directed cycle* in $\mathcal{G}$ occurs when $X \leftrightarrow Y \in \mathbf{E}$ and $Y \in \mathbf{An}_{\mathcal{G}}(X)$.

Definition 1 (Mixed and Ancestral Graph) *A graph is mixed if all of its edges are either directed or bi-directed. A mixed graph is* **ancestral** *if the graph does not contain any directed or almost directed cycles.*

Given a path $p = \langle V_0, V_1, \ldots, V_n \rangle$, node V_i, $i \in 1, 2, \ldots, n$ is a *collider* on p if both edges incident to V_i have an arrowhead towards V_i. We also say that triple (V_{i-1}, V_i, V_{i+1}) forms a collider. Otherwise V_i is called a *non-collider* on p. The criterion of *m*-separation leads to a graphical way of determining the probabilistic properties stemming from the causal semantics of the graph:

Definition 2 (*m*-connection, *m*-separation) *In a mixed graph $\mathcal{G} = (\mathbf{E}, \mathbf{V})$, a path p between A and B is* ***m-connecting*** *relative to (condition to) a set of vertices $\mathbf{Z}$, $\mathbf{Z} \subseteq \mathbf{V} \setminus \{A, B\}$ if*

1. *Every non-collider on p is not a member of $\mathbf{Z}$.*

2. *Every collider on the path is an ancestor of some member of $\mathbf{Z}$.*

*A and B are said to be m-**separated** by $\mathbf{Z}$ if there is no m-connecting path between A and B relative to $\mathbf{Z}$. Otherwise, we say they are m-**connected** given $\mathbf{Z}$. We denote the m-separation of A and B given $\mathbf{Z}$ as $MSep(A; B|\mathbf{Z})$. Non-empty sets $\mathbf{A}$ and $\mathbf{B}$ are m-separated given $\mathbf{Z}$ (symb. $MSep(\mathbf{A}; \mathbf{B}|\mathbf{Z})$) if for every $A \in \mathbf{A}$ and every $B \in \mathbf{B}$ A and B are m-separated given $\mathbf{Z}$. ($\mathbf{A}$, $\mathbf{B}$ and $\mathbf{Z}$ are disjoint). We also define the set of all m-separations as $\mathcal{J}_m(\mathcal{G})$:*

$$\mathcal{J}_m(\mathcal{G}) \equiv \{\langle \mathbf{X}, \mathbf{Y} | \mathbf{Z} \rangle, \textit{s.t. } MSep(\mathbf{X}; \mathbf{Y} | \mathbf{Z}) \textit{ and } \mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{O}\}.$$

We also define the set $\mathcal{J}$ of all conditional independencies $\mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z}$, where $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are disjoint sets of variables, in the joint distribution of $\mathcal{P}$ of $\mathbf{O}$:

$$\mathcal{J}(\mathcal{P}) \equiv \{\langle \mathbf{X}, \mathbf{Y} | \mathbf{Z} | \rangle, \text{s.}t., \mathbf{X} \perp\!\!\!\perp \mathbf{Y} | \mathbf{Z} \text{ and } \mathbf{X}, \mathbf{Y}, \mathbf{Z} \subseteq \mathbf{O}\}.$$

The set $\mathcal{J}(\mathcal{P})$ is also called the *independence model* of $\mathcal{P}$. The *m*-separation criterion is meant to connect the graph with the observed independencies in the distribution under the following assumption:

Definition 3 (Faithfulness) *We call a distribution $\mathcal{P}$ over a set of variables* $\mathbf{O}$ ***faithful*** *to a graph $\mathcal{G}$, and vice versa, iff:*

$$\mathcal{J}(\mathcal{P}) = \mathcal{J}_m(\mathcal{G}).$$

A graph is faithful iff there exists a distribution faithful to it. When the above equation holds, we say the Faithfulness Condition holds for the graph and the distribution.

When the faithfulness condition holds, every *m*-separation present in $\mathcal{G}$ corresponds to a conditional independence in $\mathcal{J}(\mathcal{P})$ and vice-versa. The following definition describes a subset of ancestral graphs in which every missing edge (non-adjacency) corresponds to at least one conditional independence:

Definition 4 (Maximal Ancestral Graph, MAG) *An ancestral graph $\mathcal{G}$ is called* maximal *if for every pair of non-adjacent vertices (X,Y), there is a (possibly empty) set* $\mathbf{Z}$*, $X,Y \notin \mathbf{Z}$ such that $\langle X,Y|\mathbf{Z}\rangle \in \mathcal{J}(\mathcal{G})$.*

Every ancestral graph can be transformed into a unique equivalent MAG (i.e., with the same independence model) with the possible addition of bi-directed edges. We denote the marginal of a distribution $\mathcal{P}$ over a set of variables $V \setminus L$ $\mathbf{L}$ as $\mathcal{P}[_{\mathbf{L}}$, and the independence model stemming from the marginalized distribution as $\mathcal{J}(\mathcal{P})[_{\mathbf{L}}$, that is,

$$\mathcal{J}(\mathcal{P}[_{\mathbf{L}}) = \mathcal{J}(\mathcal{P})[_{\mathbf{L}} \equiv \{\langle \mathbf{X},\mathbf{Y}|\mathbf{Z}\rangle \in \mathcal{J}(\mathcal{P}) : (\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})\cap\mathbf{L} = \emptyset\}.$$

Equivalently, we define the set of *m*-separations of $\mathcal{G}$ restricted on the marginal variables as:

$$\mathcal{J}_m(\mathcal{G})[_{\mathbf{L}} \equiv \{\langle \mathbf{X},\mathbf{Y}|\mathbf{Z}\rangle \in \mathcal{J}_m(\mathcal{G}) : (\mathbf{X}\cup\mathbf{Y}\cup\mathbf{Z})\cap\mathbf{L} = \emptyset\}.$$

A simple graphical transformation for a MAG $\mathcal{G}$ faithful to a distribution $\mathcal{P}$ with independence model $\mathcal{J}(\mathcal{P})$ exists that provides a unique MAG $\mathcal{G}[_{\mathbf{L}}$ that represents the causal ancestral relations and the independence model $\mathcal{J}(\mathcal{P})[_{\mathbf{L}}$ after marginalizing out variables in $\mathbf{L}$. Formally,

Definition 5 (Marginalized Graph $\mathcal{G}[_L$) *Graph $\mathcal{G}[_L$ has vertex set $\mathbf{V}\setminus\mathbf{L}$, and edges defined as follows: If X,Y are s.t. , $\forall\mathbf{Z}\subseteq\mathbf{V}\setminus(\mathbf{L}\cup\{X,Y\})$, $\langle X,Y|\mathbf{Z}\rangle \notin \mathcal{J}(\mathcal{G})$ and*

$$\begin{array}{lcll} X \notin \mathbf{An}_{\mathcal{G}}(Y); Y \notin \mathbf{An}_{\mathcal{G}}(X) & & X \leftrightarrow Y & \\ X \in \mathbf{An}_{\mathcal{G}}(Y); Y \notin \mathbf{An}_{\mathcal{G}}(X) & \textit{then} & X \rightarrow Y & \text{in } \mathcal{G}[_L. \\ X \notin \mathbf{An}_{\mathcal{G}}(Y); Y \in \mathbf{An}_{\mathcal{G}}(X) & & X \leftarrow Y & \end{array}$$

We will call $\mathcal{G}[_L$ the marginalized graph $\mathcal{G}$ over $\mathbf{L}$*.*

The following result has been proved in Richardson and Spirtes (2002):

Theorem 6 *If $\mathcal{G}$ is a MAG over* $\mathbf{V}$*, and $\mathbf{L}\subseteq\mathbf{V}$, then $\mathcal{G}[_{\mathbf{L}}$ is also a MAG and*

$$\mathcal{J}_m(\mathcal{G})[_{\mathbf{L}} = \mathcal{J}_m(\mathcal{G}[_L).$$

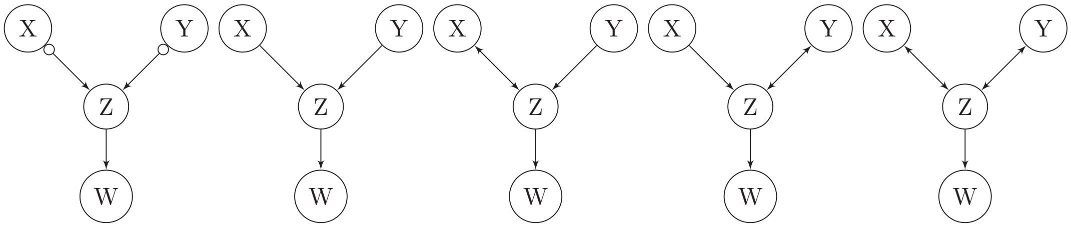


Figure 1: A PAG (left) and the MAGs of the respective equivalence class; all MAGs represent the same independence model over variables $\{X,Y,Z,W\}$.

If $\mathcal{G}$ is faithful to a distribution $\mathcal{P}$ over $\mathbf{V}$, then the above theorem implies that $\mathcal{J}(\mathcal{P})[_\mathbf{L} = \mathcal{J}(\mathcal{G})[_\mathbf{L} = \mathcal{J}(\mathcal{G}[_L)$; in other words the graph $\mathcal{G}[_\mathbf{L}$ constructed by the above process faithfully represents the marginal independence model $\mathcal{J}[_\mathbf{L}(\mathcal{P})$.

Different MAGs encode different causal information, but may share the same independence models and thus are statistically indistinguishable based on these models alone. Such MAGs define a Markov equivalence class based on the concepts of unshielded collider and discriminating path: A triple of nodes (X,Y,W) is called *unshielded* if X is adjacent to Y, Y is adjacent to W, and X is not adjacent to W. A path $p = \langle X,\dots,W,V,Y\rangle$ is called a *discriminating* path for V if X is not adjacent to Y, and every vertex between X and Y is a collider on p and an ancestor of Y. The following result has been proved in Spirtes and Richardson (1996):

Proposition 7 *Two MAGs over the same vertex set are Markov equivalent if and only if:*

1. *They share the same edges.*
2. *They share the same unshielded colliders.*
3. *If a path p is discriminating for a vertex V in both graphs, V is a collider on the path on one graph if and only if it is a collider on the path on the other.*

A *Partial Ancestral Graph* is a graph containing (up to) three kinds of endpoints: arrowhead $(>)$, tail $(-)$, and circle $(\circ)$, and represents a MAG Markov equivalence class in the following manner: It has the same adjacencies as any member of the equivalence class, and every non-circle endpoint is invariant in any member of the equivalence class. Circle endpoints correspond to uncertainties; the definitions of paths are extended with the prefix *possible* to denote that there is a configuration of the uncertainties in the path rendering the path ancestral or *m*-connecting. For example if $X \circ\!-\!\circ Y \circ\!\rightarrow W$, $\langle X,Y,W\rangle$ is a possible ancestral path from X to W, but not a possible ancestral path from W to X. An example PAG, and some of the MAGs in the respective equivalence class are shown in Figure 1. FCI (Spirtes et al., 2001; Zhang, 2008) is a sound algorithm which outputs a PAG over a set of variables $\mathbf{V}$ when given access to an independence model over $\mathbf{V}$.

The MAG formulation is a generalization of the graph of a (Causal) Bayesian Network (CBN) intended to explicitly model and reason with latent variables and particularly, latent confounding variables. The absence of such confounding variables is (often unrealistically) assumed when learning Causal Bayesian Networks, named the *Causal Sufficiency* assumption. The presence of latent confounders can be modeled in MAGs with bidirectional edges. The graph of a CBN is a MAG without bidirectional edges. Similarly, the Faithfulness Condition we define for MAGs generalizes the Faithfulness for CBNs. This work is inspired by the following scenario: *there exists an unknown causal mechanism over variables* $\mathbf{V}$*, represented by a faithful CBN* $\langle \mathcal{P}, \mathcal{G} \rangle$. Based on the theory presented in this section (Theorem 6), each marginal distribution of $\mathbf{P}$ over a subset $\mathbf{O} = \mathbf{V} \setminus \mathbf{L}$ is *faithful* to the MAG $\mathcal{G}[_{\mathbf{L}}$ described in definition 5.

3. Scope and Motivation of Integrative Causal Analysis

A general objective is to develop algorithms that are able to co-analyze data sets that are heterogeneous in various aspects, including data sets defined over different variables sets, experimental conditions, sampling methodologies (e.g., observational vs. case-control sampling) and others. In addition, cross-sectional data sets could be eventually co-analyzed with temporal data sets measuring either time-series data or repeated measurements data. Finally, the integrative analysis should also include prior knowledge about the data and their semantics. Some of the tasks of the integrative analysis can be the identification of the causal structure of the data generating mechanism, the selection of the next most promising experiment, the construction of predictive models, the prediction of the effect of manipulations, or the selection of the manipulation that best achieves a desired effect.

The work in this paper however, focuses on providing a first step towards this direction. It addresses the problem of learning the structure of the data generating process from data sets defined over different variable sets. In addition, it focuses on providing proof-of-concept experiments of the main INCA idea on the simplest cases and comparing against current alternatives. Finally, it gives methods that predict the strength of dependence between Y and Z, which can be seen as constructing a simple predictive model without having access to the joint distribution of the data.

We now make concrete some of these ideas by presenting a motivating fictitious integrative analysis scenario:

- **Study 1** (i.i.d., observational sampling, variables A, B, C, D): A scientist is studying the “relation” between contraceptives and breast cancer. In a random sample of women, he measures variables $\{A, B, C, D\}$ corresponding to quantities Suffers from *Thrombosis (Yes/No)*, *Contraceptives (Yes/No)*, *Concentration of Protein C in the Blood (numerical)* and *Develops Breast Cancer by 60 Years Old (Yes/No)*. The researcher then develops predictive models for Breast Cancer and, given that he finds B associated with D (among other associations), announces taking contraceptives as a risk-factor for developing Breast Cancer.

- **Study 2** (randomized controlled trial, variables A, B, C, D): Another scientist checks whether (variable C) *Protein C* (causally) protects against cancer. In a randomized controlled experiment she randomly assigns women into two groups and measures the same variables $\{A, B, C, D\}$. The first group is injected with high levels of the protein in their blood, while the latter is injected with enzymes that dissolve only the specific protein, effectively removing it from the blood. If C and D are negatively correlated in her data, the scientist concludes that the protein is causally protecting against the development of breast cancer. Notice that, data from Study 2 cannot be merged with Study 1 because the joint distributions of the data may be different. For example, assuming that C is caused by the disease D (e.g., the disease changes the concentration of the protein in the blood) then C will be highly associated with D in Study 1; in contrast, in Study 2 where the levels of C exclusively depend on the group

Variables / Study	A	B	C	D	E	F
	Thrombosis	Contraceptives	Protein C	Cancer	Protein Y	Protein Z
	(Yes/No)	(Yes/No)	(numerical)	(Yes/No)	(numerical)	(numerical)
1 (observational data)	Yes	No	10.5	Yes	-	-
	No	Yes	5.3	No	-	-
	No	Yes	0.01	No	-	-
2 (experimental data)	No	No	**0**(Control)	No	-	-
	Yes	No	**0**(Control)	Yes	-	-
	Yes	Yes	**5.0**(Treat.)	Yes	-	-
3 (different variables)	-	-	-	Yes	0.03	9.3
	-	-	-	No	3.4	22.2
4 (prior knowledge)	B causally affects A: $B \dashrightarrow A$					

Figure 2: Tabular depiction of the different studies (data sets). Study 1 is a random sample aiming at predicting D and identifying risk factors. Study 2 is a Randomized Controlled Trial were the levels of C for a subject are randomly decided and enforced by the experimenter, aiming at identifying a causal relation with cancer. Forced values are denoted by bold font. Study 3 is also an observational study about D, but measuring different variables than Study 1. Prior knowledge provides a piece of causal knowledge but the raw data are not available. Typically, such studies are analyzed independently of each other.

assignment, C and D are not associated. Thus, statistical inferences made based on analyzing Study 2 in isolation probably result in lower statistical power.

- **Study 3** (i.i.d., observational sampling, variables D,E,F): A biologist studies the relation of a couple of proteins in the blood, represented with variables E and F and their relation with breast cancer. She measures in a random sample of women variables $\{D,E,F\}$. As with analyzing Study 1, she develops predictive models for Breast Cancer (based on E and F instead) and checks whether the two proteins are risk factors. These data cannot be pulled together with Studies 1 or 2 because they measure different variables.

- **Prior Knowledge**: A doctor establishes a causal relation between the use of *Contraceptives* (variable B) and the development of *Thrombosis* (variable A), that is, “B causes A” denoted as $B \dashrightarrow A$.[1] Unfortunately, the raw data are not publicly available.

The three studies and prior knowledge are depicted Figure 2. Notice that, treating the empty cells as missing values is meaningless given that it is impossible for an algorithm to estimate the joint distribution between variables never measured together without additional assumptions (see Rubin 1974 for more details).

[1] We use a double arrow $\dashrightarrow$ to denote a causal relation without reference to the context of other variables. This is to avoid confusion with the use of a single arrow $\rightarrow$ in most causal models (e.g., Causal Bayesian Networks) that denotes a *direct* causal relation (or inducing path, see Richardson and Spirtes 2002), where direct causality is defined in the context of the rest of the variables in the model.

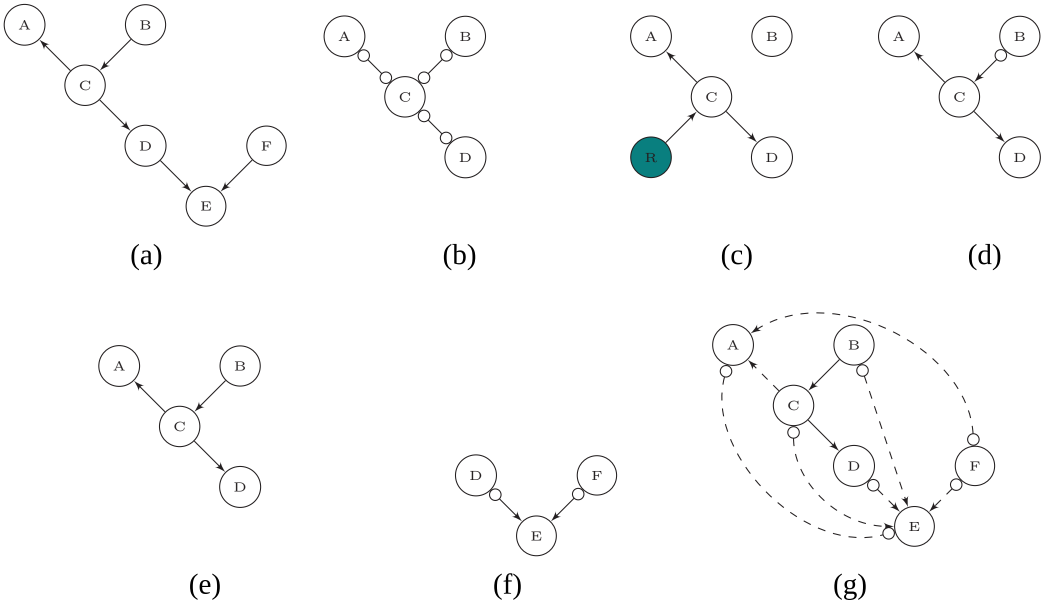


Figure 3: (a) Assumed unknown causal structure. (b) Structure induced by Study 1 alone. (c) Structure induced by Study 2 alone. (d) Structure induced by INCA of Studies 1 and 2. New inference: C is not causing B but they are associated. (e) Structure induced after incorporating knowledge “B causes A”. New inference: B causes A and D. (f) Structure induced by Study 3 alone. (g) Structure induced by all studies and knowledge. Dashed edges denote edges whose both existence and absence is consistent with the data. New inference: F and C (two proteins) are not causing each other nor do they have a latent confounder, even though we never measure them together in a study.

We now show informally the reasoning for an integrative causal analysis of the above studies and prior knowledge and compare against independent analysis of the studies. Figure 3(a) shows the presumed true, unknown, causal structure. Figure 3(b-c) shows the causal model induced (asymptotically) by an independent analysis of the data of Study 1 and Study 2 respectively using existing algorithms, such as FCI (Spirtes et al., 2001; Zhang, 2008) and assuming data generated by the true model. The R variable denotes the randomization procedure that assigns patients to control and treatment groups. Notice that it removes any causal link into C since the value of C only depends on the result of the randomization. Figure 3(d) shows the causal model that can be inferred by co-analyzing both studies together. By INCA of Study 1 and 2 it is now additionally inferred that B and C are correlated but C does not cause B: If C was causing B, we would have found the variables dependent in Study 2 (the randomization procedure would not have eliminated the causal link $C \rightarrow B$). If we also incorporate prior knowledge that “B causes A” we obtain the graph in Figure 3(e): “B causes A” implies that there has to be at least one directed (causal) path from B to A. Thus, the only possible such path $B\circ\!\rightarrow C \rightarrow A$ becomes directed $B \rightarrow C \rightarrow A$. In other words using prior knowledge we now additionally infer that “B is causing C”: the association found in Study 1 cannot be totally explained by the presence of a latent variable. Analyzing independently Study 3 we obtain the graph of Figure 3(f). In contrast INCA of Study 3 with the rest of data and knowledge results in

Figure 3(g). This type of graph is called the Pairwise Causal Graph (Triantafillou et al., 2010) and is presented in detail in Section 5. The dashed edges denote statistical indistinguishability about the existence of the edge, that is, there exist a consistent causal model with all data and knowledge having the edge, and one without the edge. Among other interesting inferences, notice that F *and* C *(two proteins) are not causing each other nor do they have a latent confounder, even though we never measure them together*. This is because if $F \rightarrow C$, or $C \leftarrow F$, or there exists latent H such that $F \leftarrow H \rightarrow C$ it would also imply an association between F and D. These two are found independent however, in Study 3.

4. Running Example

To illustrate the main ideas and concepts, as well as provide a proof-of-concept validation in real data, we have identified the smallest and simplest scenario that we could think of, that makes a testable prediction. Specifically, we identify a special case that predicts an unconditional dependence $Y \not\perp\!\!\!\perp Z|\emptyset$, as well as certain conditional dependencies $Y \not\perp\!\!\!\perp Z|\mathbf{S}$, for some $\mathbf{S} \neq \emptyset$, between two variables not measured in the same samples, based on two data sets, one measuring Y, and one measuring Z.

Example 1 *We assume two i.i.d data sets* $\mathcal{D}_1$ *and* $\mathcal{D}_2$ *are provided on variables* $\mathbf{O}_1 = \{X,Y,W\}$ *and* $\mathbf{O}_2 = \{X,Z,W\}$ *respectively. We assume that the independence models of the data sets are* $\mathcal{J}_1 = \{\langle X,W|Y\rangle\}$ *and* $\mathcal{J}_2 = \{\langle X,W|Z\rangle\}$*, in other words the one and only independence in* $\mathcal{D}_1$ *is* $X \perp\!\!\!\perp W|Y$*, and in* $\mathcal{D}_2$ *is* $X \perp\!\!\!\perp W|Z$*. Based on the input data it is possible to induce with existing causal analysis algorithms, such as FCI the following PAGs from each data set respectively:*

$$\mathcal{P}_1 : X \circ\!-\!\circ Y \circ\!-\!\circ W$$

and

$$\mathcal{P}_2 : X \circ\!-\!\circ Z \circ\!-\!\circ W.$$

These are also shown graphically in Figure 4. The problem is to identify one or all MAGs defined on $\mathbf{O} = \{X,Y,Z,W\}$ *consistent with the independence models* $\mathcal{J}_1$ *and* $\mathcal{J}_2$*, or equivalently, both PAGs* $\mathcal{P}_1$ *and* $\mathcal{P}_2$*.*

These two PAGs represent all the sound inferences possible about the structure of the data, when analyzing the data sets in isolation and independently of each other. We next develop the theory for their causal co-analysis.

5. Integrative Causal Analysis of Data Sets with Overlapping Variable Sets

In this section, we address the problem of integratively analyzing multiple data sets defined over different variable sets. Co-analyzing these data sets is meaningful (using this approach) only when these variable sets overlap; otherwise, there are no additional inferences to be made unless other information connects the two data sets (e.g., the presence of prior knowledge connecting some variables).

We assume that we are given K data sets $\{\mathcal{D}_i\}_{i=1}^K$ each with samples identically and independently distributed defined over corresponding subsets of variables $\mathbf{O}_i$. From these data we can estimate the independence models $\{\mathcal{J}_i\}_{i=1}^K$ using statistical tests of conditional independence. *A*

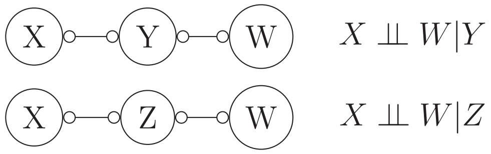


Figure 4: Definition of the co-analysis problem of Example 1: two observational i.i.d. data sets defined on variables $\mathbf{O}_1 = \{X,Y,W\}$ and $\mathbf{O}_2 = \{X,Z,W\}$ are used to identify the independence models $\mathcal{J}_1 = \{\langle X,W|Y\rangle\}$ and $\mathcal{J}_2 = \{\langle X,W|Z\rangle\}$. These models are represented by PAGs $\mathcal{P}_1$ and $\mathcal{P}_2$ shown in the figure. The problem is to identify one or all MAGs defined on $\mathbf{O} = \{X,Y,Z,W\}$ consistent with both $\mathcal{P}_1$ and $\mathcal{P}_2$.

major assumption in the theory and algorithms presented is that the independence models can be identified without any statistical errors. Section 6 discusses how we address this issue when experimenting with real data sets in the presence of statistical errors. We denote the union of all variables as $\mathbf{O} = \cup_{i=1}^{K}\mathbf{O}_i$ and also define $\overline{\mathbf{O}_i} \equiv \mathbf{O} \setminus \mathbf{O}_i$. We now define the problem below:

Definition 8 (Find Consistent MAG) *Assume the distribution of* $\mathbf{O}$ *is faithful. Given independence models* $\{\mathcal{J}(\mathbf{O}_i)\}_{i=1}^{K}$, $\mathbf{O}_i \subseteq \mathbf{O}, i = 1\ldots K$, *induce a MAG* $\mathcal{M}$ *s.t., for all* i

$$\mathcal{J}(\mathcal{M}[_{\overline{\mathbf{O}_i}}) = \mathcal{J}(P_i)$$

where P_i *is the distribution of* $\mathbf{O}_i$.

In other words, we are looking for a model (graph) $\mathcal{M}$ such that when we consider its marginal graphs over each variable set $\mathbf{O}_i$, each one faithfully represents the observed independence model of that data set. We can reformulate the problem in graph-theoretic terms. Let $\mathcal{P}_i$ be the PAG representing the Markov equivalence class of all MAGs consistent with the independence model $\mathcal{J}_i$. $\mathcal{P}_i$ can be constructed with a sound and complete algorithm such as Fast Causal Inference (FCI) (Spirtes et al., 2001). We can thus recast the problem above as identifying a MAG $\mathcal{M}$ such that,

$$M[_{\overline{\mathbf{O}_i}} \in \mathcal{P}_i, \text{for all } i$$

(abusing the notation to denote with $\mathcal{P}_i$ both the PAG and the equivalence class).

The first algorithm to solve the above problem is ION (Tillman et al., 2008), which identifies the set of PAGs (defined over $\mathbf{O}$) of all consistent MAGs. Subsequently, in Triantafillou et al. (2010), we proposed the algorithm Find Consistent MAG (FCM) that converts the problem to a satisfiability problem for improved computational efficiency. FCM returns one consistent MAG with all input PAGs. Similar ideas have been developed to learn joint structure from marginal structures in decomposable graphs such as undirected graphs (Kim and Lee, 2008) and Bayesian Networks (Kim, 2010). Going back to Example 1, Figure 5 shows all 14 consistent MAGs with the input PAGs in the scenario. The FCM algorithm arbitrarily returns one of them as the solution to the problem (of course, the algorithm can be easily modified to return all solutions). Figure 6 (right) shows the output of ION on the same problem.

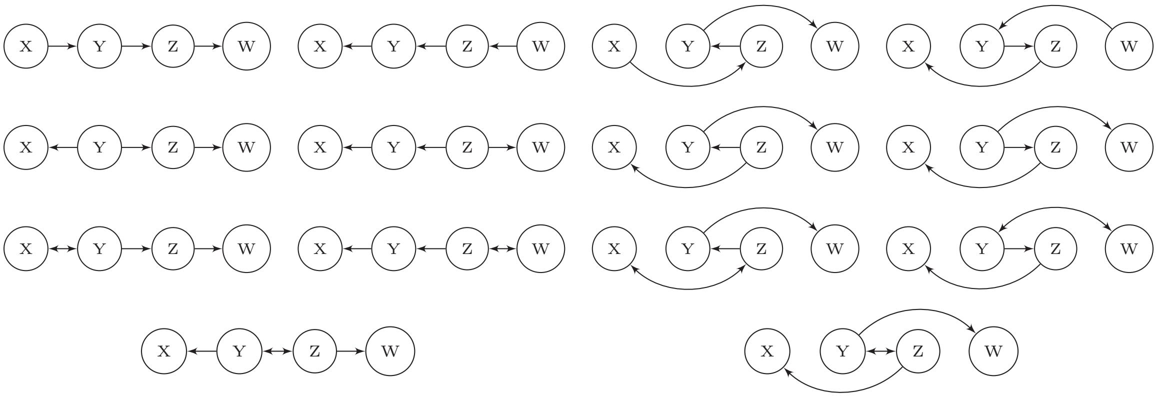


Figure 5: Solution of the co-analysis problem of Example 1: The 14 depicted MAGs are all and only the consistent MAGs with the PAGs shown in Figure 4. In all these MAGs the independencies $X \perp\!\!\!\perp W|Y$ and $X \perp\!\!\!\perp W|Z$ hold (and only them). Notice that, even though the edge $X - Y$ exists in $\mathcal{P}_1$ (Example 1), some of the consistent MAGs (the ones on the right of the figure) do not contain this edge: *adjacencies in the input PAGs do not simply transfer to the solution MAGs.* The FCM algorithm would arbitrarily output one of these MAGs as the solution of the problem of Example 1.

5.1 Representing the Set of Consistent MAGs with Pairwise Causal Graphs

The set of consistent MAGs to a set of PAGs is defined as follows:

Definition 9 (Set of Consistent MAGs) *We call the set of all MAGs $\mathcal{M}$ over variables $\mathbf{O}$ consistent with the set of PAGs $\mathbf{P} = \{\mathcal{P}_i\}_{i=1}^N$ over corresponding variable sets $\mathbf{O}_i$, where $\mathbf{O} = \cup_i \mathbf{O}_i$ as the Set of Consistent MAGs with $\mathbf{P}$ denoted with $\mathbf{M}(\mathbf{P})$.*

Unfortunately, $\mathbf{M}(\mathbf{P})$ cannot in general be represented with a single PAG: the PAG formalism represents a set of equivalent MAGs *when learning from a single data set and its independence model*. In Example 1 though, notice that the MAGs in $\mathbf{M}(\mathbf{P})$ in Figure 5 have a different skeleton (i.e., set of edges ignoring the edge-arrows), so they cannot be represented by a single PAG.

The PAG formalism allows the set of *m*-separations that entail the *m*-separations of all MAGs in the class to be read off its graph in polynomial time. Unfortunately, there is currently no known compact representation of $\mathbf{M}(\mathbf{P})$ such that the *m*-separations that hold for all members of the set can be easily identified (i.e., in polynomial time).

We have introduced (Triantafillou et al., 2010) a new type of graph called the *Pairwise Causal Graph* (PCG) that graphically represents $\mathbf{M}(\mathbf{P})$. However, PCG do not always allow the *m*-separations of each member MAG to be easily identified. A PCG focuses on representing the possible causal pair-wise relations among each pair of variables X and Y in $\mathbf{O}$.

Definition 10 (Pairwise Causal Graph) *We consider the MAGs in $\mathbf{M}(\mathbf{P})$ consistent with the set of PAGs $\mathbf{P} = \{\mathcal{P}_i\}_{i=1}^N$ defined over $\{\mathbf{O}_i\}_{i=1}^N$. A Pairwise Causal Graph $\mathcal{U}$ is a partially oriented mixed graph over $\bigcup_i \mathbf{O}_i$ with two kinds of edges dashed (- -) and solid (—) and three kinds of endpoints(>, -, ∘) with the following properties:*

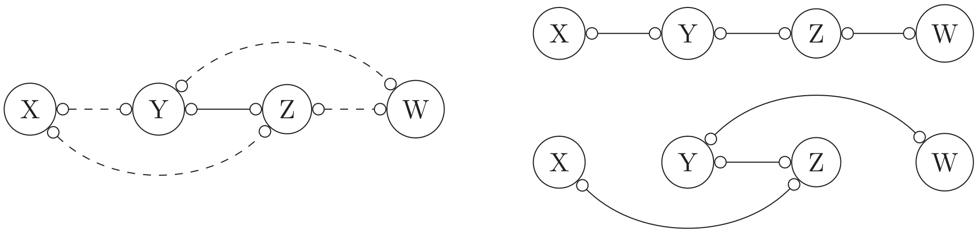


Figure 6: (left) Pairwise Causal Graph (PCG) representing the set of consistent MAGs of Example 1. This PCG is the output of the cSAT+ algorithm on the problem of Example 1. Alternatively, the set of consistent MAGs can be represented with two PAGs (right). This is the output of the ION algorithm on the same problem.

1. *X — Y in $\mathcal{U}$ iff X is adjacent to Y in every consistent $\mathcal{M} \in \mathbf{M}(\mathbf{P})$.*

2. *X -- Y in $\mathcal{U}$ iff X is adjacent to Y in at least one but not all consistent $\mathcal{M} \in \mathbf{M}(\mathbf{P})$.*

3. *X and Y are not adjacent in $\mathcal{U}$ iff they are not adjacent in any consistent $\mathcal{M} \in \mathbf{M}(\mathbf{P})$.*

4. *The right end-point of edge X -- Y is oriented as $>$, -, or $\circ$ iff X is into Y in all, none, or at least one (but not all) consistent MAG $\mathcal{M} \in \mathbf{M}(\mathbf{P})$ where X and Y are adjacent. Similarly, for the left end-point and for solid edges $X - Y$.*

Solid edges, missing edges, as well as end-points marked with"$>$" and "$-$" show invariant characteristics that hold in all consistent MAGs. Dash edges and "$\circ$"-marked end-points represent uncertainty of the presence of the edge and the type of the end-point.

The PCG of Example 1 is shown in Figure 6 (left). For computing the PCG one can employ the cSAT+ algorithm (Triantafillou et al., 2010). There are several points to notice. The invariant graph features are the solid edge Y — Z and the missing edge between X and W; these are shared by all consistent MAGs. The remaining edges are dashed showing that they are present in at least one consistent MAG. All end-points are marked with "$\circ$" showing that any type of orientation is possible for each of them. The graph fails to graphically represent certain constraints, for example, that there is no MAG that simultaneously contains edges $X - Y$ and $X - Z$; in general, the presence of an edge (or a particular end-point) in a consistent MAG may entail the absence of some other edge (or end-point). It also fails to depict the *m*-separation $X \perp\!\!\!\perp W|Z$ or the fact that any solution has a chain-like structure.

Nevertheless, the graph still conveys valuable information: *the solid edge X — Y along with the Faithfulness condition entails that Y and Z are associated given any subset of the other variables, even though Y and Z are never measured together in any input data set.* This is a testable prediction on which we base the computational experiments in Section 6. Alternatively, the set $\mathbf{M}(\mathbf{P})$ could be represented with *two* PAGs shown in 6 (right), as the set of MAGs consistent with either one them. These PAGs form the output of ION on this problem.

6. Predicting the Presence of Unconditional Dependencies

We now discuss how to implement the identification of the scenario in Example 1 to predict the presence of dependencies.

6.1 Predictions of Dependencies

Recall that, in Example 1 we assume we are given two data sets on variables $\mathbf{O}_1 = \{X,Y,W\}$ and $\mathbf{O}_2 = \{X,Z,W\}$. We then determine, if possible, whether their independence models are respectively $\mathcal{J}_1 = \{\langle X,W|Y\rangle\}$ and $\mathcal{J}_2 = \{\langle X,W|Z\rangle\}$ by a series of unconditional and conditional tests of independence. If this is the case, we predict an association between Y and Z. The details of determining the independence model are important. Let us denote the p-value of an independence test with null hypothesis $X \perp\!\!\!\perp Y|\mathbf{Z}$ as $p_{X\perp\!\!\!\perp Y|\mathbf{Z}}$. In the algorithms that follow, we make statistical decisions with the following rules:

- If $p_{X\perp\!\!\!\perp Y|\mathbf{Z}} \leq \alpha$ conclude $X \not\perp\!\!\!\perp Y|\mathbf{Z}$ (reject the null hypothesis).
- If $p_{X\perp\!\!\!\perp Y|\mathbf{Z}} \geq \beta$ conclude $X \perp\!\!\!\perp Y|\mathbf{Z}$ (accept the null hypothesis).
- Otherwise, forgo making a decision.

Algorithm 1: Predict Dependency: Full-Testing Rule (**FTR**)

```
Input: Data Sets $\mathcal{D}_1$ and $\mathcal{D}_2$ on variables $\{X,Y,W\}$ and $\{X,Z,W\}$, respectively
if in $\mathcal{D}_1$ we conclude
    // determine whether $\mathcal{J}_1 = \{\langle X,W|Y\rangle\}$
    $X \perp\!\!\!\perp W|Y$ , $X \not\perp\!\!\!\perp Y|\emptyset$ , $Y \not\perp\!\!\!\perp W|\emptyset$ , $X \not\perp\!\!\!\perp W|\emptyset$ , $X \not\perp\!\!\!\perp Y|W$ , $Y \not\perp\!\!\!\perp W|X$
  and in $\mathcal{D}_2$ we conclude
    // determine whether $\mathcal{J}_2 = \{\langle X,W|Z\rangle\}$
    $X \perp\!\!\!\perp W|Z$ , $X \not\perp\!\!\!\perp Z|\emptyset$ , $Z \not\perp\!\!\!\perp W|\emptyset$ , $X \not\perp\!\!\!\perp W|\emptyset$ , $X \not\perp\!\!\!\perp Z|W$ , $Z \not\perp\!\!\!\perp W|X$
then
    Predict $Y \not\perp\!\!\!\perp Z|\emptyset$
    Predict either ($X \circ\!-\!\circ Y \circ\!-\!\circ Z \circ\!-\!\circ W$) or ($X \circ\!-\!\circ Z \circ\!-\!\circ Y \circ\!-\!\circ W$) holds
else
    Do not make a prediction
end
```

The details are shown in Algorithm 1 named Full-Testing Rule, or FTR for short. We note a couple of observations. First, the algorithm is opportunistic. It does not produce a prediction whenever possible, but only for the case presented in Example 1. In addition, it makes a prediction only when the p-values of the tests are either too high or too low to relatively safely accept dependencies and independencies. Second, to accept an independence model, for example, that $\mathcal{J}_1 = \{\langle X,W|Y\rangle\}$ all possible conditional and unconditional tests among the variables are performed. If any of these tests is inconclusive or contradictory to $\mathcal{J}_1$, the latter is not accepted and no prediction is made. In the terminology of Spirtes et al. (2001), we test for a *detectable failure of faithfulness*. Similar ideas have also been devised in Ramsey et al. (2006) and Spanos (2006). This rule characteristic is important in case one would like to generalize these ideas to larger graphs and sets of variables:

performing all possible tests becomes quickly prohibitive, and the probability of statistical errors increases.

If however, one assumes the Faithfulness Condition holds among variables $\{X,Y,Z,W\}$, then it is not necessary to perform all such tests to determine the independence models. Algorithms for inducing graphical models from data, such as FCI and PC (Spirtes et al., 2001) are based on this observation to gain computational efficiency. The Minimal-Testing Rule, MTR for short, performs only a minimal number of tests that together with Faithfulness may entail that $\mathcal{J}_1 = \{\langle X,W|Y\rangle\}$ and $\mathcal{J}_2 = \{\langle X,W|Z\rangle\}$ and lead to a prediction. The details are shown in Algorithm 2.

Algorithm 2: Predict Dependency Minimal-Testing Rule (**MTR**)

```
Input: Data Sets $\mathcal{D}_1$ and $\mathcal{D}_2$ on variables $\{X,Y,W\}$ and $\{X,Z,W\}$, respectively
if in $\mathcal{D}_1$ we conclude
        // determine whether $\mathcal{J}_1 = \{\langle X,W|Y\rangle\}$
        $X \perp\!\!\!\perp W|Y$ , $X \not\perp\!\!\!\perp Y|\emptyset$ , $Y \not\perp\!\!\!\perp W|\emptyset$
    and in $\mathcal{D}_2$ we conclude
        // determine whether $\mathcal{J}_2 = \{\langle X,W|Z\rangle\}$
        $X \perp\!\!\!\perp W|Z$ , $X \not\perp\!\!\!\perp Z|\emptyset$ , $Z \not\perp\!\!\!\perp W|\emptyset$
then
    Predict $Y \not\perp\!\!\!\perp Z|\emptyset$
    Predict either ($X \circ\!-\!\circ Y \circ\!-\!\circ Z \circ\!-\!\circ W$) or ($X \circ\!-\!\circ Z \circ\!-\!\circ Y \circ\!-\!\circ W$) holds
else
    Do not make a prediction
end
```

6.2 Heuristic Predictions of Dependencies Based on Transitivity

Is it really necessary to develop and employ the theory presented to make such predictions? Could there be other simpler and intuitive rules that are as predictive, or more predictive? For example, a common heuristic inference people are sometimes willing to make is the transitivity rule: if Y is correlated with X and X is correlated with Z, then predict that Y is also correlated with Z. The FTR and MTR rules defined also check these dependencies: $X \not\perp\!\!\!\perp Y$ in $\mathcal{D}_1$ and $X \not\perp\!\!\!\perp Z$ in $\mathcal{D}_1$, so one could object that any success of the rules could be attributed to the transitivity property often holding in Nature. We implement the Transitivity Rule (TR), shown in Algorithm 3 to compare against the INCA-based FTR and MTR rules. Obviously, the Transitivity Rule is not sound in general,[2] but on the other hand, FTR and MTR are also based on the assumption of Faithfulness, which may as well be unrealistic. The verdict will be determined by experimentation.

6.3 Empirical Evaluation of Predicting Unconditional Dependencies

We have applied and evaluated the three rules against each-other as well as random predictions (prior probability of a pair being dependent) on real data, in a way that becomes testable. Specifically, given a data set $\mathcal{D}$ we randomly partition its samples to three data sets of equal size, $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_t$. The latter is hold out for testing purposes. In the first two data sets, we identify quadruples of

[2] The Transitivity Rule should be sound when the marginal of the three variables is faithful to a *Markov Random Field*.

```
Algorithm 3: Predict Dependency Transitivity Rule (TR)
Input: Data Sets $\mathcal{D}_1$ and $\mathcal{D}_2$ on variables $\{Y,X\}$ and $\{X,Z\}$, respectively
if in $\mathcal{D}_1$: $Y \not\perp\!\!\!\perp X|\emptyset$ and in $\mathcal{D}_2$: $X \not\perp\!\!\!\perp Z|\emptyset$ then
    Predict $Y \not\perp\!\!\!\perp Z|\emptyset$
else
    Do not make a prediction
end
```

Name	Reference	# istances	# vars	Group Size	Vars type	Scient. domain
Covtype	Blackard and Dean (1999)	581012	55	55	N/O	Agricultural
Read	Guvenir and Uysal (2000)	681	26	26	N/C/O	Business
Infant-mortality	Mani and Cooper (2004)	5337	83	83	N	Clinical study
Compactiv	Alcalá-Fdez et al. (2009)	8192	22	22	C	Computer science
Gisette	Guyon et al. (2006a)	7000	5000	50	C	Digit recognition
Hiva	Guyon et al. (2006b)	4229	1617	50	N	Drug discovering
Breast-Cancer	Wang (2005)	286	17816	50	C	Gene expression
Lymphoma	Rosenwald et al. (2002)	237	7399	50	C	Gene expression
Wine	Cortez et al. (2009)	4898	12	12	C	Industrial
Insurance-C	Elkan (2001)	9000	84	84	N/O	Insurance
Insurance-N	Elkan (2001)	9000	86	86	N/O	Insurance
p53	Danziger et al. (2009)	16772	5408	50	C	Protein activity
Ovarian	Conrads (2004)	216	2190	50	C	Proteomics
C&C	Frank and Asuncion (2010)	1994	128	128	C	Social science
ACPJ	Aphinyanaphongs et al. (2006)	15779	28228	50	C	Text mining
Bibtex	Tsoumakas et al. (2010)	7395	1995	50	N	Text mining
Delicious	Tsoumakas et al. (2010)	16105	1483	50	N	Text mining
Dexter	Guyon et al. (2006a)	600	11035	50	N	Text mining
Nova	Guyon et al. (2006b)	1929	12709	50	N	Text mining
Ohsumed	Joachims (2002)	5000	14373	50	C	Text mining

Table 1: Data Sets included in empirical evaluation of Section 6.3. N- Nominal, O - Ordinal, C - Continuous.

variables $\{X,Y,Z,W\}$ for which the Full-Testing and the Minimal-Testing Rules apply. Notice that, the two rules perform tests among variables $\{X,Y,W\}$ in $\mathcal{D}_1$ and among variables $\{X,Z,W\}$ in $\mathcal{D}_2$; *the rules do not access the joint distribution of* Y,Z. Similarly, for the Transitivity Rule we identify triplets $\{X,Y,Z\}$ where the rule applies. Subsequently, we measure the predictive performance of the rules. In more detail:

- *Data Sets*: We selected data sets in an attempt to cover a wide range of sample-sizes, dimensionality (number of variables), types of variables, domains, and tasks. The decision for inclusion depended on availability of the data, ease of parsing and importing them. *No data set was a posteriori removed out of the study, once selected.* Table 1 assembles the list of data sets and their characteristics before preprocessing. Some minimal preprocessing steps were applied to several data sets that are described in Appendix A.

- *Tests of Independence*: For discrete variables we have used the G^2-test (a type of likelihood ratio test) with an adjustment for the degrees-of-freedom used in Tsamardinos et al. (2006)

and presented in detail in Tsamardinos and Borboudakis (2010). For continuous variables we have used a test based on the Fisher z-transform of the partial correlation as described in Spirtes et al. (2001). The two tests employed are typical in the graphical learning literature. In some cases ordinal variables were treated as continuous, while in others the continuous variables were discretized (see Appendix A) so that every possible quadruple $\{X,Y,Z,W\}$ was either treated as all continuous variables or all discrete and one of the two tests above could be applied.

- *Significance Thresholds*: There are two threshold parameters: level α below which we accept dependence and level β above which we accept independence; the TR rule only employs the α parameter. For FTR these thresholds were always set to $\alpha_{FTR} = 0.05$ and $\beta_{FTR} = 0.3$ without an effort to optimize them. Some minimal anecdotal experimentation with FTR showed that the performance of the algorithm is relative insensitive to the values of α_{FTR} and β_{FTR} and the algorithm works without fine-tuning. Notice that FTR requires 10 dependencies and 2 independencies to be identified, while MTR requires 4 dependencies and 2 independencies, and TR requires 2 dependencies to be found. Thus, FTR is more conservative than MTR and TR for the same values of α and β. The Bonferroni correction for MTR dictates that $\alpha_{MTR} = \alpha_{FTR} \times \frac{4}{10} = 0.02$, while for TR we get $\alpha_{TR} = \alpha_{FTR} \times \frac{2}{10} = 0.01$ (TR however, does not require any independencies present so this adjustment may not be conservative enough). We run MTR with threshold values $\alpha_{MTR} \in \{0.05, 0.02, 0.002, 0.0002\}$, that is equal to the threshold of FTR, with the Bonferroni adjustment, and stricter than Bonferroni by one and two orders of magnitude. The β_{MTR} parameter is always set to 0.3. In a similar fashion for TR, we set $\alpha_{TR} \in \{0.05, 0.01, 0.001, 0.0001\}$.

- *Identifying Quadruples*: In low-dimensional data sets (number of variables less than 150), we check the rules on all quadruples of variables. This is time-prohibitive however, for the larger data sets. In such cases, we randomly permute the order of variables and partition them into groups of 50 and consider quadruples only within these groups. The column named “Group Size” in Table 1 notes the actual sizes of the variable groups used.

- *Measuring Performance*: The ground truth for the presence of a predicted correlation is not known. We thus seek to statistically evaluate the predictions. Specifically, for each predicted pair of variables X and Y, we perform a test of independence in the corresponding hold-out test set $\mathcal{D}_t$ and store its p-value $p_{X \perp\!\!\!\perp Y|\emptyset}$. The lower the p-value the higher the probability the pair is truly correlated. We consider as “accurate” a prediction whose p-value is less than a threshold t and we report the accuracy of each rule.

Definition 11 (Prediction Accuracy) *We denote with M_i^R and U_i^R the multiset and set respectively of p-values of the predictions of rule R applied on data set i. The p-values are computed on the hold-out test set. The accuracy of the rule on data set i at threshold t is defined as:*

$$Acc_i^R(t) = \#\{p <= t, p \in M_i^R\}/|M_i^R|.$$

We also define the average accuracy *over all data sets (each data set is weighted the same)*

$$\overline{Acc}^R(t) = \frac{1}{20}\sum_{i=1}^{20} Acc_i^R(t)$$

and the pooled accuracy *over the union of predictions (each prediction is weighted the same)*

$$\underline{Acc}^R(t) = \#\{p <= t, i = 1\ldots 20, p \in M_i^R\} / \sum_i |M_i^R|.$$

The reason M_i^R is defined as a multiset stems from the fact that a dependency $Y \not\perp\!\!\!\perp Z|\emptyset$ may be predicted multiple times if a rule applies to several quadruples $\{X_i, Y, Z, W_i\}$ or triplets $\{X_i, Y, Z\}$ (for the Transitivity Rule). The number of predictions of each rule R (i.e., $|M_i^R|$) is shown in Table 2, while Table 8 in Appendix A reports $|U_i^R|$, the number of pairs $X - Y$ predicted correlated. In some cases (e.g., data sets Read and ACPJ) the Full-Testing Rule does not make any predictions. Overall however, the rules typically make hundreds or even thousands of predictions.

Data Set	$\text{FTR}_{0.05}$	$\text{MTR}_{0.02}$	$\text{TR}_{0.01}$
Covtype	222	33277	54392
Read	0	9	4713
Infant Mortality	22	2038	3736
Compactiv	135	679	3950
Gisette	423	35824	134213
hiva	554	65967	151582
Breast-Cancer	1833	141643	470212
Lymphoma	7712	188216	394572
Wine	4	73	431
Insurance-C	1839	30569	40173
Insurance-N	226	18270	47115
p53	46647	1645476	1995354
Ovarian	539165	1604131	2015133
C&C	99241	416934	301218
ACPJ	0	219	16574
Bibtex	1	3982	25948
Delicious	856	32803	105776
Dexter	0	2	117
Nova	0	124	3473
Ohsumed	0	64	5358

Table 2: Number of predictions $|M_i^R|$ with “Bonferroni” correction for rules FTR, MTR and TR.

Overall Performance: The accuracies at $t = 0.05$, $Acc_i(t)$, $\overline{Acc}(t)$, and $\underline{Acc}(t)$ for the three rules as well as the one achieved by guessing at random are shown in Figure 7. The Bonferroni adjusted thresholds for MTR and TR were used: $\alpha_{FTR} = 0.05, \alpha_{MTR} = 0.02, \alpha_{TR} = 0.01$. Similar figures for all sets of thresholds are shown in Appendix A, Section A.3. Over all predictions, the Full-Testing Rule achieves accuracy 96%, consistently higher than guessing at random, the MTR and the TR. The same results are also depicted in tabular form in Table 3, where additionally, the statistical significance is noted. The null hypothesis is that $Acc_i^{FTR}(0.05) \leq Acc_i^R(0.05)$, for R being MTR or TR. The one-tail Fisher’s exact test (Fisher, 1922) is employed when computationally feasible, otherwise the Pearson χ^2 test (Pearson, 1900) is used instead. FTR is typically performing statistically significantly better than all other rules.

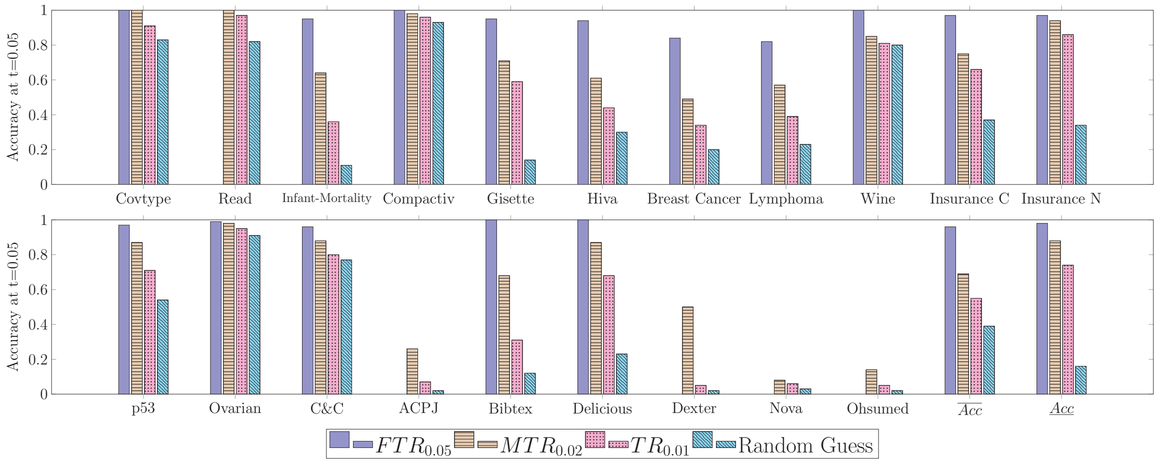


Figure 7: Accuracies Acc_i for each data set, as well as the average accuracy $\overline{Acc}$ (each data set weighs the same) and the pooled accuracy $\underline{Acc}$ (each prediction weighs the same). All accuracies are computed as threshold $t = 0.05$. FTR's accuracy is always above 80% and always higher than MTR, TR, and random guess.

Sensitivity to the α *parameter*: The results are not particularly sensitive to the significance thresholds used for α for MTR and TR. Figures 9 (a-b) show the average accuracy $\overline{Acc}$ and the pooled accuracy $\underline{Acc}$ as a function of the $alpha$ parameter used: no correction, Bonferroni correction, and stricter than Bonferroni by one and two orders of magnitude. The accuracy of MTR and TR improves as they become more conservative but never reaches the one by FTR even for the stricter thresholds of $\alpha_{MTR} = 0.0002$ and $\alpha_{TR} = 0.0001$.

Sensitivity to t: The results are also not sensitive to the particular significance level t used to define accuracy. Figure 8 graphs $Acc_i^R(t)$ over $t = [0, 0.05]$ for two typical data sets as well as $\underline{Acc}(t)$ and $\overline{Acc}(t)$. The situation is similar and consistent across all data sets considered, which are shown in Appendix A. The lines of the Full Testing Rule rise sharply, which indicates that the p-values of its predictions are concentrated close to zero.

Explaining the difference of FTR and MTR: Asymptotically and when the data distribution is faithful to a MAG, the FTR and the MTR rules are both sound (100% accurate). However, when the distribution is not faithful, the performance difference could become large because FTR tests for faithfulness violations as much as possible in an effort to avoid false predictions. This may explain the large differences in accuracies observed in the Infant Mortality, Gisette, Hiva, Breast-Cancer, and Lymphoma data sets. When the distribution is faithful, but the sample is finite, we expect some but small differences. For example when MTR falsely determines that $X \not\perp\!\!\!\perp Y|\emptyset$ due to a false positive test, the FTR rule still has a chance to avoid an incorrect prediction by additionally testing $X \not\perp\!\!\!\perp Y|W$. To support this theoretical analysis we perform experiments with simulated data where the network structure is known. Specifically, we employ the structure of the ALARM (Beinlich et al., 1989), INSURANCE (Binder et al., 1997) and HAILFINDER (Abramson et al., 1996) Bayesian Networks. We sample 20 continuous and 20 discrete pairs of data sets D_1 and D_2 from distributions faithful to the network structure using different randomly chosen parameterizations for the continuous case, and the original network parameters for the discrete case. We do the same for

Data Set	$FTR_{0.05}$	$MTR_{0.02}$	$TR_{0.01}$	Random Guess
Covtype	1.00	1.00	0.91^{**}	0.83^{**}
Read	-	1.00	0.97	0.82
Infant Mortality	0.95	0.64^{**}	0.36^{**}	$0.11^{\spadesuit}$
Compactiv	1.00	0.98	0.96^{*}	0.93^{**}
Gisette	0.95	$0.71^{\spadesuit}$	$0.59^{\spadesuit}$	$0.14^{\spadesuit}$
hiva	0.94	$0.61^{\spadesuit}$	$0.44^{\spadesuit}$	$0.30^{\spadesuit}$
Breast-Cancer	0.84	$0.49^{\spadesuit}$	$0.34^{\spadesuit}$	$0.20^{\spadesuit}$
Lymphoma	0.82	$0.57^{\spadesuit}$	$0.39^{\spadesuit}$	$0.23^{\spadesuit}$
Wine	1.00	0.85	0.81	0.80
Insurance-C	0.97	$0.75^{\spadesuit}$	$0.66^{\spadesuit}$	$0.37^{\spadesuit}$
Insurance-N	0.97	0.94^{*}	0.86^{**}	$0.34^{\spadesuit}$
p53	0.97	$0.87^{\spadesuit}$	$0.71^{\spadesuit}$	$0.54^{\spadesuit}$
Ovarian	0.99	$0.98^{\spadesuit}$	$0.95^{\spadesuit}$	$0.91^{\spadesuit}$
C&C	0.96	$0.88^{\spadesuit}$	$0.80^{\spadesuit}$	$0.77^{\spadesuit}$
ACPJ	-	0.26	0.07	0.02
Bibtex	1.00	0.68	0.31	0.12^{**}
Delicious	1.00	$0.87^{\spadesuit}$	$0.68^{\spadesuit}$	$0.23^{\spadesuit}$
Dexter	-	0.50	0.05	0.02
Nova	-	0.08	0.06	0.03
Ohsumed	-	0.14	0.05	0.02
$\overline{ACC^R}$	0.96	0.69^{**}	0.55^{**}	0.39^{**}
$\underline{ACC^R}$	0.98	$0.88^{\spadesuit}$	$0.74^{\spadesuit}$	$0.16^{\spadesuit}$

Table 3: $ACC_i^R(t)$ at $t = 0.05$ with "Bonferroni" correction for rules FTR, MTR, TR and Random Guess. Marks *, **, and ♠ denote a statistically significant difference from FTR at the levels of 0.05, 0.01, and machine-epsilon respectively.

sample sizes 100, 500, 1000. Subsequently, we apply the FTR and MTR rules with $\alpha_{FTR} = 0.05$ and $\alpha_{MTR} = 0.02$ (Bonferroni adjusted) on each pair of D_1 and D_2 and all possible quadruples of variables. The true accuracy is not computed on a test data set D_t but on the known graph instead by checking whether Y and Z are d-connected given X and W. The mean true accuracies over all samplings are reported in Figure 10. The difference in performance on the faithful, simulated data is usually below 5%. In contrast, the largest difference in performance on the real data sets is over 35% (Breast-Cancer), while the difference of the pooled accuracies is 10%. Thus, violations of faithfulness seem to be the most probable explanation for the large difference in accuracy on the real data.

6.4 Summary, Interpretation, and Conclusions

We now comment and interpret the results of this section:

- Notice that even if all predicted pairs are truly correlated, the accuracy may not reach 100% due to the presence of Type II errors (false negatives) *in the test set*.

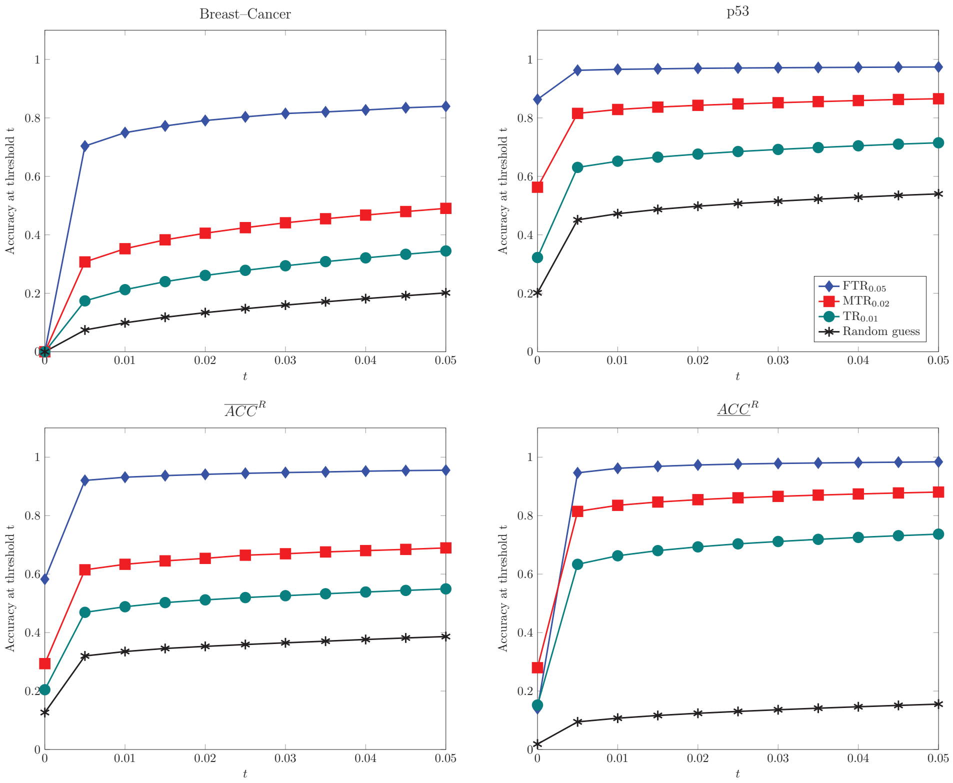


Figure 8: Accuracies $Acc_i^R(t)$ as a function of threshold t for two typical data sets along with $\overline{ACC}^R(t)$ and $\underline{ACC}^R(t)$. The remaining data sets are plot in Appendix A Section A.3. Predicted dependencies have p-values concentrated close to zero. The performance differences are insensitive to the threshold t in the performance definition.

- The FTR rule performs the test for the X-W association independently in both data sets. Given that the data in our experiments come from exactly the same distribution, they could be pooled together to perform a single test; alternatively, if this is not appropriate, the p-values of the tests could be combined to produce a single p-value (Tillman, 2009; Tsamardinos and Borboudakis, 2010).

- The results show that *the Full-Testing Rule accurately predicts the presence of dependencies*, statistically significantly better than random predictions, across all data sets, regardless of the type of data or the idiosyncracies of a domain. The rule is successful in gene-expression data, mass-spectra data measuring proteins, clinical data, images and others. The accuracy of predictions is robustly always above 0.80 and over all predictions it is 0.96; the difference with random predictions is of course more striking in data sets where the percentage of correlations (prior probability) is relatively small, as there is more room for improvement.

- *The Full-Testing Rule is noticeably more accurate than the Minimal-Testing Rule*, due to testing whether the Faithfulness Condition holds in the induced PAGs. The result is important considering that most constraint-based algorithms assume the Faithfulness Condition to induce models, *but do not check whether the induced model is Faithful*. These results indicate that when the latter is not the case, the model (and its predictions) may not be reliable. On the other hand, the FTR rule is also noticeably more conservative: the number of predictions it makes is significantly lower than the one made by MTR. In some data sets (e.g., Compactiv, Insurance-N, and Ovarian) by using the MTR vs. the FTR one sacrifices a small percentage of accuracy (less than 3% in these cases) to gain one order of magnitude more predictions. However, caution should be exercised because in certain data sets MTR is over 35% less accurate than FTR.

- *The Full-Testing Rule is more accurate than the Transitivity Rule*. Thus, the performance of the Full-Testing Rule cannot be attributed to simply performing a super-set of the tests performed by the Transitivity Rule.

- *Predictions are the norm case and not occur in contrived or rare cases only.* Even though there were few or no predictions for a couple of data sets, there are typically hundreds or thousands of predictions for each data set. This is the case despite the fact that we are only looking for a special-case structure and the search for these structures is limited within groups of 50 variables for the larger data sets. The results are consistent with the ones in Triantafillou et al. (2010), where larger structures were induced from simulated data.

- *FTR makes almost no predictions in the text data*:[3] this actually makes sense and is probably evidence for the validity of the method: it is semantically hard to interpret the presence of a word "causing" another word to be present.[4]

- FTR is an opportunistic algorithm that sacrifices completeness to increase accuracy, as well as improve computational efficiency and scalability. General algorithms for co-analyzing data over overlapping variable sets, such as ION (Tillman et al., 2008), IOD (Tillman and Spirtes, 2011) and cSAT (Triantafillou et al., 2010) could presumably make more predictions, and more general types of predictions (e.g., also predict independencies). However, their computational and learning performance on a wide range of domains and high-dimensional data sets is still an open question and an interesting future direction to pursue.

7. Predicting the Presence of Conditional Dependencies

The FTR and the MTR not only predict the presence of the dependency $Y \not\perp\!\!\!\perp Z|\emptyset$ given two data sets on $\mathbf{O}_1 = \{X,Y,W\}$ and $\mathbf{O}_2 = \{X,Z,W\}$; the rules also predict that either $X \circ\!-\!\circ Y \circ\!-\!\circ Z \circ\!-\!\circ W$ or $X \circ\!-\!\circ Z \circ\!-\!\circ Y \circ\!-\!\circ W$ is the model that generated both data sets (see Algorithms 1 and 2). Both of these models also imply the following dependencies:

$$Y \not\perp\!\!\!\perp Z|X,$$

[3] The only predictions in text data are in Bibtex (1 prediction) and in Delicious (856), which are the only text data sets that are actually not purely bag-of-words data sets but include variables corresponding to tags. 66% of the predictions made in Delicious involves tag variables, as well as the single prediction in Bibtex.

[4] However, causality between words is still conceivable in our opinion: deciding to include a word in a document may change a latent variable corresponding to a mental state of the author, which in turn causes her to include some other word.

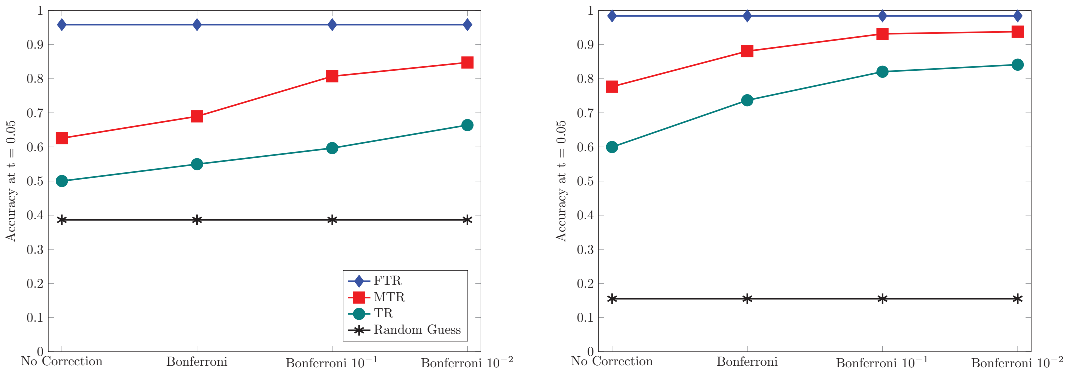


Figure 9: Average accuracy $\overline{Acc}(0.05)$ (left) and pooled accuracy $\underline{Acc}(0.05)$ (right) for each rule as a function of α thresholds used: $\alpha_{MTR} \in \{0.05, 0.02, 0.002, 0.0002\}$ and $\alpha_{TR} \in \{0.05, 0.01, 0.001, 0.0001\}$ corresponding to no correction, Bonferroni correction, and stricter than Bonferroni by one and two orders of magnitude respectively. FTR's performance is higher even when MTR and TR become quite conservative.

$$Y \not\perp\!\!\!\perp Z | W,$$

$$Y \not\perp\!\!\!\perp Z | \{X, W\}.$$

In other words, the rules predict that the dependency between Y and Z is not mediated by either X or W inclusively. To test whether all these predictions hold simultaneously at threshold t we compute:

$$p^* = \max_{\mathbf{S} \subseteq \{X,W\}} p_{Y \perp\!\!\!\perp Z | \mathbf{S}}$$

and test whether $p^* \leq t$. The above dependencies are all the dependencies that are implied by the model but not tested by the FTR given that it has no access to the joint distribution of Y and Z. Note that we forgo providing a value for p^* when any of the conditional dependencies can not be calculated, that is, when there are not enough samples to achieve large enough power, see Tsamardinos and Borboudakis (2010). The accuracy of the predictions for all dependencies in the model, named Structural Accuracy because it scores all the dependencies implied by the structure of model, is defined in a similar fashion to Acc (Definition 11) but based on p^* instead of p:

$$SAcc_i^R(t) = \#\{p^* <= t, p \in M_i^R\} / |M_i^R|.$$

The *SAcc* for each FTR, MTR (with "Bonferroni" correction) and randomly selected quadruples is shown in Figure 7.1; the remaining data sets are shown in Appendix A. There is no line for the TR as it concerns triplets of variables and makes no predictions about conditional dependencies. Both FTR and MTR have maximum p-values p^* concentrated around zero. The curves do not rise as sharp as those in Figure 8 since the p^* values are always larger than the corresponding $p_{Y \perp\!\!\!\perp Z|\emptyset}$. We also calculate the accuracy at $t = 0.05$ for all data sets (see Table 9 in Appendix A Section A.2). The results closely resemble the ones reported in Table 3, with FTR always outperforming random guess. FTR outperforms MTR on most data sets (and hence $\overline{SACC}^{FTR} > \overline{SACC}^{MTR}$; however, over all predictions their performance is quite similar.

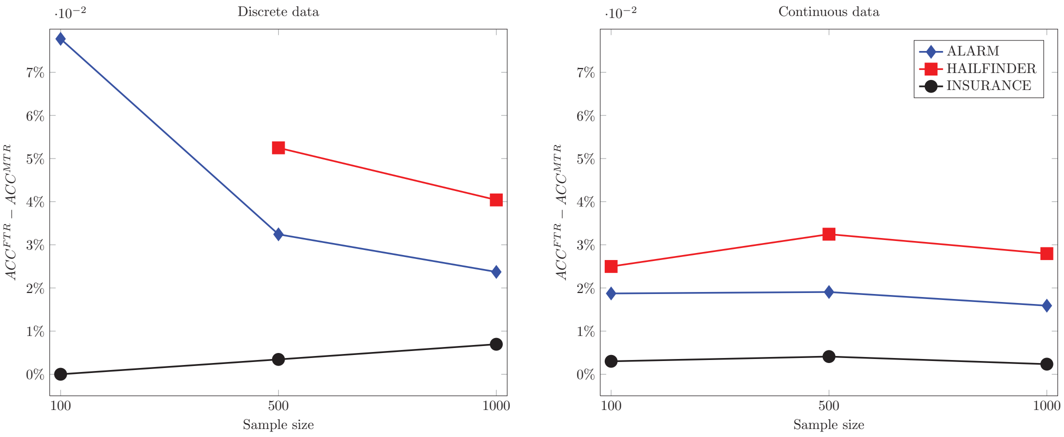


Figure 10: Difference between ACC^{FTR} and ACC^{MTR} for discrete (left) and continuous (right) simulated data sets. Results calculated using the "Bonferroni" correction (i.e., FTR$_{0.05}$ and MTR$_{0.02}$). The difference between FTR and MTR is larger than 5% only in two cases with low sample size (ALARM and HAILFINDER networks); however, the difference steeply decreases as the sample size increases. No prediction was made for HAILFINDER with discrete data and 100 samples. The difference between FTR and MTR on faithful data is relatively small.

7.1 Summary, Interpretation, and Conclusions

The results show that both the FTR and MTR rules correctly predict all the dependencies (conditional and unconditional) implied by the models involving the two variables never measured together. These results provide evidence that these rules often correctly identify the data generating structure.

8. Predicting the Strength of Dependencies

In this section, we present and evaluate ideas that turn the qualitative predictions of FTR to quantitative predictions. Specifically, for Example 1 we show *how to predict the strength of dependence* in addition to its existence. In addition to the Faithfulness Condition, we assume that when the FTR applies on quadruple $\{X,Y,Z,W\}$, all dependencies are linear with independent and normally distributed error terms. However, the results of these section could possibly be generalized to more relaxed settings, for example, when some of the error terms are non-Gaussian (Shimizu et al., 2006, 2011). When the Full-Testing Rule applies, we can safely assume the true structure is one of the MAGs shown in Figure 5. Given linear relationships among the variables, we can treat these MAGs as linear Path Diagrams (Richardson and Spirtes, 2002). We also consider normalized versions of the variables with zero mean and standard deviation of one. Let us consider one of the possible MAGs:

$$M_1 : X \xleftarrow{\rho_{XY}} Y \xrightarrow{\rho_{YZ}} Z \xrightarrow{\rho_{ZW}} W$$

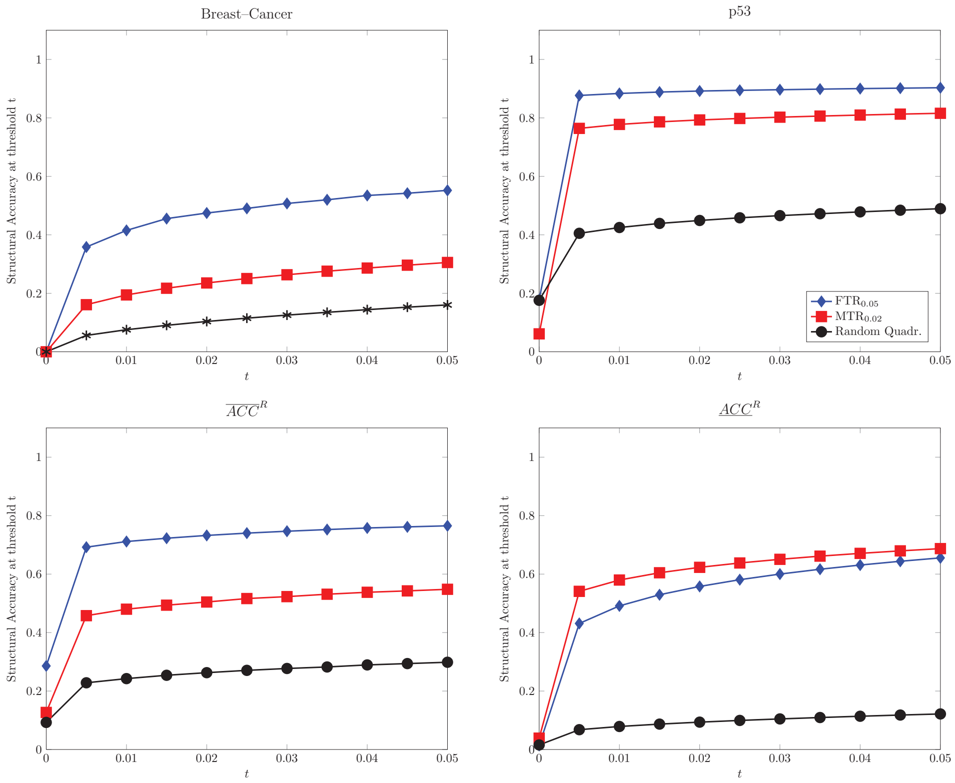


Figure 11: Structural Accuracies $SAcc_i^R(t)$ as a function of threshold t for two typical data sets along with $\overline{SACC}^R(t)$ and $\underline{SACC}^R(t)$. The remaining data sets are plot in Appendix A Section A.2. FTR outperforms MTR on most of the data sets, and thus $\overline{SACC}^{FTR}(t) > \overline{SACC}^{MTR}(t)$. However, since MTR ouperforms FTR on few data sets with a large number of predictions and so $\underline{SACC}^{MTR}(t)$ is slightly better than $\underline{SACC}^{FTR}(t)$ for $t <= 0.05$.

where ρ_{XY} is the *regression coefficient* of regressing X on Y, that is,

$$X = \rho_{XY} Y + \varepsilon$$

and ε is the error term. Since we have standardized the variables, and since the above equation is simple linear regression, ρ_{XY} coincides with the Pearson linear *correlation* between variables X and Y. Thus, there is no need to distinguish the two.[5] Now notice that in all MAGs in Figure 5 there are no colliders. Thus, as in M_1 above, all regressions are simple regressions and all standardized regression coefficients coincide with their respective correlation coefficients, and so, for the rest of the section we will not differentiate between the two.

The rules of path analysis (Wright, 1934) dictate that the correlation between two variables, for example, ρ_{XY} equals the sum of the contribution of every d-connecting path (conditioned on the

[5] If Y was a collider then it would have been regressed on multiple variables; in this case ρ_{XY} should be the partial regression coefficient which in general does not coincide with the partial correlation coefficient, even for standardized variables.

empty set); the contribution of each path is the product of the correlations on its edges. For M_1 the above rule implies (among others):

$$\rho_{XZ} = \rho_{XY} \times \rho_{YZ}$$

because from X to Z there is a single path going through Y. Recall that the 14 consistent MAGs are represented by the following PAGs:

$$P_1 : X \circ\!-\!\circ\, Y \circ\!-\!\circ\, Z \circ\!-\!\circ\, W$$

and

$$P_2 : X \circ\!-\!\circ\, Z \circ\!-\!\circ\, Y \circ\!-\!\circ\, W.$$

All MAGs consistent with P_1 entail the same constraints on the coefficients using path analysis; similarly all MAGs consistent with P_2.[6] Specifically, if P_1 is the true structure we get the constraints

$$\rho_{XZ} = \rho_{XY} \times \rho_{YZ}, \tag{1}$$

$$\rho_{YW} = \rho_{YZ} \times \rho_{ZW}. \tag{2}$$

On the other hand, if P_2 is the true structure we obtain:

$$\rho_{XY} = \rho_{XZ} \times \rho_{YZ}, \tag{3}$$

$$\rho_{ZW} = \rho_{YZ} \times \rho_{YW}. \tag{4}$$

We use ρ, $\hat{r}$, *and* r *to denote actual, predicted, and sample correlations, respectively.* The quantities that we observe are the *sample correlation coefficients*, denoted by r, for the pairs of variables measured together. Thus, we can compute the quantities $r_{XY}, r_{XZ}, r_{YW}, r_{ZW}$ from the data and we would like to predict ρ_{YZ} without available data. From Equations 1, 2, 3, 4 above we obtain four possible estimators:

$$\text{If } P_1 \text{ is true } : \hat{r}^1_{YZ} \approx \frac{r_{XZ}}{r_{XY}} \text{ from Equation 1 and } \hat{r}^2_{YZ} \approx \frac{r_{YW}}{r_{ZW}} \text{ from Equation 2,} \tag{5}$$

$$\text{if } P_2 \text{ is true } : \hat{r}^3_{YZ} \approx \frac{r_{XY}}{r_{XZ}} \text{ from Equation 3 and } \hat{r}^4_{YZ} \approx \frac{r_{ZW}}{r_{YW}} \text{ from Equation 4} \tag{6}$$

where the superscripts correspond to the equation used to produce the estimate. Notice that, each possible PAG provides two equations to predict ρ_{YZ}, that is, the parameter is overidentified. Also, the following important relation holds between the estimators:

$$\hat{r}^1_{YZ} = \frac{1}{\hat{r}^3_{YZ}} \text{ and } \hat{r}^2_{YZ} = \frac{1}{\hat{r}^4_{YZ}}.$$

This observation allows us to distinguish between PAGs P_1 and P_2: if $\hat{r}^1_{YZ}, \hat{r}^2_{YZ} \in [-1, +1]$, then their reciprocals $\hat{r}^3_{YZ}, \hat{r}^4_{YZ} \notin [-1, +1]$ and so, they are not valid estimates for a correlation. Thus, we can infer that P_1 is the true structure and employ only $\hat{r}^1_{YZ}, \hat{r}^2_{YZ}$ for estimation. Otherwise, the reverse holds $\hat{r}^3_{YZ}, \hat{r}^4_{YZ} \in [-1, +1]$, P_2 is the true structure and only $\hat{r}^3_{YZ}, \hat{r}^4_{YZ}$ should be used for estimation.

[6] In general, the consistent MAGs may disagree on the unknown correlations. In this case, these parameters may not identifiable. However, one could analyze all possible MAGs to provide bounds on the unidentifiable quantities in a similar fashion to Balke and Pearl (1997) and Maathuis et al. (2009).

Due to sampling errors it is plausible that we obtain conflicting information: $\hat{r}^1_{YZ} \in [-1,+1]$ but $\hat{r}^2_{YZ} \notin [-1,+1]$ (and so $\hat{r}^3_{YZ} \notin [-1,+1]$ and $\hat{r}^2_{YZ} \in [-1,+1]$). In that case, we forgo making any predictions.

The ramifications of the above analysis are important. In the case where all variables are jointly measured, the distribution is Faithful, the relations are linear and the error terms follow Gaussian distributions, the set of statistically indistinguishable causal graphs is determined completely by the independence model and not by the parameterization of the distribution. However, in the case of incomplete data, where some variable sets are not jointly observed, the set of indistinguishable models also depends on the parameters of the distribution, even for linear relations and Gaussian error terms. In our scenario, by analyzing the estimable parameters we can further narrow down the set of equivalent consistent MAGs.

At this point in our analysis, we are left with two valid estimators, either $\hat{r}^1, \hat{r}^2$ or $\hat{r}^3, \hat{r}^4$. All estimators are computed as ratios. We report the mean of the two valid estimators as the predicted $\hat{r}_{YZ}$ for a more robust estimation. The above procedure is formalized in Algorithm 4, named FTR-S.

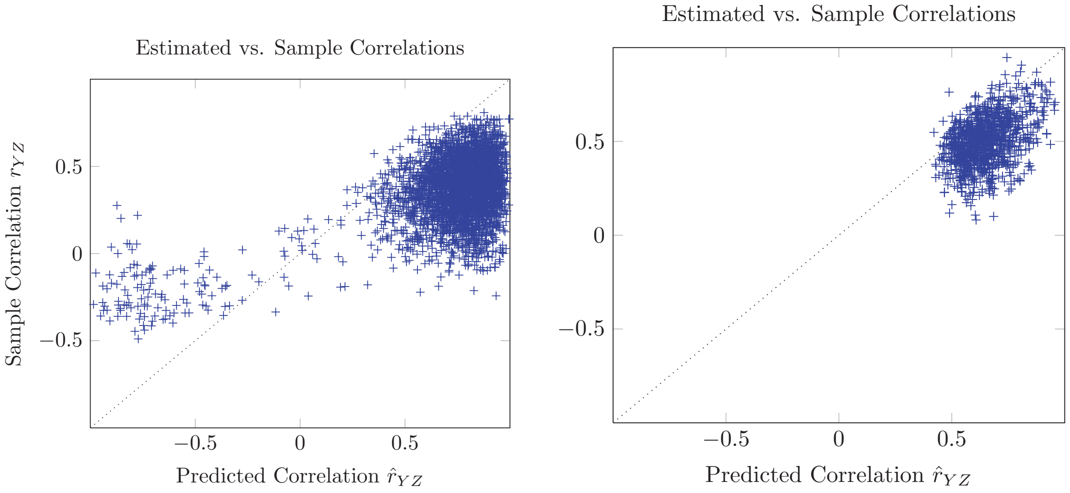
Algorithm 4: Predict Dependency Strength(**FTR-S**)

```
Input: Data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ on variables $\{X,Y,W\}$ and $\{X,Z,W\}$, respectively
if Full-Testing Rule($\mathcal{D}_1$, $\mathcal{D}_2$) does not apply then return;
In $\mathcal{D}_1$ compute $r_{XY}, r_{YW}$;
In $\mathcal{D}_2$ compute $r_{XZ}, r_{ZW}$;
$\hat{r}^1 \leftarrow \frac{r_{XZ}}{r_{XY}}$;
$\hat{r}^2 \leftarrow \frac{r_{YW}}{r_{ZW}}$;
$\hat{r}^3 \leftarrow \frac{r_{XY}}{r_{XZ}}$;
$\hat{r}^4 \leftarrow \frac{r_{ZW}}{r_{YW}}$;
if $\hat{r}^1, \hat{r}^2 \in [-1,1]$ then
    Predict $X \circ\!-\!\circ Y \circ\!-\!\circ Z \circ\!-\!\circ W$;
    Predict correlation $\hat{r}_{YZ} = \frac{1}{2}(\hat{r}^1 + \hat{r}^2)$;
end
else if $\hat{r}^3, \hat{r}^4 \in [-1,1]$ then
    Predict $X \circ\!-\!\circ Z \circ\!-\!\circ Y \circ\!-\!\circ W$;
    Predict correlation $\hat{r}_{YZ} = \frac{1}{2}(\hat{r}^3 + \hat{r}^4)$;
end
else
    Make no prediction
end
```

8.1 Empirical Evaluation of the Predictions of Correlation Strength

As in Section 6, we partition each data set with continuous variables to three data sets $\mathcal{D}_1$, $\mathcal{D}_2$, and a test set $\mathcal{D}_t$. We then apply Algorithm 4 and predict the strength of correlation $\hat{r}_{YZ}$ for various pairs of variables; we compare the predictions with the sample correlation r_{YZ} as estimated in $\mathcal{D}_t$. The results for one representative data set (Lymphoma) are shown in Figure 12(a): there is an apparent trend to overestimate the absolute value of the sample correlation.

There are several possible explanations for the bias of the method, including violations of normality, linearity, faithfulness, and even the known bias in the estimation of sample correlation coefficients (Zimmerman et al., 2003) that are used for making the predictions in Algorithm 4. In order to pinpoint the culprit, we generated data where all assumptions hold from the model M_1 shown in the beginning of this section, where we set the correlations $\rho_{XY}, \rho_{YZ}, \rho_{ZW}$ and the noise terms are independently and normally distributed. We used the entire spectrum of positive correlation coefficients for all three correlations to examine how the bias varies as a function of these correlations. We generated 1000 data sets of different sample sizes of 50, 70 and 100 samples. We then used Equation 1 to estimate r_{YZ} in each experiment. *This set of experiments revealed no significant bias for any of the experimental settings* (results are not shown for brevity).



(a) Lymphoma Data Set

(b) Simulated Data where FTR Rule Applies

Figure 12: (a) Predicted ($\hat{r}_{YZ}$) vs sample (r_{YZ}) correlation for the Lymphoma data set. There is an obvious trend to over-estimate the correlation in absolute value. (b) Simulated results from model $\mathcal{M}_1$ when ρ_{XZ} and ρ_{YW} are lower than 0.4 and observed correlations *are found significant* (FTR applies). The FTR constraint that the observed correlations are significant reproduces a similar behavior in the simulated data, explaining the bias.

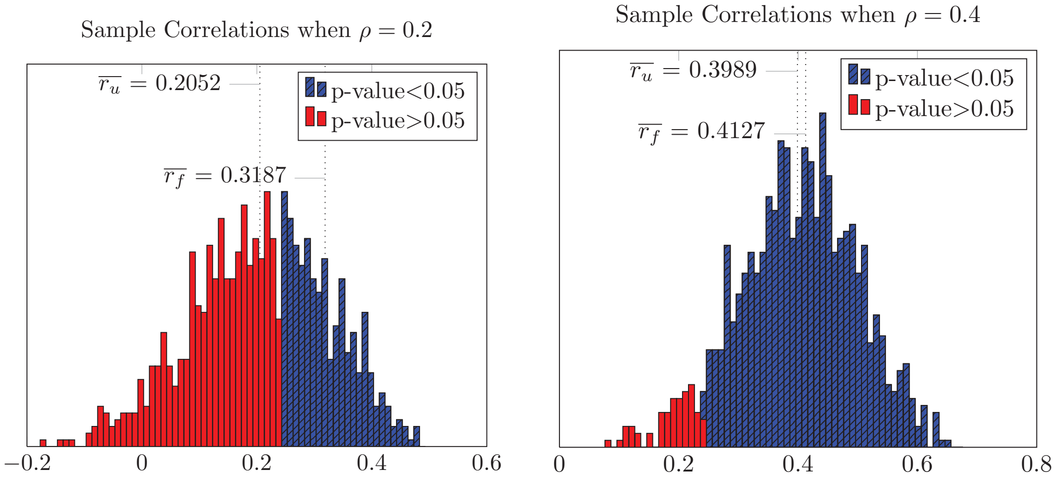
We next tested whether the bias is an artifact of the filtering by the FTR at Line 1 of the FTR-S algorithm. We re-run this procedure, but this time we kept only the predicted correlations that passed the FTR. By comparing Figure 12(a) produced on real data, and 12(b) on simulated data, we observe a similar behavior, indicating that FTR filtering seems a reasonable explanation for the bias.

An explanation of this phenomenon now follows. Suppose $M_1 : X \xleftarrow{\rho_{XY}} Y \xrightarrow{\rho_{YZ}} Z \xrightarrow{\rho_{ZW}} W$ is the data generating MAG. We expect that $\hat{r}_{YZ} = \frac{r_{XZ}}{r_{XY}}$ (the equality $\hat{r}_{YZ} = \frac{r_{YW}}{r_{ZW}}$ also holds but we ignore it to simplify the discussion). When sample correlations among $\{X,Y,Z,W\}$ pass the FTR, this means that both r_{XZ} and r_{XY} are above a cut-off threshold, as given by the Fisher test. For example, for a data set with 70 samples, two variables are considered dependent ($\rho \neq 0$) if their sample correlation is more that 0.2391 (in absolute value), whereas for a data set with 50 samples, this threshold is 0.2852.

Filtering with the Fisher test introduces a bias in the estimation of r. The bias of the estimation without filtering, r_u is $B_{r_u} = E[r_u - \rho] = \overline{r_u} - \rho$, while the bias of the estimation with filtering r_f is $B_{r_f} = E[r_f - \rho] = \overline{r_f} - \rho$, where $|r_f| \geq t$. The threshold t, as mentioned above, is the threshold determined by the Fisher test and depends on sample size. *The lower the sample size, the higher the threshold t, and so the higher the introduced bias B_{r_f}. In addition, the lower the $|\rho|$ the higher the bias B_{r_f}.*

Figure 13 illustrates these points pictorially. In this example, the distribution of the sample correlation r of two variables for sample size 70 when the true correlation is $\rho \in \{0.2, 0.4\}$. For unfiltered estimations, the bias is B_{r_u} is 0.0052 and -0.0011 for ρ equal to 0.2 and 0.4 respectively, whereas for filtered estimations the corresponding values B_{r_f} are 0.1187 and 0.0127.

Going back to the prediction $\hat{r}_{YZ} = \frac{r_{XZ}}{r_{XY}}$ notice that the numerator is always lower (in absolute value) than the denominator. Therefore, when filtered, it is, on average more overestimated than the denominator. This implies that, on average, the fraction leads to overestimating the absolute value of ρ_{YZ}. The lower the values of $|r_{XZ}|$ and $|r_{XY}|$, the larger we expect this bias to be. The situation is similar for all fractions involved in Equations 5 and 6. This hypothesis is confirmed in the data as illustrated in Figure 14 where the predictions are grouped by the mean absolute values of the denominators used in their computation.



(a) (b)

Figure 13: Histograms of the sample correlations for (a) $\rho = 0.2$ and (b) $\rho = 0.4$ for sample size 70. Red bars correspond to cases where the Fisher test returns a p-value > 0.05, whereas blue bars correspond to p-values < 0.05. The dashed lines indicate the mean sample correlation for filtered and unfiltered correlations. The lower the ρ , the more overestimated the sample correlations that pass the Fisher test, therefore the difference between the two means is larger.

The bias should be a function of sample size, the absolute value of the correlations employed for its computation, and the significance thresholds of the FTR rule. However, a full theoretical treatment of the bias is out of the scope of the paper. In the experiments that follow we remove the linear trend to over-estimate (*calibrate*) by regressing the sample correlations r_{YZ} on the predicted $\hat{r}_{YZ}$: the final calibrated prediction is $s \times \hat{r}_{YZ} + i$. For each data set the intercept i and slope s of the regression are estimated by training on the remaining data sets (leave-one-data-set-out validation). The effect of this calibration is shown in Figure 15. To avoid repetition, the detailed set of results is presented in the comparative evaluation to statistical matching in Section 9.

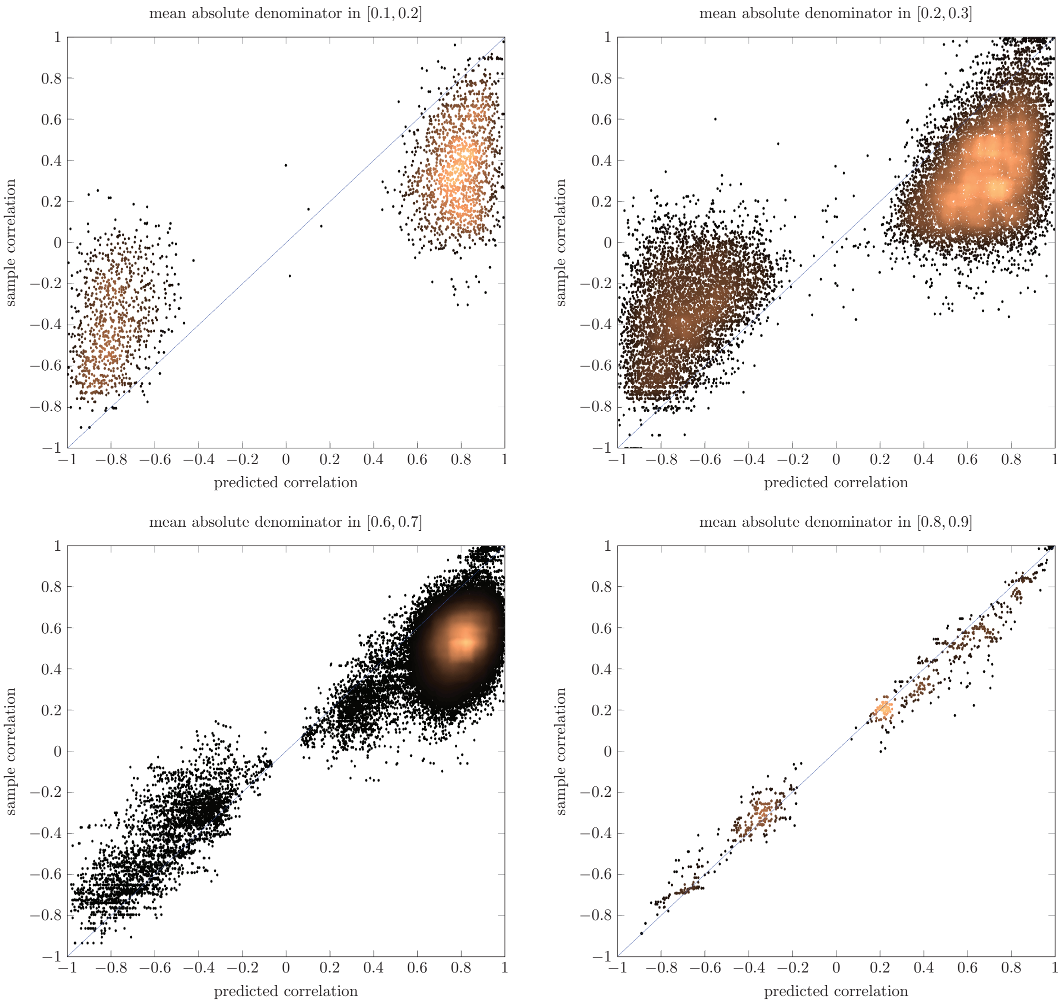
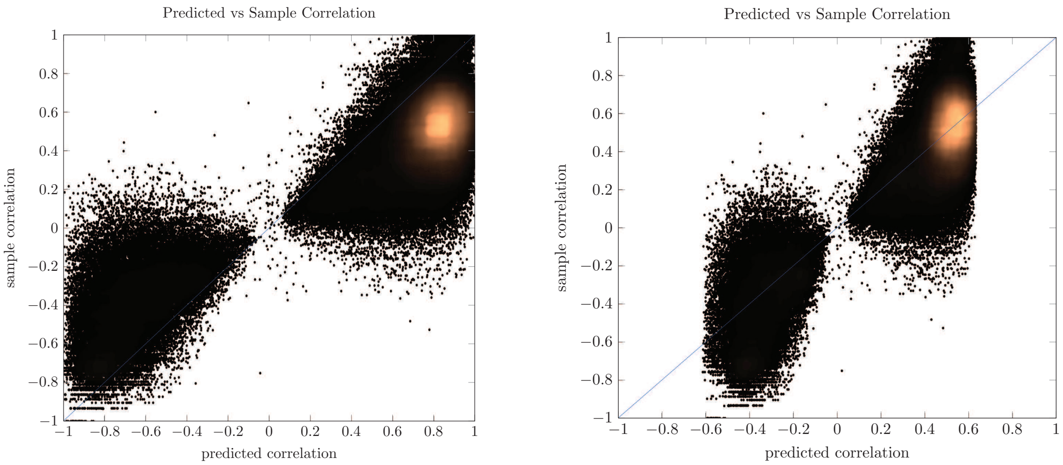


Figure 14: Predicted vs sample correlations over all data sets, grouped by the mean absolute values of the denominators used in their computation: predictions computed based on large correlations have reduced bias. Red regions correspond to higher density areas.

8.2 Summary, Interpretation, and Conclusions

We now comment and interpret the results of this section:



(a) (b)

Figure 15: Predicted vs sample correlations on all data sets (a) before, and (b) after calibration.

- FTR coupled with parametric assumptions can be used to predict the strength of dependency (correlation), providing quantitative predictions. This is equivalent to constructing a prediction model for variables not jointly observed.

- In the case of incomplete data, where some variable sets are not jointly observed, the set of indistinguishable models also depends on the parameters of the distribution, even for linear relations and Gaussian error terms. In contrast, in the case where all variables are jointly measured and the distribution is Faithful the set of statistically indistinguishable causal graphs is completely determined by the independence model (again, also assuming linearity and Gaussian error terms).

- In our simple scenario, *given the correct structure*, path analysis of the induced MAGs provides easy solutions for predicting the strength of dependence. However, *searching for the correct MAG models* by applying the FTR incurs bias on the predictions that should be taken into account.

9. Comparison Against Statistical Matching

Statistical Matching (D'Orazio et al., 2006) is a integrative analysis procedure for data sets defined over overlapping variable sets. Statistical matching addresses two main tasks named the *micro approach* and *the macro approach*. The micro approach aims to impute the missing values and construct a complete synthetic file, whereas the macro approach aims to identify some characteristics of the joint probability distribution of the variables not jointly observed. Naturally, construction of the synthetic data set premises the estimation of the parameters of the joint distribution. We focus on the macro approach as it presents an alternative to the FTR and MTR.

The problem set up is as follows: variables $\mathbf{Y} \cup \mathbf{X}$ are measured in data set $\mathcal{D}_1$, while variables $\mathbf{Z} \cup \mathbf{X}$ are measured in data set $\mathcal{D}_2$. Thus $\mathbf{X}$ are the commonly measured variables. The goal is to estimate the variances and covariances of $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$. The problem cannot be solved without additional assumptions (Rubin, 1974; D'Orazio et al., 2006). Depending on nature of the assumptions, statistical matching is able to produce either intervals or point-estimates for the covariances between $\mathbf{Y}$ and $\mathbf{Z}$. The most typical assumption in the literature able to produce point estimates is the Conditional Independence Assumption: $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}$. This is an arbitrary assumption that has been long debated. Alternatively, one can limit the shape of the distribution by imposing parametric forms, such as multivariate normality. The latter type of assumptions, for the typical distributions, do not lead to identifiable estimations, but instead provide bounds on the missing covariances. Other approaches do exist that require prior knowledge, for example, Vantaggi (2008) assumes knowledge of structural zeros and Cudeck (2000) of the structure of latent factors; such approaches however, are not directly comparable with FTR and MTR on this task. In this section we briefly present the main theory and techniques used in statistical matching, and then attempt to empirically compare against FTR.

9.1 Statistical Matching Based on the Conditional Independence Assumption

The most common assumption that allows identification of the unknown parameters is the **conditional independence assumption** (CIA): $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z}|\mathbf{X}$. The conditional independence assumption is usually paired with some parametric assumption. The most common assumption for the shape of a continuous distribution of the variables involved in the model is multivariate normality. In this case, the parameters of the jpd are the mean vector and the covariance matrix. The covariance Matrix for $\mathbf{X} \cup \mathbf{Y} \cup \mathbf{Z}$ can be written as:

$$\Sigma = \begin{bmatrix} \Sigma_{\mathbf{XX}} & \Sigma_{\mathbf{XY}} & \Sigma_{\mathbf{XZ}} \\ \Sigma_{\mathbf{YX}} & \Sigma_{\mathbf{YY}} & \Sigma_{\mathbf{YZ}} \\ \Sigma_{\mathbf{ZX}} & \Sigma_{\mathbf{ZY}} & \Sigma_{\mathbf{ZZ}} \end{bmatrix}$$

where the unknown parameter is $\Sigma_{\mathbf{YZ}}$. The CIA assumption imposes that the covariance matrix of $\mathbf{Y}$ and $\mathbf{Z}$ given $\mathbf{X}$ is null, thus,

$$\Sigma_{\mathbf{YZ}} = \Sigma_{\mathbf{YX}} {\Sigma_{\mathbf{XX}}}^{-1} \Sigma_{\mathbf{XZ}}.$$

In case we have standardized variables, and $\mathbf{Y} = \{Y\}$ and $\mathbf{Z} = \{Z\}$, the covariance matrix becomes

$$\Sigma = \begin{bmatrix} \rho_{\mathbf{XX}} & \rho_{\mathbf{X}Y} & \rho_{\mathbf{X}Z} \\ \rho_{Y\mathbf{X}} & 1 & \rho_{YZ} \\ \rho_{Z\mathbf{X}} & \rho_{ZY} & 1 \end{bmatrix}$$

and so

$$\rho_{YZ} = \rho_{Y\mathbf{X}} {\rho_{\mathbf{XX}}}^{-1} \rho_{\mathbf{X}Z}.$$

This formula can be used to produce a prediction $\hat{r}_{YZ}$ for the correlation coefficient of the not commonly observed variables Y and Z. Recall that, we assume we are given a data set $\mathcal{D}_1$ on variables $\mathbf{X} \cup Y$ and a data set $\mathcal{D}_2$ on $\mathbf{X} \cup Z$. The parameters $\rho_{\mathbf{X}Y}$ and $\rho_{\mathbf{X}Z}$ can be estimated from $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively, while the parameters $\rho_{\mathbf{XX}}$ can be estimated from either or both data sets.

In an applied setting, there is usually also a preprocessing step attempting to identify a subset of the common variables to be used in the matching process. This step serves mainly computational efficiency and interpretability purposes and does not affect the asymptotic properties of the procedure. The main method suggested in D'Orazio et al. (2006) is to disregard all variables in $\mathbf{X}$ that are *independent* with both Y and Z. The details are described in Algorithm 5.

Even though the conditional independence assumption seems quite arbitrary, it is intuitively justified in certain cases. When the number of common variables is large it is unlikely that Y

```
Algorithm 5: Predict Correlation: Statistical Matching Rule (SMR)
  Input: Data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ on variables $\{\mathbf{V} \cup Y\}$ and $\{\mathbf{V} \cup Z\}$, respectively
  $\psi_1 \leftarrow \{V \in \mathbf{V} : V \perp\!\!\!\perp Y | \emptyset\}$ in $\mathcal{D}_1$
  $\psi_2 \leftarrow \{V \in \mathbf{V} : V \perp\!\!\!\perp Z | \emptyset\}$ in $\mathcal{D}_2$
  $\mathbf{X} \leftarrow \mathbf{V} \setminus (\psi_1 \cap \psi_2)$
  Predict $\hat{r}_{YZ} = \hat{\Sigma_{Y\mathbf{X}}} \hat{\Sigma_{\mathbf{XX}}}^{-1} \hat{\Sigma_{\mathbf{X}Z}}$
```

provides *additional* information for Z, than what $\mathbf{X}$ already provides. In other words, we expect $\mathbf{Y} \perp\!\!\!\perp \mathbf{Z} | \mathbf{X}$ to hold or hold approximately. Using graphical model theory one can better formalize this intuition:

Theorem 12 *Consider a Bayesian Network of maximum degree k faithful to a distribution defined over a set of variables $\mathbf{V} = \mathbf{X} \cup Y \cup Z$, $|\mathbf{V}| = N$. Then, the CIA $Y \perp\!\!\!\perp Z | \mathbf{X}$ holds if and only if $Y \notin Mb(Z)$, where $Mb(Z)$ is the Markov Boundary of Z in the context of variables $\mathbf{V}$; if Y and Z are chosen at random the probability of the CIA being violated is upper bounded by k^2/N.*

Proof In a faithful distribution over $\mathbf{V}$, each variable Y has a unique Markov Boundary $Mb(Y)$ (Pearl, 2000) that coincides with the parents, children, and parents of children (spouses) of Y in any network faithful to the distribution. It is also easy to see that $Y \in Mb(Z) \Leftrightarrow Z \in Mb(Y)$. Finally, the $Mb(Y)$ and any of its supersets d-separates Y from any other node Z. Thus, when $Z \notin Mb(Y)$, then conditioned on the remaining variables (superset of $Mb(Y)$) Y becomes d-separated and independent of Z. Thus, the CIA holds. Conversely, if $Z \in Mb(Y)$ then it is either a neighbor of Y or a spouse. If it is a neighbor it cannot be made independent of Y conditioned on any subset of the variables (Spirtes et al., 2001). If it is a spouse of Y, then conditioned on the remaining variables (which includes the common children) it is d-connected to Y and thus dependent. Thus, the CIA does not hold.

Now, the Markov Boundary of Y is a subset of the nodes that are reachable from Y within two edges. If the network has degree at most k the probability that a randomly chosen Y belongs to the Markov Boundary of Z is less than k^2/N. ■

Thus, when the sparsity remains the same, the probability of a violation of the CIA between two randomly selected variables decreases with the number of participating variables N. The theoretical results is illustrated in Figure 16 on simulated data. The figure shows the results of the statistical matching procedure described in Algorithm 5 for simulated continuous data from a network based on the ALARM Network (Beinlich et al., 1989).[7] To recreate the scenario above we generated two data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ of 1000 samples each from the distribution of the network. We then applied the statistical matching rule described in Algorithm 5 for each pair of variables, considering that the rest of the variables in the network are jointly measured in both data sets. Finally, we generated a third data set to test the predictions of the method. The pairs of variables are partitioned in two categories: pairs of variables that belong to each other's Markov Boundary, and pairs of variables that do not belong to each other's Markov Boundary. As expected, the results are poorer for the

[7] The ALARM network a well-known network with 37 variables. We used the skeleton of ALARM to simulate a conditional linear gaussian network with random parameters.

pairs of variables that belong to each other's Markov Boundary, with a mean absolute error of 0.1649 ± 0.1088, compared to a mean absolute error of 0.0326 ± 0.0271 for pairs that do not belong to each other's Markov Boundary.

In the context of Maximal Ancestral Graphs, defining the Markov Boundary is more complicated and its cardinality cannot be likewise bounded (Pellet and Elisseeff, 2008). Nevertheless, we still expect that, in a sparse network containing a large number of jointly measured variables, the probability that $Y \in Mb(Z)$ is low. We therefore expect that, when the number of common variables is large, the CIA will often hold for randomly-chosen pairs of variables that have not been observed together. If, however, the set of variables measured in common is small, we have no good reason to expect that the conditional independence assumption holds.

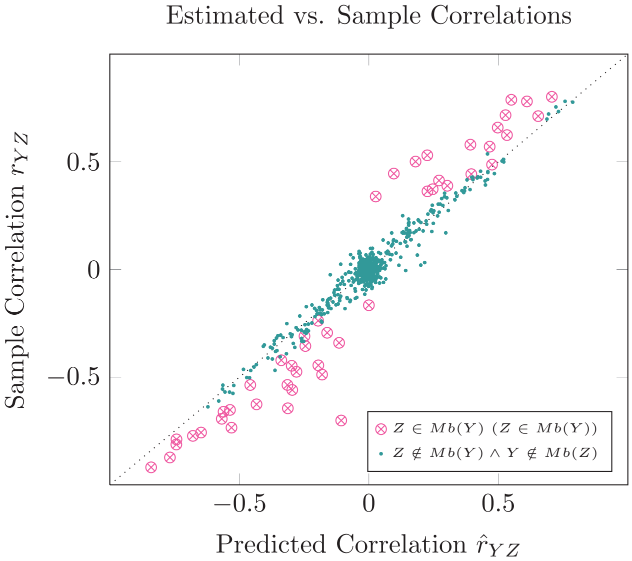


Figure 16: Predicted vs actual sample correlations using the Statistical Matching Rule for simulated data from the ALARM network. For each pair of variables, prediction is based upon the subset of the remaining 35 variables that are determined significantly correlated with either Y or Z at level 0.05 . The CIA holds when $Y \notin Mb(Z)$ in which case the mean absolute error is 0.0326 ± 0.0271; in contrast, when $Y \in Mb(Z)$ the CIA does not hold and the mean absolute error is 0.1649 ± 0.1088.

9.2 Empirical Evaluation of SMR and FTR-S

In this section, we empirically compare the SMR and FTR-S methods for predicting the correlation $\hat{r}_{YZ}$ between two variables Y and Z never jointly observed. Both SMR and FTR-S procedures provide such predictions, however, they follow different approaches that makes their comparison not straightforward:

- SMR provides a prediction for all cases. FTR-S provides a prediction given it identifies a specific structure that entails a significant correlation.

Data Sets	SMR$_G$	SMR$_Q$	FTR-S
ACPJ	445121	509000	0
Breast-Cancer	436093	356000	1005
C&C	5050	1000	70367
Compactiv	231	1000	108
Insurance-C	3486	1000	1372
Lymphoma	180074	147000	3897
Ohsumed	124505	122000	0
Ovarian	52675	43000	273456
Wine	66	495	4
p53	132299	108000	33934

Table 4: Number of predictions

- SMR can be applied to sets X with more than two commonly measured variables and get leverage from all available information. FTR-S on the other hand is applicable only when the number of common variables is two.

We applied the SMR method on all continuous data sets, simulating two scenarios. In the first scenario, SMR is applied on two data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ defined over a quadruple of variables $\{X,Y,Z,W\}$, where only X,W are jointly measured in both. The pairs of $\mathcal{D}_1$, $\mathcal{D}_2$ are simulated by considering randomly chosen variable quadruples from each variable group of each data set of Table 1; as in all experiments, $\mathcal{D}_1$ and $\mathcal{D}_2$ contain a disjoint third of the original samples. This scenario simulates a case where SMR is applied on low dimensional data; we denote it as *SMR$_Q$*. In this case, *SMR has the same information available for making predictions as FTR-S*. Since the number of possible quadruples is computationally prohibitive, we apply *SMR$_Q$* on 1000 randomly chosen quadruples from each variable group of each data set.[8] In the second scenario, SMR is applied to data sets of higher-dimensionality. Specifically, we apply SMR to all pairs of variables in the same group (see Section 6), considering the remaining 48 variables in the group as the common variables $\mathbf{X}$. We name this case *SMR$_G$*. The same leave-one-data-set-out calibration method was used for both SMR cases and FTR-S. Figures 17, 18, 19 and 20 plot the predicted vs. the sample estimates of the correlations for *SMR$_G$*, *SMR$_Q$* and FTR-S for all the continuous data sets used in the study. The figures also present the coefficient of determination R^2, the percentage of variance explained by the predictions. R^2 is also interpreted as the reduction in uncertainty obtained by using a linear function of $\hat{r}$ to predict r vs. predicting r by its expected value $E(r)$. Table 5 shows the correlation between predicted and sample estimates for all methods and data sets. Notice that R^2 is simply computed as the square of the correlation. Other metrics of performance (Mean Absolute Error and Mean Relative Absolute Error) are also presented in the Appendix A, Tables 10, 11.

[8] Notice that FTR is typically executed much more efficiently than SMR$_Q$, because of the possible pruning of the search space, for example, if X and Y are independent, there is no need to test whether the rule applies on any quadruples of the form $\langle X,Y,Z,W\rangle$. For the SMR$_Q$ rule instead, one needs to exhaustively consider all quadruples.

Data Sets	SMR_G	SMR_Q	FTR-S
ACPJ	0.05 [0.04; 0.05]	0.00 [0.00; 0.01]	-
Breast-Cancer	0.55 [0.55; 0.55]	0.25 [0.24; 0.25]	0.88 [0.87; 0.90]
C&C	0.99 [0.99; 0.99]	0.68 [0.65; 0.71]	0.91 [0.91; 0.91]
Compactiv	0.97 [0.96; 0.98]	0.49 [0.44; 0.54]	0.88 [0.83; 0.92]
Insurance-C	0.83 [0.82; 0.84]	0.47 [0.42; 0.51]	0.90 [0.89; 0.91]
Lymphoma	0.60 [0.60; 0.60]	0.32 [0.31; 0.32]	0.50 [0.47; 0.52]
Ohsumed	0.02 [0.01; 0.03]	0.01 [0.00; 0.01]	-
Ovarian	0.62 [0.62; 0.63]	0.50 [0.50; 0.51]	0.14 [0.14; 0.14]
Wine	0.83 [0.74; 0.90]	0.58 [0.52; 0.64]	0.99 [0.47; 1.00]
p53	0.91 [0.91; 0.91]	0.45 [0.44; 0.45]	0.87 [0.87; 0.87]
Mean over data sets	0.64 [0.62; 0.65]	0.38 [0.35; 0.40]	0.76 [0.68; 0.77]
On all predictions	0.73 [0.73; 0.73]	0.58 [0.57; 0.58]	0.89 [0.89; 0.89]

Table 5: Correlations among predicted $\hat{r}_{YZ}$ and sample-estimated r_{YZ}; the 95% confidence intervals are shown in brackets.

9.3 Summary, Interpretation, and Conclusions

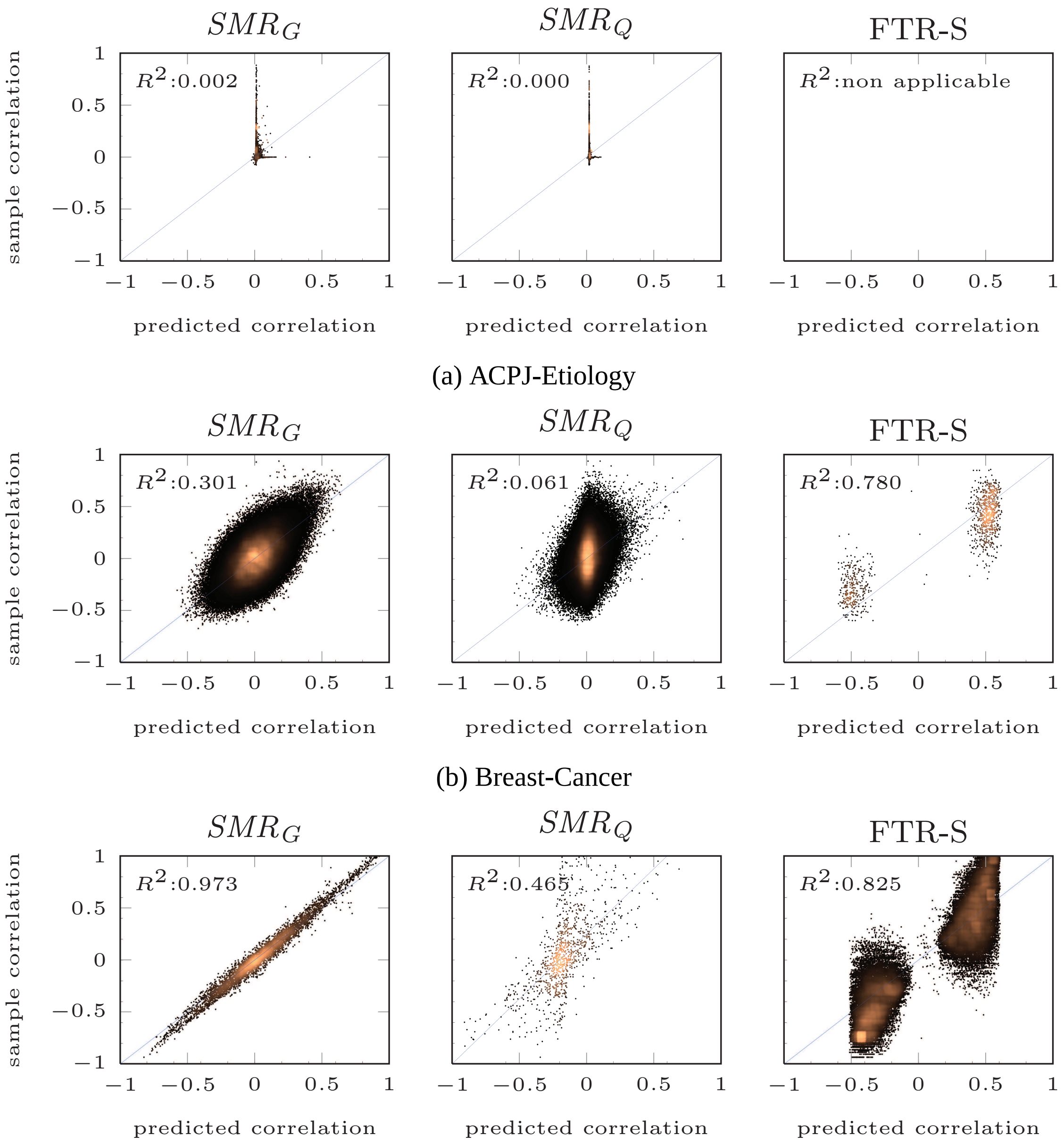
The CIA assumption is the most common assumption in statistical matching to produce point-estimates of the unknown distribution parameters. In comparison to FTR-S, we note the following:

- When predictions are based on only 2 common variables, statistical matching based on the CIA (SMR_Q) is unreliable in several data sets and particularly the text categorization ones: the correlation of predicted vs. sample estimates in ACPJ, Breast-Cancer, and Ohsumed is less than 0.3 (Table 4). In general, SMR tends to predict a zero correlation between the two variables Y and Z: the point-clouds in Figures 17, 18, 19 and 20 are vertically oriented around zero. While SMR gives a prediction in every case, it is too liberal in its predictions and the CIA is often violated, as expected by Theorem 12. Over all predictions, the correlation of predicted vs. sample estimates is 0.58.

- When predictions are based on larger sets of common variables statistical matching based on the CIA (SMR_G) is more successful. Over all predictions, the correlation of predicted vs. sample estimates is 0.73. The method still fails however, on the text data (ACPJ, Ohsumed) where the predictions are not correlated at all with the sample estimates. On the other hand, FTR-S does not make any predictions on these data sets.

- FTR-S's predictions are highly correlated with sample estimates (0.89 correlation), which is the highest correlation achieved by any of the three methods. However, we point out that these metrics are computed on different sets of predictions and their comparative interpretation is not straightforward (see Appendix A, Section A.2 for more metrics and discussion).

- FTR-S is a novel alternative to statistical matching based on the CIA. FTR-S predictions are better correlated with the sample estimates of the unknown parameters, particularly when the number of common variables is low; we thus recommend that FTR-S should be preferred than existing statistical matching alternatives making the CIA in such cases.



(a) ACPJ-Etiology

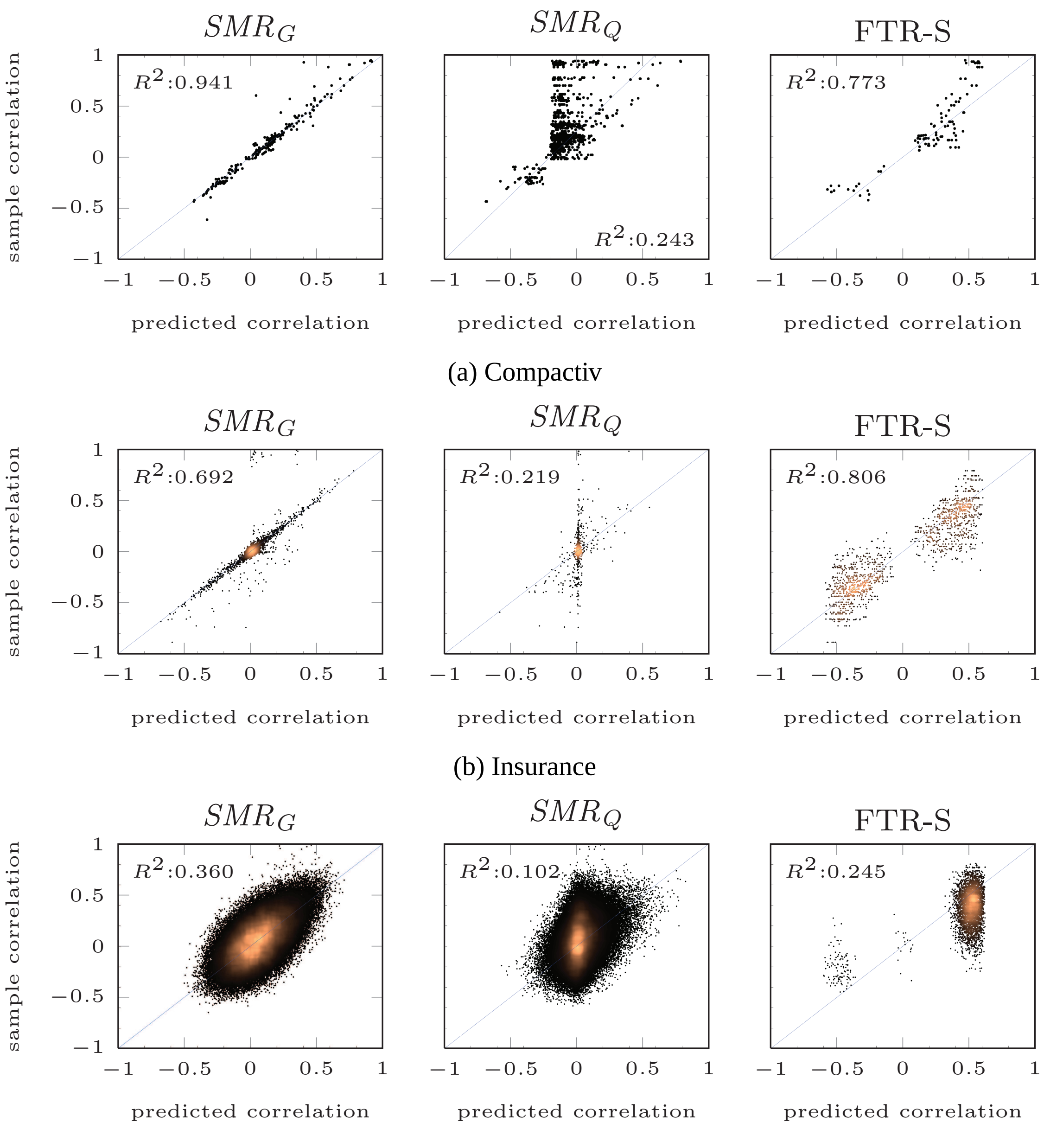
(b) Breast-Cancer

(c) C&C

Figure 17: Predicted vs Sample Correlations for SMR_Q, SMR_G, FTR-S

9.4 Statistical Matching Based on the Assumption of Multivariate Normality

The conditional independence assumption attempts to overcome the lack of joint information of the variables of interest. However, it can often be a misspecified assumption as pointed out in the literature (D'Orazio et al., 2006) and our simulated results above. An alternative approach, is to limit oneself to an assumption involving only the shape of the distribution. The most common distributional assumption adopted by statistical matching techniques for continuous variables is



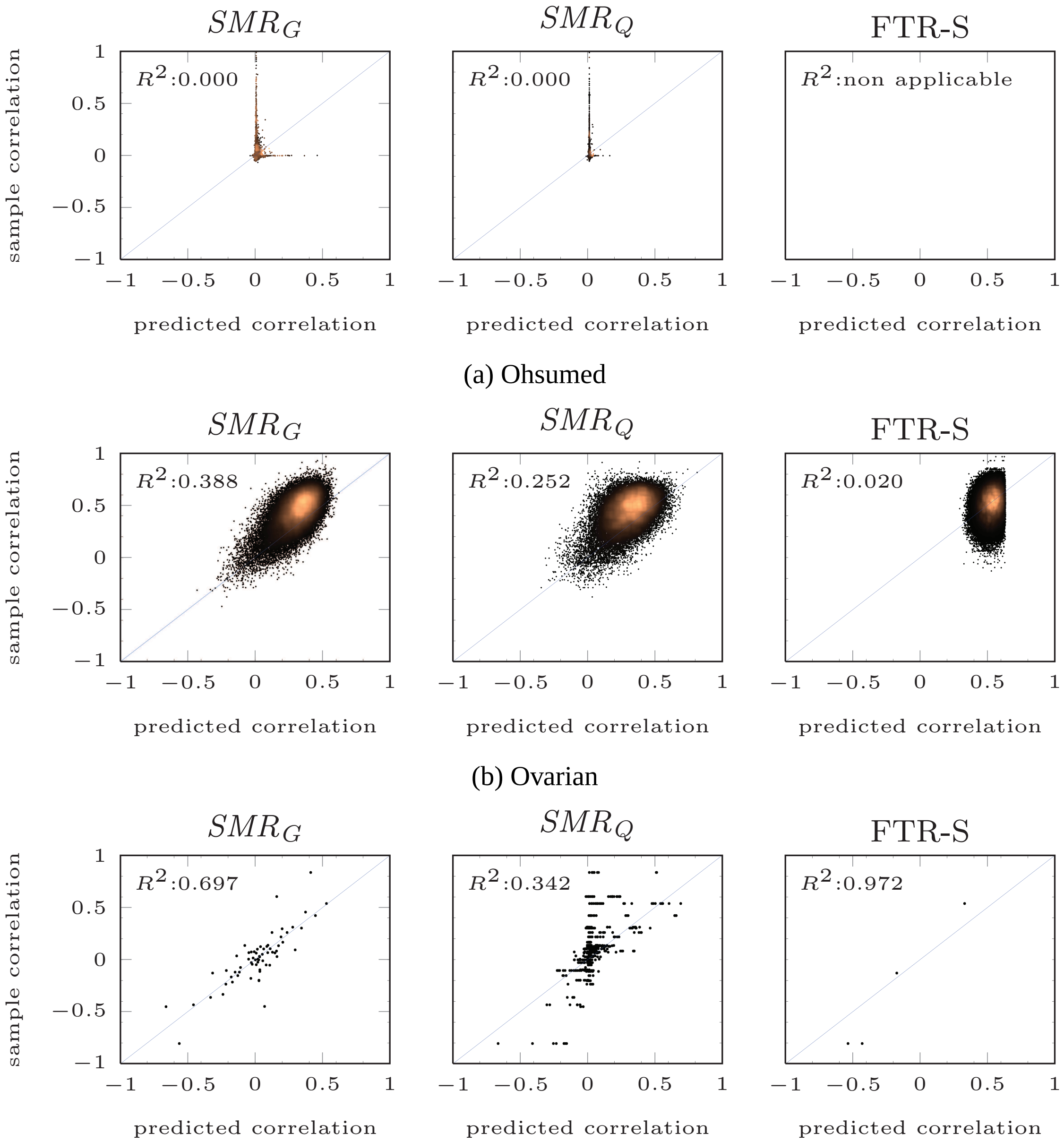
(a) Compactiv

(b) Insurance

(c) Lymphoma

Figure 18: Predicted vs Sample Correlations for SMR_Q, SMR_G, FTR-S

multivariate normality. Of course, multivariate normality alone does not allow the estimation of the parameters of the model. It does, however, impose some constraints on the parameters. These constraints stem from the positive semi-definiteness of the covariance matrix in multivariate normal distributions, thus, they naturally apply to any distribution with a positive semi-definite covariance matrix.



(a) Ohsumed

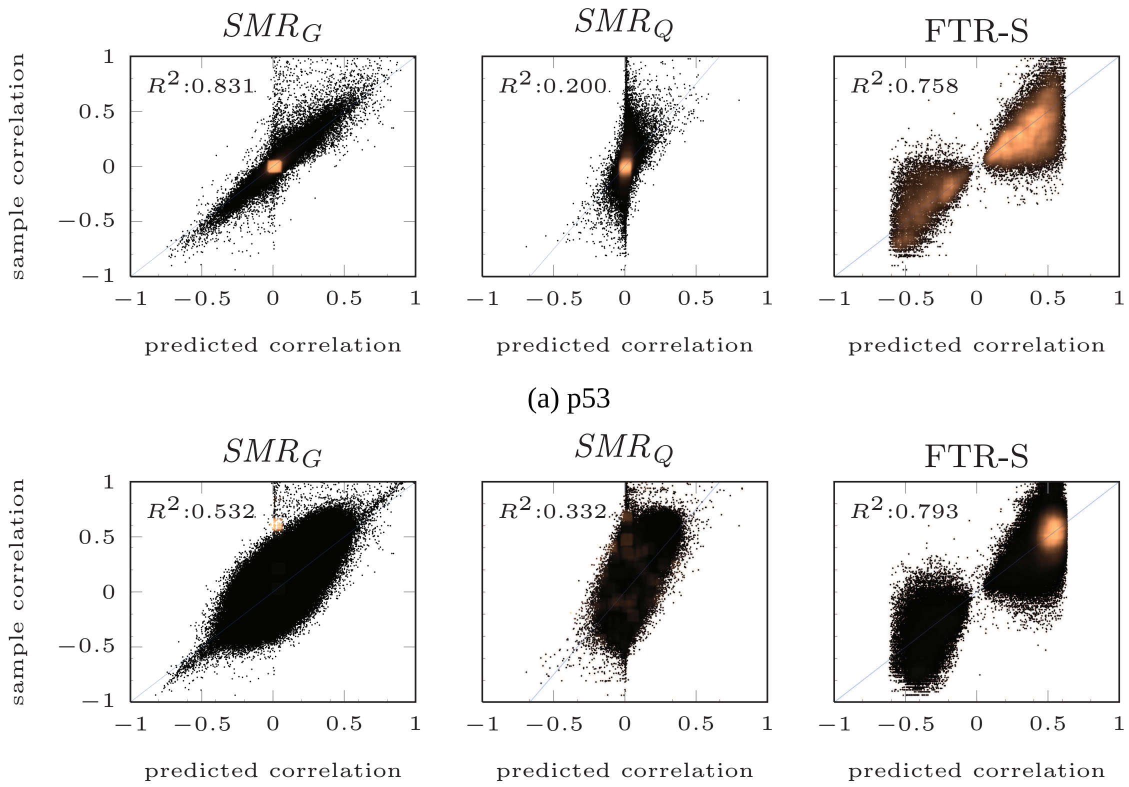
(b) Ovarian

(c) Wine

Figure 19: Predicted vs Sample Correlations for *SMR*$_Q$, *SMR*$_G$, FTR-S

Let us consider again *standardized* variables $\{\mathbf{X}, Y, Z\}$ and assume their joint is distributed as multivariate normal with correlation / covariance matrix Σ (which is symmetric)

$$\Sigma = \begin{bmatrix} \rho_{\mathbf{X}\mathbf{X}} & \rho_{\mathbf{X}Y} & \rho_{\mathbf{X}Z} \\ \rho_{Y\mathbf{X}} & 1 & \rho_{YZ} \\ \rho_{Z\mathbf{X}} & \rho_{ZY} & 1 \end{bmatrix}.$$



(a) p53

(b) All predictions

Figure 20: Predicted vs Sample Correlations for SMR_Q, SMR_G, FTR-S

The unknown quantity in the problem is parameter ρ_{YZ}. One can start from the requirement that Σ must be positive semi-definite to prove that ρ_{YZ} must lie within the interval $C \pm \sqrt{(D)}$ (Moriarity and Scheuren, 2001), where

$$C = \sum_{i=1}^{p} \sum_{j=1}^{p} \rho_{YX_i} \times B^{i,j} \times \rho_{ZX_j}$$

and

$$D = [1 - \sum_{i=1}^{p} \sum_{j=1}^{p} \rho_{YX_i} \times B^{i,j} \times \rho_{YX_j}] \times [1 - \sum_{i=1}^{p} \sum_{j=1}^{p} \rho_{ZX_i} \times B^{i,j} \times \rho_{ZX_j}]$$

where p is the cardinality of set $\mathbf{X}$, and B is the inverse of $\rho_{\mathbf{XX}}$, and $B^{i,j}$ is B's i,j element. This constraint is equivalent stating that the partial correlation $\rho_{YZ|\mathbf{X}}$ parameter can range freely in the interval [-1, 1]. Instead, the CIA specifies that $\rho_{YZ|\mathbf{X}} = 0$, that is, the mid-point of the interval.

The formula above can be applied to quadruples of variables to produce bounds for the unknown parameter ρ_{YZ}. The usefulness of such a prediction depends, of course, on the length of the predicted interval. In case the interval does not include 0, we may also say that the method *predicts an unconditional independence for Y and Z*. This procedure is described in Algorithm 6. In practice, we apply Algorithm 6 using the sample estimates $\hat{r}$ in place of the unknown population parameters ρ. The sample estimates are the maximum likelihood ones. The uncertainty of the estimation could be considered in the computation of the intervals by considering the worst case over all correlation estimates $\hat{r}$ that belong in the 95% confidence interval of their corresponding ρ. However, in this case the algorithm would produce wider intervals and thus fewer predictions.

Algorithm 6: Predict Dependency and Its Strength: Multivariate Normality Rule (**MNR**)

```
Input: Data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ on variables $\{X,Y,W\}$ and $\{X,Z,W\}$, respectively
Compute sample correlation matrix $\Sigma$ (except unknown quantity $\rho_{YZ}$) ;
$MNI \leftarrow [C - \sqrt{(D)}, C + \sqrt{(D)}]$;
if $0 \notin MNI$ then
    Predict $Y \not\perp\!\!\!\perp Z | \emptyset$ ;
end
Predict $\hat{r}_{XY} \in MNI$
```

9.5 Empirical Evaluation and Comparison of MNR and FTR

In order to evaluate how often MNR provides a prediction, we applied Algorithm 6 on real data. Applying Algorithm 6 on all possible combinations of four variables is prohibitive. Thus, to evaluate the MNR we randomly sampled 1000 quadruples from each group of 50 variables in each data set, for all data sets with continuous variables; For the Wine data set we generated all possible 495 quadruples out of its 12 variables.

Table 6 reports MNR performances on the randomly chosen quadruples. The columns of the table present the total number of randomly chosen quadruples (1000 × the number of chunks, except for the Wine data set), the number of predictions made by MNR on these random quadruples, the accuracies Acc^{MNR} and Acc^{FTR} at threshold $t = 0.05$. We then calculate (project) the *expected* number of predictions by the MNR rule, had it been applied on all possible quadruples. The final column presents the ratio of the number of predictions by the FTR rule over the *expected* number of predictions made by the MNR rule on all possible quadruples.

First, notice that MNR, similarly to FTR, does not provide any predictions for the text data sets ACPJ and Ohsumed data sets. Second, the rule is in general, highly accurate and on par with FTR. The most important observation however, is that the MNR does not outperform FTR in the number of predictions. The number of predictions made by FTR ranges from about 25% to 50% of those made by MNR (in four out of eight data sets) to 4 to 6 times more than MNR in the remaining data sets.

To examine whether the predictions of MNR rule overlap with those of FTR, we applied the MNR rule on the quadruples where FTR makes a prediction. The comparison is shown in Table 7. *MNR is able to predict a dependence only for* 1% *to* 25% *of FTR predictions.* The results in both Tables 6 and 7 clearly indicate that the two methods share only a small subset of common predictions, and thus neither method subsumes the other.

9.6 Summary, Interpretation, and Conclusions

We now comment and interpret the results of this section:

- It is possible to predict the presence of dependencies and bound their strength with distributional assumptions other than Faithfulness, such as multivariate normality.

Data Set	# rand. quads sampled	#MNR predictions on sampled quads	ACC^{MNR}	ACC^{FTR}	#FTR predictions / #expected MNR predictions on all quads
Breast-Cancer	356000	2	0.50	0.84	3.98
C&C	1000	45	1.00	0.96	0.02
Compactiv	1000	30	1.00	1.00	0.62
Insurance-C	1000	4	0.75	0.97	0.24
Lymphoma	147000	12	0.67	0.82	2.79
Ovarian	43000	391	0.99	0.99	5.99
p53	108000	39	1.00	0.97	5.19
Wine	495	7	1.00	1.00	0.57

Table 6: A comparison between FTR vs. MNR in predicting unconditional dependencies on randomly sampled quadruples. The columns are: the data set name, the total number of randomly sampled quadruples (1000 × the number of chunks, except for the Wine data set), the number of predictions made by MNR on those, the accuracies Acc^{MNR} and Acc^{FTR} at threshold $t = 0.05$. The final column presents the ratio of the number of predictions by the FTR rule over the *expected* number of predictions made by the MNR rule on all possible quadruples. The number of predictions made by FTR ranges from about 25% to 50% of those made by MNR to 4 to 6 times more than MNR.

Data Set	#FTR predictions	#MNR predictions restricted to cases FTR makes a prediction	% common predictions	*ACC* of both MNR and FTR
Breast-Cancer	1833	32	0.02	1.00
C&C	99241	10640	0.11	1.00
Compactiv	135	28	0.21	1.00
Insurance-C	1839	15	0.01	1.00
Lymphoma	7712	681	0.09	0.97
Ovarian	539165	59327	0.11	1.00
p53	46647	413	0.01	1.00
Wine	4	1	0.25	1.00

Table 7: A comparison between FTR vs. MNR in predicting unconditional dependencies on the cases where both rules apply.

- The sets of predictions entailed by assuming Faithfulness (FTR) and multivariate normality (MNR) do not overlap to a significant degree and neither method subsumes the other and they could be considered complementary. For example, the MNR makes a prediction only in the 1% to 25% of cases where FTR applies. In addition, in some data sets MNR makes only 2% of the number of FTR predictions, while in others MNR makes 6 times more predictions.

10. Related Work

Whole sub-fields have been developed to address the problem of integrative analysis, that we review briefly. Statistical matching has been reviewed, presented, and compared against in Section 9. Meta-Analysis focuses on the co-analysis of studies with similar sampling and experimental design characteristics with the purpose of making inferences about a single association. Meta-Analysis in Statistics (O'Rourke, 2007) combines the results of several studies to address a set of related research hypotheses. While meta-analysis focuses on a pair-wise association of a variable with an outcome of interest, a recent interesting extension addresses the problem of estimating the multivariate associations (for example, in the form of a regression model) with the target variable (Samsa et al., 2005); such methods often appear under the names of meta-regression and univariate synthesis (Zhou et al., 2009). The main idea of the latter is to assume a parametric form of the regression model and estimate the sufficient statistics from several homogeneous (in terms of being conducted on the same population, experimental conditions, sampling, etc.) studies that may not measure all variables (risk factors in this context). Both statistical matching and meta-analysis's scope does not extend to other sources of heterogeneity of the data sets, such as different experimental conditions.

In Computer Science and Machine Learning, the field of Transfer Learning (Pan and Yang, 2010) represents a main effort in integrative analysis. In Transfer Learning, successful search control strategies, model priors, and other characteristics transfer among different domains and/or tasks. When the task (target) is the same but the domains (populations) are different, this type of Transfer Learning is called Domain Adaptation. In this case, typically one would like to translate the estimated conditional distribution $P_s(Y|X)$ used for prediction in a source distribution to a target distribution $P_t(Y|X)$ that may be different (e.g., has a different marginal class distribution). Given that such methods are typically non-causal based, they cannot transfer to data sets where manipulations have been performed (causal methods could transfer predictive models to manipulated distributions as we show in Tsamardinos and Brown 2008, also shown in Maathuis et al. 2010). In addition, the input space for the predictors X has to be common. When the domain is the same (same distribution), but the tasks (target variables) are different, the type of Transfer Learning is called *Multi-Task Learning*. This type of learning attempts to simultaneously build models for several tasks in an effort to use one for leveraging the performance on the others. Typically this is performed by using a shared representation and learning common induced features. Again, these inferences are limited as they can only combine studies under the same sampling and experimental conditions on the same sets of variables.

Other fields may seem related in a first glance, but are orthogonal to the proposed research. The field of Relational Learning (Getoor and Taskar, 2007) does not really address the problem of learning from different data sets/studies over different samples, rather than a single data set (the one stemming from implicitly propositionalizing the database) in the form of relational tables. Similarly, the field of Distributed Learning (Cannataro et al., 2002) is restricted to designing time and communication-efficient analysis of what is essentially a single data set stored in different locations.

Other related work includes efforts to combine models (that may be developed from different data sets) on the same system but on different scales (Gennari et al., 2008). Typically, such methods involve mechanical models using differential equations and are not concerned with statistical models. In addition, these methods concern vertical integration at different temporal or spatial scales, while INCA proposes a horizontal integration of studies.

11. Discussion and Conclusions

We presented the basic idea and concept behind Integrative Causal Analysis (INCA), an approach for co-analyzing data sets that are heterogeneous in several aspects, such as in terms of measured variables and experimental conditions in the context of available prior knowledge. In this approach, one attempts to identify one or all causal models that are consistent with all available data and pieces of prior knowledge, and reason with them. Depending on the assumptions connecting causality with estimable quantities, co-analysis may lead to more inferences than independent analysis of the data sets.

In this paper, we focus on the problem of analyzing data sets over different variable sets. We employ Maximal Ancestral Graphs (MAGs) to model independencies in the data distributions and assume the latter are faithful to some MAG. As a proof-of-concept, we identify the simplest scenario where the INCA idea provides testable predictions, and specifically it predicts the presence and strength of an unconditional dependence, and a chain-like causal structure (entailing several additional conditional dependencies). The idea is implemented in the following algorithms: the Full-Testing Rule (FTR), the Minimal-Testing Rule (MTR) and FTR-S that additionally predicts the strength of the dependence.

The empirical results show that FTR and MTR are able to accurately predict the presence and strength of unconditional dependencies, as well as all the conditional dependencies entailed by the causal model. These predictions are better than chance and cannot be explained by the transitivity of dependencies often holding in Nature. Against typical statistical matching algorithms, FTR-S's predictions are better correlated with sample estimates particularly when the number of common variables is low.

Inducing causal models from observational data has been long debated (Pearl, 2000; Spirtes et al., 2001; Pearl, 2009). In our experiments, we do not employ the causal semantics of the models to predict the effect of manipulations but their ability to represent independencies, based on the assumption of Faithfulness. The results support that graphical models and the assumption of Faithfulness can make testable predictions and can be exploited for novel statistical inferences. While this is not a direct proof in favor of the causal semantics of the models, we do note that both Faithfulness and MAGs have been inspired by theories of probabilistic causality.

The empirical results show that the proposed algorithms' predictions are abundant, indicating the potential of the approach. Extending the theory and algorithms for increased efficiency, statistical robustness, range of tasks, data types, types of prior knowledge, and settings seems a promising direction that may allow the co-analysis of a large part of available studies and data sets.

Acknowledgments

We would like to thank the anonymous reviewers and the Editor for their constructive comments and feedback; We would like to thank Giorgos Borboudakis for his comments and help with proofreading the text; we would also like to thank our affiliated institutions for hosting this research.

Appendix A. Supplementary Material

In this appendix we provide supplementary information for the data sets used in the experiments presented in this paper, as well as some additional results.

A.1 Data Sets Preprocessing

Missing data imputation and discretization were *separately* performed, when necessary, on each sub-data-set $\mathcal{D}_1$, $\mathcal{D}_2$ and $\mathcal{D}_t$. Continuous variables X were discretized in three intervals:

- $]-\text{inf};\ mean(X) - std(X)]$
- $[mean(X) + std(X);\ \text{inf}[$
- remaining values.

Missing data were substituted with mean values (continuous, ordinal variables) or encoded as a distinct value (nominal variables). Our implementation of the G^2 test requires that nominal variables with n distinct values are econded as $0 \dots n-1$. When necessary we re-encoded nominal variables for respecting this convention.

A.1.1 ACPJ

Preprocessing steps: 2765 variables were found constant in at least one sub-data-set and were consequently eliminated from the analysis.
Download information: Aliferis et al. (2010) kindly provided us with the data.

A.1.2 BIBTEX

Preprocessing steps: No particular preprocessing steps.
Download information: The data set is freely available from the MULAN project website:
http://sourceforge.net/projects/mulan/ (checked on February 10, 2011).

A.1.3 C&C

Preprocessing steps: The first five attributes were eliminated because they do not carry relevant information. Columns with more than 80% of missing values were removed.
Download information: The data set is freely available from the UCI Machine Learning repository:
http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime (checked on February 10, 2011).

A.1.4 COMPACTIV

Preprocessing steps: No particular preprocessing steps.
Download information: The data set is freely available from the KEEL software web site:
http://sci2s.ugr.es/keel/dataset.php?cod=49 (checked on February 10, 2011).

A.1.5 COVTYPE

Preprocessing steps: Attributes $1 \dots 10$ were discretized.
Download information: The data set is freely available from the UCI Machine Learning repository:
http://archive.ics.uci.edu/ml/datasets/Covertype
(checked on February 10, 2011).

A.1.6 DELICIOUS

Preprocessing steps: No particular preprocessing steps.

Download information: The data set is freely available from the MULAN project website: http://sourceforge.net/projects/mulan/ (checked on February 10, 2011).

A.1.7 HIVA

Preprocessing steps: No particular preprocessing steps.

Download information: The data set is freely available from the web site: http://www.causality.inf.ethz.ch/al_data/HIVA.html (checked on February 10, 2011).

A.1.8 INSURANCE-C

Preprocessing steps: All variables were considered as continuous; nominal variables (namely, attributes 1 and 5) were eliminated.

Download information: The data set is freely available from the UCI Machine Learning repository: http://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+(COIL+2000) (checked on February 10, 2011).

A.1.9 INSURANCE-N

Preprocessing steps: All variables were considered as nominal.

Download information: The data set is freely available from the UCI Machine Learning repository: http://archive.ics.uci.edu/ml/datasets/Insurance+Company+Benchmark+(COIL+2000) (checked on February 10, 2011).

A.1.10 P53

Preprocessing steps: Samples with missing values were eliminated from the analysis (180 rows in total).

Download information: The data set is freely available from the UCI Machine Learning repository: http://archive.ics.uci.edu/ml/datasets/p53+Mutants (checked on February 10, 2011).

A.1.11 READ

Preprocessing steps: Continuous variables (namely attributes 24, 25 and 26) were discretized.

Download information: The data set is freely available from the web site: http://funapp.cs.bilkent.edu.tr/DataSets/ (checked on February 10, 2011).

A.1.12 WINE

Preprocessing steps: Two different data sets are available, respectively about red and white wines. For our experimentation we used only the white wines data set (the one with more samples).

Download information: The data set is freely available from the UCI Machine Learning repository: http://archive.ics.uci.edu/ml/datasets/Wine+Quality (checked on February 10, 2011).

Data Set	$\text{FTR}_{0.05}$	$\text{MTR}_{0.02}$	$\text{TR}_{0.01}$
Covtype	59	810	1431
Read	0	9	260
Infant Mortality	10	427	1170
Compactiv	69	193	231
Gisette	330	12340	31648
hiva	366	16174	34977
Breast-Cancer	1371	68077	228610
Lymphoma	4473	51794	122857
Wine	3	44	66
Insurance-C	394	2212	3264
Insurance-N	95	1002	2527
p53	15181	95195	129372
Ovarian	41600	48376	52646
C&C	4168	5048	5050
ACPJ	0	190	15994
Bibtex	1	1858	16087
Delicious	524	6042	21351
Dexter	0	2	116
Nova	0	115	3280
Ohsumed	0	60	5227

Table 8: Number of unique predictions $|U_i^R|$ with “Bonferroni” correction for rules FTR, MTR, TR and Random Guess

A.1.13 BREAST-CANCER, DEXTER, GISETTE, INFANT-MORTALITY, LYMPHOMA, NOVA, OHSUMED, OVARIAN

Preprocessing steps: No particular preprocessing steps.
Download information: Aliferis et al. (2010) kindly provided us with the data.

A.2 Supplementary Tables

Table 10 presents the performance of the algorithms as measured by the Mean Absolute Error (MAE) of the predictions $\hat{r}_{YZ}$ and the sample-estimates r_{YZ}: $1/N \cdot \sum |\hat{r}^i - r^i|$, where N is the total number of predictions of an algorithm. This metric may favor algorithms that often predict zero correlations on data sets where the number of dependencies is low. This is the case of SMR_G and SMR_Q on the ACPJ data set (see Figure 17a). SMR_G and SMR_Q achieve an MAE of *only* 0.01 and 0.02 respectively because they always predict values close to zero, while failing to detect any high correlation. The corresponding correlations between predictions and sample-estimates on the same data set are low: 0.05 and 0.00 respectively.

Table 11 presents the performance of the algorithms as measured by the Mean Relative Absolute Error (MRAE) of the predictions $\hat{r}_{YZ}$ and the sample-estimates r_{YZ}: $1/N \cdot \sum |\hat{r}^i - r^i|/|r^i|$, where N is the total number of predictions of an algorithm. This metric penalizes more algorithms that attempt predictions of small correlations (such as *SMR*) because even a small absolute error may lead to a high relative error. For example, SMR on the Ovarian data set has a high MRAE (on the order of 10^9 despite a correlation between predictions and sample-estimates of 0.62 .

Data Set	$FTR_{0.05}$	$MTR_{0.02}$	Random Quadruple
Covtype	1.00	0.99	$0.74^{\spadesuit}$
Read	-	-	-
Infant Mortality	0.60	0.44	0.10^{**}
Compactiv	0.87	0.93^{*}	0.83
Gisette	0.80	$0.59^{\spadesuit}$	$0.11^{\spadesuit}$
hiva	0.71	$0.47^{\spadesuit}$	$0.22^{\spadesuit}$
Breast-Cancer	0.55	$0.31^{\spadesuit}$	$0.16^{\spadesuit}$
Lymphoma	0.46	$0.34^{\spadesuit}$	$0.18^{\spadesuit}$
Wine	1.00	0.70	0.73
Insurance-C	0.86	$0.65^{\spadesuit}$	$0.42^{\spadesuit}$
Insurance-N	0.57	0.50	0.17^{**}
p53	0.90	$0.82^{\spadesuit}$	$0.49^{\spadesuit}$
Ovarian	0.61	$0.62^{\spadesuit}$	$0.50^{\spadesuit}$
C&C	0.78	$0.73^{\spadesuit}$	$0.66^{\spadesuit}$
ACPJ	-	0.26	0.02
Bibtex	1.00	0.55	0.08^{**}
Delicious	0.99	$0.81^{\spadesuit}$	$0.19^{\spadesuit}$
Dexter	-	0.50	0.02
Nova	-	0.07	0.03
Ohsumed	-	0.14	0.02
$\overline{SACC^R}$	0.78	0.55^{*}	0.30^{**}
$\underline{SACC^R}$	0.66	$0.69^{\spadesuit}$	$0.12^{\spadesuit}$

Table 9: $SACC_i^R(t)$ at $t = 0.05$ with “Bonferroni” correction for rules FTR, MTR and Random Quadruple. Marks *, **, and ♠ denote a statistically significant difference from FTR at the levels of 0.05, 0.01, and machine-epsilon respectively.

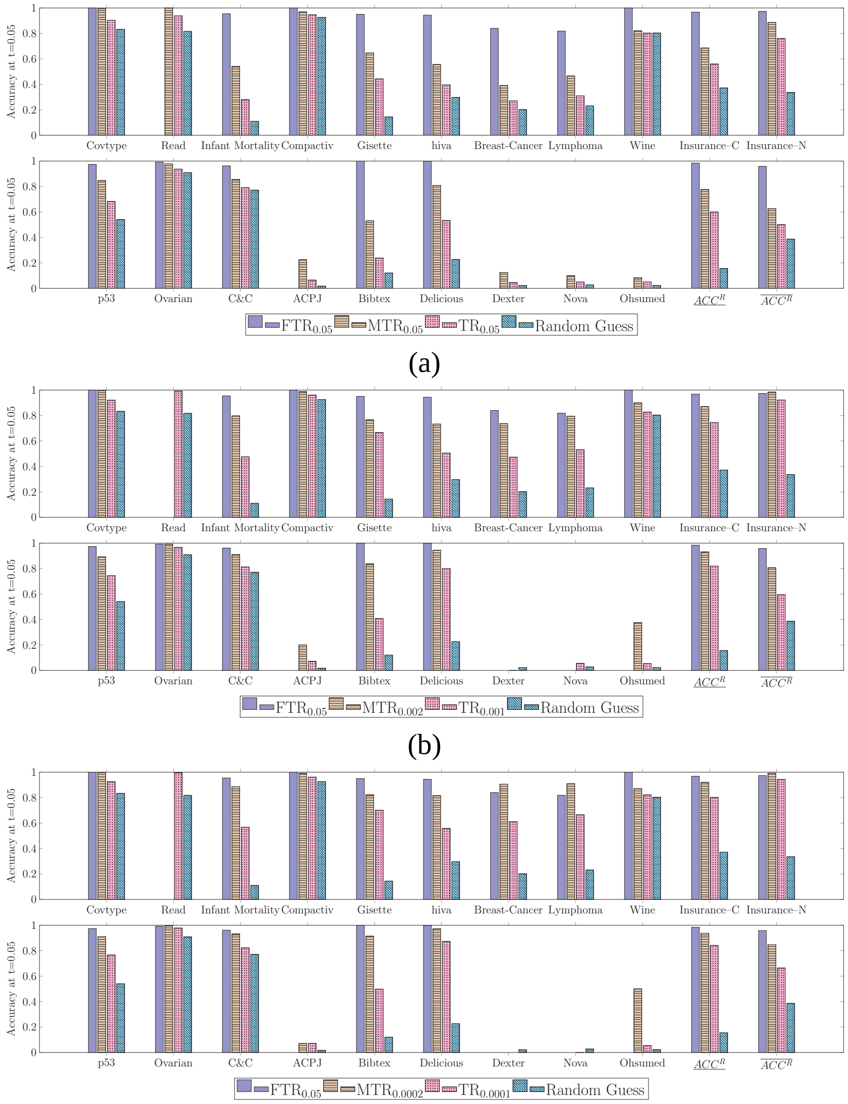
Data Sets	SMR_G	SMR_Q	FTR-S
ACPJ	0.01 ± 0.01	0.02 ± 0.01	-
Breast-Cancer	0.11 ± 0.08	0.13 ± 0.10	0.18 ± 0.13
C&C	0.05 ± 0.03	0.19 ± 0.18	0.18 ± 0.13
Compactiv	0.04 ± 0.06	0.19 ± 0.20	0.14 ± 0.12
Insurance-C	0.03 ± 0.08	0.09 ± 0.14	0.14 ± 0.12
Lymphoma	0.12 ± 0.09	0.14 ± 0.11	0.17 ± 0.14
Ohsumed	0.01 ± 0.02	0.02 ± 0.02	-
Ovarian	0.15 ± 0.10	0.16 ± 0.11	0.09 ± 0.07
Wine	0.09 ± 0.10	0.15 ± 0.17	0.22 ± 0.14
p53	0.03 ± 0.05	0.07 ± 0.10	0.14 ± 0.12
Over data sets	0.06 ± 0.06	0.12 ± 0.11	0.16 ± 0.12
Over predictions	0.07 ± 0.08	0.07 ± 0.09	0.11 ± 0.10

Table 10: Mean Absolute Error (MAE) between the calibrated predictions $\hat{r}_{YZ}$ and sample-estimated r_{YZ} (average value ± standard deviation). SMR_G refers to the Statistical Matching Rule applied on all pairs of variables in the same group, considering the remaining 48 variables in the group as common variables. SMR_Q is the Statistical Matching Rule applied on quadruples of variables randomly chosen from the same group. Finally, FTR-S consists in the Full Testing Rule modified for estimating the strength of the dependency, see Algorithm 4.

Data Sets	SMR$_G$	SMR$_Q$	FTR-S
ACPJ	13.17 ± 87.17	27.22 ± 141.50	-
Breast-Cancer	5.74 ± 624.51	2.79 ± 90.41	1.39 ± 4.51
C&C	1.52 ± 39.16	3.53 ± 44.98	1.30 ± 16.80
Compactiv	0.39 ± 1.43	1.79 ± 9.39	0.46 ± 0.53
Insurance-C	2.79 ± 11.04	2.10 ± 5.15	2.44 ± 18.04
Lymphoma	4.51 ± 182.18	3.66 ± 181.90	5.77 ± 145.88
Ohsumed	4.62 ± 30.53	7.72 ± 8.95	-
Ovarian	$7.32 \times 10^9 \pm 1.68 \times 10^{13}$	0.58 ± 5.51	0.20 ± 0.44
Wine	1.31 ± 2.24	1.78 ± 5.65	0.38 ± 0.06
p53	34.95 ± 7982.92	19.86 ± 4544.32	4.76 ± 290.58
Over data sets	$7.32 \times 10^9 \pm 1.68 \times 10^{13}$	7.10 ± 503.78	2.09 ± 59.61
Over predictions	$2.79 \times 10^9 \pm 3.28 \times 10^{12}$	14.36 ± 1320.98	0.87 ± 87.92

Table 11: Mean Relative Absolute Error (MRAE) between the calibrated predictions $\hat{r}_{YZ}$ and sample-estimated r_{YZ} (average value ± standard deviation) SMR$_G$ refers to the Statistical Matching Rule applied on all pairs of variables in the same group, considering the remaining 48 variables in the group as common variables. SMR$_Q$ is the Statistical Matching Rule applied on quadruples of variables randomly chosen from the same group. Finally, FTR-S consists in the Full Testing Rule modified for estimating the strength of the dependency, see Algorithm 4. For the Ovarian data set the SMR$_G$ rule provides predictions for cases with nearby-zero sample estimated r_{YZ}, and these predictions generate extremely high MRAE values. Once excluded such cases, the SMR$_G$ MRAE on the Ovarian data set is 0.54 ± 12.16, while the MRAE averaged over all data sets and over all predictions is 6.95 ± 897.33 and 10.45 ± 2498.28, respectively.

A.3 Supplementary Figures



(a)

(b)

(c)

Figure 21: Accuracies Acc_i for each data set, as well as the average accuracy $\overline{Acc}$ (each data set weighs the same) and the pooled accuracy $\underline{Acc}$ (each prediction weighs the same). (a) All rules are applied without any correction of significance threshold and all accuracies are computed at $t = 0.05$ (b) Accuracies Acc_i calculated with the “Bonferroni 10^{-1}” significance threshold correction. (c) Accuracies Acc_i calculated with the “Bonferroni 10^{-2}” significance threshold correction.

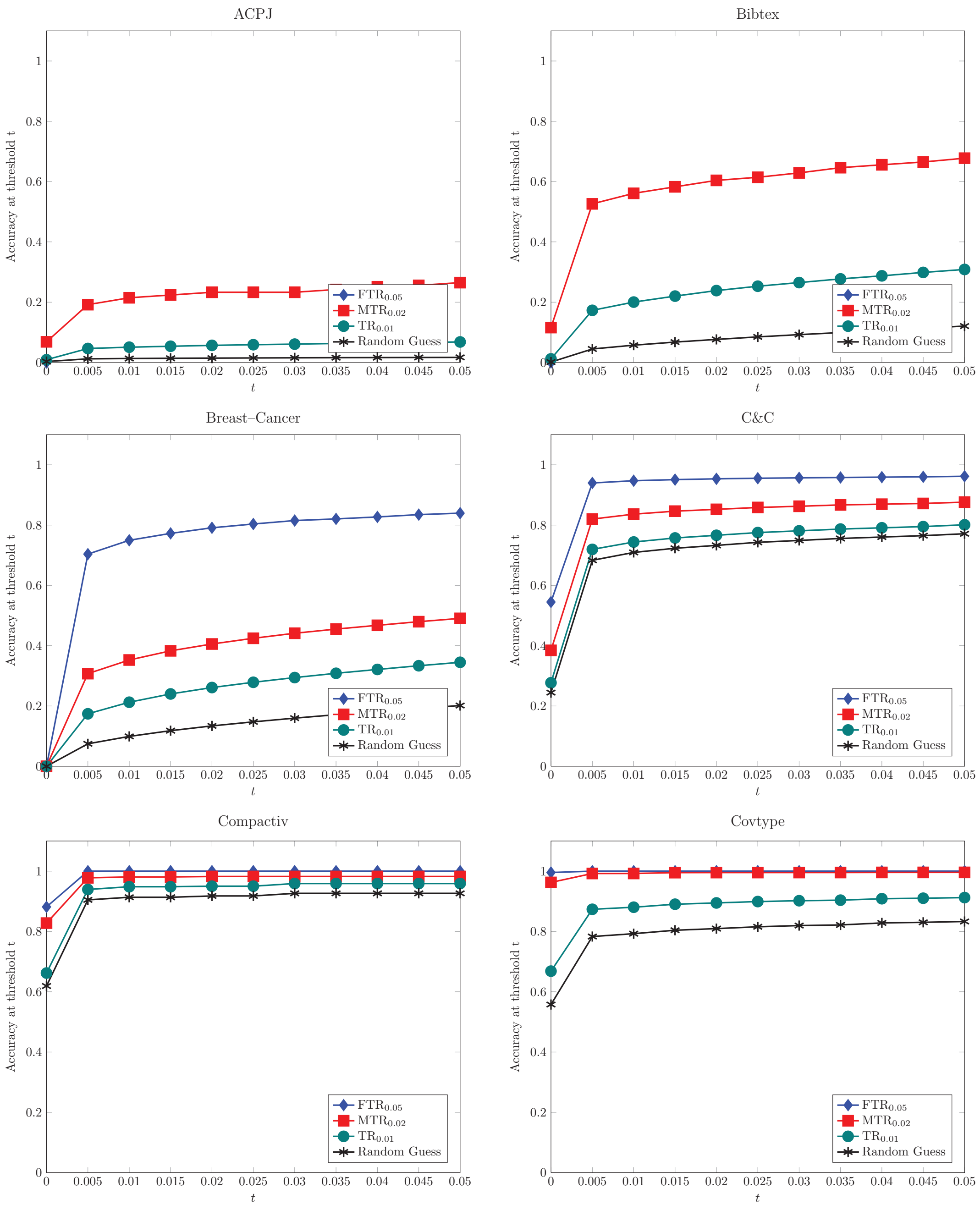


Figure 22: Accuracy at threshold t for data sets ACPJ-Etiology, Bibtex, Breast Cancer and Communities and Crime, Compactiv and Covtype for different rules

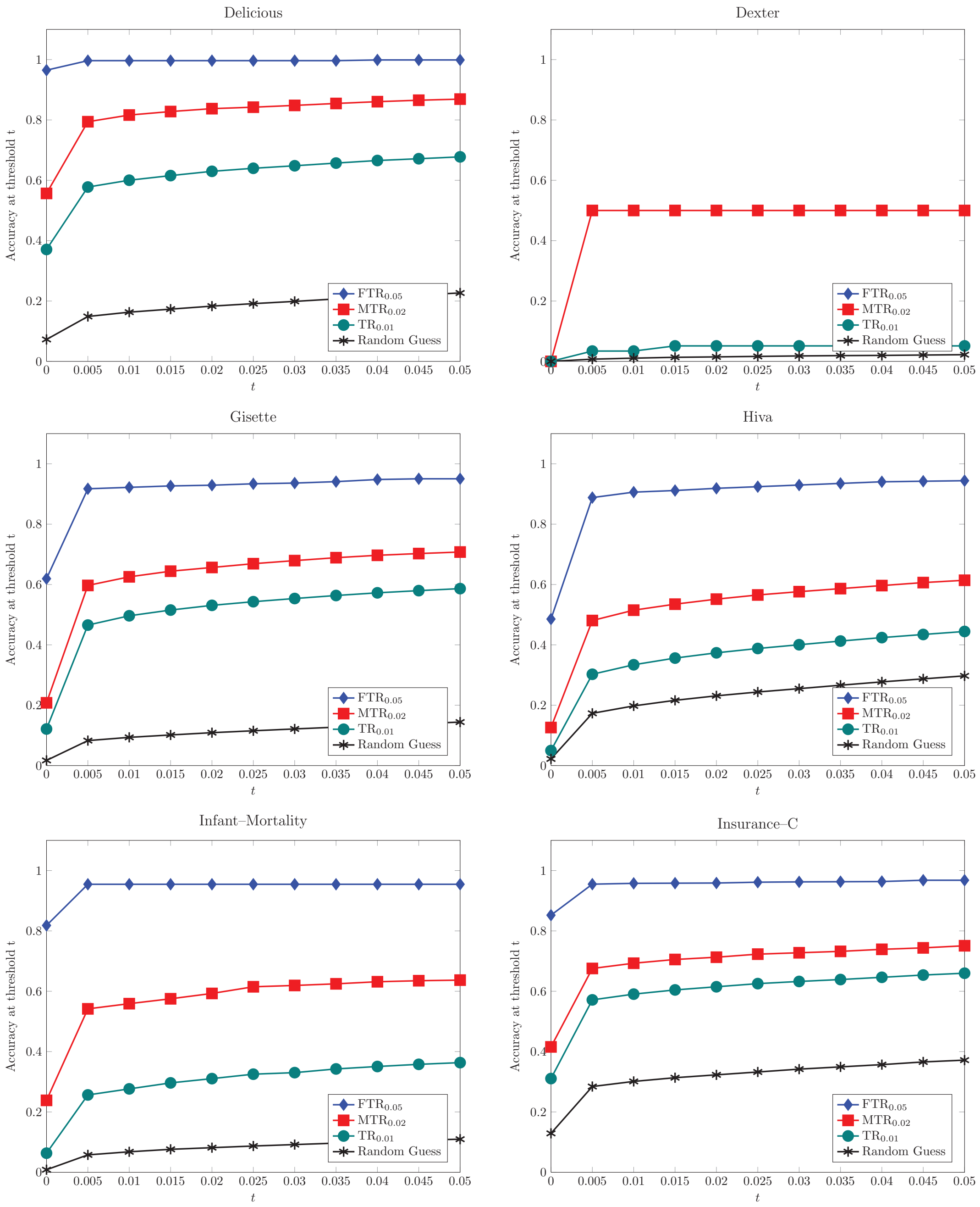


Figure 23: Accuracy at threshold t for data sets Delicious, Dexter, Gisette, Hiva, Infant Mortality and Insurance-C for different rules

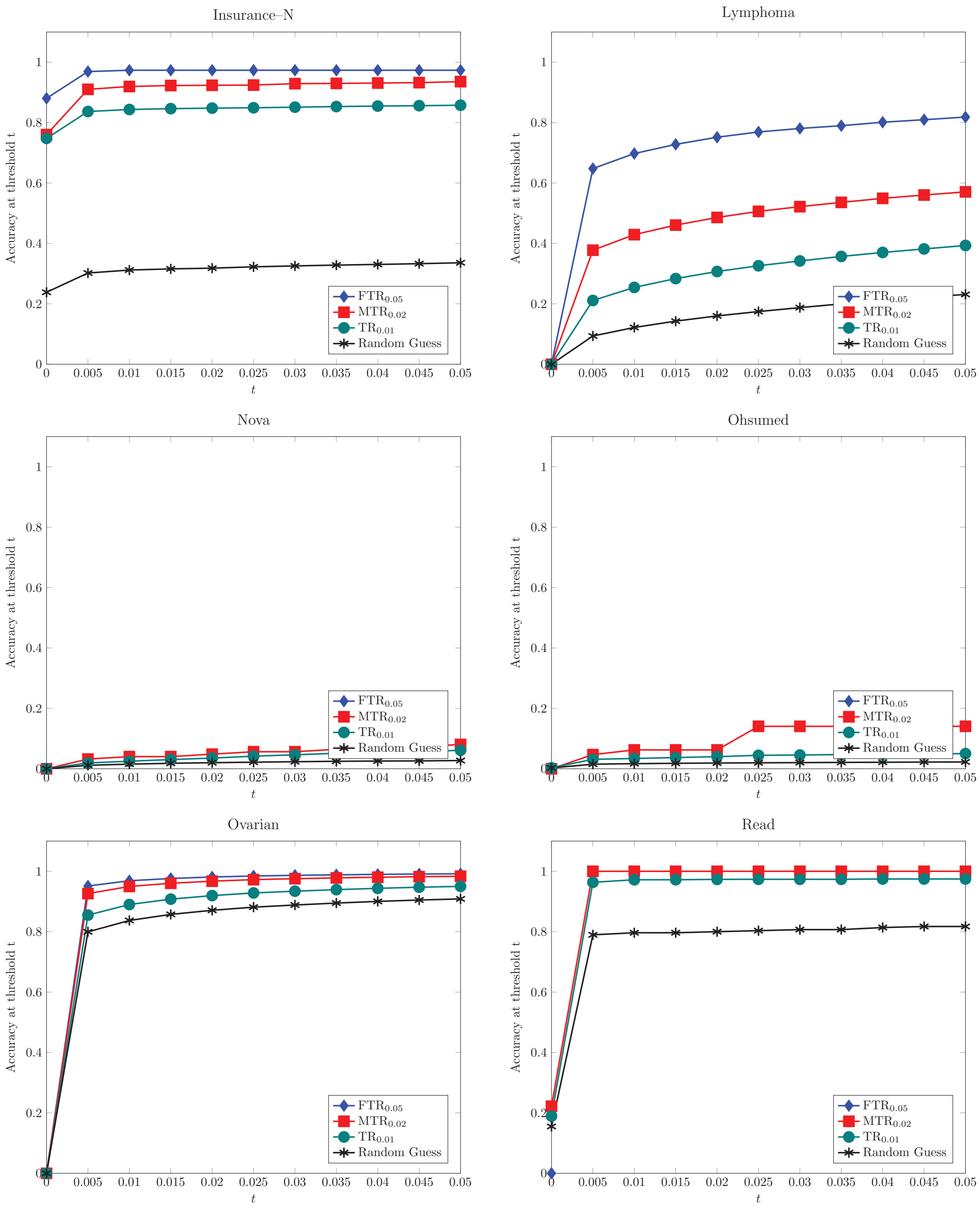


Figure 24: Accuracy at threshold t for data sets Insurance-N, Lymphoma, Nova, Ohsumed, Ovarian, Read and for different rules

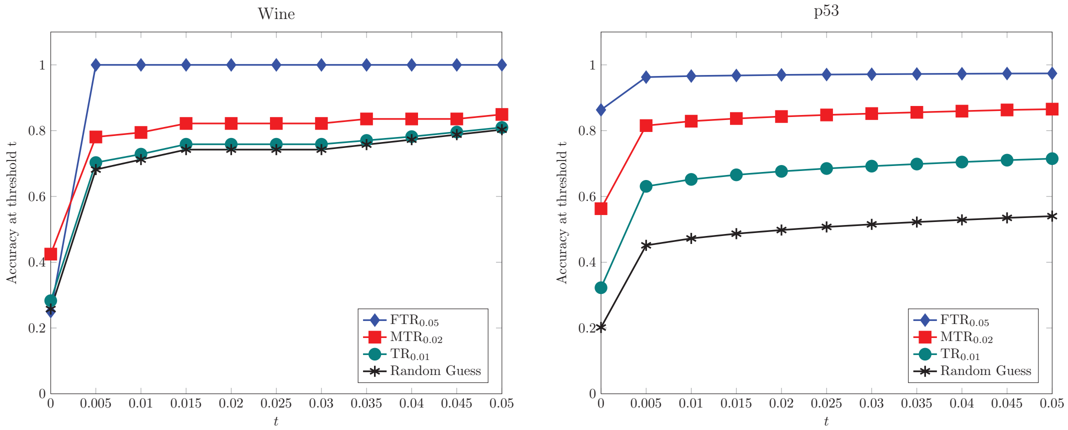


Figure 25: Accuracy at threshold t for data sets Wine, p53

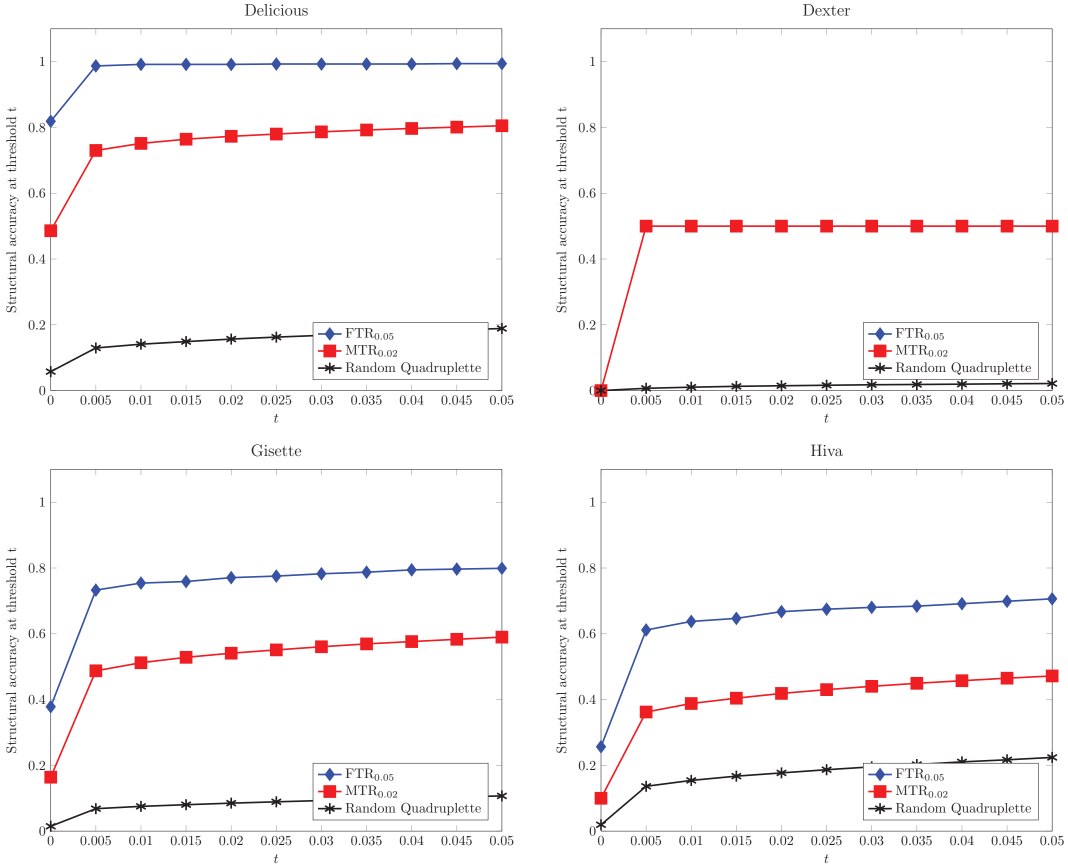


Figure 26: Structural accuracy at threshold t for data sets Delicious, Dexter, Gisette and Hiva for different rules

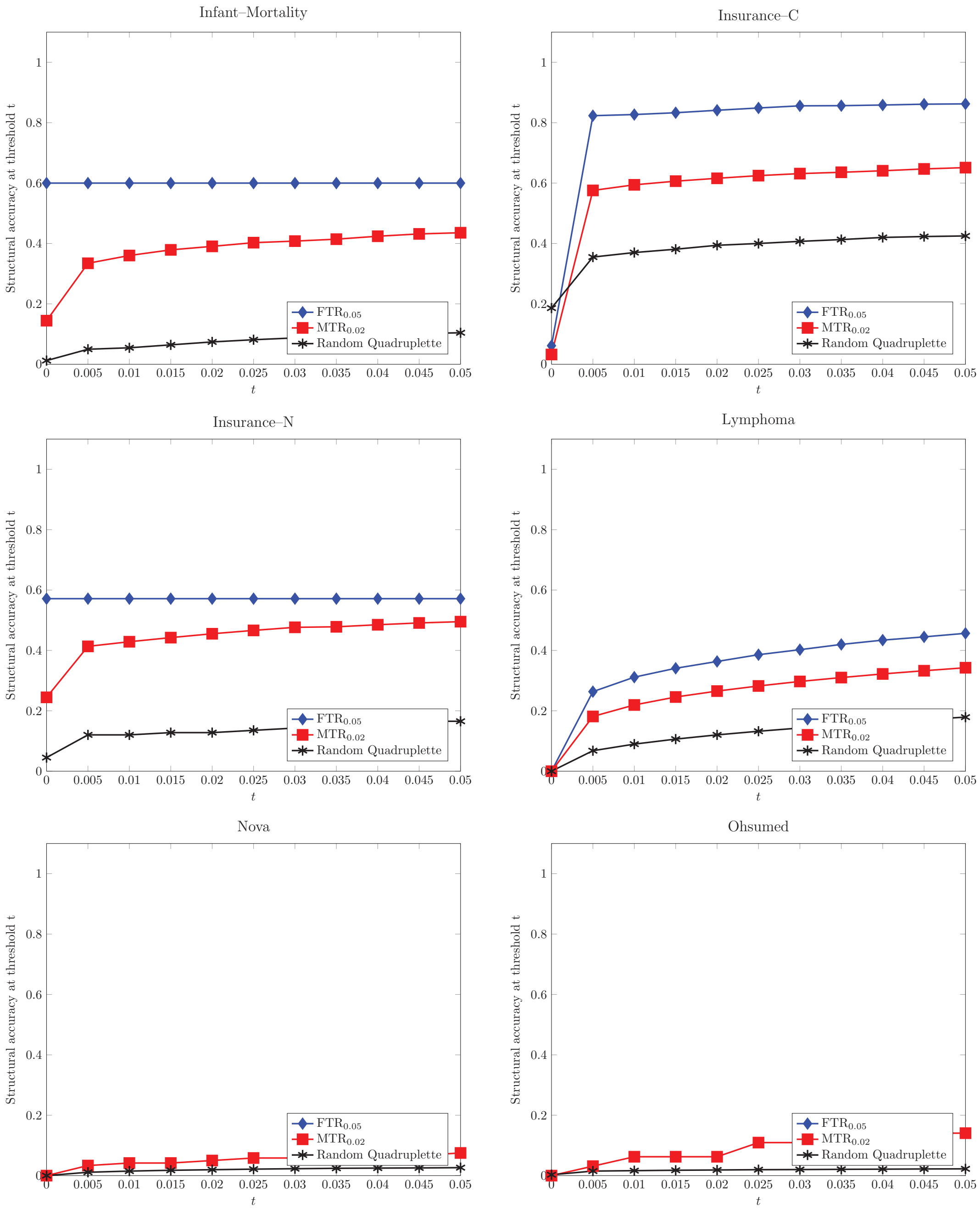


Figure 27: Structural accuracy at threshold t for data sets Infant Mortality, Insurance-C, Insurance-N, Lymphoma, Nova and Ohsumed for different rules

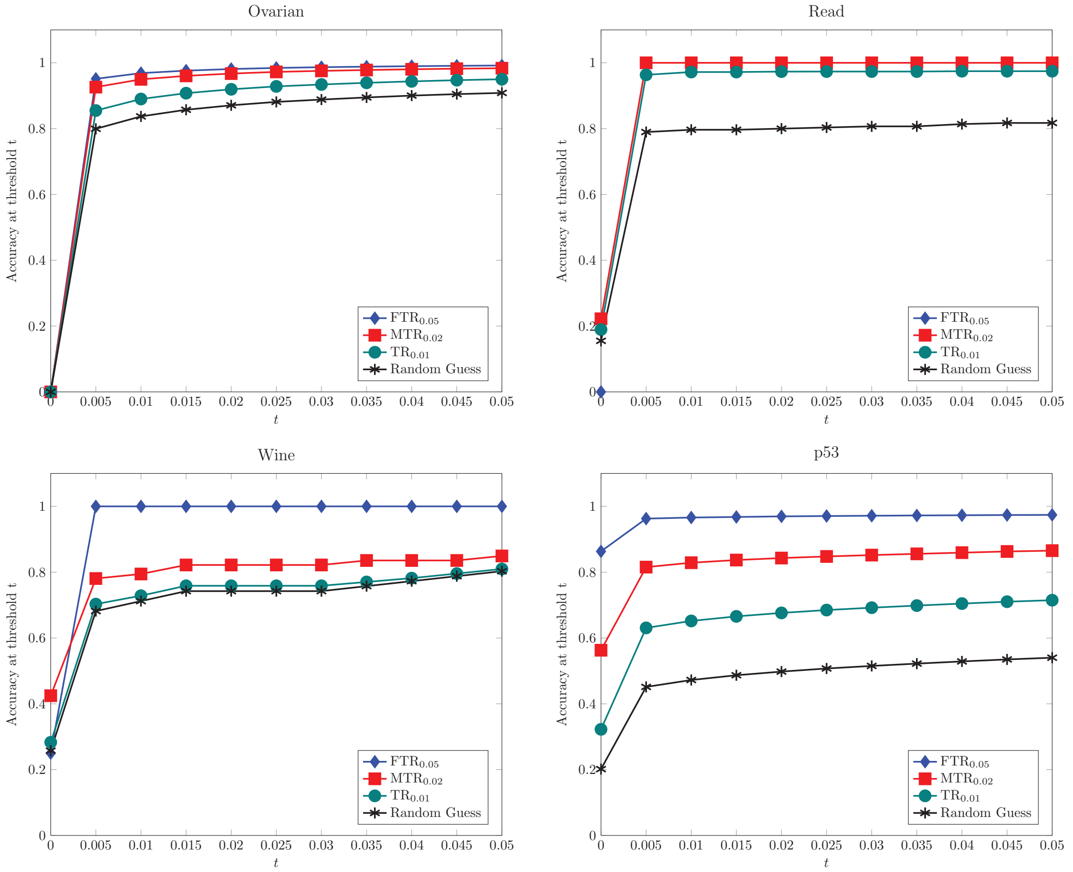


Figure 28: Structural accuracy at threshold t for data sets Ovarian, Read, Wine and p53 for different rules

References

B Abramson, J Brown, W Edwards, A Murphy, and RL Winkler. Hailfinder: A Bayesian system for forecasting severe weather. *International Journal of Forecasting*, 12(1):57–71, 1996.

J Alcalá-Fdez, L Sánchez, S García, M J Jesus, S Ventura, J M Garrell, J Otero, C Romero, J Bacardit, V M Rivas, J C Fernández, and F Herrera. KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing*, 13(3):307–318, 2009.

CF Aliferis, A Statnikov, I Tsamardinos, S Mani, and X Koutsoukos. Local causal and Markov blanket induction for causal discovery and feature selection for classification part ii : Analysis and extensions. *Journal of Machine Learning Research*, 11:235–284, 2010.

N Angelopoulos and J Cussens. Bayesian learning of Bayesian networks with informative priors. *Annals of Mathematics and Artificial Intelligence*, 54(1-3):53–98, 2008.

Y Aphinyanaphongs, AR Statnikov, and CF Aliferis. A comparison of citation metrics to machine learning filters for the identification of high quality MEDLINE documents. *JAMIA*, 13(4):446–455, 2006.

A Balke and J Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1172–1176, 1997.

IA Beinlich, HJ Suermondt, RM Chavez, and GF Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Second European Conference on Artificial Intelligence in Medicine*, volume 38, pages 247–256. Springer-Verlag, Berlin, 1989.

J Binder, D Koller, S Russell, and K Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29, 1997.

JA Blackard and DJ Dean. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24(3):131–151, 1999.

G Borboudakis, S Triantafillou, V Lagani, and I Tsamardinos. A constraint-based approach to incorporate prior knowledge in causal models. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2011.

M Cannataro, D Talia, and P Trunfio. Distributed data mining on the grid. *Future Gener. Comput. Syst.*, 18:1101–1112, October 2002.

T Claassen and T Heskes. Causal discovery in multiple models from different experiments. In *Advances in Neural Information Processing Systems (NIPS 2010)*, volume 23, pages 1–9, 2010.

TP et al. Conrads. High-resolution serum proteomic features for ovarian cancer detection. *Endocrine-Related Cancer*, 11(2):163–78, 2004.

GF Cooper and Ch Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of Uncertainty in Artificial Intelligence (UAI 1999)*, volume 10, pages 116–125, 1999.

P Cortez, Ao Cerdeira, F Almeida, T Matos, and J Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, November 2009. ISSN 01679236.

R Cudeck. An estimate of the covariance between variables which are not jointly observed. *Psychometrika*, 65(4):539–546, 2000.

SA Danziger, R Baronio, L Ho, L Hall, K Salmon, GW Hatfield, P Kaiser, and RH Lathrop. Predicting positive p53 cancer rescue regions using most informative positive (MIP) active learning. *PLoS Computational Biology*, 5(9):12, 2009.

M D'Orazio, MD Zio, and M Scanu. *Statistical Matching: Theory and Practice*. Wiley, 2006.

F Eberhardt. A sufficient condition for pooling data. *Synthese*, 163(3):433–442, February 2008.

F Eberhardt, PO Hoyer, and R Scheines. Combining experiments to discover linear cyclic models with latent variables. In *Proceedings of Artificial Intelligence anf Statistics 2010*, volume 9, pages 185–192, 2010.

C Elkan. Magical thinking in data mining: lessons from CoIL challenge 2000. *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 426–431, 2001.

RA Fisher. On the interpretation of $\chi 2$ from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

A Frank and A Asuncion. UCI machine learning repository, 2010. URL `http://archive.ics.uci.edu/ml`.

JH Gennari, ML Neal, BE Carlson, and DL Cook. Integration of multi-scale biosimulation models via light-weight semantics. *Pacific Symposium On Biocomputing*, 425:414–25, 2008.

L Getoor and B Taskar. *Introduction to Statistical Relational Learning*, volume L. The MIT Press, 2007.

HA Guvenir and I Uysal. Bilkent University function approximation repository, 2000. URL `http://funapp.cs.bilkent.edu.tr`.

I Guyon, S Gunn, M Nikravesh, and L Zadeh. *Feature Extraction, Foundations and Applications*. Springer–Verlag, Berlin, Germany, 2006a.

I Guyon, A Saffari, G Dror, and J Buhmann. Performance prediction challenge. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 1649–1656, 2006b.

A Hyttinen, F Eberhardt, and PO Hoyer. Causal discovery for linear cyclic models with latent variables. In *Proccedings of the 5th European Workshop on Probabilistic Graphical Models*, 2010.

A Hyttinen, F Eberhardt, and PO Hoyer. Noisy-OR models with latent confounding. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, 2011.

Th Joachims. *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms. (The Kluwer International Series in Engineering and Computer Science)*. Springer, 2002.

SH Kim. Markovian combination of subgraphs of DAGs. In *Proceedings of The 10th IASTED International Conference on Artificial Intelligence and Applications*, pages 90–95, 2010.

SH Kim and S Lee. *New Developments in Robotics, Automation and Control*. In-Tech, Vienna, Austria, 2008.

MH Maathuis, M Kalisch, and P Bühlmann. Estimating high-dimensional intervention effects from observational data. *Annals of Statistics*, 37:3133–3164, 2009.

MH Maathuis, D Colombo, M Kalisch, and P Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248, 2010. ISSN 15487105.

S Mani and GF Cooper. Causal discovery using a Bayesian local causal discovery algorithm. *Medinfo 2004*, 11:731–735, 2004.

C Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th Conference of Uncertainty in Aritficial Intelligence*, pages 403–410, 1995.

C Moriarity and F Scheuren. Statistical matching: a paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics*, 17:407–422, 2001.

RT O'Donnell, AE Nicholson, B. Han, KB Korb, MJ Alam, and LR Hope. Incorporating Expert Elicited Structural Information in the CaMML Causal Discovery Program. Technical report, Clayton School of Information Technology, Monash University, Melbourne, 2006.

K O'Rourke. An historical perspective on meta-analysis: dealing quantitatively with varying study results. *Journal of the Royal Society of Medicine*, 100(12):579–582, 2007.

SJ Pan and Q Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

J Pearl. *Causality: Models, Reasoning and Inference*, volume 113 of *Hardcover*. Cambridge University Press, 2000.

J Pearl. Causal inference in statistics: an overview. *Statistics Surveys*, 3:96–146, 2009.

K Pearson. On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophical Magazine*, 50(50):157–175, 1900.

JP Pellet and A Elisseeff. Finding latent causes in causal networks: an efficient approach based on Markov blankets. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems*, 2008.

J Ramsey, P Spirtes, and J Zhang. Adjacency faithfulness and conservative causal inference. In *Proceedings of Uncertainty in Artificial Intelligence*, 2006.

T Richardson and P Spirtes. Ancestral graph markov models. *The Annals of Statistics*, 30(4): 962–1030, 2002.

A Rosenwald et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma. *N Engl J Med*, 346(25):1937–1947, 2002.

DG Rubin. Characterizing the estimation of parameters in incomplete-data problems. *Journal of the American Statistical Association*, 69:467–474, 1974.

G Samsa, G Hu, and M Root. Combining information from multiple data sources to create multivariable risk models: Illustration and preliminary assessment of a new method. *Journal of Biomedicine and Biotechnology*, 2005(2):113–123, 2005.

S Shimizu, PO Hoyer, A Hyvärinen, and A Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(2):2003–2030, 2006.

S Shimizu, T Inazumi, Y Sogawa, A Hyvarinen, Y Kawahara, T Washio, PO Hoyer, and K Bollen. DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

A Spanos. Revisiting the omitted variables argument: Substantive vs. statistical adequacy. *Journal of Economic Methodology*, 13(2):179–218, 2006.

P Spirtes and TS Richardson. A polynomial time algorithm for determining DAG equivalence in the presence of latent variables and selection bias. In *Proceedings of the 6th International Workshop on Artificial Intelligence and Statistics*, pages 489–500, 1996.

P Spirtes, C Glymour, and R Scheines. *Causation, Prediction, and Search*. The MIT Press, second edition, January 2001.

J Tian and J Pearl. Causal Discovery from Changes. *Proceedings of UAI*, pages 512–521, 2001.

RE Tillman. Structure learning with independent non-identically distributed data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1041–1048. ACM, 2009.

RE 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 14th International Conference on Artificial Intelligence and Statistics*, volume 15, pages 3–15, 2011.

RE Tillman, David Danks, and Clark Glymour. Integrating locally learned causal structures with overlapping variables. In *Advances in Neural Information Processing Systems (NIPS*, 2008.

S Triantafillou, I Tsamardinos, and IG Tollis. Learning causal structure from overlapping variable sets. In *Proceedings of Artificial Intelligence and Statistics*, volume 9, 2010.

I Tsamardinos and G Borboudakis. Permutation testing improves Bayesian network learning. In *ECML PKDD*, pages 322–337, 2010.

I Tsamardinos and LE Brown. Bounding the false discovery rate in local Bayesian network learning. In *Proceedings of the 23rd Conference on Artificial Intelligence (AAAI)*, pages 1100–1105, 2008.

I Tsamardinos and S Triantafillou. The possibility of integrative causal analysis: Learning from different datasets and studies. *Journal of Engineering Intelligent Systems*, 17(2/3):163–175, 2009.

I Tsamardinos, LE Brown, and CF Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

G Tsoumakas, I Katakis, and I Vlahavas. Mining multi-label data. *Data Mining and Knowledge Discovery Handbook*, pages 1–20, 2010.

B Vantaggi. Statistical matching of multiple sources: A look through coherence. *Int. J. Approx. Reasoning*, 49(3):701–711, 2008.

Y et al. Wang. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365(9460):671–679, 2005.

AV Werhli and D Husmeier. Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article15, 2007.

S Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5(3):161–215, 1934.

J Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

XH Zhou, N Hu, G Hu, and M Root. Synthesis analysis of regression models with a continuous outcome. *Statistics in Medicine*, 28(11):1620–1635, 2009.

DW Zimmerman, BD Zumbo, and RH Williams. Bias in estimation and hypothesis testing of correlation. *Psicoliogica*, 24:133–158, 2003.