# Estimation and Control of the False Discovery Rate in Bayesian Network Skeleton Identification, with Application to Biological Data

Angelos Armen
Computer Science Department,
University of Crete,
Heraklion, Crete, Greece

March 2011

University of Crete
Computer Science Department

**Estimation and Control of the False Discovery Rate in Bayesian Network Skeleton Identification, with Application to Biological Data**

Thesis submitted by
Angelos Armen
in partial fulfillment of the requirements for the
Master of Science degree in Computer Science

THESIS APPROVAL

Author: Angelos Armen

Committee approvals:

Ioannis Tsamardinos
Assistant Professor, Thesis Supervisor

Panayiota Poirazi
Principal Researcher

Panagiotis Tsakalides
Professor

Director of Graduate Studies:

Angelos Bilas
Associate Professor

Heraklion, March 2011

# Contents

# List of Figures

# List of Tables

# Acknowledgements

I would like to thank my supervisor, Dr. Ioannis Tsamardinos, for the support and guidance he showed me throughout my thesis writing. In addition, I would like to express my gratitude to my colleagues in Bio-Informatics Laboratory and especially Sofia Triantafilou for providing assistance in numerous ways. I am also grateful to FORTH-ICS for its financial support. Finally, I would like to thank my family, especially my brother Georgios, and my friends for their support.

Angelos Armen

**Abstract**

Bayesian networks are graphical models that represent probabilistic relationships among variables with extensive applications including biological data analysis. In this work, we focus on the problem of estimating and controlling the False Discovery Rate (FDR) in learning the skeleton (set of edges without regard of direction) of a network. We present a unified approach to FDR estimation and control in Bayesian network skeleton identification and experimentally evaluate the performance of the most common FDR estimator in both tasks over several networks and sample sizes. We employ simulated data as well as real flow cytometry measurements of proteins and phospholipids in our evaluation. We demonstrate that estimation in some cases is not conservative and strong control is not achieved, while in other cases estimation is overly conservative. After identifying the possible causes of this lack of accuracy, we evaluate several approaches to deal with them. The results of these evaluations indicate that the goal of accurately estimating and controlling the FDR in all cases using the common FDR estimators may be unrealistic. Thus, we pursue the more realistic goal of accurately estimating and controlling the FDR according to a relaxed definition of false discovery. Our work opens new directions in the utilization of the FDR in learning Bayesian network structure and in estimating structural uncertainty in general.

**Περίληψη**

Τα Μπεϋσιανά δίκτυα είναι γραφικά μοντέλα που αναπαριστούν πιθανοτικές σχέσεις μεταξύ μεταβλητών με εκτενείς εφαρμογές συμπελαμβανομένης της ανάλυσης βιολογικών δεδομένων. Σε αυτή τη δουλειά, εστιάζουμε στο πρόβλημα της εκτίμησης και του ελέγχου του Ρυθμού Ψευδών Ανακαλύψεων (False Discovery Rate, FDR) στην εκμάθηση του σκελετού (σύνολο ακμών ανεξαρτήτως κατεύθυνσης) ενός δικτύου. Παρουσιάζουμε μια ενοποιημένη προσέγγιση στην εκτίμηση και τον έλεγχο του FDR της ταυτοποίησης σκελετού Μπεϋσιανών δικτύων και αποτιμούμε την επίδοση της πιο κοινής εκτιμήτριας του FDR και στα δυο προβλήματα σε αρκετά δίκτυα και μεγέθη δείγματος. Χρησιμοποιούμε εξομοιωμένα δεδομένα καθώς και πραγματικές μετρήσεις κυτταρομετρίας ροής πρωτεϊνών και φωσφολιπιδίων στην αποτίμηση μας. Δείχνουμε ότι η εκτίμηση σε μερικές περιπτώσεις δεν είναι συντηρητική και ισχυρός έλεγχος δεν επιτυγχάνεται, ενώ σε άλλες περιπτώσεις η εκτίμηση είναι υπερβολικά συντηρητική. Έχοντας ταυτοποιήσει τις πιθανές αιτίες αυτής της έλλειψης ακρίβειας, αποτιμούμε αρκετές προσεγγίσεις που επιλαμβάνονται αυτών. Τα αποτελέσματα αυτών των αποτιμήσεων υποδηλώνουν ότι ο στόχος της ακριβούς εκτίμησης και ελέγχου του FDR σε όλες τις περιπτώσεις χρησιμοποιώντας τις κοινές εκτιμήτριες του FDR ίσως να είναι μη ρεαλιστικός. Συνεπώς επιδιώκουμε τον πιο ρεαλιστικό στόχο της ακριβούς εκτίμησης και ελέγχου του FDR σύμφωνα με έναν χαλαρωμένο ορισμό της ψευδούς ανακάλυψης. Η δουλειά μας ανοίγει νέες κατευθύνσεις στη χρησιμοποίηση του FDR στην εκμάθηση δομής Μπεϋσιανών δικτύων και γενικά στην εκτίμηση της δομικής αβεβαιότητας.

# Chapter 1

# Introduction

Bayesian networks are graphical models that represent probabilistic relationships among variables with extensive applications including biological data analysis. A Bayesian network is comprised of a directed acyclic graph (DAG) whose nodes are the variables and the conditional probability distributions of each variable given values of its parents in the DAG. The DAG is called the structure of the Bayesian network. Structure learning is concerned with learning structure from data, and constraint-based algorithms are a class of algorithms for this purpose. These algorithms work in two phases, skeleton identification and edge orientation. Skeleton identification is concerned with identifying the set of links, i.e., edges without regard of direction, of the DAG or, in other words, its skeleton. Edge orientation is concerned with orienting these edges.

This thesis focuses on the problem of estimating and controlling the False Discovery Rate (FDR) in skeleton identification. In this context, FDR is the expected proportion of false links in the output skeleton. FDR is useful when it is of interest that the output skeleton is comprised of mostly true links. This is the case in biological applications such as learning gene networks from gene expression data, where the experimental verification of the learned connections among the genes is expensive.

We present a unified approach to FDR estimation and control in Bayesian network skeleton identification and experimentally evaluate the performance of the most common FDR estimator in both tasks over several networks and sample sizes. To assess performance on real data, a special network is included in our experiments: Structure learning is applied to flow cytometry measurements of proteins and phospholipids, as if the measurements were generated by a Bayesian network whose structure is the presumed protein-signaling network involving the measured molecules, as described in the biological literature.

We demonstrate that estimation in some cases is not conservative and strong control is not achieved, while in other cases estimation is overly conservative. After identifying the possible causes of this lack of accuracy, we evaluate several approaches to deal with them. The results of these evaluations indicate that the goal of accurately estimating and controlling the FDR in all cases using the common FDR estimators may be unrealistic. Thus, we pursue the more realistic goal of accurately estimating and controlling the FDR according to a relaxed definition of false discovery.

The rest of this thesis is organized as follows: Chapter 2 provides the necessary background on Bayesian network skeleton identification. Chapter 3 discusses the utilization of the FDR in skeleton identification, introduces a unified approach to FDR estimation and control in skeleton identification and presents an experimental evaluation of the performance of the most common FDR estimator in both tasks. Chapter 4 discusses and presents experimental evaluations of possible improvements over the results presented in the previous chapter. Finally, chapter 5 introduces and presents and experimental evaluation of a relaxed definition of false discovery.

# Chapter 2

# Bayesian network skeleton identification

In the first part of this chapter we present basic Bayesian network theory, while in the second part we present Bayesian network skeleton identification.

## 2.1 Bayesian networks

*Bayesian inference*, i.e., the act of computing conditional probabilities by applying Bayes' theorem, becomes too complex when it involves many related variables [26]. *Bayesian networks* are graphical models that address the problems of: [26]

1. Encoding probabilistic relationships among a large set of variables.
2. Performing probabilistic inference with those variables.

Bayesian networks do so by exploiting the so-called *Markov condition*, a relationship between graphs and probability distributions:

**Definition 2.1.** *(Markov condition) Suppose a joint probability distribution $P$ of the variables in some set $\mathbf{V}$ and a directed acyclic graph (DAG) $\mathbb{G} = (\mathbf{V}, \mathbf{E})$, where $\mathbf{E}$ is a set of edges. The pair $(\mathbb{G}, P)$ satisfies the Markov condition if every variable $X \in \mathbf{V}$ is conditionally independent of the set $\mathbf{ND}_X$ of its nondescendants given the set $\mathbf{PA}_X$ of its parents [26]:*

$$I_P(\{X\}, \mathbf{ND}_X \,|\, \mathbf{PA}_X)$$

Because the variables in $P$ are the nodes in $\mathbb{G}$, the terms variable and node are used interchangeably in the context of Bayesian networks. The following theorem states how the Markov condition enables the *factorization* of $P$:

**Theorem 2.1.** *If the pair $(\mathbb{G}, P)$ of a DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ and a joint probability distribution $P$ of the variables in some set $\mathbf{V} = \{X_1, X_2, \ldots, X_n\}$ satisfies the Markov condition, then $P$ is equal to the product of its conditional distributions of all nodes given values of its parents [26]:*

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | \, \mathbf{pa}_{X_i})$$

*where $x_1, x_2, \ldots, x_n$ are the values of $X_1, X_2, \ldots, X_n$ and $\mathbf{pa}_{X_i}$ are the values of the parents of $X_i$.*

*Proof.* The proof can be found in [26]. □

A Bayesian network is a pair $(\mathbb{G}, P)$ that satisfies the Markov condition. For categorical variables, a Bayesian network can be constructed by starting with a DAG and specifying the conditional probability distribution of each node given values of its parents of the node in the DAG:

**Theorem 2.2.** *Suppose a DAG $\mathbb{G}$ whose nodes are categorical variables and that the conditional probability distribution of each node given values of its parents in $\mathbb{G}$ is specified. The product of these conditional probability distributions is the joint probability distribution $P$ of the variables and $(\mathbb{G}, P)$ is a Bayesian network [26].*

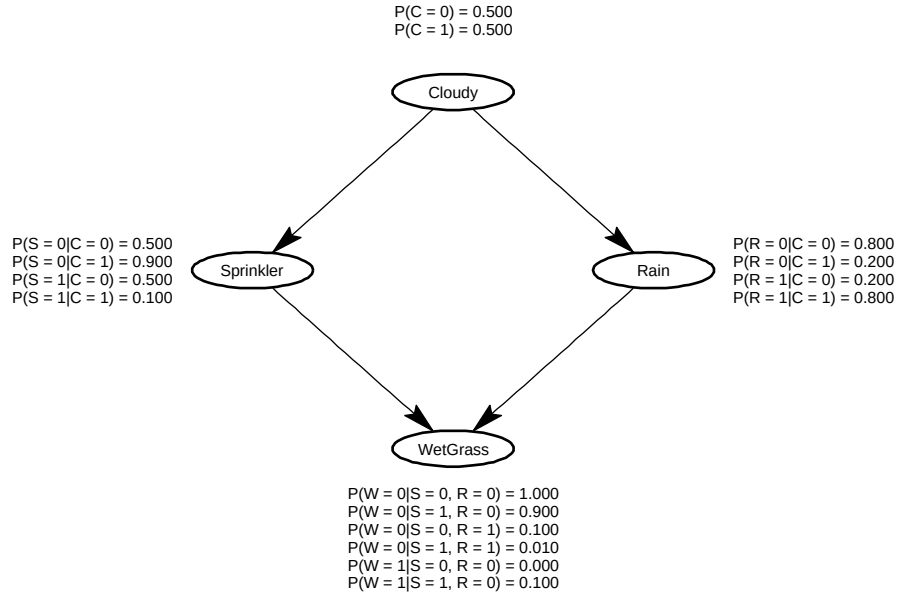



Figure 2.1: A binary Bayesian network [31] with four variables.

*Proof.* The proof can be found in [26]. □

In a Bayesian network $(\mathbb{G}, P)$, $\mathbb{G}$ is called the *structure* and the values in the conditional probability distributions in $P$ are called the *parameters* [26].

In this thesis, we refer to a Bayesian network whose all variables are categorical as a *categorical Bayesian network*. Moreover, we refer to a (categorical) Bayesian network whose all variables are binary as a *binary Bayesian network*.

**Example 2.1.** *Fig. 2.1 shows a binary Bayesian network [31] with four variables. According to the statement of the Markov condition, the following two conditional independencies hold:*

$$I_P(\{Sprinkler\}, \{Rain\}|\{Cloudy\})$$

$$I_P(\{WetGrass\}, \{Cloudy\}|\{Sprinkler, Rain\})$$

*That is, when we know whether is cloudy, knowledge of whether the sprinkler is on gives us no further information about whether it rains and vice-versa. Also, when we know whether the sprinkler is on and whether it rains, knowing whether is cloudy gives us no further information about whether the grass is wet.*

### 2.1.1 d-separation

When the Markov condition is satisfied, each node is conditionally independent from its nondescendants given its parents. However, these conditional independencies *entail* even more conditional independencies:

**Definition 2.2.** *(Entailment) Let* $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ *be a DAG where* $\mathbf{V}$ *is a set of random variables. We say that, based on the Markov condition,* $\mathbb{G}$ *entails conditional independence* $I_P(\mathbf{A}, \mathbf{B}|\mathbf{C})$ *for* $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C} \subseteq \mathbf{V}$ *if*

$$I_P(\mathbf{A}, \mathbf{B}|\mathbf{C}) \text{ holds for every } P \in \mathbf{P}$$

*where* $\mathbf{P}$ *is the set of all probability distributions* $P$ *such that* $(\mathbb{G}, P)$ *satisfies the Markov condition [26].*

The Markov condition specifically entails all and only those conditional independencies in $P$ that are *identified* in $\mathbb{G}$ by a graphical criterion called *d-separation*. Before discussing d-separation, let us discuss *chains*.

In a DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$, a chain is the set of edges connecting the nodes of a set $\{X_1, X_2, \ldots, X_k\}$, where $k > 2$, such that $(X_{i-1}, X_i) \in \mathbf{E}$ or $(X_i, X_{i-1}) \in \mathbf{E}$ for $2 \leq i \leq k$ [26].

- For a chain $X \rightarrow Z \rightarrow Y$, we say that the edges meet *head-to-tail* at $Z$.
- For a chain $X \leftarrow Z \rightarrow Y$, we say that the edges meet *tail-to-tail* at $Z$.
- For a chain $X \rightarrow Z \leftarrow Y$, we say that the edges meet *head-to-head* at $Z$.

We call a chain $X$–$Z$–$Y$ such that $X$ and $Y$ are not adjacent an *uncoupled meeting* [26].

**Definition 2.3.** *(Chain blocking) Let* $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ *be a DAG,* $\mathbf{A} \subseteq \mathbf{V}$, $X$ *and* $Y$ *be distinct nodes in* $\mathbf{V} \setminus \mathbf{A}$ *and* $\rho$ *be a chain between* $X$ *and* $Y$. $\rho$ *is* blocked *by* $\mathbf{A}$ *if one of the following holds [26]:*

1. *There is a node* $Z \in \mathbf{A}$ *on the chain* $\rho$*, and the edges incident to* $Z$ *on* $\rho$ *meet head-to-tail at* $\mathbf{Z}$*.*
2. *There is a node* $Z \in \mathbf{A}$ *on the chain* $\rho$*, and the edges incident to* $Z$ *on* $\rho$ *meet tail-to-tail at* $\mathbf{Z}$*.*
3. *There is a node* $Z$*, such that* $Z$ *and all of* $Z$*'s descendants are not in* $\mathbf{A}$*, on the chain* $\rho$*, and the edges incident to* $Z$ *on* $\rho$ *meet head-to-head at* $\mathbf{Z}$*.*

**Example 2.2.** *In Fig. 2.1, the chain* $[Cloudy, Sprinkler, WetGrass]$ *is blocked by* $\{Sprinkler\}$ *because the edges on the chain incident to Sprinkler meet head-to-tail at Sprinkler, while* $[Sprinkler, Cloudy, Rain]$ *is blocked by* $\{Cloudy\}$ *because the edges on the chain incident to Cloudy meet tail-to-tail at Cloudy and by* $\{WetGrass\}$ *because* $WetGrass \in \{WetGrass\}$ *is a descendant of Cloudy. The chain* $[Sprinkler, WetGrass, Rain]$ *is blocked by* $\emptyset$ *because the edges on the chain incident to* $WetGrass$ *meet head-to-head at* $WetGrass$*,* $WetGrass \notin \emptyset$ *and* $WetGrass$ *has no descendents. However,* $[Sprinkler, WetGrass, Rain]$ *is* not *blocked by* $WetGrass$ *because* $WetGrass \in \{WetGrass\}$*.*

Now let us now present the definition of d-separation, first for nodes and then for sets of nodes:

**Definition 2.4.** *(d-separation of nodes) Let* $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ *be a DAG, and* $X$ *and* $Y$ *be distinct nodes in* $\mathbf{V} \setminus \mathbf{A}$*. We say* $X$ *and* $Y$ *are* d-separated *by* $\mathbf{A}$ *in* $\mathbb{G}$ *if every chain between* $X$ *and* $Y$ *is blocked by* $\mathbf{A}$ *[26].*

**Definition 2.5.** *(d-separation of sets of nodes) Let* $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ *be a DAG, and* $\mathbf{A}$*,* $\mathbf{B}$*, and* $\mathbf{C}$ *be mutually disjoint subsets of* $\mathbf{V}$*. We say* $\mathbf{A}$ *and* $\mathbf{B}$ *are d-separated by* $\mathbf{C}$ *in* $\mathbb{G}$ *if for every* $X \in \mathbf{A}$ *and* $X \in \mathbf{B}$*,* $X$ *and* $Y$ *are d-separated by* $\mathbf{C}$ *[26]. We write* $I_{\mathbb{G}}(\mathbf{A}, \mathbf{B}|\mathbf{C})$*.*

**Example 2.3.** *In Fig. 2.1,* $I_{\mathbb{G}}(\{WetGrass\}, \{Cloudy\}|\{Sprinkler, Rain\})$ *holds because both chains between* $X$ *and* $Y$*, namely* $[WetGrass, Sprinkler, Cloudy]$ *and* $[WetGrass, Rain, Cloudy]$*, are blocked by* $\{Sprinkler, Rain\}$*.*

We say that conditional independency $\mathrm{I}_P(\mathbf{A}, \mathbf{B}|\mathbf{C})$ is identified by d-separation in $\mathbb{G}$ when $\mathrm{I}_{\mathbb{G}}(\mathbf{A}, \mathbf{B}|\mathbf{C})$ holds. The following theorem states that a d-separation in $\mathbb{G}$ implies a conditional independency in $P$:

**Theorem 2.3.** *Based on the Markov condition, a DAG* $\mathbb{G}$ *entails all and only those conditional independencies that are identified by d-separation in* $\mathbb{G}$*:*

$$I_{\mathbb{G}}(\mathbf{A}, \mathbf{B}|\mathbf{Z}) \implies I_P(\mathbf{A}, \mathbf{B}|\mathbf{Z})$$

*Proof.* The proof can be found in [26]. □

### 2.1.2 Markov equivalence

Many DAGs have the same d-separations. We refer to two DAGs having the same d-separations as being *Markov equivalent*:

**Definition 2.6.** *(Markov equivalence) Let* $\mathbb{G}_1 = (\mathbf{V}, \mathbf{E_1})$ *and* $\mathbb{G}_2 = (\mathbf{V}, \mathbf{E_1})$ *be two DAGs containing the same set of nodes* $\mathbf{V}$*. Then* $\mathbb{G}_1$ *and* $\mathbb{G}_2$ *are called Markov equivalent if for every three mutually disjoint subsets* $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$*,* $\mathbf{A}$ *and* $\mathbf{B}$ *are d-separated by* $\mathbf{C}$ *in* $\mathbb{G}_1$ *if and only if* $\mathbf{A}$ *and* $\mathbf{B}$ *are d-separated by* $\mathbf{C}$ *in* $\mathbb{G}_2$ *[26]. That is*

$$I_{\mathbb{G}_1}(\mathbf{A}, \mathbf{B}|\mathbf{C}) \iff I_{\mathbb{G}_2}(\mathbf{A}, \mathbf{B}|\mathbf{C})$$

**Theorem 2.4.** *Two DAGs are Markov equivalent if and only if, based on the Markov condition, they entail the same conditional independencies [26].*

*Proof.* The proof follows immediately from Theorem 2.3. □

The following lemma relates adjacency to d-separation in a DAG and is of great importance in learning the structure of a Bayesian network from data, as we will see in Section 2.2:

**Lemma 2.1.** *Let* $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ *be a DAG and* $X, Y \in \mathbf{V}$*. Then* $X$ *and* $Y$ *are adjacent in* $\mathbb{G}$ *if and only if they are not d-separated by some set in* $\mathbb{G}$ *[26]:*

$$Adj_{\mathbb{G}}(X, Y) \iff \nexists \mathbf{S}_{XY} \subseteq \mathbf{V} \setminus \{X, Y\} \text{ s.t. } I_{\mathbb{G}}(\{X\}, \{Y\}|\mathbf{S}_{XY})$$

*where* $Adj_{\mathbb{G}}(X, Y)$ *denotes that* $X$ *and* $Y$ *are adjacent in* $\mathbb{G}$*.*

*Proof.* The proof can be found in [26]. □

The following corollary of Lemma 2.1 is also of great importance in learning the structure of a Bayesian network from data:

**Corollary 2.1.** *Let $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ be a DAG and $X, Y \in \mathbf{V}$. Then if $X$ and $Y$ are d-separated by some set, they are d-separated either by the set of the parents of $X$ or the set of the parents of $Y$ [26]:*

$$\exists \mathbf{Z} \subseteq \mathbf{V} \setminus \{X, Y\} \text{ s.t. } I_{\mathbb{G}}(\{X\}, \{Y\} | \mathbf{Z}) \implies$$
$$I_{\mathbb{G}}(\{X\}, \{Y\} | \mathbf{PA}_X) \text{ or } I_{\mathbb{G}}(\{X\}, \{Y\} | \mathbf{PA}_Y)$$

*Proof.* The proof follows from the proof of Lemma 2.1 [26]. □

The next lemma states that Markov equivalent DAGs have the same *links* (edges without regard of direction) or, in other words, the same *skeleton*.

**Lemma 2.2.** *If $\mathbb{G}_1$ and $\mathbb{G}_2$ are Markov equivalent, then $X$ and $Y$ are adjacent in $\mathbb{G}_1$ if and only if they are adjacent in $\mathbb{G}_2$. That is, Markov equivalent DAGs have the same links:*

$$\mathbb{G}_1 \text{ and } \mathbb{G}_2 \text{ Markov equivalent} \implies \left[ Adj_{\mathbb{G}_1}(X, Y) \iff Adj_{\mathbb{G}_2}(X, Y) \right]$$

*Proof.* The proof can be found in [26]. □

The next theorem establishes a criterion for recognizing Markov equivalence:

**Theorem 2.5.** *Two DAGs $\mathbb{G}_1$ and $\mathbb{G}_2$ are Markov equivalent if and only if they have the same links (skeleton) and the same set of uncoupled head-to-head meetings.*

*Proof.* The proof can be found in [26]. □

This criterion allows us to represent a Markov equivalence class with a single graph, called the *DAG pattern*:

**Definition 2.7.** *(DAG pattern) A DAG pattern for a Markov equivalence class is the graph that has the same links (skeleton) as the DAGs in the equivalence class and has oriented all and only the edges common to all of the DAGs in the equivalence class [26].*

Since all DAGs in a Markov equivalence class have by definition the same d-separations, we can define d-separation for DAG patterns:

**Definition 2.8.** *(d-separation for DAG patterns) Let gp be a DAG pattern whose nodes are the elements of $\mathbf{V}$ and $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, and let $\mathbf{C}$ be a mutually disjoint subsets of $\mathbf{V}$. We say $\mathbf{A}$ and $\mathbf{B}$ are d-separated by $\mathbf{C}$ in gp (and denote it with $I_{gp}(\mathbf{A}, \mathbf{B} | \mathbf{C})$) if $\mathbf{A}$ and $\mathbf{B}$ are d-separated by $\mathbf{C}$ in every DAG $\mathbb{G}$ in the Markov equivalence class represented by gp [26].*

*Proof.* The proof follows from Lemma 2.1. □

The following lemma relates adjacency to d-separation in a DAG pattern:

**Lemma 2.3.** *Let gp be a DAG and $X$ and $Y$ be nodes in gp. Then $X$ and $Y$ are adjacent in gp if and only if they are not d-separated by some set in gp [26]:*

$$Adj_{gp}(X, Y) \iff \nexists \mathbf{S}_{XY} \subseteq \mathbf{V} \setminus \{X, Y\} \text{ s.t. } I_{gp}(\{X\}, \{Y\} | \mathbf{S}_{XY})$$

*where $Adj_{gp}(X, Y)$ denotes that $X$ and $Y$ are adjacent in gp.*

### 2.1.3 Faithfulness

In general, there may be conditional independencies in $P$ that are not identified by d-separation in $\mathbb{G}$. The criterion of *faithfulness* requires that this is not the case:

**Definition 2.9.** *(Faithfulness) Suppose we have a joint probability distribution $P$ of the random variables in some set $\mathbf{V}$ and a DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$. We say that $(\mathbb{G}, P)$ satisfies the* faithfulness *condition if, based on the Markov condition, $\mathbb{G}$ entails all and only conditional independencies in $P$. That is, the following two conditions hold [26]:*

1. *$(\mathbb{G}, P)$ satisfies the Markov condition*
2. *All conditional independencies in $P$ are entailed by $\mathbb{G}$, based on the Markov condition.*

When $(\mathbb{G}, P)$ satisfies the faithfulness condition, we say $P$ and $\mathbb{G}$ are *faithful* to each other [26]. The following theorem establishes a necessary and sufficient criterion for faithfulness:

**Theorem 2.6.** *Suppose we have a joint probability distribution $P$ of the random variables in some set $\mathbf{V}$ and a DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$. Then $(\mathbb{G}, P)$ satisfies the faithfulness condition if and only if all and only conditional independencies in $P$ are identified by d-separation in $\mathbb{G}$ [26].*

*Proof.* The proof follows immediately from Theorem 2.3. □

If $P$ is faithful to some DAG then $P$ is faithful to an equivalence class of DAGs, as stated by the following theorem:

**Theorem 2.7.** *(Faithfulness for DAG patterns) If $(\mathbb{G}, P)$ satisfies the faithfulness criterion, then $P$ satisfies the criterion with all and only those DAGs in the Markov equivalence class $gp$ of $\mathbb{G}$. In addition, the d-separations in $gp$ identify all and only conditional independences in $P$. We say that $gp$ and $P$ are* faithful *to each other [26].*

*Proof.* The proof follows immediately from Theorem 2.6. □

We say that $P$ *admits a faithful DAG representation* if $P$ is faithful to some DAG (and therefore to some DAG pattern) [26]. Almost all categorical Bayesian networks are faithful [25].

## 2.2 Bayesian network skeleton identification

Historically, the structure of a Bayesian network was manually constructed by a domain expert and then the parameters were either specified by the expert or learned from data using methods for *parameter learning* [26]. The difficulties of manually constructing large Bayesian networks lead the researchers to devise methods for *structure learning* from data [26].

The goal of structure learning is to find a DAG $\mathbb{G}$ or DAG pattern $gp$ faithful to a distribution $P$ given a sample from $P$, assuming $P$ admits a faithful DAG representation. There are three main approaches to structure learning [46]:

1. *Search-and-score* methods search for the DAG or DAG pattern that maximizes some score function. Typically, the score function is the posterior probability of the DAG or DAG pattern given the data. If a single DAG (DAG pattern) is not found to be overwhelmingly probable, then *Bayesian model averaging* over DAGs (DAG patterns) may be performed [26].

2. *Constraint-based* methods involve two phases: first the d-separations in *gp* are identified and then they are used as *constraints* in generating *gp*. The first phase is called *constraint* or *skeleton identification* because it corresponds to learning the skeleton of *gp*. The second phase is called *edge orientation* because it corresponds to orienting the undirected edges of the skeleton identified by the first phase.

3. *Hybrid* algorithms combine the previous two methods.

The basic skeleton identification algorithm template (Algorithm 2.1) is based on Lemma 2.3, according to which lemma two nodes are adjacent in a DAG pattern if and only if they are not d-separated. For each pair $(X, Y)$ of nodes in some set $\mathbf{V}$, a search for a subset $\mathbf{S}_{XY}$ of $\mathbf{V}\backslash\{X, Y\}$ such that $\mathrm{I}_P(\{X\}, \{Y\}|\mathbf{S}_{XY})$ is identified takes place. If such subset is found, the pair is no longer considered, otherwise, link $X - Y$ is discovered.

**Algorithm Template 2.1** Basic skeleton identification [26]

**for** each pair of nodes $(X, Y) \in \mathbf{V}$ **do**
  search for a subset $\mathbf{S}_{XY} \subseteq \mathbf{V} \setminus \{X, Y\}$ s.t. $\mathrm{I}_P(\{X\}, \{Y\}|\mathbf{S}_{XY})$ is identified
  **if** no such set can be found **then**
    discover $X - Y$
  **end if**
**end for**

**Theorem 2.8.** *Suppose that the pair $(\mathbb{G}, P)$ of a DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ and a joint probability distribution $P$ of the variables in some set $\mathbf{V}$ is a Bayesian network. Suppose further that an algorithm instantiating Algorithm Template 2.1 is applied on a sample from $P$. If*

1. *$\mathbb{G}$ and $P$ are faithful to each other and*

2. *conditional independencies in $P$ are correctly identified,*

*then the algorithm discovers all links in $\mathbb{G}$.*

*Proof.* Owing to assumption 2, the algorithm discovers link $X - Y$ if and only if there is no $\mathbf{V} \setminus \{X, Y\}$ such that $\mathrm{I}_P(\{X\}, \{Y\}|\mathbf{S}_{XY})$, which, due to Lemma 2.1, is the case if and only $X$ and $Y$ are adjacent in $\mathbb{G}$. □

Algorithms based on the previous template are inefficient because they search all $2^{n-2}$ subsets of $\mathbf{V} \setminus \{X, Y\}$ to determine whether $X$ and $Y$ are adjacent in *gp* [26]. According to Corollary 2.1, $X$ and $Y$ are d-separated if and only if they are d-separated either by the parents of $X$ or the parents of $Y$. The parents are, of course, unknown, so algorithms based on Corollary 2.1 check the conditional independency of $X$ and $Y$ given all subsets of supersets of the

neighbors $\mathbf{ADJ}_X$ of $X$ and $\mathbf{ADJ}_X$ of $Y$. These algorithms are instantiations of Algorithm Template 2.2.

First, for each $X \in \mathbf{V}$, the current estimate $\widehat{\mathbf{ADJ}}_X$ of $\mathbf{ADJ}_X$ is initialized to a subset of $\mathbf{V} \setminus \{X\}$. Moreover, the set $\mathbf{R}_X$ ("rest") of the nodes not yet considered for inclusion in $\widehat{\mathbf{ADJ}}_X$ is initialized to $\mathbf{V} \setminus (\widehat{\mathbf{ADJ}}_X \cup \{X\})$. Then, for each $X \in \mathbf{V}$, the following two steps are repeated until $\mathbf{R}_X$ is empty:

1. For each each $Y \in \widehat{\mathbf{ADJ}}_X$, a search for a subset $\mathbf{S}_{XY}$ of $\widehat{\mathbf{ADJ}}_X \setminus \{Y\}$ such that $\mathrm{I}_P(\{X\}, \{Y\} | \mathbf{S}_{XY})$ takes place. If such subset is found, $Y$ and $X$ are removed from $\widehat{\mathbf{ADJ}}_X$ and $\widehat{\mathbf{ADJ}}_Y$ respectively.

2. A subset $\mathbf{W}$ of $\mathbf{R}_X$ is moved from $\mathbf{R}_X$ to $\widehat{\mathbf{ADJ}}_X$.

**Algorithm Template 2.2** Efficient skeleton identification

**for** each $X \in \mathbf{V}$ **do**
  $\widehat{\mathbf{ADJ}}_X \leftarrow \mathbf{U}$ s.t. $\mathbf{U} \subseteq \mathbf{V} \setminus \{X\}$
  $\mathbf{R}_X \leftarrow \mathbf{V} \setminus (\widehat{\mathbf{ADJ}}_X \cup \{X\})$
**end for**
**for** each $X \in \mathbf{V}$ **do**
  **repeat**
    **for** each $Y \in \widehat{\mathbf{ADJ}}_X$ **do**
      search for $\mathbf{S}_{XY} \subseteq \widehat{\mathbf{ADJ}}_X \setminus \{Y\}$ s.t. $\mathrm{I}_P(\{X\}, \{Y\} | \mathbf{S}_{XY})$ is identified
      **if** such a set is found **then**
        $\widehat{\mathbf{ADJ}}_X \leftarrow \widehat{\mathbf{ADJ}}_X \setminus \{Y\}$
        $\widehat{\mathbf{ADJ}}_Y \leftarrow \widehat{\mathbf{ADJ}}_Y \setminus \{X\}$
      **end if**
    **end for**
    $\widehat{\mathbf{ADJ}}_X \leftarrow \widehat{\mathbf{ADJ}}_X \cup \mathbf{W}$ s.t. $\mathbf{W} \subseteq \mathbf{R}_X$
    $\mathbf{R}_X \leftarrow \mathbf{R}_X \setminus \mathbf{W}$
  **until** $\mathbf{R}_X = \emptyset$
**end for**

Some constraint-based structure learning algorithms instantiating Algorithm Template 2.2 in their skeleton identification phase are *PC* [35] and algorithms belonging to the *Local to Global Learning* (LGL) class of constraint-based algorithms [3, 4], such as *MMHC* [46].

**Lemma 2.4.** *Suppose that the pair $(\mathbb{G}, P)$ of a DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ and a joint probability distribution $P$ of the variables in some set $\mathbf{V}$ is a Bayesian network. Suppose further that an algorithm instantiating Algorithm Template 2.2 is applied on a sample from $P$. If*

1. *$\mathbb{G}$ and $P$ are faithful to each other and*

2. *conditional independencies in $P$ are correctly identified,*

*then the algorithm discovers all links in $\mathbb{G}$.*

*Proof.* For each $X \in \mathbf{V}$ and $Y \in \mathbf{V} \setminus \{X\}$, the algorithm eventually inserts $Y$ in $\widehat{\mathbf{ADJ}}_X$. Owing to the assumption 2, the algorithm removes $Y$ from $\widehat{\mathbf{ADJ}}_X$ if

and only if there is a subset $\mathbf{S}_{XY}$ of $\widehat{\mathbf{ADJ}}_X$ such that $\mathrm{I}_P(\{X\},\{Y\}|\mathbf{S}_{XY})$ holds, which, due to Lemma 2.3, is not the case if $Y \in \mathbf{PA}_X$. Thus, $\mathbf{PA}_X$ is never removed from $\widehat{\mathbf{ADJ}}_X$. That is, the algorithm discovers all links in $\mathbb{G}$. □

**Lemma 2.5.** *Suppose a Bayesian network $(\mathbb{G}, P)$ and that an algorithm instantiating Algorithm Template 2.2 is applied on a sample from $P$. If*

1. *$\mathbb{G}$ and $P$ are faithful to each other and*
2. *conditional independencies in $P$ are correctly identified,*

*then the algorithm does not discover links not in $\mathbb{G}$.*

*Proof.* Suppose that nodes $X$ and $Y$ are adjacent in $\mathbb{G}$. The algorithm eventually inserts $Y$ in $\widehat{\mathbf{ADJ}}_X$ and $X$ in $\widehat{\mathbf{ADJ}}_Y$ and, owing to the assumption 2, it removes them if and only if there is a subset $\mathbf{S}_{XY}$ of $\widehat{\mathbf{ADJ}}_X$ or $\widehat{\mathbf{ADJ}}_Y$ such that $\mathrm{I}_P(\{X\},\{Y\}|\mathbf{S}_{XY})$ holds. Owing to Lemma 2.4, $\mathbf{PA}_X$ and $\mathbf{PA}_Y$ are never removed from $\widehat{\mathbf{ADJ}}_X$ and $\widehat{\mathbf{ADJ}}_Y$ respectively, and, due to Corollary 2.1, either $\mathrm{I}_P(\{X\},\{Y\}|\,\mathbf{PA}_X)$ or $\mathrm{I}_P(\{X\},\{Y\}|\,\mathbf{PA}_Y)$ holds. Thus, the algorithm eventually removes $Y$ from $\widehat{\mathbf{ADJ}}_X$ and $X$ from $\widehat{\mathbf{ADJ}}_Y$. That is, the algorithm does not discover $X - Y$. □

**Theorem 2.9.** *Suppose a Bayesian network $(\mathbb{G}, P)$ and that an algorithm instantiating Algorithm Template 2.2 is applied on a sample from $P$. If*

1. *$\mathbb{G}$ and $P$ are faithful to each other and*
2. *conditional independencies in $P$ are correctly identified,*

*then the algorithm discovers all and only links in $\mathbb{G}$.*

*Proof.* The proof follows immediately from Lemmas 2.4 and 2.5. □

Conditional independencies are identified from a sample from $P$ by performing *hypothesis tests* of conditional independence. Before discussing these tests, let us first review hypothesis testing.

### 2.2.1 Hypothesis testing

Let $\theta \in \Omega$ be a parameter (e.g. the mean) of some probability distribution. Suppose that the value of $\theta$ is unknown and we would like to prove the hypothesis that $\theta \in \Omega_1 \subset \Omega$. The hypothesis that the opposite is true, i.e., the hypothesis that $\theta \in \Omega_0 = \Omega \setminus \Omega_1$, is called the *null* hypothesis and denoted by $H_0$. The hypothesis that $\theta \in \Omega_1$ is called the *alternative* hypothesis and denoted by $H_1$. The problem of deciding whether to accept $H_0$ or $H_1$ given a random sample from the distribution is a *hypothesis testing* problem [9]. A hypothesis test is a procedure for making this decision.

Hypothesis tests are usually performed using a *test statistic* $T$. A test statistic is a function that summarizes the sample. $H_0$ is rejected and, subsequently, $H_1$ is accepted, if the value $t$ of the statistic lies in some region $C$ called the *critical region* of the test. A *Type I Error* or *false positive* occurs when $H_0$ is rejected when $H_0$ is true. The probability of a Type I error is $\Pr(T \in C|\theta \in \Omega_0)$ and called the *False Positive Rate* (FPR) of the test. A *Type II Error* or *false*

*negative* occurs when $H_0$ is accepted when $H_1$ is true. The probability of a Type II error is $1 - \Pr(T \in C|\theta \in \Omega_1)$ and called the *False Negative Rate* (FNR), while $\pi = \Pr(T \in C|\theta \in \Omega_1)$ is called the *power* of the test.

One would like both the FPR and the FNR to be low. However, these two goals act in competition to each other, in general [9]. For example, suppose a test that always accepts $H_0$. The FPR of this test is 0, but the FNR is 1. On the other hand, suppose a test that always rejects $H_0$. The FNR of this test is 0, but the FPR is 1. These two goals have to be balanced somehow. It is usually required that the FPR is below some threshold $\alpha$ called the *significance level* of the test. The most common value of $\alpha$ is 0.05. Then, among all tests with FPR $\leq \alpha$, the one with that maximizes power is chosen. If we reject $H_0$ when $T \geq c$ for some constant $c$, then the FPR of the test is $\Pr(T \geq c|\theta \in \Omega_0)$ and the power is $\Pr(T \geq c|\theta \in \Omega_1)$. Both $\Pr(T \geq c|\theta \in \Omega_0)$ and $\Pr(T \geq c|\theta \in \Omega_1)$ are nonincreasing functions of $c$. Thus, in order for $\Pr(T \geq c|\theta \in \Omega_0) \leq \alpha$ and $\Pr(T \geq c|\theta \in \Omega_1)$ to be as large as possible, we find the smallest $c$ such that $\Pr(T \geq c|\theta \in \Omega_0) \leq \alpha$.

Typically, $\alpha$ is not chosen beforehand and then the acceptance or rejection of $H_0$ at level $\alpha$ simply reported. The *p-value* of the test is reported instead [9]. The p-value is the smallest significance level $\alpha$ such that $H_0$ would still be rejected. If we reject $H_0$ when $T \geq c$ for some threshold $c$, then the p-value of the test is $\Pr(T \geq t|\theta \in \Omega_0)$ [9]. If $c$ is the smallest $c$ such that $\Pr(T \geq c|\theta \in \Omega_0) \leq \alpha$, then rejecting $H_0$ if $T \geq c$ corresponds to rejecting $H_0$ if the p-value $\leq \alpha$.

**Example 2.4.** *[9] Let $X_1, ..., X_{10}$ be the a random sample from a Bernoulli distribution with parameter p. Suppose that we would like to test the hypothesis $H_0 : p \leq 3$ versus the alternative $H_1 : p > 3$. Let $Y = \sum_{i=1}^{10} X_i$ be our statistic. Y follows the binomial distribution with parameters $n = 10$ and p. Suppose that we choose to reject $H_0$ if $Y \geq c$ for some threshold c. Also suppose that we want to control the FPR $\leq 0.1$ while maximizing the power of the test. Thus, c should be the smallest c such that $Pr(Y \geq c|p = 3) = \sum_{y=c}^{10} Pr(Y = y|p = 0.3) \leq 0.1$. From a table of the binomial distribution or using statistical software we can find out that the smallest such integer c is 6, for which $\sum_{y=6}^{10} Pr(Y = y|p = 0.3) = 0.0473$; 0.0473 is the p-value of the test.*

### 2.2.2 Testing conditional independence

To determine whether $\mathrm{I}_P(\{X\}, \{Y\}|\mathbf{Z})$, a test $\mathrm{test}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ of the null hypothesis $\mathrm{H}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ that $X$ and $Y$ are conditionally independent given $Z$ versus the alternative hypothesis $\mathrm{H}_{\neg \mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ that they are not, is performed. First, the value $t_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ of a test statistic $\mathrm{T}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ is calculated and then the p-value $\mathrm{p}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ is computed. If $\mathrm{p}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})} > \alpha$, then $\mathrm{H}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ is accepted. Subsequently, $(X, Y)$ is no longer considered and $X - Y$ is not discovered. If $\mathrm{p}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})} \leq \alpha$, $\mathrm{H}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ is rejected and the search for a d-separating set continues.

Typically, the $G$ test is employed when all variables in $\mathbf{V}$ are categorical [46, 3]. The G test uses the $G$ statistic. Before discussing this statistic, let us first discuss *contingency tables*.

The contingency table of a sample from a multivariate categorical distribution is a table that contains the number of occurrences of each possible obser-

vation of the distribution in the sample [2]. In a test $\text{test}_{\text{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ involving only categorical variables, the contingency table has dimensions $|\mathbf{D}_X| \times |\mathbf{D}_Y| \times |\mathbf{D}_{Z_1}| \times \cdots \times |\mathbf{D}_{Z_{|Z|}}|$, where $\mathbf{D}_X$ is the domain of variable $X$. An empty cell in which observations are impossible is called a *structural zero* [2].

The G statistic is defined as follows:

$$G \triangleq \sum_{a,b,\mathbf{c}} N_{i,j,k}^{a,b,\mathbf{c}} \ln \frac{N_{i,j,k}^{a,b,\mathbf{c}} N_k^{\mathbf{c}}}{N_{i,k}^{a,\mathbf{c}} N_{j,k}^{b,\mathbf{c}}}$$

where $N_{i,j,k}^{a,b,\mathbf{c}}$ is the number of observations with $X = a$, $Y = b$ and $\mathbf{Z} = \mathbf{c}$ in the sample and $N_k^{\mathbf{c}}$, $N_{i,k}^{a,\mathbf{c}}$ and $N_{j,k}^{b,\mathbf{c}}$ are defined accordingly. When $\text{I}_P(\{X\},\{Y\}|\mathbf{Z})$ holds, the G statistic asymptotically (i.e., as the sample size $n \to \infty$) follows the $\chi^2$ distribution with *df* degrees of freedom. Assuming no structural zeros, *df* is given by the following equation:

$$df = (|\mathbf{D}_X| - 1)(|\mathbf{D}_Y| - 1) \prod_{k:Z_k \in \mathbf{Z}} |\mathbf{D}_{Z_k}| \tag{2.1}$$

As a heuristic, in their implementation of PC, Spirtes, Glymour and Scheines [35] reduce *df* by one for each *sampling zero*, i.e., for each cell of the contingency table that is zero. On the other hand, in their implementation of MMHC, Tsamardinos, Brown and Aliferis [46] calculate *df* according to Steck and Jaakkola [36].

A test is typically attempted only if it is reliable according to a *reliability criterion*, otherwise it is ignored [46]. A reliable test both (a) meets the distributional assumptions of the statistic used and (b) has sufficient power [12].

For categorical variables, Fienberg [14] recommends that there at least five observations per cell of the contingency table of the test, on average, for the test to be reliable. Following the practice of Aliferis et. al. [3], we refer to the lower limit on the average number of observations per cell as the *heuristic power size* and denote it with *h-ps*. We refer to the corresponding reliability criterion as the *heuristic power rule*.

In general, as conditioning set cardinality $|\mathbf{Z}|$ increases, reliability decreases [46]. Thus, another possible reliability criterion is to consider a test reliable only if $|\mathbf{Z}|$ is below some upper limit. Following the practice of Aliferis et. al. [3], we denote this upper limit with *max-k*. We refer to the corresponding reliability criterion as the *conditioning cardinality rule*. The conditioning cardinality rule with *max-k* $> |\mathbf{V}| - 2$ is the same as with *max-k* $= |\mathbf{V}| - 2$. For categorical variables, the conditioning cardinality rule is usually used in conjunction with the heuristic power rule as a way to reduce the execution time of the algorithm [3].

Even if a test is attempted, it may not be completed because the computation of its p-value is not possible. This is the case when *df* calculated according to either Spirtes, Glymour and Scheines [35] or Steck and Jaakkola [36] is negative. In this case also the test is ignored.

A Type I Error or false positive occurs when $\neg\,\text{I}_P(\{X\},\{Y\}|\mathbf{Z})$ is concluded while $\text{I}_P(\{X\},\{Y\}|\mathbf{Z})$ holds. A Type II Error or false negative occurs when $\text{I}_P(\{X\},\{Y\}|\mathbf{Z})$ is concluded while $\neg\,\text{I}_P(\{X\},\{Y\}|\mathbf{Z})$ holds. Since a link is not discovered once a set that renders its ends conditionally independent is found, a single Type II Error results in a true link missing, while multiple Type I Errors result in a false link in the identified skeleton.

**Theorem 2.10.** *Suppose a Bayesian network $(\mathbb{G}, P)$ and that a hypothesis test based algorithm instantiating either Algorithm Template 2.1 or Algorithm Template 2.2 is applied on a sample from $P$. If*

1. *$\mathbb{G}$ and $P$ are faithful to each other,*
2. *all tests considered by the algorithm are reliable according to the employed reliability criterion,*
3. *all tests attempted by the algorithm are completed and*
4. *all tests completed yield the correct result,*

*then the algorithm discovers all and only links in $\mathbb{G}$.*

*Proof.* Owing to assumptions 2-4, the algorithm correctly identifies conditional independencies in $P$. Thus, the proof for Algorithm Template 2.1 follows immediately from Theorem 2.8, while the proof for Algorithm Template 2.2 follows immediately from Theorem 2.9. □

In the theorem above we assumed that all tests considered by the algorithm are reliable according to the employed reliability criterion, thus attempted. In the following lemmas and theorem, presented for the first time in this work, we drop this assumption:

**Lemma 2.6.** *Suppose that the pair $(\mathbb{G}, P)$ of a DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ and a joint probability distribution $P$ of the variables in some set $\mathbf{V}$ is a Bayesian network. Suppose further that a hypothesis test based algorithm instantiating Algorithm Template 2.2 is applied on a sample from $P$. If*

1. *$\mathbb{G}$ and $P$ are faithful to each other and*
2. *completed tests do not yield a false negative result*

*then the algorithm discovers all links in $\mathbb{G}$.*

*Proof.* For each $X \in \mathbf{V}$ and $Y \in \mathbf{V} \setminus \{X\}$, the algorithm eventually inserts $Y$ in $\widehat{\mathbf{ADJ}}_X$. The algorithm removes $Y$ from $\widehat{\mathbf{ADJ}}_X$ if and only if there is a subset $\mathbf{S}_{XY}$ of $\widehat{\mathbf{ADJ}}_X$ or $\widehat{\mathbf{ADJ}}_Y$ such that $\mathrm{I}_P(\{X\}, \{Y\}|\mathbf{S}_{XY})$ is concluded. Owing to assumption 2, the algorithm concludes $\mathrm{I}_P(\{X\}, \{Y\}|\mathbf{S}_{XY})$ if and only if $\mathrm{test}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{S}_{XY})}$ is attempted, completed and $\mathrm{I}_P(\{X\}, \{Y\}|\mathbf{S}_{XY})$ holds. The latter is not the case if $Y \in \mathbf{PA}_X$, due to Lemma 2.3. Thus, $\mathbf{PA}_X$ is never removed from $\widehat{\mathbf{ADJ}}_X$. That is, the algorithm discovers all links in $\mathbb{G}$. □

**Lemma 2.7.** *Suppose that the pair $(\mathbb{G}, P)$ of a DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ and a joint probability distribution $P$ of the variables in some set $\mathbf{V}$ is a Bayesian network. Suppose further that a hypothesis test based algorithm instantiating Algorithm Template 2.2 is applied on a sample from $P$ and discovers a link $X - Y$. If*

1. *$\mathbb{G}$ and $P$ are faithful to each other and*
2. *completed tests do not yield a false negative result*

*then the algorithm considers $\mathrm{test}_{I_P(\{X\},\{Y\}|\,\mathbf{PA}_X^0)}$ and $\mathrm{test}_{I_P(\{X\},\{Y\}|\,\mathbf{PA}_Y^0)}$, where:*

- $\mathbf{PA}_X^0 = \mathbf{PA}_X \setminus \{Y\}$ *and* $\mathbf{PA}_Y^0 = \mathbf{PA}_Y$*, if* $Y \in \mathbf{PA}_X$
- $\mathbf{PA}_X^0 = \mathbf{PA}_X$ *and* $\mathbf{PA}_Y^0 = \mathbf{PA}_Y \setminus \{X\}$*, if* $X \in \mathbf{PA}_Y$
- $\mathbf{PA}_X^0 = \mathbf{PA}_X$ *and* $\mathbf{PA}_Y^0 = \mathbf{PA}_Y$*, if* $X$ *and* $Y$ *are not adjacent in* $\mathbb{G}$

*Proof.* The algorithm eventually inserts $Y$ in $\widehat{\mathbf{ADJ}}_X$ and $X$ in $\widehat{\mathbf{ADJ}}_Y$ and considers $\text{test}_{\text{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ for each subset $\mathbf{Z}$ of $\widehat{\mathbf{ADJ}}_X$ and $\widehat{\mathbf{ADJ}}_Y$. $\mathbf{PA}_X^0$ and $\mathbf{PA}_Y^0$ are never removed from $\widehat{\mathbf{ADJ}}_X$ and $\widehat{\mathbf{ADJ}}_Y$ respectively due to Lemma 2.6. Thus, the algorithm eventually considers $\text{test}_{\text{I}_P(\{X\},\{Y\}|\,\mathbf{PA}_X^0)}$ and $\text{test}_{\text{I}_P(\{X\},\{Y\}|\,\mathbf{PA}_Y^0)}$. □

**Lemma 2.8.** *Suppose that the pair* $(\mathbb{G}, P)$ *of a DAG* $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ *and a joint probability distribution* $P$ *of the variables in some set* $\mathbf{V}$ *is a Bayesian network. Suppose further that a hypothesis test based algorithm instantiating Algorithm Template 2.2 is applied on a sample from* $P$ *and discovers a link* $X - Y$ *not in* $\mathbb{G}$*. If*

1. $\mathbb{G}$ *and* $P$ *are faithful to each other,*
2. *all tests attempted by the algorithm are completed and*
3. *all tests completed yield the correct result,*

*then either* $\text{test}_{I_P(\{X\},\{Y\}|\,\mathbf{PA}_X)}$ *or* $\text{test}_{I_P(\{X\},\{Y\}|\,\mathbf{PA}_Y)}$ *is not attempted.*

*Proof.* The algorithm considers $\text{test}_{\text{I}_P(\{X\},\{Y\}|\,\mathbf{PA}_X)}$ and $\text{test}_{\text{I}_P(\{X\},\{Y\}|\,\mathbf{PA}_Y)}$ due to Lemma 2.7 and either $\text{I}_P(\{X\},\{Y\}|\,\mathbf{PA}_X)$ or $\text{I}_P(\{X\},\{Y\}|\,\mathbf{PA}_Y)$ holds due to Corollary 2.1. If both tests were attempted, then they would be completed due to assumption 2, they would yield the correct result due to assumption 3 and $X - Y$ would not be discovered. Therefore, either of the tests is not attempted. □

**Theorem 2.11.** *Suppose a Bayesian network* $(\mathbb{G}, P)$ *and that an algorithm instantiating Algorithm Template 2.2 is applied on a sample from* $P$*. If*

1. $\mathbb{G}$ *and* $P$ *are faithful to each other,*
2. *all tests attempted by the algorithm are completed and*
3. *all tests completed yield the correct result,*

*then the algorithm discovers all links in* $\mathbb{G}$*, as well as links* $X - Y$ *for which either* $\text{test}_{I_P(\{X\},\{Y\}|\,\mathbf{PA}_X)}$ *or* $\text{test}_{I_P(\{X\},\{Y\}|\,\mathbf{PA}_Y)}$ *is not attempted.*

*Proof.* The proof follows immediately from Lemmas 2.6 and 2.8. □

The falsely discovered links in the identified skeleton have an intuitive interpretation if and only if the reliability criterion is intuitive. The most intuitive reliability criterion presented is, undoubtedly, the conditioning cardinality rule. When employing this criterion, a falsely discovered link $X - Y$ implies that either $X$ or $Y$ has more than *max-k* parents.

In the following lemma and theorem, also presented for the first time here, we assume that tests completed do not yield a false negative result:

**Lemma 2.9.** *Suppose that the pair $(\mathbb{G}, P)$ of a DAG $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ and a joint probability distribution $P$ of the variables in some set $\mathbf{V}$ is a Bayesian network. Suppose further that a hypothesis test based algorithm instantiating Algorithm Template 2.2 is applied on a sample from $P$ and discovers a link $X - Y$ not in $\mathbb{G}$. If*

1. *$\mathbb{G}$ and $P$ are faithful to each other,*
2. *all tests attempted by the algorithm are completed and*
3. *completed tests do not yield a false negative result,*

*then either $test_{I_P(\{X\},\{Y\}|\,\mathbf{PA}_X)}$ or $test_{I_P(\{X\},\{Y\}|\,\mathbf{PA}_Y)}$ is either not attempted or yields the incorrect result.*

*Proof.* The algorithm considers $\text{test}_{\text{I}_P(\{X\},\{Y\}|\,\mathbf{PA}_X)}$ and $\text{test}_{\text{I}_P(\{X\},\{Y\}|\,\mathbf{PA}_Y)}$ due to Lemma 2.7. If both tests were attempted, completed and yielded the correct result, then $X - Y$ would not be discovered; this is because either $\text{I}_P(\{X\},\{Y\}|\,\mathbf{PA}_X)$ or $\text{I}_P(\{X\},\{Y\}|\,\mathbf{PA}_Y)$ holds due to Corollary 2.1 and the algorithm does not discover $X - Y$ if there is a subset $\mathbf{S}_{XY}$ of $\widehat{\mathbf{ADJ}}_X$ or $\widehat{\mathbf{ADJ}}_Y$ such that $\text{I}_P(\{X\},\{Y\}|\mathbf{S}_{XY})$ is concluded. Therefore, either of the tests is either not attempted or yielded the incorrect result; the possibility that it is attempted but not completed is ruled out due to assumption 2. □

**Theorem 2.12.** *Suppose a Bayesian network $(\mathbb{G}, P)$ and that an algorithm instantiating Algorithm Template 2.2 is applied on a sample from $P$. If*

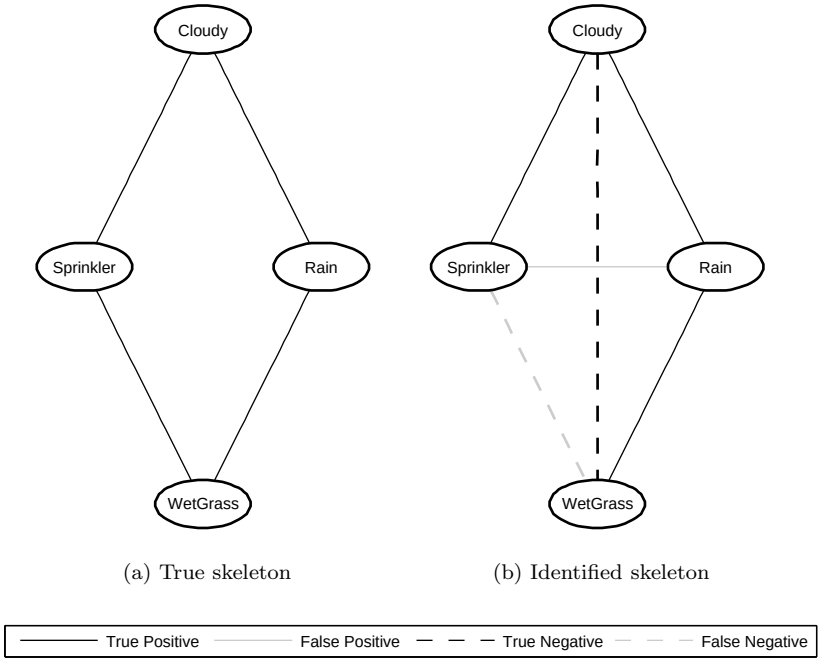
1. *$\mathbb{G}$ and $P$ are faithful to each other,*
2. *all tests attempted by the algorithm are completed and*
3. *completed tests do not yield a false negative result,*

*then the algorithm discovers all links in $\mathbb{G}$, as well as links $X - Y$ for which either $test_{I_P(\{X\},\{Y\}|\,\mathbf{PA}_X)}$ or $test_{I_P(\{X\},\{Y\}|\,\mathbf{PA}_Y)}$ is either not attempted or yields the incorrect result.*

*Proof.* The proof follows immediately from Lemmas 2.6 and 2.9. □

Finally, let us we give an example of skeleton identification:

**Example 2.5.** *We applied the skeleton identification part of MMHC with $\alpha = 0.05$ on a sample of size 100 from the Bayesian network of Example 2.1 (Fig. 2.1). The skeleton of the network is shown in Fig. 2.2a. The identified skeleton is shown in 2.2b. 3/4 = 75% of the true links are correctly identified, while 1/4 = 25% of the identified links are false.*




(a) True skeleton (b) Identified skeleton

Figure 2.2: Example of skeleton identification.

# Chapter 3

# Estimation and control of the False Discovery Rate in skeleton identification

In addition to identifying the skeleton of a Bayesian network, we would like to assess *confidence* on the identified skeleton. To this end, we view skeleton identification as *multiple hypothesis testing*. In multiple hypothesis testing, a reasonable error rate to utilize is the *False Discovery Rate* (FDR). We introduce a unified approach to estimation and control of the FDR in skeleton identification. Then we present an experimental evaluation of the performance of the most common FDR estimator in both tasks. Finally, we discuss other approaches to assessing confidence in structure learning and their relationship to FDR in skeleton identification.

## 3.1 Multiple hypothesis testing

*Multiple hypothesis testing* refers to testing several hypotheses simultaneously [34, 38]. Suppose that $m$ hypotheses are being tested with corresponding p-values $p_1, p_2, \ldots, p_m$. Typically, a p-value threshold $t$ is chosen and hypotheses with corresponding p-value $\leq t$ are rejected [38]. Table 3.1 summarizes the outcomes when applying this approach. The threshold $t$ can be either fixed beforehand or selected by a procedure that controls some error rate while maximizing power.

In single hypothesis testing, FPR is controlled at some level $\alpha$ while maximizing power. In multiple hypothesis testing, however, there are many possible error rates to control. Table 3.2 lists the most common ones. Note that, while in single hypothesis testing the FPR, the FNR and power are the *probability* of a false positive, a false negative and a true positive, respectively, in multiple hypothesis testing they are the *expected proportion* of false positives, false negatives and true positives, respectively. It is not hard to see that, when the p-values are independent, the probability of a single result being a true positive or a false negative is equal to the expected proportion of false positives and false negatives, respectively.

The traditional approach to multiple hypothesis testing is to control the *Familywise Error Rate* (FWER), which is defined as the probability of making *at least one* Type I error:

$$\text{FWER} \triangleq \Pr(V > 0)$$

where $V$ is the number of rejected true null hypotheses. FWER is controlled at level $t$ by rejecting hypotheses with p-value $\leq t/m$; this is known as the *Bonferroni correction*. FWER is useful when it is of interest to prevent *any* single false positive from occurring. However, controlling the FWER at level $t$ results in a substantial loss of power compared to when controlling the FPR at the same level [6]. It is usually of greater interest to reject as many hypotheses

Table 3.1: Outcomes from $m$ hypotheses tests when rejecting all hypotheses with corresponding p-value $\leq t$

| | Null accepted | Null rejected | Total |
|---|---|---|---|
| Null is true | $U$ | $V$ | $m_0$ |
| Alternative is true | $T$ | $S$ | $m_1$ |
| | W | R | m |

Table 3.2: Common multiple hypothesis testing error rates

| Name | Abbreviation | Definition |
|---|---|---|
| False Positive Rate | FPR | $\mathrm{E}[V/m0]$ |
| False Negative Rate | FNR | $\mathrm{E}[T/m1]$ |
| Familywise Error Rate | FWER | $\Pr(V > 0)$ |
| False Discovery Rate | FDR | $\mathrm{E}[V/R\|R > 0] \Pr(R > 0)$ |
| Positive False Discovery Rate | pFDR | $\mathrm{E}[V/R\|R > 0]$ |

as possible while keeping the proportion of false positives among the rejected hypotheses low. The *False Discovery Rate* (FDR) is an error rate designed to this end [6, 39, 38].

## 3.2 False Discovery Rate

False Discovery Rate is a multiple hypothesis testing error rate introduced by Benjamini and Hochberg [6], loosely defined as the expected proportion of false positives among the rejected hypotheses ("discoveries") and useful when one is interested having mostly true positives among our discoveries. Its precise definition is

$$\mathrm{FDR} \triangleq \mathrm{E}\left[\frac{V}{R \vee 1}\right] = \mathrm{E}\left[\frac{V}{R}\middle| R > 0\right] \Pr(R > 0)$$

where $R$ is the number of rejections and $R \vee 1$ corresponds to setting $V/R$ to 0 when $R = 0$. There are two approaches to utilizing FDR, namely *control* and *estimation*.

### 3.2.1 Control of the False Discovery Rate

The FDR control approach is to set an FDR threshold (level) $q$ and find a p-value threshold $t$ such that *strong control*[1] of the FDR under $q$ is achieved, i.e., $\mathrm{FDR}(t) \leq q$, where $\mathrm{FDR}(t)$ is the FDR when rejecting all hypotheses with corresponding p-value $\leq t$. Procedure 3.1, referred to as the *Benjamini-Hochberg* (BH) procedure, is proven to achieve strong control, assuming independent p-values [6] or what is called *positive regression dependence* of the p-values on each of the null p-values [7]:

**Procedure 3.1** Benjamini-Hochberg (BH) procedure

Let $\mathrm{p}_{(1)} \leq \mathrm{p}_{(2)} \cdots \leq \mathrm{p}_{(m)}$ be the ordered p-values
Let $k = \arg\max_i \{\mathrm{p}_{(i)} \leq \frac{i}{m} q\}$
Reject hypotheses corresponding to $\mathrm{p}_{(i)} : i = 1 \ldots k$ if $k$ exists, otherwise none

[1] In the statistical literature, the term "strong control" refers to the case that an error rate is controlled under any configuration of the null hypotheses, while the term "weak control" refers to the case that an error rate is controlled when all null hypothesis are true.

**Example 3.1.** *Consider the following 15 ordered p-values [6]:*

$$0.0001, 0.0004, 0.0019, 0.0095, 0.0201, 0.0278, 0.0298, 0.0344, 0.0459, 0.324$$

$$0.4262, 0.5719, 0.6528, 0.759, 1.000$$

*Controlling the FPR at level* 0.05 *results in rejecting the hypotheses corresponding to the first 9 p-values. Controlling the FWER at the same level with the Bonferroni correction results in rejecting the hypotheses corresponding to the first 3 p-values, because* $0.05/15 = 0.0033$*. Controlling the FDR at the same level with the BH procedure results in rejecting the hypotheses corresponding to the first 4 p-values, because* $p_{(4)} = 0.0095$ *is the first p-value, from the end, for which* $p_{(i)} \leq (i/15) \cdot 0.05$ *holds:*

$$p_{(4)} = 0.0095 \leq \frac{i}{15} 0.05 = 0.013$$

### 3.2.2 Estimation of the False Discovery Rate

The FDR estimation approach is to set a p-value threshold $t$ and estimate $\mathrm{FDR}(t)$ in a conservative manner, i.e., $\mathrm{E}[\widehat{\mathrm{FDR}}(t)] \geq \mathrm{FDR}(t)$. $\mathrm{E}[\widehat{\mathrm{FDR}}(t)] - \mathrm{FDR}(t)]$ is called the *bias* of $\widehat{\mathrm{FDR}}(t)$. $\widehat{\mathrm{FDR}}(t)$ is conservative if and only if its bias is non-negative. The estimation approach is introduced by Storey [39], along with a family of estimators proven to be conservative when the p-values are independent:[2]

$$\widehat{\mathrm{FDR}}_{\mathrm{BH}}(t) \triangleq \frac{m \cdot t}{\mathrm{R}(t) \vee 1}$$

FDR estimators can be used to define valid FDR controlling procedures [41]. It is not hard to see that taking the largest $t$ such that $\widehat{\mathrm{FDR}}_{\mathrm{BH}}(t) \leq q$ corresponds to applying the BH procedure [39]; this is why we subscripted the previous estimator with BH.

Let $\widehat{\mathrm{FDR}}$ denote a general FDR estimator. Throughout this thesis, we refer to the act of taking the largest $t$ such that $\widehat{\mathrm{FDR}}_0(t) \leq q_0$ as the *FDR controlling procedure* with $\widehat{\mathrm{FDR}}_0$ and $q = q_0$. We refer to the FDR controlling procedure with $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ simply as the BH procedure. Following the practice of Storey [39], we denote the p-value threshold returned by the FDR controlling procedure with $\mathrm{t_q}(\widehat{\mathrm{FDR}})$.

It is not hard to see that the FDR controlling procedure strongly controls the FDR at $q$ if and only if $\widehat{\mathrm{FDR}}[\mathrm{t_q}(\widehat{\mathrm{FDR}})]$ is conservative:

$$\mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}})] \leq q \iff \mathrm{E}\{\widehat{\mathrm{FDR}}[\mathrm{t_q}(\widehat{\mathrm{FDR}})]\} \geq \mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}})]$$

Throughout this thesis, we refer to $q - \mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}})]$ as the bias of the FDR controlling procedure. The FDR controlling procedure strongly controls the FDR if and only if its bias is non-negative.

[2] Storey's [39] estimators also include a $\hat{\pi}_0(\lambda)$ term, an estimator of the proportion of true null hypotheses. Because $\hat{\pi}_0(\lambda)$ is not applicable in this context, we use $\hat{\pi}_0(\lambda) = 1$ instead.

### 3.2.3 The q-value

As mentioned in Section 2.2.1, researchers usually do not simply report the acceptance or rejection of a hypothesis according to a predetermined significance level (FPR threshold) $\alpha$ but report the p-value corresponding to the hypothesis instead. Recall that the p-value is the smallest significance level such that the hypothesis would still be rejected [9].

In multiple hypothesis testing with FDR control, the analogue of the p-value is the *q-value*, introduced by Storey [39]. The q-value corresponding to a hypothesis is the smallest FDR threshold such that the hypothesis would still be rejected. When rejecting hypotheses with corresponding p-value $\leq t$ for some threshold $t$, the q-value q-value$(p)$ corresponding to p-value $p$ is $\min_{t \geq p} \mathrm{FDR}(t)$ and estimated by $\hat{\text{q}}\text{-value}(p) = \min_{t \geq p} \widehat{\mathrm{FDR}}(t)$. It is easy to see that applying the FDR controlling procedure with FDR threshold $q$ is equivalent to rejecting hypotheses with corresponding $\hat{\text{q}}$-value $\leq q$.

## 3.3 Utilizing the False Discovery Rate in skeleton identification

### 3.3.1 Skeleton identification as multiple hypothesis testing

In order to utilize FDR, skeleton identification is viewed as a multiple hypothesis testing, the null hypotheses being the absence of links. To perform the test $\mathrm{test}_{\neg\,\mathrm{Adj}(X,Y)}$ of the hypothesis $\mathrm{H}_{\neg\,\mathrm{Adj}(X,Y)}$ of absence of a link between nodes $X$ and $Y$, a skeleton identification algorithm completes the tests $\mathrm{test}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ of the hypotheses $\mathrm{H}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ and obtain test statistics $\mathrm{t}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ and p-values $\mathrm{p}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ for some set $\mathbf{C}_{XY}$ of subsets $\mathbf{Z}$ of $\mathbf{V}$. When we reject $\mathrm{H}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ when $\mathrm{t}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})} \geq c$ for some threshold $c$, $\mathrm{p}_{\neg\,\mathrm{Adj}(X,Y)}$ is

$$\mathrm{p}_{\neg\,\mathrm{Adj}(X,Y)} = \Pr\left(\bigcap_{\mathbf{Z}\in\mathbf{C}_{XY}} \mathrm{T}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})} \geq \mathrm{t}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})} \,\middle|\, \neg\,\mathrm{Adj}(X,Y)\right)$$

Unfortunately, $\mathrm{p}_{\neg\,\mathrm{Adj}(X,Y)}$ is unavailable. However, the p-values $\mathrm{p}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ for $\mathbf{Z} \in \mathbf{C}_{XY}$ can be used to upper-bound $\mathrm{p}_{\neg\,\mathrm{Adj}(X,Y)}$ thanks to the following lemmas and theorems.

The first lemma is found in related work by Tsamardinos and Brown [45] and concerns the upper-bounding of the p-value corresponding to a hypothesis related to $\mathrm{H}_{\neg\,\mathrm{Adj}(X,Y)}$:[3]

**Lemma 3.1.** *Let $H_{\exists \mathbf{S}_{XY}\in\mathbf{C}\ s.t.\ I_P(\{X\},\{Y\}|\mathbf{S}_{XY})}$ be the hypothesis that there is a set $\mathbf{S}_{XY}$ in a set of sets $\mathbf{C}$ such that $I_P(\{X\},\{Y\}|\mathbf{S}_{XY})$. The corresponding p-value $p_{\exists \mathbf{S}_{XY}\in\mathbf{C}\ s.t.\ I_P(\{X\},\{Y\}|\mathbf{S}_{XY})}$ is upper-bounded by the maximal among the p-values $p_{I_P(\{X\},\{Y\}|\mathbf{Z})}$ for $\mathbf{Z} \in \mathbf{C}$:*

$$p_{\exists \mathbf{S}_{XY}\in\mathbf{C}\ s.t.\ I_P(\{X\},\{Y\}|\mathbf{S}_{XY})} \leq \max_{\mathbf{Z}\in\mathbf{C}} p_{I_P(\{X\},\{Y\}|\mathbf{Z})}$$

[3] For clarity, a slightly different version is stated here. The proof is the same.

*Proof.* Let $A_0$, $A_0^i$ and $A_1^i$ denote the event that $\mathrm{H}_{\exists \mathbf{S}_{XY} \in \mathbf{C} \text{ s.t. } \mathrm{I}_P(\{X\},\{Y\}|\mathbf{S}_{XY})}$, $\mathrm{H}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ and $\mathrm{H}_{\neg\, \mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ is true, respectively. When $A_0$ occurs, at least one $A_0^i$ occurs. Without loss of generality, assume that for $i \leq k$, $A_0^i$ occurs when $A_0$ occurs. Let further $p = \mathrm{p}_{\exists \mathbf{S}_{XY} \in \mathbf{C} \text{ s.t. } \mathrm{I}_P(\{X\},\{Y\}|\mathbf{S}_{XY})}$, $T_i = \mathrm{T}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$, $t_i = \mathrm{t}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$, $p_i = \mathrm{p}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$.

$$
\begin{aligned}
p_S &= \Pr\left(\bigcap_i T_i \geq t_i \middle| A_0\right) = \prod_i \Pr\left(T_i \geq t_i \middle| \left\{\bigcap_{j<i} T_j \geq t_j\right\} \cap A_0\right) \\
&= \prod_i \Big[ \Pr\left(T_i \geq t_i \middle| \left\{\bigcap_{j<i} T_j \geq t_j\right\} \cap A_0 \cap A_0^i\right) + \\
&\qquad \Pr\left(T_i \geq t_i \middle| \left\{\bigcap_{j<i} T_j \geq t_j\right\} \cap A_0 \cap A_1^i\right) \Big] \\
&\leq \prod_{i \leq k} \Pr\left(T_i \geq t_i \middle| \left\{\bigcap_{j<i} T_j \geq t_j\right\} \cap A_0^i\right) \leq \Pr\left(T_1 \geq t_1 \middle| A_0^1\right) = p_1 \leq \max_i p_i
\end{aligned}
$$

The second equality is due to the chain rule for random variables. The third equality is due to the law of total probability. The first inequality is due to $A_0^i \subseteq A_0$ for $i \leq k$. □

The next four lemmas and theorems appear for the first time in this work and establish sufficient criteria for hypothesis test based skeleton identification algorithms to upper-bound $\mathrm{p}_{\neg\, \mathrm{Adj}(X,Y)}$. The following theorem concerns all hypothesis test based skeleton identification algorithms:

**Theorem 3.1.** *Suppose a Bayesian network $(\mathbb{G}, P)$ and that a hypothesis test based skeleton identification algorithm is applied on a sample from $P$. If*

1. *$\mathbb{G}$ and $P$ are faithful to each other and*
2. *the test $test_{I_P(\{X\},\{Y\}|\mathbf{Z})}$ of a hypothesis $H_{I_P(\{X\},\{Y\}|\mathbf{Z})}$ that would be true if $H_{\neg\, Adj(X,Y)}$ was true is completed*

*then $p_{\neg\, Adj(X,Y)}$ is upper-bounded by the maximal among $p_{I_P(\{X\},\{Y\}|\mathbf{Z})}$ for $\mathbf{Z} \in \mathbf{C}_{XY}$:*

$$p_{\neg\, Adj(X,Y)} \leq \max_{\mathbf{Z} \in \mathbf{C}_{XY}} p_{I_P(\{X\},\{Y\}|\mathbf{Z})}$$

*Proof.* Owing to assumption 2 and Lemma 2.3, $\mathrm{H}_{\neg\, \mathrm{Adj}(X,Y)}$ is equivalent to $\mathrm{H}_{\exists \mathbf{S}_{XY} \in \mathbf{C}_{XY} \text{ s.t. } \mathrm{I}_P(\{X\},\{Y\}|\mathbf{S}_{XY})}$. Therefore:

$$\mathrm{p}_{\neg\, \mathrm{Adj}(X,Y)} = \mathrm{p}_{\nexists \mathbf{S}_{XY} \in \mathbf{C}_{XY} \text{ s.t. } \mathrm{I}_P(\{X\},\{Y\}|\mathbf{S}_{XY})} \leq \max_{\mathbf{Z} \in \mathbf{C}_{\mathbf{XY}}} \mathrm{p}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$$

The last inequality is due to Lemma 3.1. □

The following two lemmas and one theorem concern algorithms instantiating Algorithm Template 2.2:

**Lemma 3.2.** *Suppose a Bayesian network $(\mathbb{G}, P)$ and that an algorithm instantiating Algorithm Template 2.2 is applied on a sample from $P$ and discovers a link $X - Y$. If*

1. $\mathbb{G}$ *and* $P$ *are faithful to each other,*

2. $test_{I_P(\{X\},\{Y\}|\,\mathbf{PA}^0_X)}$ *and* $test_{I_P(\{X\},\{Y\}|\,\mathbf{PA}^0_Y)}$*, where*

   - $\mathbf{PA}^0_X = \mathbf{PA}_X \setminus \{Y\}$ *and* $\mathbf{PA}^0_Y = \mathbf{PA}_Y$*, if* $Y \in \mathbf{PA}_X$
   - $\mathbf{PA}^0_X = \mathbf{PA}_X$ *and* $\mathbf{PA}^0_Y = \mathbf{PA}_Y \setminus \{X\}$*, if* $X \in \mathbf{PA}_Y$
   - $\mathbf{PA}^0_X = \mathbf{PA}_X$ *and* $\mathbf{PA}^0_Y = \mathbf{PA}_Y$*, if* $X$ *and* $Y$ *are not adjacent in* $\mathbb{G}$

   *are reliable according to the employed reliability criterion,*

3. *all tests attempted by the algorithm are completed and*

4. *completed tests do not yield a false negative result,*

*then* $p_{\neg\, Adj(X,Y)}$ *is upper-bounded by the maximal among* $p_{I_P(\{X\},\{Y\}|\mathbf{Z})}$ *for* $\mathbf{Z} \in \mathbf{C}_{XY}$*:*

$$p_{\neg\, Adj(X,Y)} \leq \max_{\mathbf{Z}\in\mathbf{C}_{XY}} p_{I_P(\{X\},\{Y\}|\mathbf{Z})}$$

*Proof.* The algorithm considers $\text{test}_{\text{I}_P(\{X\},\{Y\}|\,\mathbf{PA}^0_X)}$ and $\text{test}_{\text{I}_P(\{X\},\{Y\}|\,\mathbf{PA}^0_Y)}$ due to Lemma 2.8. Both tests are attempted due to assumption 1 and completed due to assumption 2. Owing to Corollary 2.1, either $\text{I}_P(\{X\},\{Y\}|\,\mathbf{PA}^0_X)$ or $\text{I}_P(\{X\},\{Y\}|\,\mathbf{PA}^0_Y)$ would hold if $X$ and $Y$ were not adjacent in $\mathbb{G}$. Therefore, a test $\text{test}_{\text{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ of a hypothesis $\text{H}_{\text{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ that would be true if $\text{H}_{\neg\,\text{Adj}(X,Y)}$ was true is completed and the proof concludes due to Theorem 3.1. □

**Lemma 3.3.** *Suppose a Bayesian network* $(\mathbb{G}, P)$ *and that an algorithm instantiating Algorithm Template 2.2 is applied on a sample from* $P$ *and does not discover a link* $X - Y$*. If*

1. $\mathbb{G}$ *and* $P$ *are faithful to each other*

2. *completed tests do not yield a false negative result,*

*then* $p_{\neg\, Adj(X,Y)}$ *is upper-bounded by the maximal among* $p_{I_P(\{X\},\{Y\}|\mathbf{Z})}$ *for* $\mathbf{Z} \in \mathbf{C}_{XY}$*:*

$$p_{\neg\, Adj(X,Y)} \leq \max_{\mathbf{Z}\in\mathbf{C}_{XY}} p_{I_P(\{X\},\{Y\}|\mathbf{Z})}$$

*Proof.* Owing to Lemma 2.6, $X - Y$ is not in $\mathbb{G}$, otherwise the algorithm would discover it. The algorithm does not discover $X - Y$ if and only if there is a subset $\mathbf{S}_{XY}$ of $\widehat{\mathbf{ADJ}}_X$ or $\widehat{\mathbf{ADJ}}_Y$ for which $\text{I}_P(\{X\},\{Y\}|\mathbf{S}_{XY})$ is concluded. Owing to assumption 2, the algorithm concludes $\text{I}_P(\{X\},\{Y\}|\mathbf{Z})$ if and only if $\text{test}_{\text{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ is attempted, completed and $\text{I}_P(\{X\},\{Y\}|\mathbf{Z})$ holds. Therefore, a test $\text{test}_{\text{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ of a hypothesis $\text{H}_{\text{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ that would be true if $\text{H}_{\neg\,\text{Adj}(X,Y)}$ was true is completed and the proof concludes due to Theorem 3.1. □

**Theorem 3.2.** *Suppose that the pair* $(\mathbb{G}, P)$ *of a DAG* $\mathbb{G} = (\mathbf{V}, \mathbf{E})$ *and a joint probability distribution* $P$ *of the variables in some set* $\mathbf{V}$ *is a Bayesian network and* $X$ *and* $Y$ *are distinct nodes in* $\mathbf{V}$*. Suppose further that a hypothesis test based algorithm instantiating Algorithm Template 2.2 is applied on a sample from* $P$*. If*

1. *$\mathbb{G}$ and $P$ are faithful to each other*

2. *all tests considered by the algorithm are reliable according to the employed reliability criterion,*

3. *all tests attempted by the algorithm are completed and*

4. *completed tests do not yield a false negative result,*

*then $p_{\neg\, Adj(X,Y)}$ is upper-bounded by the maximal among $p_{I_P(\{X\},\{Y\}|\mathbf{Z})}$ for $\mathbf{Z} \in \mathbf{C}_{XY}$:*

$$p_{\neg\, Adj(X,Y)} \leq \max_{\mathbf{Z}\in\mathbf{C}_{XY}} p_{I_P(\{X\},\{Y\}|\mathbf{Z})}$$

*Proof.* The proof follows immediately from Lemmas 3.2 and 3.3. □

Skeleton identification can be viewed as a multiple testing procedure that rejects link absence hypotheses with corresponding p-value $\leq \alpha$: $\mathrm{H}_{\neg\,\mathrm{Adj}(X,Y)}$ is accepted once $\mathrm{p}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})} > \alpha$ for some $\mathbf{Z} \in \mathbf{V} \setminus \{X, Y\}$, or equivalently, if $\max_{\mathbf{Z}\in\mathbf{C}_{XY}} \mathrm{p}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})} > \alpha$ [22]. Assuming independent link absence p-values, skeleton identification controls the FPR at level $\alpha$.

### 3.3.2 Related work

Tsamardinos and Brown [45] follow the *estimation* approach in *local* Bayesian network learning. Local learning is concerned with learning the set $\mathbf{PC}_X$ of parents and children of a target node $X$ or, in other words, the neighbors of $X$. In order to estimate FDR, local learning is viewed as multiple hypothesis testing, the null hypotheses being the absence of nodes from $\mathbf{PC}_X$. Since hypotheses with corresponding p-value $\leq \alpha$ are rejected, the FDR of local learning is FDR($\alpha$). Tsamardinos and Brown [45] apply local learning on samples of size $n \in \{1000, 5000, 10000\}$ of five networks from real decision support systems, targeting all or some of the nodes of each network. Then they estimate FDR($\alpha$) with $\widehat{\mathrm{FDR}}_{\mathrm{BH}}(\alpha)$. For $n \in \{5000, 10000\}$, FDR($\alpha$) is conservatively estimated, in general. However, for $n = 1000$, there are many cases where this is not the case. The situation is improved using a relaxed definition of false discovery (see Chapter 5).

Li and Wang [22] follow the *control* approach in skeleton identification (*global* Bayesian network learning). They modify the skeleton identification phase of the PC algorithm [35] and end up with $\mathrm{PC}_{\mathrm{FDR}}$–skeleton, a skeleton identification algorithm with *embedded* FDR control. $\mathrm{PC}_{\mathrm{FDR}}$–skeleton is proven to strongly control the FDR at a given level $q$ under assumptions similar to those of Theorem 3.2. $\mathrm{PC}_{\mathrm{FDR}}$–skeleton does not accept a conditional independence hypothesis (and subsequently, a link absence hypothesis) when its p-value is above some significance level $\alpha$ but instead applies the BH procedure with $q$ to the *up-to-date* maximal conditional independence p-values after an up-to-date maximal conditional independence p-value is updated. Given that, if an FDR control procedure strongly controls FDR given some p-values, it continues to do so given upper bounds on those p-values, $\mathrm{PC}_{\mathrm{FDR}}$–skeleton strongly controls the FDR when it terminates. Li and Wang [22] apply $\mathrm{PC}_{\mathrm{FDR}}$–skeleton with $q = 0.05$ to samples of size $n = 500$ of 48 randomly generated networks with number of nodes $|\mathbf{V}| \in \{15, 20, 25, 30\}$ and varying characteristics. $\mathrm{PC}_{\mathrm{FDR}}$–skeleton achieves an

FDR noticeably lower than $q$, while a heuristic modification of the algorithm achieves an FDR around $q$. However, this modification is not theoretically proven to strongly control the FDR.

Schäfer and Strimmer [33] introduce a framework for learning *Graphical Gaussian Models* (GMMs) from small samples. In contrast to Bayesian networks, GMMs are undirected graphical models. A GMM consists of an undirected graph and a partial correlation matrix for some multivariate normal variables. An edge in the graph corresponds to a non-zero partial correlation coefficient for the ends of the edge given the rest nodes. GMM learning is viewed as multiple hypothesis testing, the null hypotheses being the absence of edges. Once the p-values corresponding to the hypotheses are calculated, the BH procedure is applied to control the FDR at level $q$. Schäfer and Strimmer [33] apply their framework with $q = 0.05$ to samples of size $n \in \{10, 20, \ldots, 210\}$ of randomly generated networks with 100 variables and report the FPR, power and the FDR. Although the FPR is close to 0 for all $n$ and power reaches 0.6 for $n = 210$, the FDR is around $q$ only after about $n = 100$.

### 3.3.3 A unified approach to estimation and control of the False Discovery Rate in skeleton identification

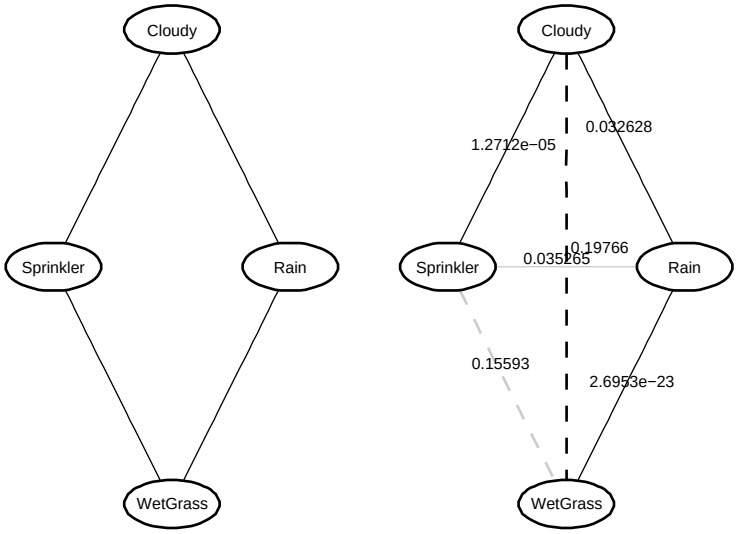
In this work, we

1. adapt the FDR estimation approach of Tsamardinos and Brown [45] to skeleton identification,
2. extend it to any p-value threshold $t \leq \alpha$ and
3. unify it with FDR control at any FDR threshold $q$.

Our unified approach to estimation and control of the FDR in skeleton identification is as follows: first apply skeleton identification with significance level $\alpha$ and then either estimate the FDR at a p-value threshold $t \leq \alpha$ or control the FDR at an FDR threshold $q$.

There is no point in estimating FDR($t$) for $t > \alpha$, since skeleton identification, viewed as a multiple testing procedure, has already accepted link absence hypotheses with corresponding p-value $> \alpha$. For the same reason, when applying the FDR controlling procedure, there is no point in retaining links with p-value in the interval $(\alpha\ \mathrm{t_q}(\widehat{\mathrm{FDR}})]$ if $\mathrm{t_q}(\widehat{\mathrm{FDR}}) > \alpha$. However, as demonstrated in the next section, $\mathrm{t_q}(\widehat{\mathrm{FDR}}) \ll \alpha$ for the values of $q \in [10^{-3}, 10^{-1}]$ we tried.

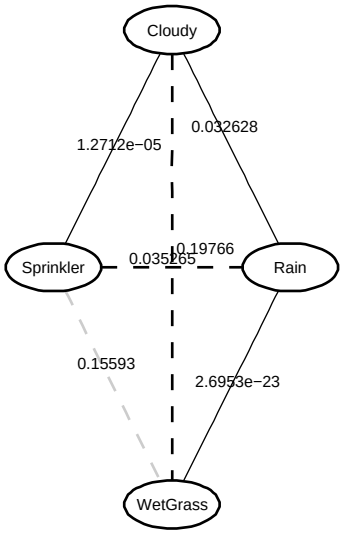
**Example 3.2.** *We applied the skeleton identification part of MMHC with* $\alpha = 0.05$ *on a sample of size 100 of the Bayesian network of Example 2.1 (Fig. 2.1). The skeleton of the network is shown in Fig. 3.1a. The identified skeleton is shown in 3.1b. Table 3.3 lists the values of FDR-related quantities corresponding to each pair of nodes in the identified skeleton, in ascending order of p-values. By applying a p-value threshold* $t = 0.035$ *we get the skeleton of Fig. 3.1c and realized FDR*$(t) = 0$*. We estimate FDR*$(0.035)$ *by* $\widehat{FDR}_{BH}(0.035) = 6{\cdot}0.035/3 = 0.07$*. By applying the BH procedure with* $q = 0.05$ *we get the skeleton of Fig. 3.1d and realized FDR 0.*



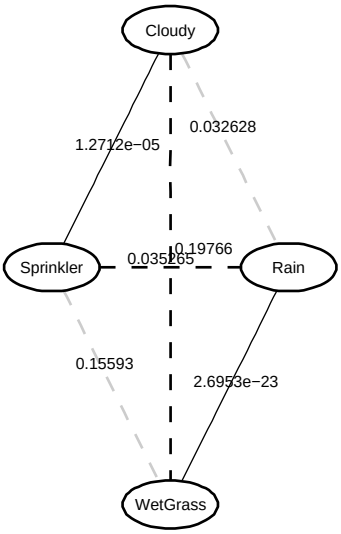

(a) True skeleton




(b) Identified skeleton




(c) Identified skeleton after applying a p-value threshold $t = 0.035$




(d) Identified skeleton after applying the Benjamini-Hochberg FDR controlling procedure with FDR threshold $q = 0.05$

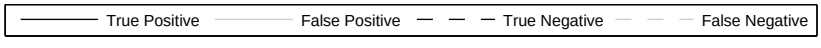



Figure 3.1: Example of the unified approach to estimation and control of the False Discovery Rate (FDR) in skeleton identification. The number by each line is the p-value corresponding to the pair of nodes the line connects.

Table 3.3: Values of False Discovery Rate (FDR) - related quantities corresponding to each pair of nodes (link absence hypothesis) in the identified skeleton of Example 3.2, in ascending order of p-values. Only values corresponding to p-values $< \alpha = 0.05$ are listed. $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ is the value of Storey's [39] FDR estimator $\widehat{\mathrm{FDR}}_{\mathrm{BH}}(t)$ at p-value threshold $t$ equal to the corresponding p-value. The realized FDR corresponding to a p-value is the realized FDR when rejecting link absence hypotheses with smaller or equal corresponding p-value.

| Pair of nodes | p-value ↓ | $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ | realized FDR |
|---|---|---|---|
| $(Rain, WetGrass)$ | 2.695e-23 | 1.6172e-22 | 0 |
| $(Cloudy, Sprinkler)$ | 1.271e-05 | 3.8137e-05 | 0 |
| $(Cloudy, Rain)$ | 0.03263 | 0.065256 | 0 |
| $(Sprinkler, Rain)$ | 0.03527 | 0.052898 | 0.25 |
| $(Sprinkler, WetGrass)$ | 0.1559 | | |
| $(Cloudy, WetGrass)$ | 0.1977 | | |

Let us compare our approach to Li and Wang's [22] $\mathrm{PC}_{\mathrm{FDR}}$–skeleton. $\mathrm{PC}_{\mathrm{FDR}}$–skeleton is a *single-stage* FDR controlling skeleton identification algorithm. Our approach allows for both estimation and control of the FDR in skeleton identification. FDR control using our approach is a *two-stage* process that allows for FDR control at any level $q$ after skeleton identification takes place, while $q$ must be fixed in advance with $\mathrm{PC}_{\mathrm{FDR}}$–skeleton. On the other hand, $\mathrm{PC}_{\mathrm{FDR}}$–skeleton does not require an FPR threshold $\alpha$ to be specified.

Finally, our approach concerns Bayesian networks while the approach of Schäfer and Strimmer [33] concerns GMMs. Bayesian networks are richer models than GMMs, although the latter impose no acyclicity constraint to the graph like the former do.

## 3.4 Experimental results

We evaluated $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ in the tasks of estimation and control of the FDR in skeleton identification using five Bayesian networks obtained from online repositories (see Appendix), as well as flow cytometry measurements of the proteins and phospholipids in a known protein-signaling network [32].

Regarding estimation, we would like to check if $\widehat{\mathrm{FDR}}(t)$ is conservative for each p-value threshold $t$, network and sample size. Because the expectations $\mathrm{E}[\widehat{\mathrm{FDR}}(t)]$ and $\mathrm{FDR}(t) = \mathrm{E}\{\mathrm{V}(t)/[\mathrm{R}(t) \vee 1]\}$ are unavailable, we estimate them by the respective means over 100 random samples of the network. We do the same for $\mathrm{FPR}(t)$ and $\pi(t)$, the False Positive Rate and power of $t$ respectively.

Regarding control, we would like to check if the BH procedure achieves strong control at each FDR threshold $q$ for each network and sample size. Again, we estimate $\mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BH}})] = \mathrm{E}\{\mathrm{V}(\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BH}}))/[\mathrm{R}(\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BH}})) \vee 1]\}$ and $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BH}})]$ by the respective means over the 100 samplings.

We generated 100 random samples of size 10000 from the *Alarm*, *Barley*, *Hailfinder*, *Hepar II* and *Insurance* categorical Bayesian networks obtained from online repositories. The networks are assumed to be faithful.

Let us now describe the real biological data used. Sachs et. al. [32] use a technique called *flow cytometry* to obtain 5400 observational and interventional

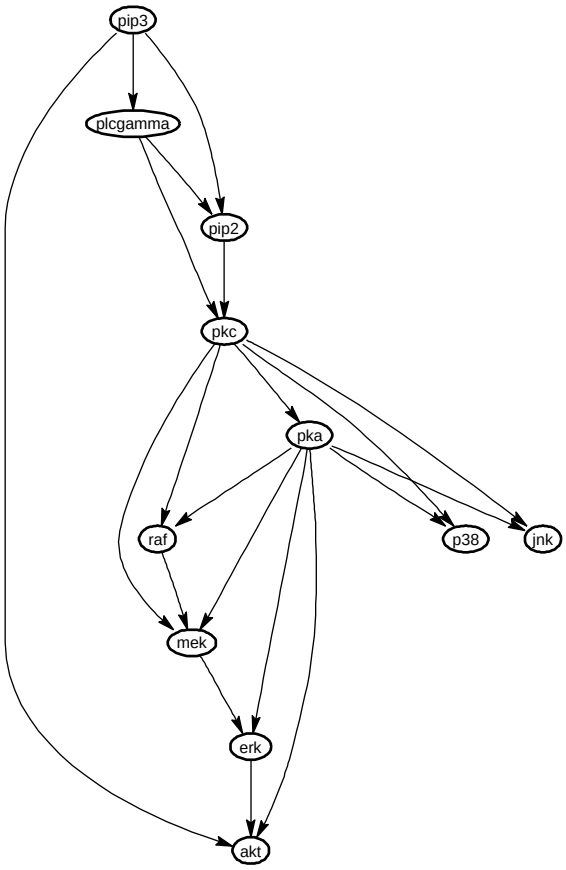



Figure 3.2: The presumed protein-signaling network involving the 11 proteins and phospholipids measured by Sachs et. al. [32].

measurements of 11 proteins and phospholipids. They discretize the data into three levels and learn a Bayesian network structure using Bayesian model averaging. The learned structure is compared to the presumed protein-signaling network involving these molecules (Fig. 3.2), as described in the biological literature. The true network contains 20 edges, while the learned structure contains 17 true links (among of which 16 are correctly oriented) and misses 3 links; no links are falsely discovered. That is, the *realized* power is $17/20 = 0.85$ and the *realized* FDR is 0.

We generated 100 bootstrap samples of size 1000 from the 1800 discretized observational measurements made by Sachs et. al. [32]. We did not use interventional measurements because, as far as we know, there is currently no methodology for applying constraint-based algorithms to a mixture of experimental and observational data. Skeleton identification was applied to the bootstrap samples assuming that they were generated by a faithful categorical Bayesian network whose structure is the protein-signaling network described in the literature; we refer to this Bayesian network as *Protein*.

We applied the skeleton identification phase of MMHC [46], which phase we call MMPC-skeleton,[4] on each of the 100 samples of each network, each time using the first $n = \{100, 1000, 10000\}$ observations ($n = \{100, 1000\}$ for Protein). We used the G test with $\alpha = 0.05$ and the heuristic power rule and calculate the degrees of freedom according to Steck and Jaakkola [36].

In order to evaluate $\widehat{\text{FDR}}_{\text{BH}}$ in the task of FDR estimation, we computed $\widehat{\text{FDR}}_{\text{BH}}(t)$ and $\text{FDR}(t)$ for 50 logarithmically spaced in $[10^{-8}, 10^{-1}]$ p-value thresholds $t$ for all samples of each network.

- $\text{FDR}(t)$ (Fig. 3.3) varies greatly among networks and increases as $t$ increases or $n$ decreases on all but two networks.[5]
- $\text{FPR}(t)$ (Fig. 3.4), $\pi(t)$ (Fig. 3.5) and $\pi[\text{t}_\text{q}(\widehat{\text{FDR}}_{\text{BH}})]$ (Fig. 3.8) also vary greatly among networks.
- $\widehat{\text{FDR}}_{\text{BH}}(t)$ is getting more conservative as $t$ increases, in general (Fig. 3.6). For $n \in \{1000, 10000\}$ $\widehat{\text{FDR}}_{\text{BH}}(t)$ is conservative (or almost conservative) on four networks but also not conservative for $n = 100$ on three networks. The estimator $\widehat{\text{FDR}}_{\text{BH}}(\alpha)$ of the FDR of skeleton identification is conservative (or almost conservative) in most cases. However, it is overly conservative for $n \in \{1000, 10000\}$ on half of the networks.

In order to evaluate $\widehat{\text{FDR}}_{\text{BH}}$ in the task of FDR control, we applied the BH procedure with 50 logarithmically spaced in $[10^{-3}, 10^{-1}]$ FDR thresholds $q$ for each sample from each network. The results are the following:

- The BH procedure achieves (or almost achieves) tight strong control of the FDR for $n \in \{1000, 10000\}$ on four networks but also fails on the other two; it also fails for $n = 100$ on all but two networks. Recall from Section 3.2 that the BH procedure strongly controls the FDR at level $q$ if and only if $\widehat{\text{FDR}}(\text{t}_\text{q}(\widehat{\text{FDR}}))$ is conservative. For $q$ such that tight strong control is achieved (Fig. 3.7), $\text{E}[\text{t}_\text{q}(\widehat{\text{FDR}}_{\text{BH}})]$ belongs to the lower part of the range of $t$ in Fig. 3.6, where $\widehat{\text{FDR}}_{\text{BH}}(t)$ is slightly conservative.
- $\pi[\text{t}_\text{q}(\widehat{\text{FDR}}_{\text{BH}})]$ (Fig. 3.8) is a reflection of $\pi(t)$ (Fig. 3.5).

The results on both tasks are summarized as follows:

- Estimation is conservative and strong control is achieved in some cases and thresholds but not in others.
- Estimation of the FDR of skeleton identification is in general conservative, although overly conservative in half of the cases.
- When strong control is achieved, it is tight.

[4] We call it this way because it is based on the MMPC local learning algorithm by Tsamardinos, Brown and Aliferis [46].

[5] We also considered an alternative definition of the FDR, called the *positive* FDR (pFDR) [39, 40] (See Section 3.5.5). We found $\text{FDR}(t) = \text{pFDR}(t)$ for all networks so we did not consider pFDR any further.

Given these results, we would recommend controlling the FDR at level $q \in [10^{-3}, 10^{-1}]$ with $\widehat{\text{FDR}}_{\text{BH}}$ only on networks that are known to be similar to the ones for which we demonstrated that strong control of the FDR is achieved. We would not recommend estimating the FDR of skeleton identification with $\widehat{\text{FDR}}_{\text{BH}}$ when using $\alpha = 0.05$ as the significance level of the tests, since when estimation is conservative, it is overly conservative.

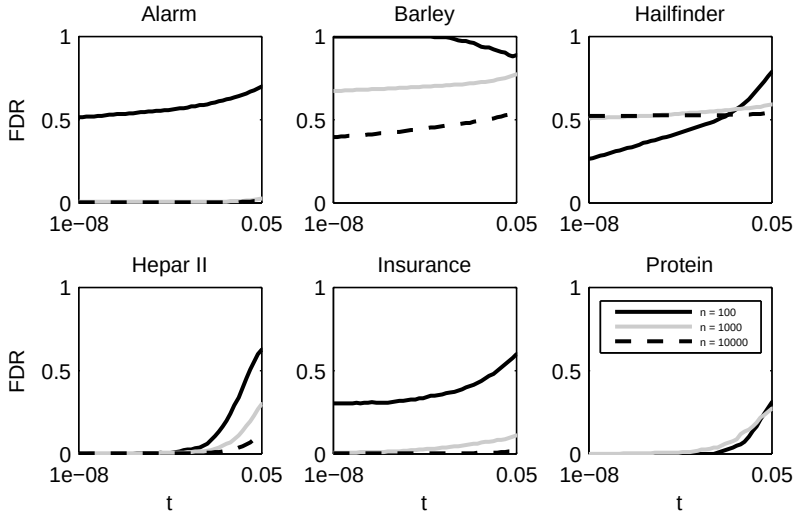



Figure 3.3: FDR($t$): the False Discovery Rate (FDR) of each p-value threshold $t$ for each network and sample size $n$ (for Protein, $n \in \{100, 1000\}$). X-axes are in logarithmic-10 scale. FDR varies greatly among networks and increases as $t$ increases or $n$ decreases, except on Barley and Hailfinder. FDR(0.05) is the FDR of skeleton identification.

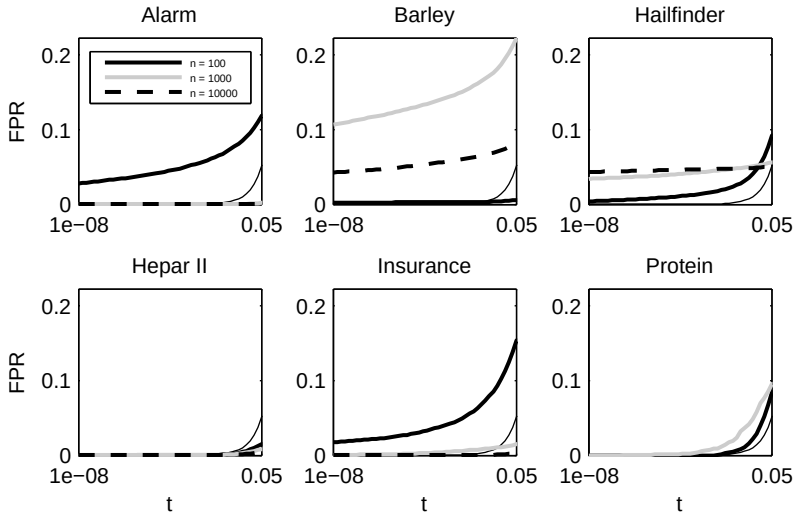



Figure 3.4: FPR($t$): the False Positive Rate (FPR) of each p-value threshold $t$ for each network and sample size $n$ (for Protein, $n \in \{100, 1000\}$). X-axes are in logarithmic-10 scale. FPR varies greatly among networks and increases as $t$ increases. FPR(0.05) is the FPR of skeleton identification. The thin black curve is FPR($t$) = $t$. Curves below FPR($t$) = $t$ indicate strong control of the FPR: This happens for $n = 100$ on Barley and Hepar II, for $n = 1000$ and large $t$ on Alarm, Hepar II and Insurance and for $n = 10000$ and large $t$ on Alarm, Hepar II and Insurance.

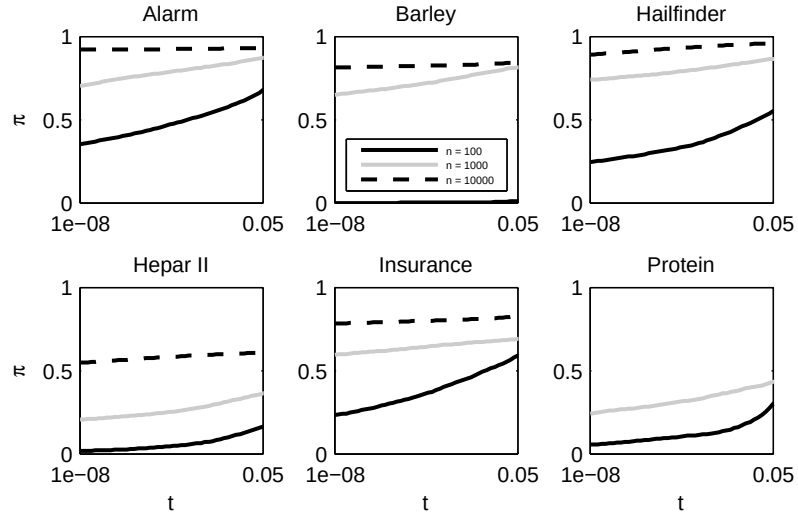



Figure 3.5: $\pi(t)$: power of each p-value threshold $t$, network and sample size $n$ (for Protein, $n \in \{100, 1000\}$). X-axes are in logarithmic-10 scale. Power varies greatly among networks and increases as $t$ or $n$ increases. $\pi(0.05)$ is the power of skeleton identification.

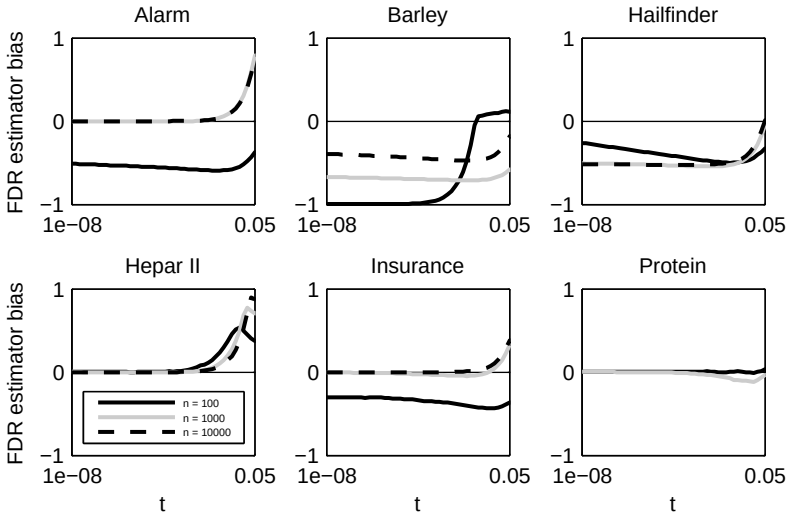



Figure 3.6: $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BH}}(t)] - \mathrm{FDR}(t)$: bias of Storey's [39] False Discovery Rate (FDR) estimator $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ of each p-value threshold $t$, for each network and sample size $n$ (for Protein, $n \in \{100, 1000\}$). X-axes are in logarithmic-10 scale. Bias increases as $t$ increases except for $n = 100$ and small $t$ on Alarm, Hailfinder and Insurance, for large $t$ on Hepar II and on Protein. Bias is non-negative or slightly negative for $n = 100$ on Hepar II, for $n = 1000$ on Alarm, Hepar II, Insurance and Protein and for $n \in \{1000, 10000\}$ on Alarm, Hepar II and Insurance, while it is noticeably negative for $n = 100$ on Alarm, Hailfinder and Insurance, for $n = 1000$ on Barley and Hailfinder and for $n = 10000$ on Barley. $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BH}}(0.05)] - \mathrm{FDR}(0.05)$ is the bias in estimating the FDR of skeleton identification and is non-negative or slightly negative for $n = 100$ on Barley, Hepar II and Protein, for $n = 1000$ on Alarm, Hepar II, Insurance and Protein and for $n = 10000$ on Alarm, Hailfinder, Hepar II and Insurance. However, it is too positive for $n = 100$ on Hepar II and for $n \in \{1000, 10000\}$ on Alarm, Hepar II and Insurance.

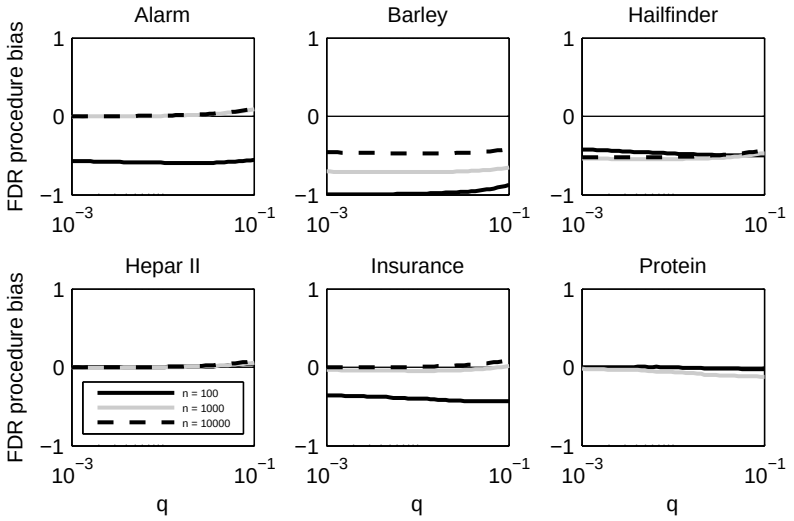



Figure 3.7: $q - \mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BH}})]$: bias of the Benjamini-Hochberg False Discovery Rate (FDR) controlling procedure [7] at FDR threshold $q$ for each network and sample size $n$ (for Protein, $n \in \{100, 1000\}$). X-axes are in logarithmic-10 scale. Bias is close to 0 for $n \in \{1000, 10000\}$ on Alarm, Hepar II and Insurance and on Protein. Bias is noticeably negative for $n = 100$ on Alarm, Barley, Hailfinder and Insurance and for $n \in \{1000, 10000\}$ on Barley and Hailfinder.

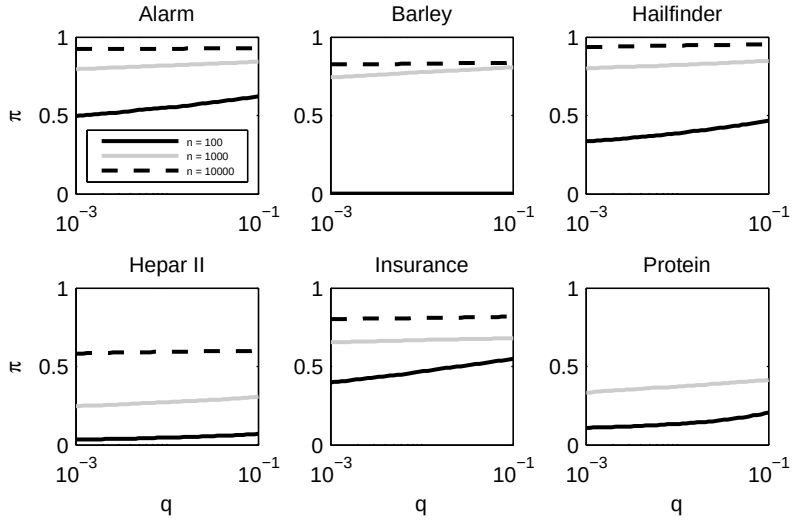



Figure 3.8: $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BH}})]$: power of the Benjamini-Hochberg False Discovery Rate (FDR) controlling procedure [7] with FDR threshold $q$ for each network and sample size $n$ (for Protein, $n \in \{100, 1000\}$). X-axes are in logarithmic-10 scale. Power varies greatly among networks and increases as $q$ or $n$ increases.

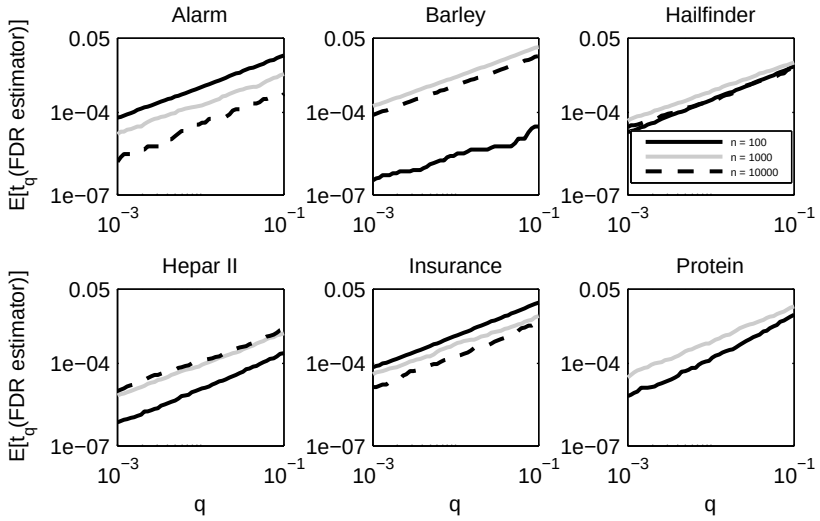



Figure 3.9: $\mathrm{E}[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BH}})]$: expected p-value threshold $t$ returned by the False Discovery Rate (FDR) controlling procedure with FDR threshold $q$ for each network and sample size $n$ (for Protein, $n \in \{100, 1000\}$). All axes are in logarithmic-10 scale. For $q$ such that $q - \mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BH}})]$ is close to 0 (Fig. 3.7), $\mathrm{E}[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BH}})]$ belongs to the lower part of the range of $t$ in Fig. 3.6, where $\widehat{\mathrm{FDR}}(t)$ bias is close to 0.

There are two possible causes for the observed lack of accuracy of $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ in FDR estimation and control. Either

1. the dependence of the maximal conditional independence p-values is not supported by $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ (which assumes independence or positive regression dependence for the p-values) or

2. there are maximal conditional independence p-values that are not upper bounds on the link absence p-values

or both. The next chapter deals with these issues.

Note that these are issues regarding the process of skeleton identification and the employed FDR estimators. They are not issues of our unified approach to utilizing FDR. The assumptions of the single-stage FDR control approach of Li and Wang [22] are similar to ours and $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ is also used there. $\mathrm{PC}_{\mathrm{FDR}-}$ skeleton was evaluated with randomly generated networks and with $q = 0.05$ only [22]. An empirical comparison of our two-stage approach to FDR control to the single-stage approach of Li and Wang [22] using networks from online repositories is yet to be made.

## 3.5 Other approaches to assessing confidence in structure learning

Confidence can be assessed on entire structures or certain structural *features* (e.g. links). Skeleton identification, viewed as multiple hypothesis testing, as-

sesses confidence on links (the link absence p-values or q-values) and entire skeletons (the FPR or FDR). Search-and-score methods (Section 2.2) assess confidence on entire structures (the score). Now let us further discuss Bayesian model averaging and present the use of bootstrap in assessing confidence on structural features.

### 3.5.1 Bayesian model averaging

As mentioned in Section 2.2, the search-and-score approach to structure learning is to search for the DAG (DAG pattern) $\mathbb{G}$ that maximizes some score function, which typically is the posterior probability $\Pr(\mathbb{G}|d)$ of $\mathbb{G}$ given the data $d$. $\Pr(\mathbb{G}|d)$ is called the *Bayesian score* of $\mathbb{G}$. Exhaustive consideration of possible structures is computationally infeasible when the number of nodes is not small, because the number of possible DAGs is super-exponential to the number of nodes [30]. The number of DAG patterns is smaller but it is still large [37]. For this reason, algorithms that perform a heuristic search over the space of DAGs or DAG patterns have been devised [26].

When the number of variables is small and the sample size is large, a single structure can be orders of magnitude more probable [18]. However, when the sample size is small relative to the number of variables, there are often many structures that are equally probable and choosing one would be arbitrary [26]. In such cases, one would perform Bayesian model averaging to compute the probability $\Pr(\text{Present}_f\,|d)$ of the event $\text{Present}_f$ that a certain feature $f$ of the DAG or DAG pattern $\mathbb{G}$ is present, given the data $d$:

$$\Pr(\text{Present}_f\,|d) = \sum_{\mathbb{G}} 1\{\text{Present}_f^{\mathbb{G}}\} \cdot P(\mathbb{G}|d)$$

where $1\{\cdot\}$ is the indicator function which is 1 when its input occurs and 0 otherwise and $\text{Present}_f^{\mathbb{G}}$ denotes the event that $f$ is present in $\mathbb{G}$.

The exact computation of $\Pr(\text{Present}_f\,|d)$ by averaging over all possible structures is computationally infeasible when the the number of nodes is not small for the reasons discussed above. In these cases, $\Pr(\text{Present}_f\,|d)$ can be approximated by searching for highly probable structures and then average over them [26]. This is usually done using *Monte Carlo Markov Chain* (MCMC) based methods [24, 16, 10, 17], which are very computationally expensive.

$\Pr(\text{Present}_f\,|d)$, where $f$ is a subnetwork (e.g. a single edge) can be computed exactly for moderately sized networks (about 25 nodes or less) using dynamic programming [20, 19, 43]. Application to larger networks is currently prohibited due to space requirements. Recent work [28, 29] is targeted on reducing these requirements.

Listgarten and Heckerman [23] present a Bayesian and a frequentist approach to estimating the number of false edges in a DAG $\hat{\mathbb{G}}$ learned by any structure learning method. Their frequentist approach is to estimate FDR by a particular permutation-based FDR estimator and is discussed in Section 4.1.2. Their Bayesian approach is to estimate the expected number $\mathrm{E}[S|d]$ of true edges given data $d$ by averaging over all possible structures:

$$\mathrm{E}[S|d] = \sum_{\mathbb{G}} \mathrm{S}(\mathbb{G}, \hat{\mathbb{G}}) \cdot P(\mathbb{G}|d)$$

where $\mathrm{S}(\mathbb{G}, \hat{\mathbb{G}})$ is the number of edges that are present both in $\hat{\mathbb{G}}$ and $\mathbb{G}$.

### 3.5.2 The bootstrap

Friedman et. al. [15] use the bootstrap to estimate the probability $\Pr(\mathrm{Present}_f^{\hat{\mathbb{G}}} \mid |D| = n)$ of the event $\mathrm{Present}_f^{\hat{\mathbb{G}}}$ that a certain feature $f$ of the DAG or DAG pattern $\hat{\mathbb{G}}$ learned from a sample of size $n$ is present. If the structure learning algorithm is *consistent*, then we can expect that, as $n$ increases, $\Pr(\mathrm{Present}_f^{\hat{\mathbb{G}}} \,||D| = n) \to 1$ if $\mathrm{Present}_f^{\mathbb{G}}$ occurs and $\Pr(\mathrm{Present}_f^{\hat{\mathbb{G}}} \,||D| = n) \to 0$ if $\mathrm{Present}_f^{\mathbb{G}}$ does not occur [15].

*Non-parametric* bootstrap works as follows: First, $B$ bootstrap samples (i.e., samples with replacement) are generated from the original sample $d$. Then structure learning is applied to each bootstrap sample. Finally, $\Pr(\mathrm{Present}_f^{\hat{\mathbb{G}}} \,||D| = n)$ is estimated by

$$\frac{1}{B} \sum_{i=1}^{B} 1\{\mathrm{Present}_f^{\hat{\mathbb{G}}_i}\}$$

where $\hat{\mathbb{G}}_i$ denotes the structure learned from the i-th bootstrap sample. The learned structures from the non-parametric bootstrap can be also used as the highly probable structures in the estimation of $\Pr(\mathrm{Present}_f \,|d)$ in Bayesian model averaging.

*Parametric* bootstrap is similar to the non-parametric except that we generate $B$ samples from the network learned from the original sample.

### 3.5.3 Classification of pairs of variables

The scores in the search-and-score approach are not directly comparable to the error rates in the constraint-based approach. However, methods estimating or exactly computing $\mathrm{E}[S|d]$ can be compared to FDR estimators and methods estimating or exactly computing $\Pr(\mathrm{Present}_{X \_ Y} \,|d)$ or $\Pr(\mathrm{Present}_{X \_ Y}^{\hat{\mathbb{G}}} \,||D| = n)$ for each pair $(X, Y)$ of variables in $\mathbf{V}$ can be compared to skeleton identification in the context of *classification*.

A *classifier* is a mapping from instances to predicted classes [13]. We consider only two classes, namely *positive* and *negative*. *Binary* classifiers output only a binary class label which indicates the predicted class of the instance. *Scoring* classifiers output a probability or a *score* in general, representing the degree to which the instance belongs to a class [13]. A scoring classifier can be used with a *threshold* to produce a binary classifier: if the score of an instance is above the threshold classify the instance as a positive, otherwise classify it as a negative [13].

Skeleton identification, Bayesian model averaging for estimating or exactly computing $\Pr(\mathrm{Present}_{X \_ Y} \,|d)$ and the bootstrap for estimating $\Pr(\mathrm{Present}_{X \_ Y}^{\hat{\mathbb{G}}} \,||D| = n)$ can be viewed as scoring classification of pairs of variables as *links* or *non-links*. The score is $\Pr(\mathrm{Present}_{X \_ Y} \,|d)$, $\Pr(\mathrm{Present}_{X \_ Y}^{\hat{\mathbb{G}}} \mid |D| = n)$ and $-\,\mathrm{p}_{\neg\,\mathrm{Adj}(X,Y)}$ (the negative link absence p-value) or $-q_{\neg\,\mathrm{Adj}(X,Y)}$ (the negative link absence q-value) for each method respectively.

*Receiver Operating Characteristic* (ROC) graphs (Fig. 3.10) are a way to visualize classifier performance. An ROC graph plots the TPR vs. the FPR of

one or more classifiers. Thus, such a graph depicts the relative trade-off between these two quantities for each classifier [13]. A binary classifier is characterized by a single (FPR, TPR) pair, i.e., a single point in the graph. The point $(0, 1)$ corresponds to the perfect classifier. The closer a point to $(0, 1)$ is the better the corresponding classifier is. The points $(0, 0)$ and $(1, 1)$ correspond to the classifier that always predicts negative and positive respectively. The point $(p, p)$ on the diagonal line $y = x$ corresponds to the random classifier that predicts positive with probability $p$. Points below the diagonal correspond to classifiers that perform worse than random. However, if the output of the latter classifiers is reversed, then they correspond to points above the diagonal. For scoring classifiers, each threshold results in a different (FPR, TPR) pair, i.e., a different point in the ROC graph. If we vary the threshold from $\infty$ to $-\infty$, we produce a *curve* from (0,0) to (1,1). This curve shows the ability of the classifier to rank positive instances relative to negative instances [13]. The area under the curve (AUC) is usually used as a summary of classifier performance. The AUC of a classifier also equals the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [13].

ROC analysis has been extensively used for comparing methods for estimating or exactly computing $\Pr(\text{Present}_{X \to Y} | d)$ or $\Pr(\text{Present}^{\hat{\mathbb{G}}}_{X \to Y} | |D| = n)$ to each other [16, 19, 10, 17]. An experimental comparison of skeleton identification to methods for estimating or exactly computing $\Pr(\text{Present}_{X \_ Y} | d)$ or $\Pr(\text{Present}^{\hat{\mathbb{G}}}_{X \_ Y} | |D| = n)$ for each pair $(X, Y)$ of variables in $\mathbf{V}$ is yet to be conducted.

The *Positive Predictive Value* (PPV) of a classifier on a dataset is defined as the ratio $S/R$ of the number $S$ of true positives to the number $R$ of predicted positives. It is easy to see (Tables 3.1 and 3.2) that the PPV of multiple hypothesis testing, viewed as binary classification, is equal to 1 - realized FDR. Thus, $1 - \widehat{\text{FDR}}$ can be used an estimator of E[PPV]. $\text{E}[S|d]/R$, where $R$ is the number of links in $\hat{\mathbb{G}}$, can be also used as an estimator of E[PPV]. Listgarten and Heckerman [23] compare their two approaches in terms of PPV estimation error in the context of classification of possible edges as edges or non-edges. An experimental comparison of Listgarten and Heckerman's Bayesian approach to FDR estimators in the context of classification of pairs of variables as links or non-links is yet to be conducted.

### 3.5.4 Skeleton identification as classification of pairs of variables

We generated the ROC curve (Fig. 3.10) of skeleton identification and calculated the AUC (Fig. 3.11) for each network and sample size. Using either the negative link absence p-values or the negative link absence q-values as link scores yields the same ROC curve because p-values and q-values are analogous quantities; we used the p-values. We set link absence p-values $> \alpha$ to 1 because we are not interested in the ability of skeleton identification in scoring discarded links (see Section 3.3.3).

Classification performance is improved as $n$ increases and is high even in some cases conservative estimation (Fig. 3.6) and strong control (Fig. 3.7) of the FDR is not achieved. Thus, even in these cases, negative maximal conditional independence p-values are good relative link scores.

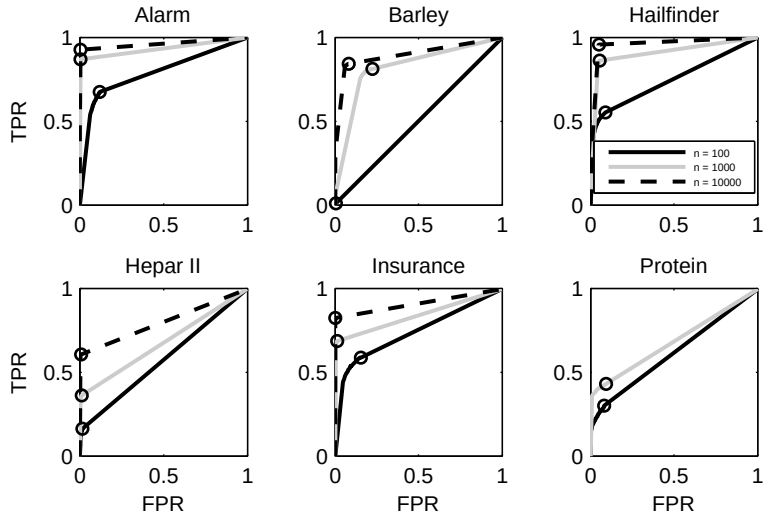



Figure 3.10: Receiver Operating Characteristic (ROC) curve for each network and sample size $n$ (for Protein, $n \in \{100, 1000\}$). The point of each curve that corresponds to p-value threshold $\alpha = 0.05$ is surrounded by a circle. Curves tend to the point $(1, 1)$ as $n$ increases. Curves are not far from $(1, 1)$ for $n \in \{1000, 10000\}$ on Barley and Hailfinder, even if conservative estimation (Fig. 3.6) and strong control (Fig. 3.7) of the FDR is not achieved. The curve for $n = 100$ on Barley almost coincides with $x = y$.

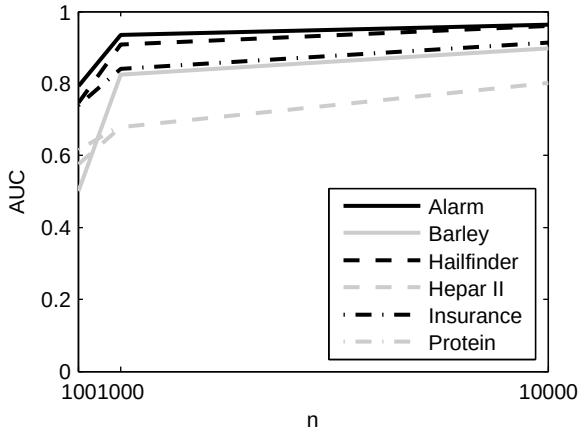



Figure 3.11: Area under the ROC curve (AUC) for each sample size $n$ and network (for Protein, $n \in \{100, 1000\}$. AUC increases as $n$ increases. AUC is large for $n \geq 1000$ on all networks except Hepar II. It is large even on Barley and Hailfinder, for which FDR conservative estimation (Fig. 3.6) and strong control (Fig. 3.7) is not achieved.

### 3.5.5 A Bayesian interpretation of the False Discovery Rate in skeleton identification

Storey [40] introduces a modified version of the FDR called the *positive False Discovery Rate* (pFDR), and discusses its advantages over the FDR. The pFDR and the FDR are asymptotically ($m \to \infty$) equivalent for a fixed p-value threshold [39]; we confirmed this for the experiments of Section 3.4. The definition of pFDR is the following:

$$\text{pFDR} \triangleq \text{E}\left[\frac{V}{R}\middle| R > 0\right]$$

An interesting property of the pFDR is that, if we assume that the test statistics come from a mixture distribution, then the pFDR can be expressed as a Bayesian posterior probability [40]:

**Theorem 3.3.** *Suppose $m$ identical hypothesis tests are performed with the statistics $T_1, \ldots, T_m$ and critical region $C$. Assume that $(T_i, H_i)$ are i.i.d. random variables, $T_i|H_i$ $(1 - H_i) \cdot F_0 + H_i \cdot F_1$ for some null distribution $F_0$ and alternative distribution $F_1$, and $H_i$ Bernoulli$(\pi_1)$ for $i = 1, \ldots, m$. Then [40]*

$$pFDR(C) = Pr(H = 0|T \in C) \tag{3.1}$$

*where $pFDR(C)$ is the pFDR obtained when rejecting hypotheses with corresponding statistic $\in C$ and $\pi_0 = 1 - \pi_1$ is the implicit prior probability used in the above posterior probability.*

*Proof.* The proof can be found in [40]. □

Note that Eq. 3.1 does not depend on $m$ and $\Pr(H_i = 0|T_i \in C)$ is the same for $i = 1, \ldots, m$.

As mentioned in Sections 2.2 and 3.5.1, the Bayesian approach to structure learning assumes a prior probability distribution of DAGs or DAG patterns. If we assume for any pair of nodes $(X, Y)$ that $\Pr(\neg \text{Adj}(X, Y)) = \pi_0$ and $\Pr(X - Y) = \pi_1$ such that $\pi_0 = 1 - \pi_1$, $\Pr(P_{\neg \text{Adj}(X,Y)} \leq t|\neg \text{Adj}(X, Y)) = F_0$ and $\Pr(P_{\neg \text{Adj}(X,Y)} \leq t|X - Y) = F_1$ instead, then:

$$\text{pFDR}(t) = \Pr(\neg \text{Adj}(X, Y)|P_{\neg \text{Adj}(X,Y)} \leq t)$$

That is, pFDR is the probability of a link in the output skeleton being false.

# Chapter 4

# Improving estimation and control

In the previous chapter, we demonstrated that, in some cases, $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ is not conservative and, subsequently, the BH procedure does not achieve strong control of the FDR. We identified the two possible causes of this: either (1) the dependence between the link absence p-values is not supported by $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ or (2) there are maximal conditional independence p-values that are not upper bounds on the link absence p-values or both. In this chapter we address these issues.

The first issue is addressed by using a modification of $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ which is proven to be conservative under any kind of dependence. After finding out that, even with this modification, conservative estimation and strong control are still not achieved in some cases, we can deduce that there are maximal conditional independence p-values that are not upper bounds.

Owing to Theorem 3.2, we expect that (1) an increase of the tests that are reliable according to the employed reliability criterion and (2) an increase in the power of the tests should result in an increase of the maximal conditional independence p-values that are upper bounds and, therefore, in less biased estimation and control of the FDR. Thus, we evaluate, with respect to the resulting bias of the FDR estimators, several reliability criteria and values of their parameters as well as several approaches of increasing the power of the tests.

## 4.1 Dealing with p-value dependence

### 4.1.1 The Benjamini-Yekutieli conservative modification

Benjamini and Yekutieli [7] prove that the BH procedure with $q/\left(\sum_{i=1}^{m}\frac{1}{i}\right)$ in place of $q$ achieves strong control for any form of dependence of the p-values; we refer to this modified procedure as the BY procedure. The BY procedure with $q$ corresponds to taking the largest $t$ such that the estimator below is $\leq q$:

$$\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t) \triangleq \frac{m \cdot t \cdot \left(\sum_{i=1}^{m}\frac{1}{i}\right)}{\mathrm{R}(t) \vee 1}$$

Therefore, $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)$ is conservative for any form of dependence.

First, we computed $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)$ for the runs of Section 3.4 at the same p-value thresholds $t$. As $t$ increases, $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)$ is getting more conservative at faster rate than $\widehat{\mathrm{FDR}}_{\mathrm{BH}}(t)$ (Fig. 4.1). $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(\alpha)$ is everywhere conservative, while $\widehat{\mathrm{FDR}}_{\mathrm{BH}}(\alpha)$ is not. $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(\alpha)$ is, however, overly conservative while $\widehat{\mathrm{FDR}}_{\mathrm{BH}}(\alpha)$ is just (or almost) conservative in some cases. $\widehat{\mathrm{FDR}}_{\mathrm{BH}}(t)$ is about as conservative as $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)$ for small $t$.

Then, we applied the BY procedure to the same runs and at the same FDR thresholds $q$ as in Section 3.4. The differences between the BY and the BH procedure are neglectable because the p-value thresholds returned by either procedure (Fig. 4.4) belong to the lower part of the range of Fig. 4.1, where their bias is about the same.

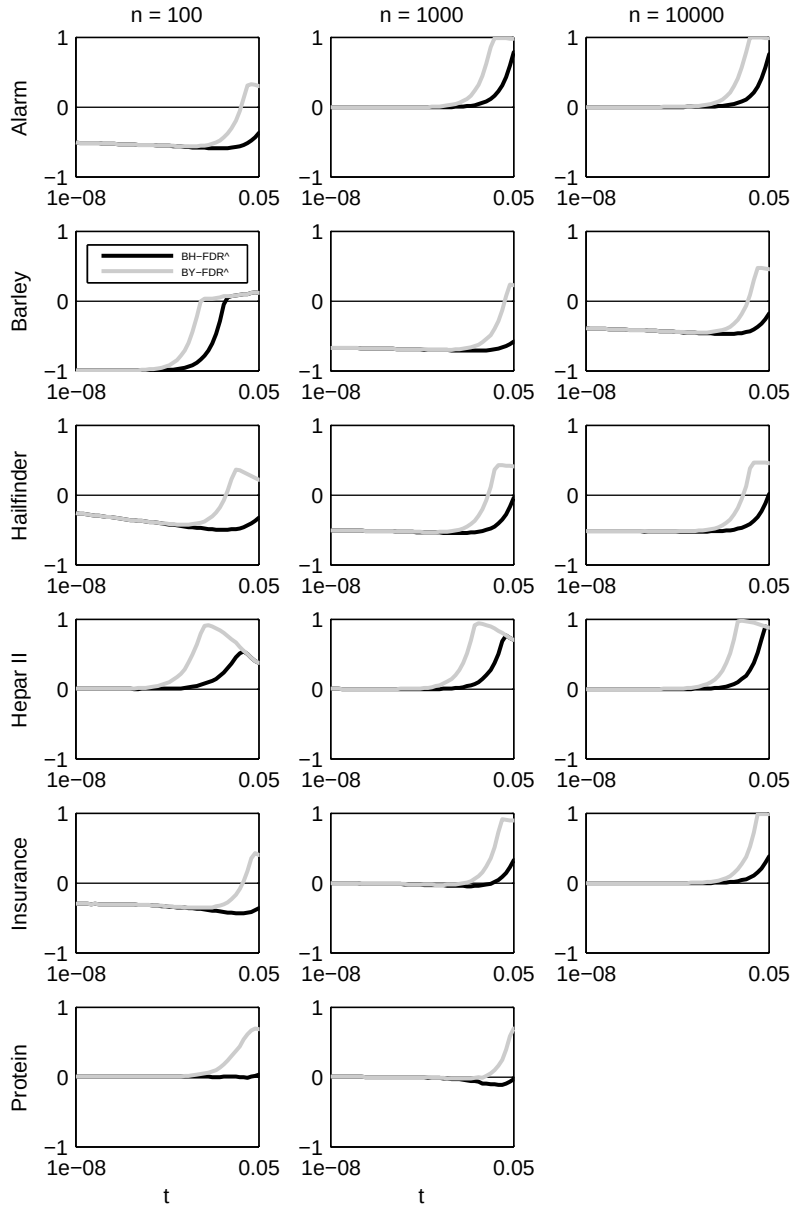



Figure 4.1: $\mathrm{E}[\widehat{\mathrm{FDR}}(t)] - \mathrm{FDR}(t)$: bias of the False Discovery Rate (FDR) estimator $\widehat{\mathrm{FDR}}$ of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and $\widehat{\mathrm{FDR}}$. X-axes are in logarithmic-10 scale. $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ (Benjamini and Yekutieli's [7] FDR estimator) curves are steeper than the respective $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ (Storey's [39] FDR estimator) ones. $\mathrm{E}[\widehat{\mathrm{FDR}}(0.05)] - \mathrm{FDR}(0.05)$ is the bias of $\widehat{\mathrm{FDR}}$ when estimating the FDR of skeleton identification and it is positive for $\widehat{\mathrm{FDR}} = \widehat{\mathrm{FDR}}_{\mathrm{BY}}$ in every case while it is not for $\widehat{\mathrm{FDR}} = \widehat{\mathrm{FDR}}_{\mathrm{BH}}$. However, $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(0.05)$ is overly conservative while $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ is just (or almost) conservative for $n = 100$ on Protein, for $n = 1000$ on Insurance and Protein and for $n = 10000$ on Hailfinder and Insurance. For large $t$ bias drops on Hepar II and is the same for both estimators because $\widehat{\mathrm{FDR}}_{\mathrm{BH}} = \widehat{\mathrm{FDR}}_{\mathrm{BY}} = 1$ (Figure not shown). $\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ and $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ bias is about the same for small $t$.

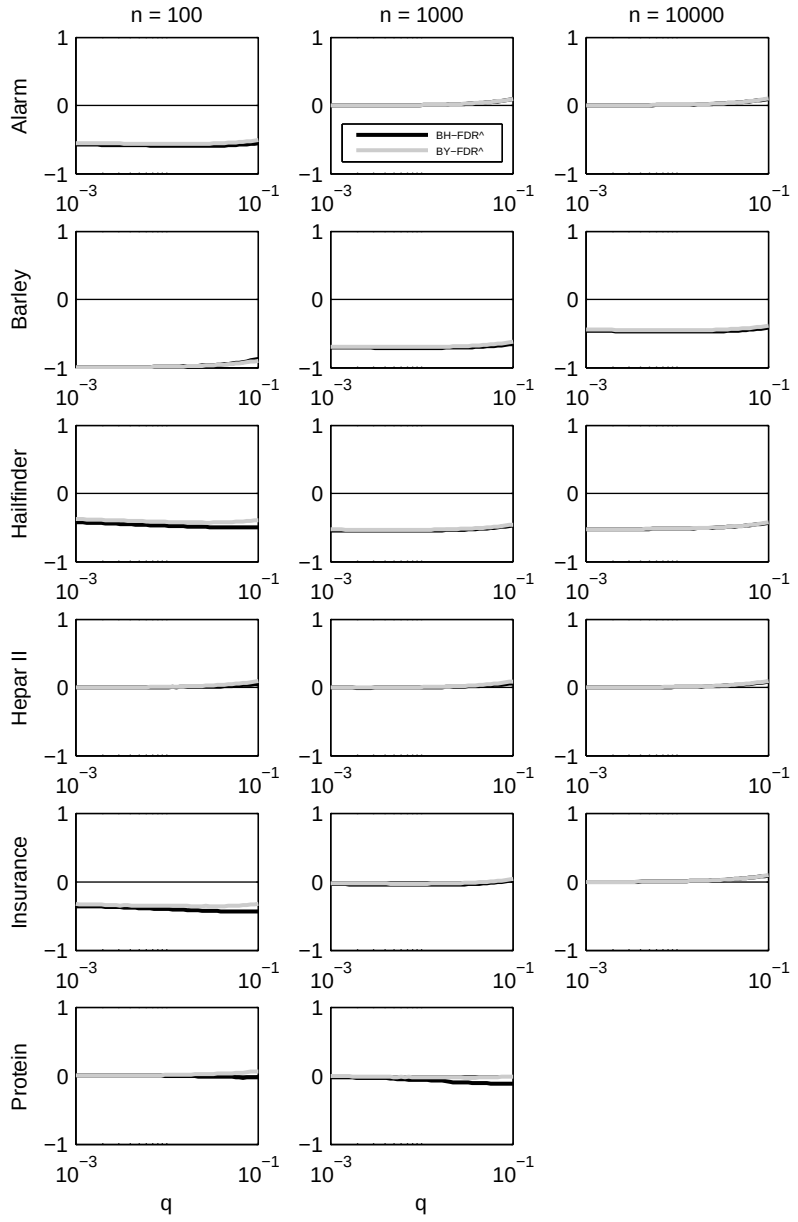



Figure 4.2: $q - \mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}})]$: bias of the False Discovery Rate (FDR) controlling procedure with each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and $\widehat{\mathrm{FDR}}$. X-axes are in logarithmic-10 scale. The differences between $\widehat{\mathrm{FDR}} = \widehat{\mathrm{FDR}}_{\mathrm{BH}}$ (Storey's [39] FDR estimator) and $\widehat{\mathrm{FDR}} = \widehat{\mathrm{FDR}}_{\mathrm{BY}}$ (Benjamini and Yekutieli's [7] FDR estimator) are neglectable because the p-value thresholds returned by the FDR controlling procedure with either estimator (Fig. 4.3) belong to the lower part of the range of Fig. 4.1, where the bias of the estimators is about the same.

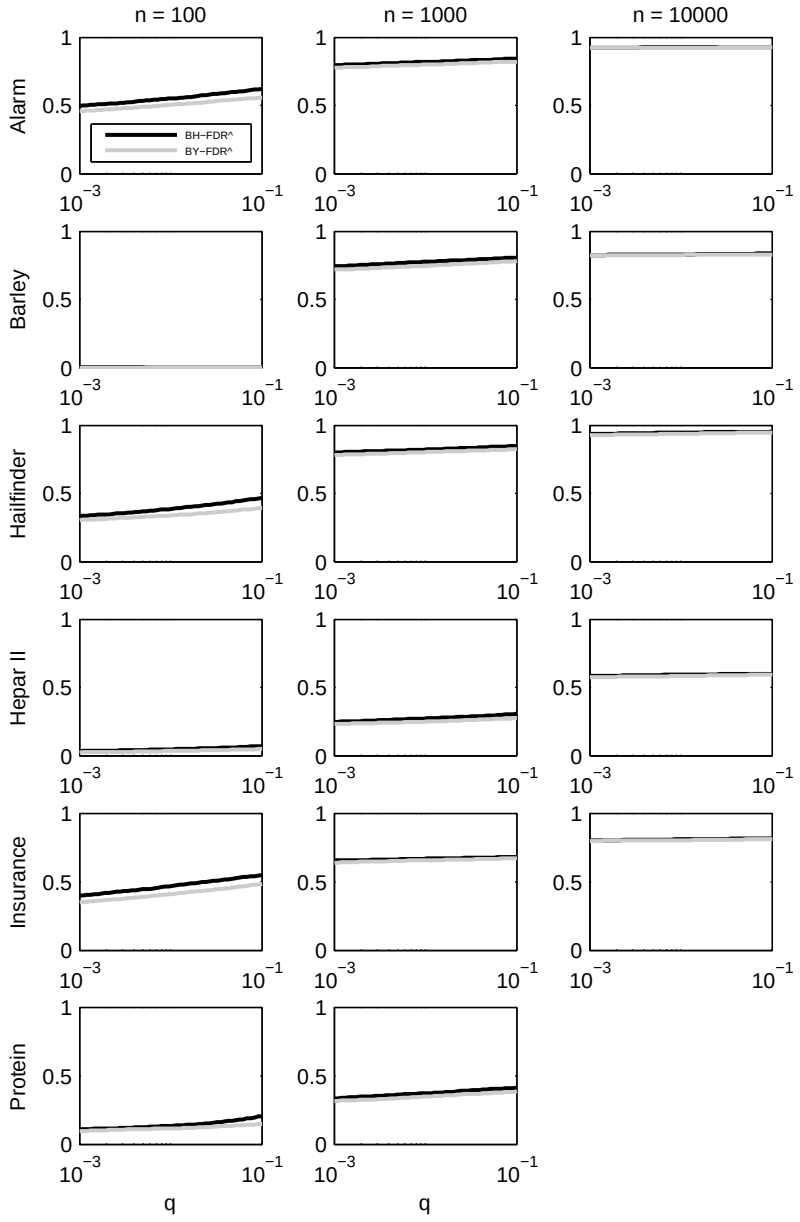



Figure 4.3: $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}})]$: power of the False Discovery Rate (FDR) controlling procedure with each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and $\widehat{\mathrm{FDR}}$. X-axes are in logarithmic-10 scale. Applying the procedure with $\widehat{\mathrm{FDR}} = \widehat{\mathrm{FDR}}_{\mathrm{BY}}$ (Benjamini and Yekutieli's [7] FDR estimator) results in a small drop in power compared to when using $\widehat{\mathrm{FDR}} = \widehat{\mathrm{FDR}}_{\mathrm{BH}}$ (Storey's [39] FDR estimator). This is because the p-value thresholds $t$ returned by the procedure are smaller in the first case (Fig. 4.3) and power decreases as $t$ decreases (Fig. 3.5).

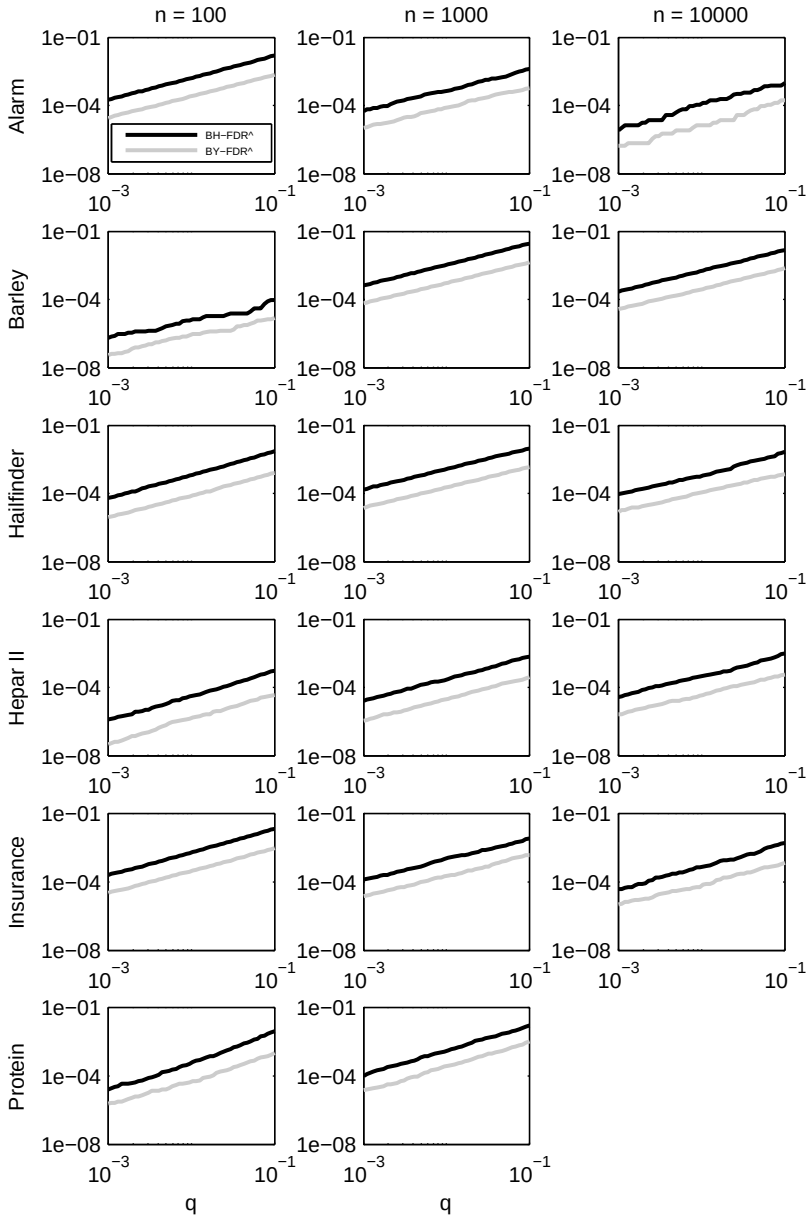



Figure 4.4: $\mathrm{E}[\mathrm{t_q}(\widehat{\mathrm{FDR}})]$: expected p-value threshold returned by the False Discovery Rate (FDR) controlling procedure with each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and $\widehat{\mathrm{FDR}}$. X-axes are in logarithmic-10 scale. $\mathrm{E}[\mathrm{t_q}(\widehat{\mathrm{FDR}})]$ is smaller with $\widehat{\mathrm{FDR}} = \widehat{\mathrm{FDR}}_{\mathrm{BY}}$ (Benjamini and Yekutieli's [7] FDR estimator) than with $\widehat{\mathrm{FDR}} = \widehat{\mathrm{FDR}}_{\mathrm{BH}}$ (Storey's [39] FDR estimator).

In summary, the main results regarding $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ are the following:

- $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(\alpha)$ is everywhere conservative, while $\widehat{\mathrm{FDR}}_{\mathrm{BH}}(\alpha)$ is not.
- $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(\alpha)$ is overly conservative when $\widehat{\mathrm{FDR}}_{\mathrm{BH}}(\alpha)$ is just conservative.
- The bias of the FDR procedure with either estimator is similar.

Thus, we would recommend the use of the BY over the BH controlling procedure with $q \in [10^{-3}, 10^{-1}]$ because the BY procedure ensures that p-value dependence is not a problem, while its bias is similar to that of the BH procedure. We still would not recommend estimating the FDR of skeleton identification since $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(\alpha)$ is even more conservative than $\widehat{\mathrm{FDR}}_{\mathrm{BH}}(\alpha)$.

Since there are cases where even $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ fails to achieve conservative estimation and strong control, we can deduce that there are maximal conditional independence p-values that are not upper bounds to the link absence p-values.

### 4.1.2 Estimating the False Discovery Rate via simulation of null statistics

Storey and Tibshirani [42] propose the following estimator for FDR[1] for any kind of dependence between the hypotheses, that can be used with general statistics and not only p-values:

$$\widehat{\mathrm{FDR}}_{\mathrm{ST}}(C) \triangleq \frac{\mathrm{E}[\mathrm{R}^0(C)]}{\mathrm{R}(C) \vee 1}$$

$\widehat{\mathrm{FDR}}_{\mathrm{ST}}(C)$ is the FDR when rejecting hypotheses with corresponding statistic $\in C$. $\mathrm{R}(C)$ is the number of rejected hypotheses. $\mathrm{E}[\mathrm{R}^0(C)]$ is the expected number of rejected true null hypotheses if all null hypotheses were true. Storey and Tibshirani [42] propose a method for estimating $\mathrm{E}[\mathrm{R}^0(C)]$ via simulation of null statistics. Under general dependence, $\widehat{\mathrm{FDR}}_{\mathrm{ST}}(C)$ is proven to be conservative under some conditions (see [42] for details). When using p-values and they are independent, $\mathrm{E}[\mathrm{R}^0(t)] = m \cdot t$ and $\widehat{\mathrm{FDR}}_{\mathrm{ST}}(t) = \widehat{\mathrm{FDR}}_{\mathrm{BH}}(t)$, which is always conservative. It is not hard to see that only the p-values of the *falsely discovered* links need to be upper-bounded for $\widehat{\mathrm{FDR}}_{\mathrm{ST}}(t)$ to be conservative. If this is the case, then $m \cdot t \geq \mathrm{E}[\mathrm{R}^0(t)]$.

As mentioned in Section 3.5.1, Listgarten and Heckerman [23] present a Bayesian and a frequentist approach to estimating the proportion of false *edges* (in contrast to links) in a DAG $\hat{\mathbb{G}}$ learned by any structure learning method. Their Bayesian approach is discussed in Section 3.5.1. Their frequentist approach essentially utilizes $\widehat{\mathrm{FDR}}_{\mathrm{ST}}$ to estimate the *realized* FDR. It is assumed that the structure learning algorithm can be decomposed into independent searches for the parents of each node. Then the distribution under the hypothesis that node $X_j$ is a not parent of $X_i$ is simulated by randomly permuting the values $x_i$ of $X_i$ in the sample from $P$.

[1] They actually propose it as an estimator for pFDR, but it can also be used to conservatively estimate FDR since FDR $\leq$ pFDR. This estimator also includes a $\hat{\pi}_0(\lambda)$ term, which is an estimator of the proportion of true null hypotheses. Because $\hat{\pi}_0(\lambda)$ is not applicable in the context of Bayesian network skeleton identification, we use $\hat{\pi}_0(\lambda) = 1$ instead.

The permutation performed by Listgarten and Heckerman [23] is not theoretically correct. The probability distribution under the hypothesis that node $X_j$ is a not parent of $X_i$ is

$$P^{ji0}(x_1, x_2, \ldots, x_n) = \prod_{k=1}^{n} P^{ji0}(x_k | \, \mathbf{pa}_{X_k}^{ji0})$$

where $\mathbf{pa}_{X_i}^{ji0} = \mathbf{pa}_{X_i} \setminus \{x_j\}$ and $\mathbf{pa}_{X_k}^{ji0} = \mathbf{pa}_{X_k}$ for $k \neq i$. To sample from this distribution one needs to know the parents of each node, i.e., the real DAG $\mathbb{G}$. Then one would randomly permute $x_i$'s while keeping $\mathbf{pa}_{X_i}^{ji0}$'s fixed. However, this results in incorrect values for the descendants of $X_i$. One should, therefore, randomly permute both $x_i$'s and the values of $X_i$'s descendants while keeping both $\mathbf{pa}_{X_i}^{ji0}$'s and the values of the parents (that are not descendants of $X_i$) of each descendant of $X_i$ fixed. $\mathbb{G}$ is, of course, unknown; however, we could use $\hat{\mathbb{G}}$ as an approximation.

The permutation method just described could be also applied to the estimation of the FDR in skeleton identification. Constraint-based algorithms belonging to the *Local to Global Learning* (LGL) class [4] (MMHC is such an algorithm) first learn the links ending to each node, i.e., they apply local learning. Then they combine the links to form the skeleton, thus completing the skeleton identification phase. Finally, they proceed with the edge orientation phase. To generate the distribution of the maximal conditional independence p-value $\max_{\mathbf{Z} \in \mathbf{C}_{\mathrm{Adj}(X,Y)}} P_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$ when $\mathrm{H}_{\mathrm{Adj}(X,Y)}$ is true, we would randomly permute $x$'s or $y$'s when the edge $X \leftarrow Y$ or $X \rightarrow Y$ is in $\hat{\mathbb{G}}$, respectively. If $\neg \, \mathrm{Adj}_{\hat{\mathbb{G}}}(X, Y)$ is the case, then we could permute either of them. Finally, we would apply local learning targeting either $X$ or $Y$ (which one of them, it does not matter) to obtain $\max_{\mathbf{Z} \in \mathbf{C}_{\mathrm{Adj}(X,Y)}} \mathrm{p}_{\mathrm{I}_P(\{X\},\{Y\}|\mathbf{Z})}$. This method is yet to be implemented and evaluated.

## 4.2 Improving upper bounds

Owing to Theorem 3.2, we expect that

1. an increase of the tests that are reliable according to the employed reliability criterion and

2. an increase in the power of the tests

should result in an increase of the maximal conditional independence p-values that are upper bounds and, therefore, in less biased estimation and control of the FDR. Thus, we evaluate, with respect to the resulting bias of the FDR estimators, several reliability criteria and values of their parameters as well as several approaches of increasing the power of the tests. In all evaluations we use $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ because it solves the p-value dependence issue (whether it is actually an issue or not), so we can focus on improving upper bounds.

### 4.2.1 Varying the reliability criterion

The POWER correction is a novel approach by Fast et. al [11] for controlling the power of the tests above a specified threshold $1-\beta$, where $\beta$ is the threshold on the FNR of the tests. The power of a test of conditional independence using a statistic that follows the $\chi^2$ distribution is a function of the sample size $n$, the significance level $\alpha$, the degrees of freedom *df* and the *effect size* $w$ in the data. POWER is the first approach to address all these four factors [12]. The $w$ parameter can either be specified in advance or estimated via cross-validation.

Fast [12] first chooses values for $w$ for sample size $n \in \{500, 1000, 2000, 5000\}$ using cross-validation with the *Diabetes*, *Hailfinder*, *Barley* and *Insurance* networks, randomly selected from the Bayesian Network Repository (see Appendix). Then, using these values of $w$ and power threshold $1 - \beta = 0.95$, he evaluates POWER along with three other reliability criteria (including a baseline criterion) with the *Alarm*, *Mildew*, *Pathfinder*, *Water* and *Win95pts* networks. He demonstrates that POWER is the only one to result in significance decreases (compared to the others, minus the baseline) in false negatives, accompanied, however, with a significance increase in false positives.

Among the other reliability criteria evaluated is the so-called "rule of thumb", supposedly used by many structure learning algorithms. According to the rule of thumb, a test is considered reliable if there are at least five observations per *degree of freedom*, on average. This is, however, different from the reliability criterion actually used in most constraint-based algorithms, which is the heuristic power rule. According to the latter criterion a test is considered reliable if there are at least *h-ps* (usually 5) observations per *cell of the contingency table*, on average. Both criteria do not maintain a constant level of power for all sample sizes, as the POWER correction does [12].

We applied MMPC-skeleton on the 100 samples of the Alarm, Hepar II and Protein networks using the first $n \in \{500, 1000, 5000\}$ observations and the heuristic power rule, the rule of thumb and the POWER reliability criteria. We used $\beta = 0.05$ with POWER. For the rest parameters we used the same values as in Section 3.4. We did not use Hailfinder, Barley and Insurance this time because they were used in the calculation of $w$ values by Fast, as mentioned above. To compensate for this loss of networks, we used *Win95pts* in this experiment and in the experiment of the next section. We used $n = 500$ and $n = 5000$ this time instead of $n = 100$ and $n = 10000$ because Fast did not calculate $w$ for the latter sample sizes. The runs on Alarm for $n = 1000$ using POWER were aborted because they were taking too much time.

- Compared to the heuristic power rule, the rule of thumb decreases FDR($t$) in one case and POWER increases FDR($t$) in some cases. On Win95pts and Protein all criteria achieve the same FDR.

- Compared to the heuristic power rule, the rule of thumb decreases $\pi(t)$ on Alarm and POWER decreases $\pi(t)$ on Alarm even more. This is in contrary to Fast's [12] results, where POWER increases power on Alarm for $n = 500$, compared to the rule of thumb. This may be due to Fast [12] performing the experiments with the PC algorithm and calculating the degrees of freedom according to Spirtes et al. [35] for the rule of thumb and according to Eq. (2.1) for POWER. On Hepar II however, POWER increases power compared to the heuristic power rule. On Win95pts and

Protein all criteria achieve the same power. This is also demonstrated for Win95pts by Fast [12]. As he explains, the tests are considered reliable by all criteria due to the special structure of Win95pts and the fact that it consists of binary only variables.

- Regarding FDR estimation: For POWER, $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)$ (Fig. 4.7) is not conservative for small $t$ in two cases, while it is or almost is for the heuristic power rule. The bias of $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)$ is the same on Win95pts and Protein for all criteria.
- Regarding FDR control: $q-\mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ (Fig. 4.8) and $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ (Fig. 4.9) are reflections of $\pi(t)$ and the bias of $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ respectively.

To sum up, compared to the heuristic power rule:

- The rule of thumb improves estimation and control of the FDR in one case but also decreases the power in that case.
- The POWER correction with $\beta = 0.05$:
  - Increases FDR in some cases.
  - Increases power on one network but decreases it on another.
  - Worsens estimation and control of the FDR in two cases.

Thus, we do not recommend using the rule of thumb or POWER with $\beta = 0.05$ over the heuristic power rule as the reliability criterion.

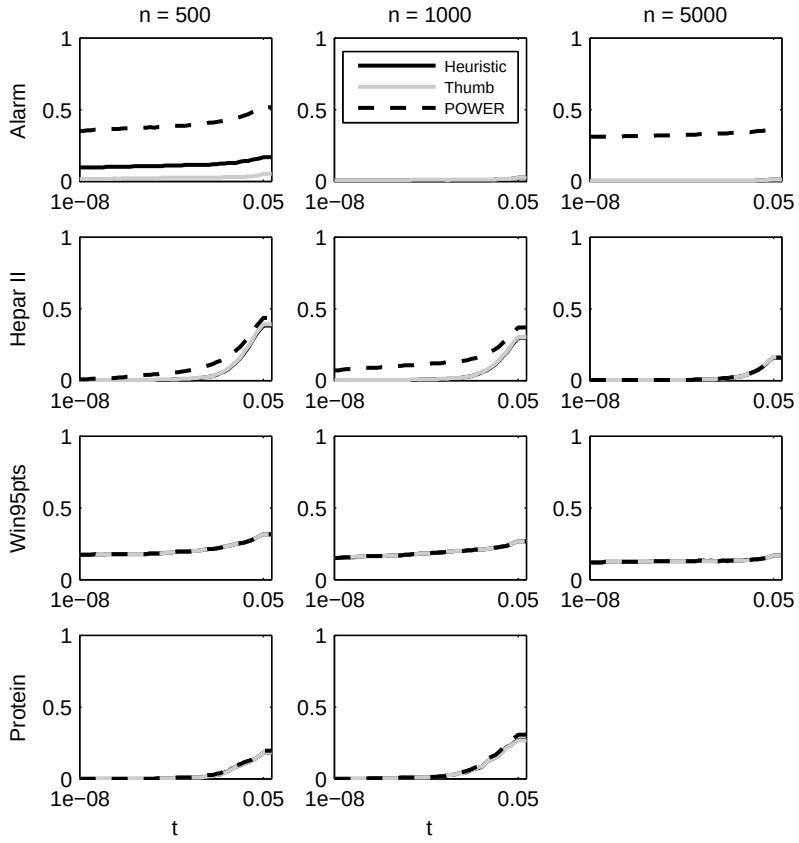



Figure 4.5: FDR($t$): FDR of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and reliability criterion. X-axes are in logarithmic-10 scale. *Heuristic*, *Thumb* and *POWER* denote the heuristic power rule, the rule of thumb and the POWER correction reliability criterion respectively. FDR on Alarm for $n = 1000$ and POWER is not shown. Compared to the heuristic power rule, the rule of thumb decreases FDR($t$) for $n = 500$ on Alarm and POWER increases FDR($t$) for $n = 500$ on Alarm, for $n = 1000$ on Hepar II and for $n = 5000$ on Alarm. On Win95pts and Protein all criteria achieve the same FDR.

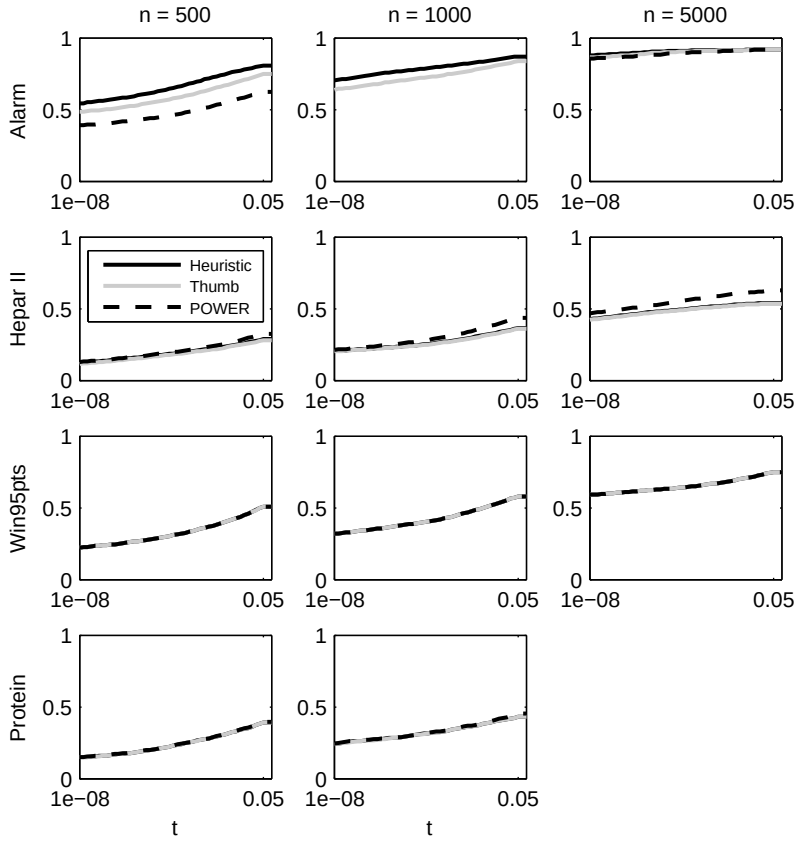



Figure 4.6: $\pi(t)$: power of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and reliability criterion. X-axes are in logarithmic-10 scale. *Heuristic*, *Thumb* and *POWER* denote the heuristic power rule, the rule of thumb and the POWER correction reliability criterion respectively. Power on Alarm for $n = 1000$ and POWER is not shown. Compared to the heuristic power rule, the rule of thumb decreases $\pi(t)$ on Alarm and POWER decreases $\pi(t)$ even more on Alarm but increases $\pi(t)$ on Hepar II. On Win95pts and Protein all criteria achieve the same power.

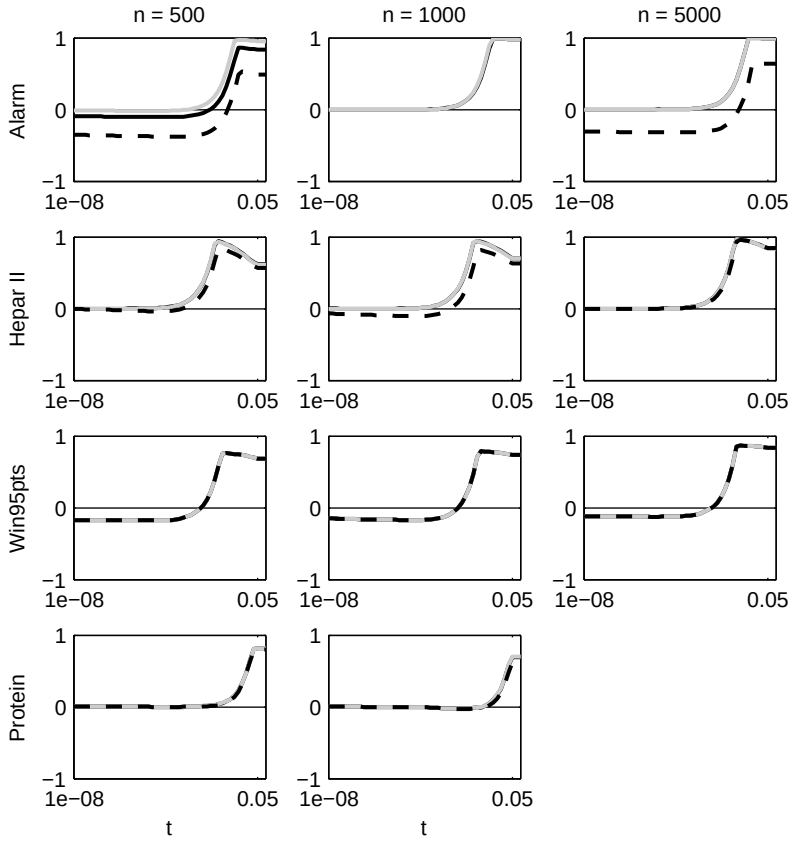



Figure 4.7: $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)] - \mathrm{FDR}(t)$: bias of the False Discovery Rate (FDR) estimator $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ (Benjamini and Yekutieli [7]) of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and reliability criterion. X-axes are in logarithmic-10 scale. The legend is the same as in Fig. 4.5. *Heuristic*, *Thumb* and *POWER* denote the heuristic power rule, the rule of thumb and the POWER correction reliability criterion respectively. Bias on Alarm for $n = 1000$ and POWER is not shown. For POWER, the $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ bias of small $t$ is noticeably negative on Alarm, while it is not or not that much for the heuristic power rule. Bias is the same on Win95pts and Protein for all criteria.

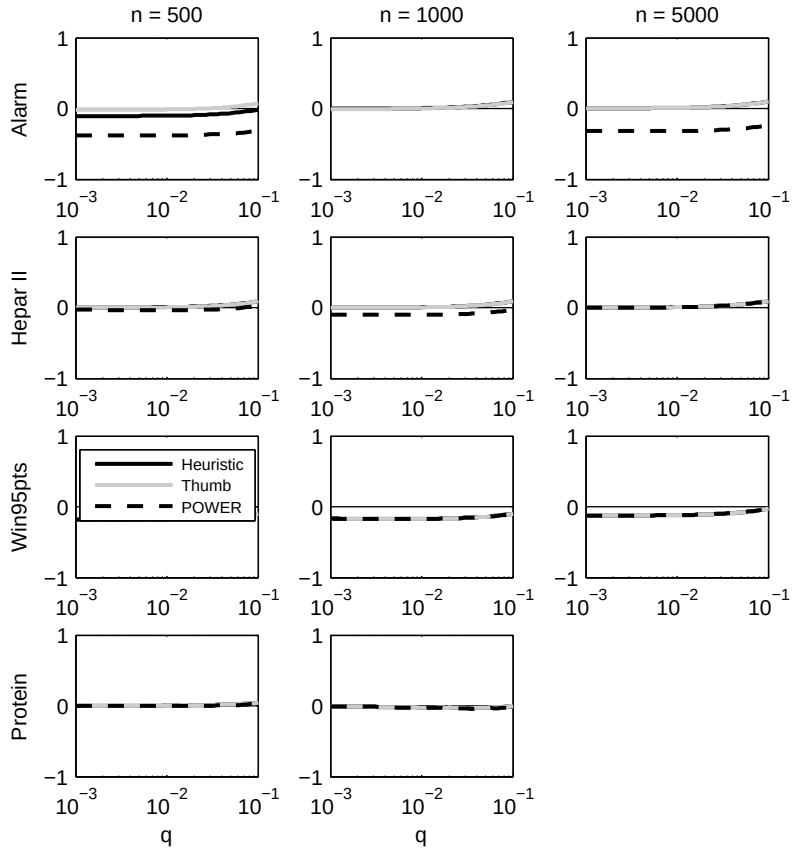



Figure 4.8: $q - \mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$: bias of the Benjamini-Yekutieli [7] False Discovery Rate (FDR) controlling procedure with each FDR threshold for each network (rows), sample size $n$ (columns) and reliability criterion. X-axes are in logarithmic-10 scale. $\mathrm{FDR}(q)$ is the FDR after applying the procedure. *Heuristic*, *Thumb* and *POWER* denote the heuristic power rule, the rule of thumb and the POWER correction reliability criterion respectively. $\pi(t)$ on Alarm for $n = 1000$ and POWER is not shown. For POWER, $q - \mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}})]$ is noticeably negative on Alarm, while it is not or not that much for the heuristic power rule. $q - \mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}})]$ is the same on Win95pts and Protein for all criteria.

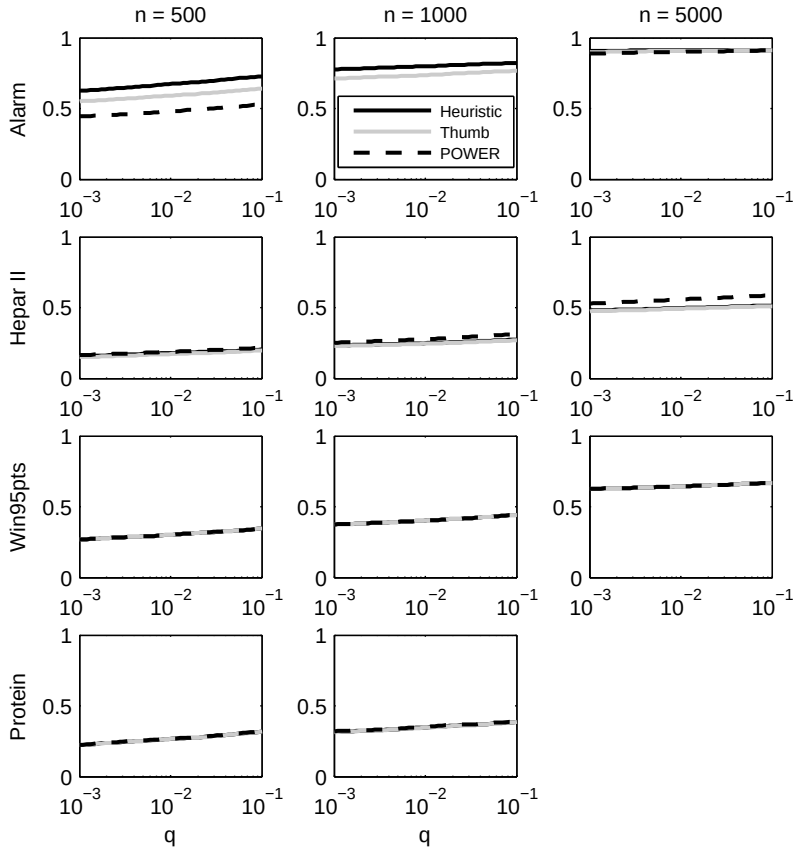



Figure 4.9: $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$: power of the Benjamini-Yekutieli [7] False Discovery Rate (FDR) controlling procedure with each FDR threshold for each network (rows), sample size $n$ (columns) and reliability criterion. X-axes are in logarithmic-10 scale. *Heuristic*, *Thumb* and *POWER* denote the heuristic power rule, the rule of thumb and the POWER correction reliability criterion respectively. $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ on Alarm for $n = 1000$ and POWER is not shown. Compared to the heuristic power rule, the rule of thumb decreases $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ for $n = 500$ on Alarm while POWER decreases $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ for $n = 500$ on Alarm but increases $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ on Hepar II. On Win95pts and Protein all criteria achieve the same power.

### 4.2.2 Varying the power threshold

Using the rule of thumb, Fast [12] demonstrates that decreasing the observations per degree of freedom threshold (thus decreasing the power threshold) decreases the false negatives (i.e., increases the power) but also increases false positives at a much faster rate, resulting in more errors overall. These effects are more pronounced for the smallest sample size used, $n = 500$. Fast [12] does not try varying the power threshold $1 - \beta$ using POWER.

We applied MMPC-skeleton on the 100 samples of the Alarm, Hepar II, Win95pts and Protein networks using the first $n \in \{500, 1000, 5000\}$ observa-

tions and POWER with $\beta \in \{0.05, 0.1, 0.2\}$. For the rest parameters we used the same values as in Section 3.4. The runs on Alarm with $\beta = 0.05$ and $n = 1000$ were aborted because they were taking too much time.

- On three networks, as $\beta$ increases FDR($t$) (Fig. 4.10) decreases. Compared to the heuristic power rule, $\beta = 0.05$ increases FDR($t$) in some cases, $\beta = 0.1$ increases FDR($t$) in one case and $\beta = 0.2$ decreases FDR($t$) in another.
- As $\beta$ increases, $\pi(t)$ (Fig. 4.11) increases on one network, decreases on another and stays the same or about the same on two networks. Compared to the heuristic power rule, $\beta \in \{0.05, 0.1, 0.2\}$ increases $\pi(t)$ on one network but decreases $\pi(t)$ on another.
- Regarding FDR estimation, as $\beta$ increases, $\widehat{\text{FDR}}_{\text{BY}}$ bias (Fig. 4.12) increases on two networks. On Win95pts there is no difference and on Protein almost no difference. For $\beta = 0.05$, $\widehat{\text{FDR}}_{\text{BY}}(t)$ (Fig. 4.7) is not conservative for small $t$ in two cases, while it is or almost is for the heuristic power rule.
- Regarding FDR control: $\pi[\text{t}_\text{q}(\widehat{\text{FDR}}_{\text{BY}})]$ (Fig. 4.14) and $q - \text{FDR}[\text{t}_\text{q}(\widehat{\text{FDR}})]$ (Fig. 4.13) are reflections of $\pi(t)$ and $\widehat{\text{FDR}}_{\text{BY}}$ bias respectively.

The above results are summarized as follows:

- Increasing $\beta$ from 0.05 to 0.2 when using the POWER correction:
  - Decreases FDR on three networks.
  - Increases power on one network but decreases it on another.
- Compared to the heuristic power rule, POWER with $\beta = 0.05$:
  - Increases FDR in some cases.
  - Increases power on one network but decreases it on another.
  - Worsens estimation and control of the FDR in two cases.
- Compared to the heuristic power rule, POWER with $\beta = 0.1$:
  - Increases FDR in one case.
  - Increases power on one network but decreases it on another.
- Compared to the heuristic power rule, POWER with $\beta = 0.2$:
  - Decreases FDR in one case.
  - Increases power on one network but decreases it on another.

Thus, we would not recommend using POWER with $\beta \in \{0.05, 0.1, 0.2\}$ over the heuristic power rule as the reliability criterion.

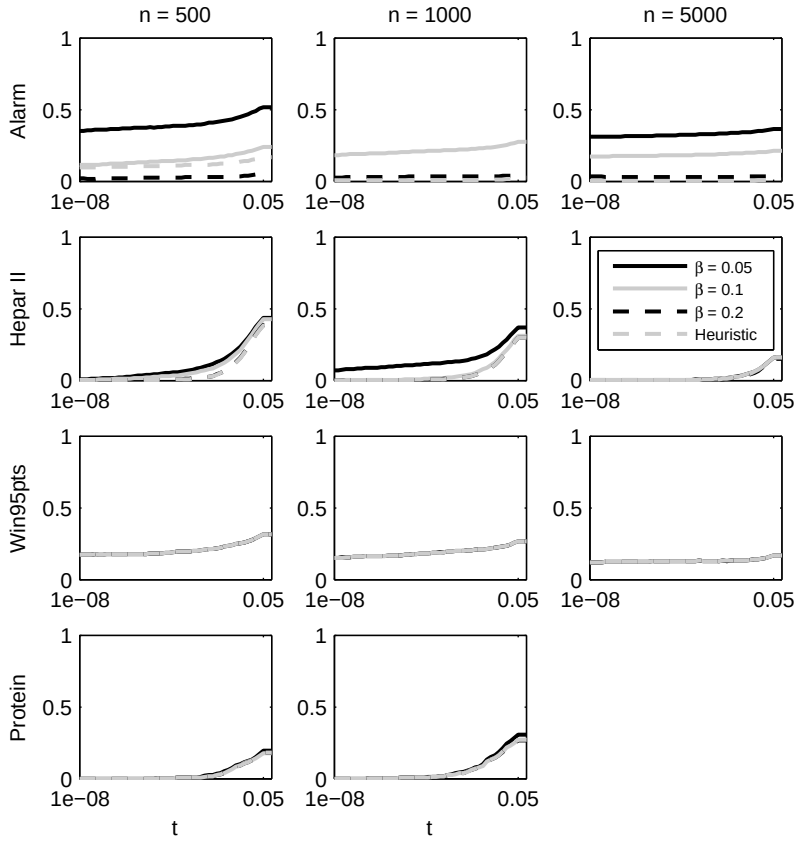



Figure 4.10: FDR($t$): False Discovery Rate (FDR) of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and power threshold $1 - \beta$. X-axes are in logarithmic-10 scale. FDR for Alarm, $\beta = 0.05$ and $n = 1000$ is not shown. *Heuristic* denotes the heuristic power rule. On Alarm, and Hepar II and Protein, as $\beta$ increases FDR($t$) decreases; the effect is more pronounced on Alarm. On Win95pts there is no difference. Compared to the heuristic power rule, $\beta = 0.05$ increases FDR($t$) for $n = 500$ on Alarm, for $n = 1000$ on Hepar II and for $n = 5000$ on Alarm, $\beta = 0.1$ increases FDR($t$) for $n = 1000$ on Alarm and $\beta = 0.2$ decreases FDR($t$) for $n = 500$ on Alarm.

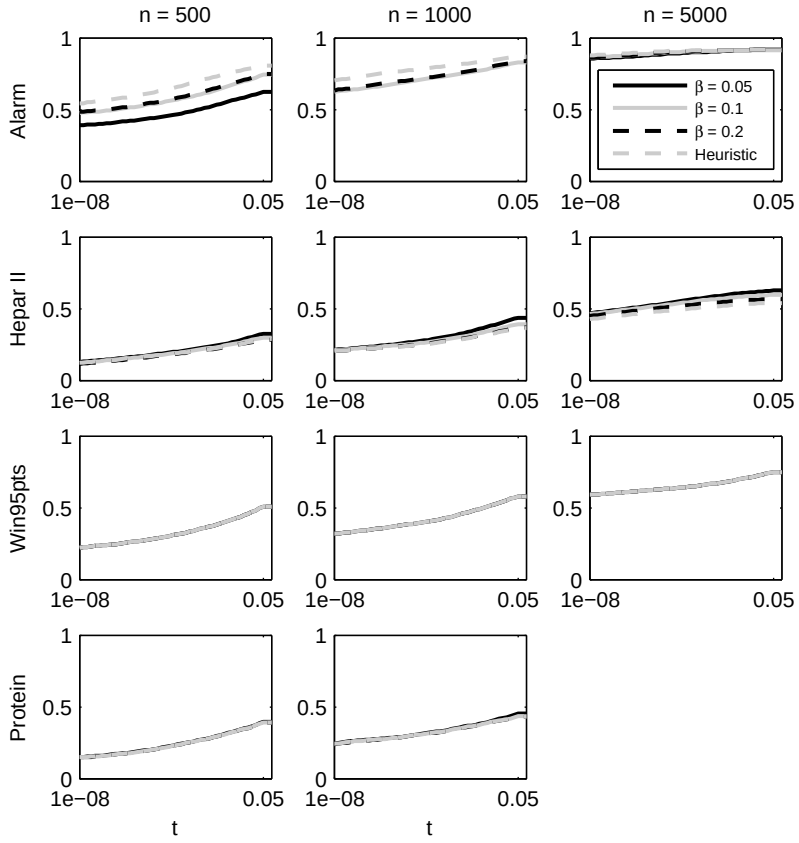



Figure 4.11: $\pi(t)$: power of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and power threshold $1-\beta$. X-axes are in logarithmic-10 scale. Power for Alarm, $\beta = 0.05$ and $n = 1000$ is not shown. *Heuristic* denotes the heuristic power rule. As $\beta$ increases, $\pi(t)$ increases on Alarm but decreases on Hepar II. On Win95pts there is no difference and on Protein almost no difference. Compared to the heuristic power rule, $\beta \in \{0.05, 0.1, 0.2\}$ increases $\pi(t)$ on Hepar II but decreases $\pi(t)$ on Alarm.

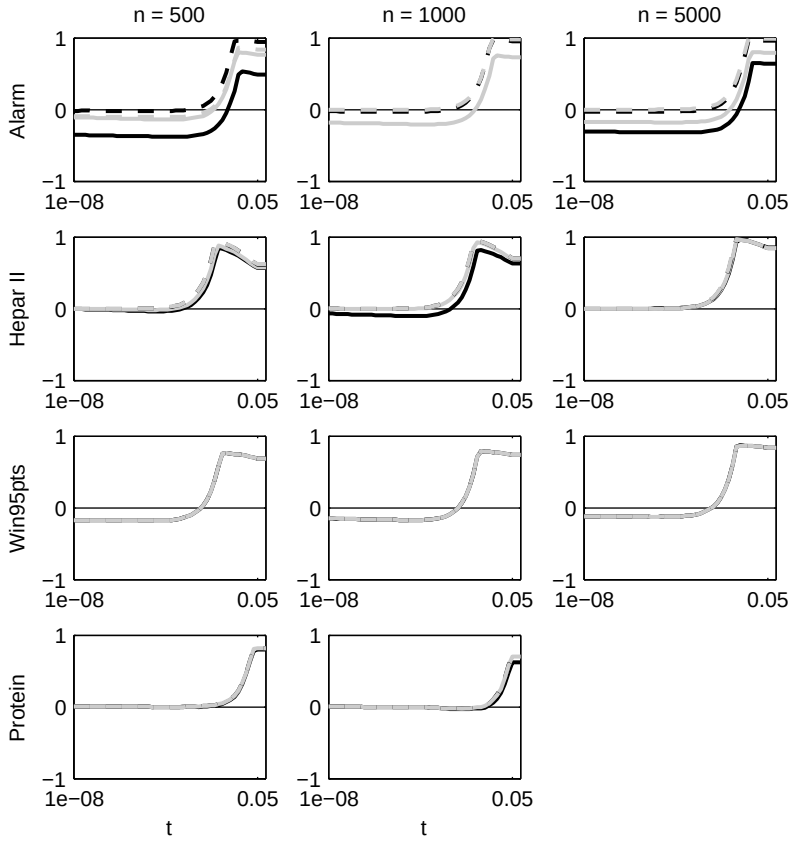



Figure 4.12: $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)] - \mathrm{FDR}(t)$: bias of the False Discovery Rate (FDR) estimator $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ (Benjamini and Yekutieli [7]) of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and power threshold $1-\beta$. X-axes are in logarithmic-10 scale. The legend is the same as in Fig. 4.10. Bias on Alarm, $\beta = 0.05$ and $n = 1000$ is not shown. *Heuristic* denotes the heuristic power rule. As $\beta$ increases, $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ bias increases on Alarm and Hepar II. On Win95pts there is no difference and on Protein almost no difference. For $\beta = 0.05$, the bias of small $t$ is noticeably negative on Alarm, while it is not or not that much for the heuristic power rule.

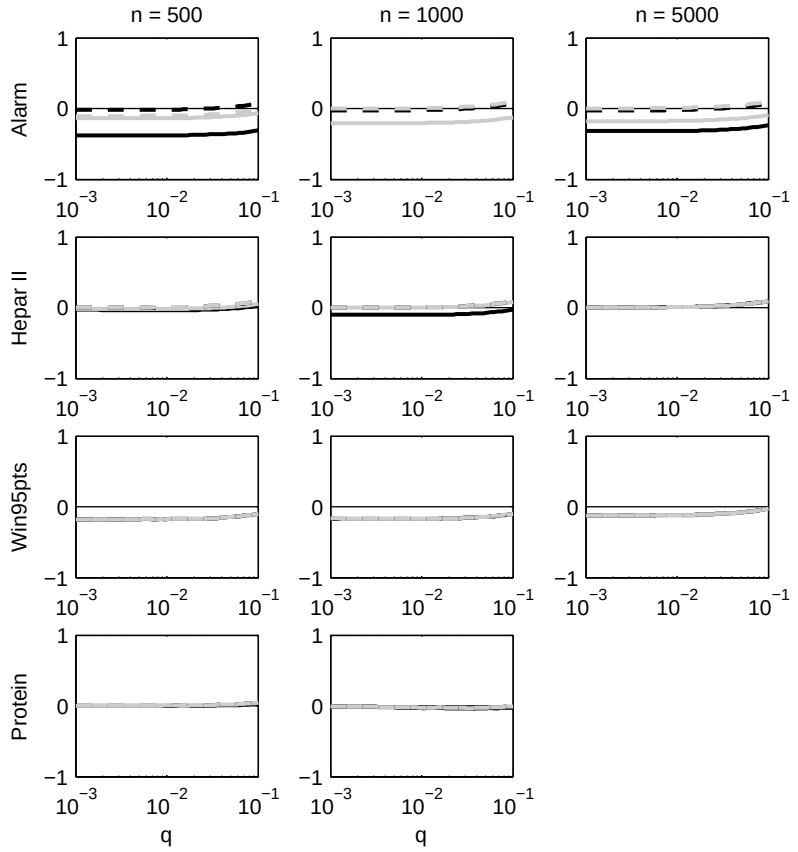



Figure 4.13: $q - \mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}})]$: bias of the Benjamini-Yekutieli [7] False Discovery Rate (FDR) controlling procedure with each FDR threshold for each network (rows), sample size $n$ (columns) and power threshold $1 - \beta$. X-axes are in logarithmic-10 scale. The legend is the same as in Fig. 4.14. $q - \mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}})]$ on Alarm, $\beta = 0.05$ and $n = 1000$ is not shown. As $\beta$ increases, $q - \mathrm{FDR}[\mathrm{t_q}(\widehat{\mathrm{FDR}})]$ increases on Alarm and Hepar II. For $\beta = 0.2$, bias is close to 0 for small $t$, while this is not the case for smaller $\beta$. On Win95pts and Protein there is no difference. For $\beta = 0.05$, bias is negative on Alarm, while it is not or not that much for the heuristic power rule. For $\beta = 0.2$, the bias for $n = 500$ on Alarm is (almost) non-negative while it is negative for the heuristic power rule.

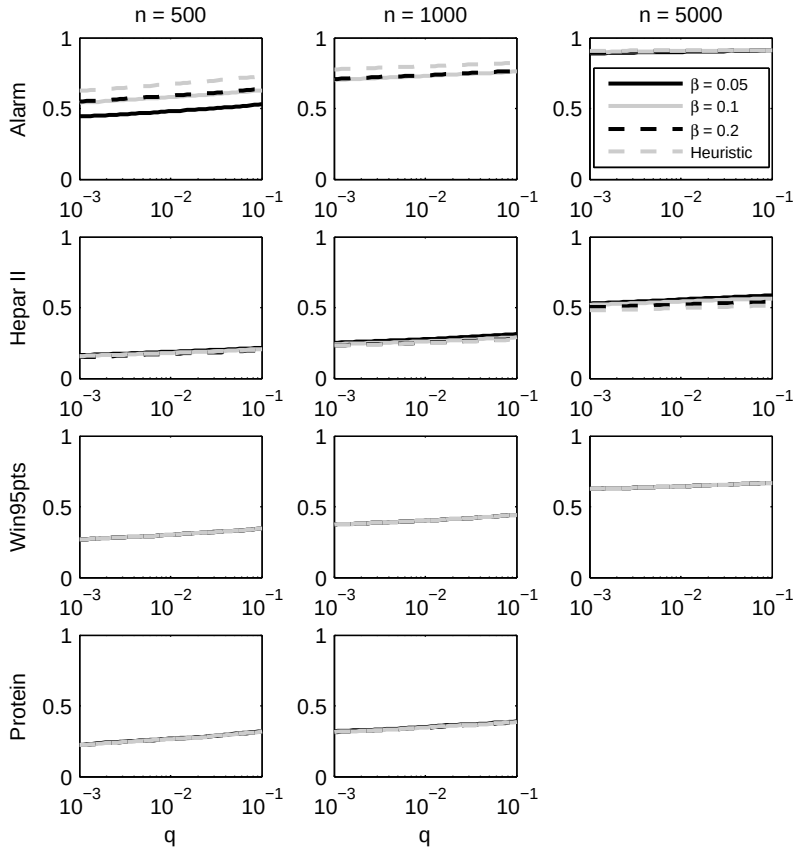



Figure 4.14: $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$: power of the Benjamini-Yekutieli [7] False Discovery Rate (FDR) controlling procedure with each FDR threshold for each network (rows), sample size $n$ (columns) and power threshold $1 - \beta$. X-axes are in logarithmic-10 scale. Power for Alarm, $\beta = 0.05$ and $n = 1000$ is not shown. *Heuristic* denotes the heuristic power rule. As $\beta$ increases, $\pi(t)$ increases on Alarm but decreases on Hepar II. On Win95pts and Protein there is no difference. Compared to the heuristic power rule, $\beta \in \{0.05, 0.1, 0.2\}$ increases $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ on Hepar II but decreases $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ on Alarm.

### 4.2.3 Varying the upper limit on conditioning set cardinality

Using the conditioning cardinality rule alone is a coarse approach to control the power of the tests. Thus, increasing the upper limit *max-k* on conditioning set cardinality corresponds to increasing the power threshold.

We applied MMPC-skeleton on each of the 100 samples of the Alarm, Barley, Hailfinder, Hepar II, Insurance and Protein networks, using the first $n \in \{100, 1000, 10000\}$ observations ($n \in \{100, 1000\}$ for Protein) and the conditioning cardinality rule with *max-k* $\in \{0, 1, 2, 3\}$. For the rest parameters we used the same values as in Section 3.4.

- As *max-k* increases from 0 to 3, FDR($t$) (Fig. 4.15) decreases at a decreasing rate. Compared to the heuristic power rule: *max-k* $= 0$ noticeably decreases FDR($t$) in one case but noticeably increases FDR($t$) in most cases; *max-k* $= 1$ noticeably decreases FDR($t$) in some cases but noticeably increases FDR($t$) in some others; *max-k* $\in \{2, 3\}$ noticeably decreases FDR($t$) in some cases.

- As *max-k* increases from 0 to 3, $\pi(t)$ (Fig. 4.16) decreases at a decreasing rate. Compared to the heuristic power rule: *max-k* $= 0$ increases $\pi(t)$ in all cases; *max-k* $= 1$ noticeably increases $\pi(t)$ in some cases, but also noticeably decreases $\pi(t)$ in some others; *max-k* $\in \{2, 3\}$ noticeably increases $\pi(t)$ in two cases, but also noticeably decreases $\pi(t)$ in some others.

- Regarding FDR estimation: As *max-k* increases from 0 to 3, $\widehat{\text{FDR}}_{\text{BY}}$ is getting more conservative at a decreasing rate (Fig. 4.17). For *max-k* $\in \{0, 1\}$, bias of small or all $t$ is negative in some cases, while bias with the heuristic power rule is non-negative these cases.

- Regarding FDR control: $q-\text{FDR}[\text{t}_{\text{q}}(\widehat{\text{FDR}}_{\text{BY}})]$ (Fig. 4.18) and $\pi[\text{t}_{\text{q}}(\widehat{\text{FDR}}_{\text{BY}})]$ (Fig. 4.19) are reflections of $\pi(t)$ and the bias of $\widehat{\text{FDR}}_{\text{BY}}$ respectively.

The results are summarized as follows:

- When using the conditioning cardinality rule, as *max-k* increases from 0 to 3:
  - FDR and power decrease at a decreasing rate.
  - FDR estimation and control are getting more conservative at a decreasing rate.

- Compared to the heuristic power rule, the conditioning cardinality rule with *max-k* $= 0$:
  - Increases FDR in most cases.
  - Increases power in all cases.
  - Worsens FDR estimation and control in some cases.

- Compared to the heuristic power rule, the conditioning cardinality rule with *max-k* $= 1$:
  - Decreases FDR in some cases but increases it in some others.
  - Increases power in some cases but decreases it in some others.
  - Worsens FDR estimation and control in some cases.

- Compared to the heuristic power rule, the conditioning cardinality rule with *max-k* $\in \{2, 3\}$:
  - Decreases FDR in some cases.
  - Increases power in some cases but decreases it in some others.

Thus, we would not recommend using the conditioning cardinality rule with $max\text{-}k \in \{0, 1, 2, 3\}$ over the heuristic power rule as the reliability criterion, even if the former is more intuitive than the heuristic power rule.

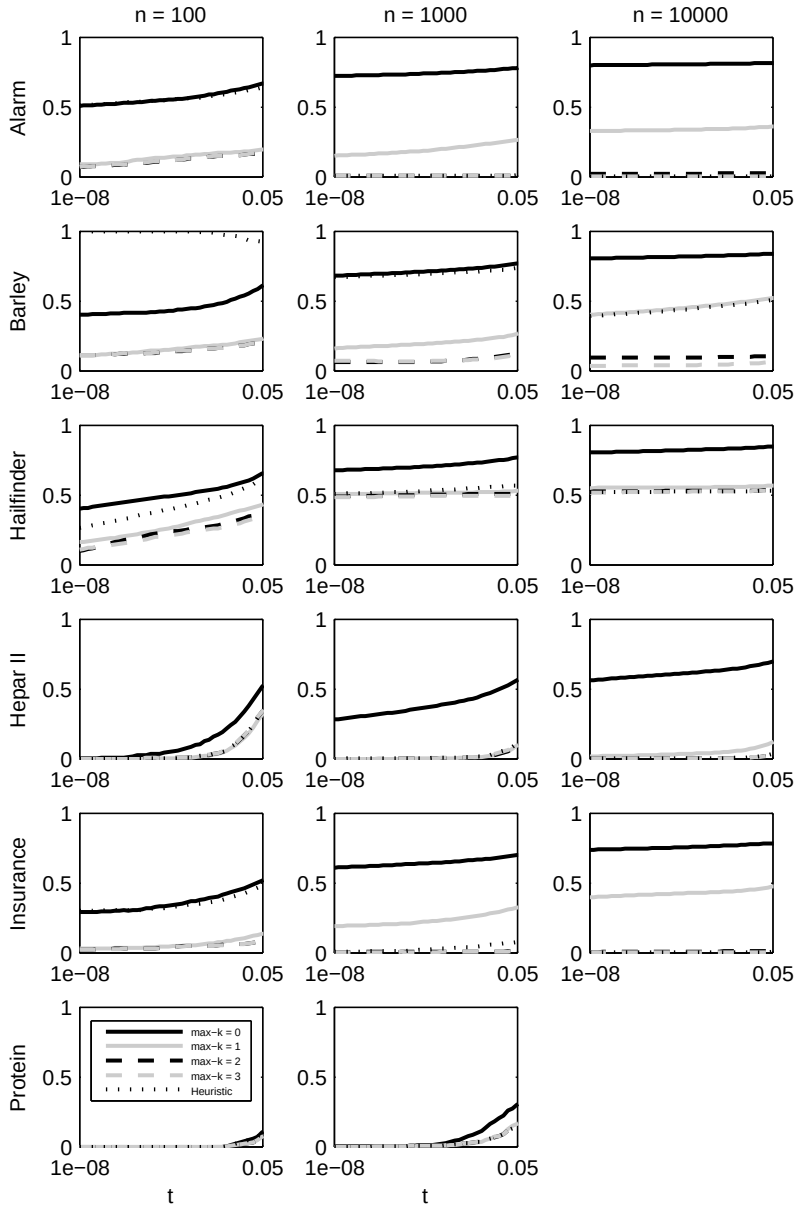



Figure 4.15: FDR($t$): False Discovery Rate (FDR) of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and upper limit $max$-$k$ on conditioning set cardinality. X-axes are in logarithmic-10 scale. *Heuristic* denotes the heuristic power rule. As $max$-$k$ increases from 0 to 3, FDR($t$) decreases at a decreasing rate. Compared to the heuristic power rule: $max$-$k = 0$ noticeably decreases FDR($t$) for $n = 100$ on Barley but noticeably increases FDR($t$) for $n = 100$ on Hailfinder and Hepar II, for $n = 1000$ on all networks except Barley and for $n = 10000$ on all networks except Protein; $max$-$k = 1$ noticeably decreases FDR($t$) for $n = 100$ on Alarm, Barley, Hailfinder and Insurance and for $n = 1000$ on Barley, but noticeably increases FDR($t$) for $n \in \{1000, 10000\}$ on Alarm and Insurance; $max$-$k \in \{2, 3\}$ noticeably decreases FDR($t$) for $n = 100$ on Alarm, Barley, Hailfinder and Insurance and for $n \in \{1000, 10000\}$ on Barley.

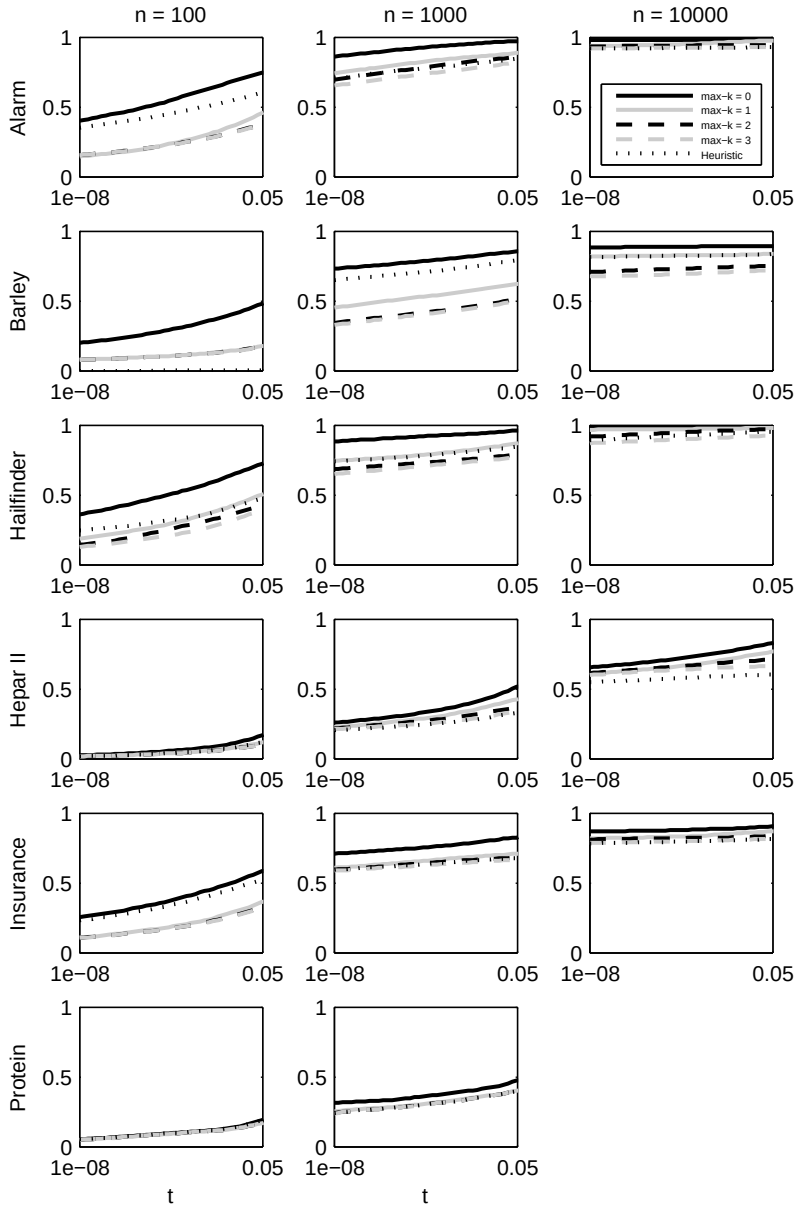



Figure 4.16: $\pi(t)$: power of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and upper limit *max-k* on conditioning set cardinality. X-axes are in logarithmic-10 scale. *Heuristic* denotes the heuristic power rule. As *max-k* increases from 0 to 3, $\pi(t)$ decreases at a decreasing rate. Compared to the heuristic power rule: *max-k* $= 0$ increases $\pi(t)$ in all cases; *max-k* $= 1$ noticeably increases $\pi(t)$ for $n = 100$ on Barley, for $n = 1000$ on Hepar II and for $n = 10000$ on Hepar II and Insurance, but also noticeably decreases $\pi(t)$ for $n = 100$ on Alarm and Insurance and for $n = 1000$ on Barley; *max-k* $\in \{2, 3\}$ noticeably increases $\pi(t)$ for $n = 100$ on Barley and $n = 10000$ on Hepar II, but also noticeably decreases $\pi(t)$ for $n = 100$ on Alarm, Hailfinder and Insurance, for $n = 1000$ on Barley and Hailfinder and for $n = 10000$ on Barley.

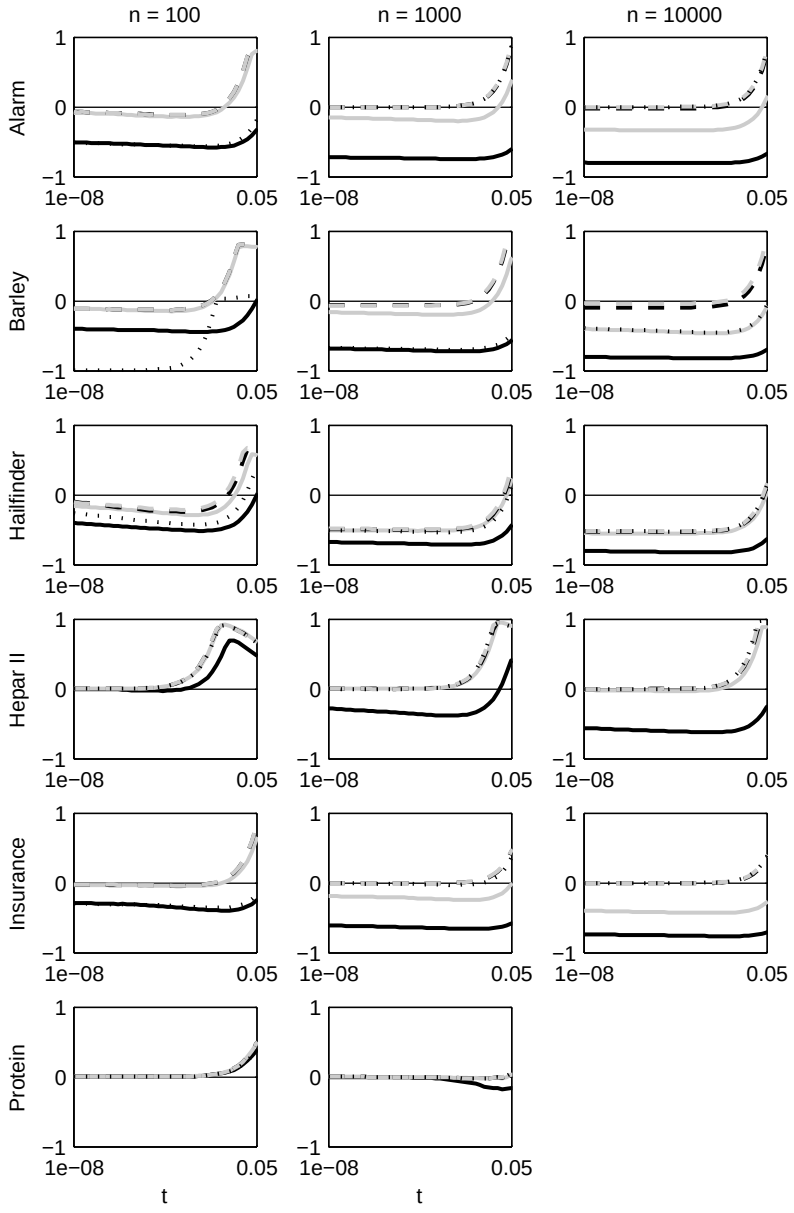



Figure 4.17: $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)] - \mathrm{FDR}(t)$: bias of the False Discovery Rate (FDR) estimator $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ (Benjamini and Yekutieli [7]) at each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and upper limit *max-k* on conditioning set cardinality. X-axes are in logarithmic-10 scale. The legend is the same as in Fig. 4.15. *Heuristic* denotes the heuristic power rule. As *max-k* increases from 0 to 3, $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ bias increases at a decreasing rate. For *max-k* $= 0$, bias of small or all $t$ is negative for $n = 100$ on Barley, for $n = 1000$ on Protein and for $n \in \{1000, 10000\}$ on Alarm, Hepar II and Insurance; for *max-k* $= 1$, bias of small or all $t$ is negative for $n \in \{1000, 10000\}$ on Alarm and Insurance; bias with the heuristic power rule is non-negative in all of the aforementioned cases.

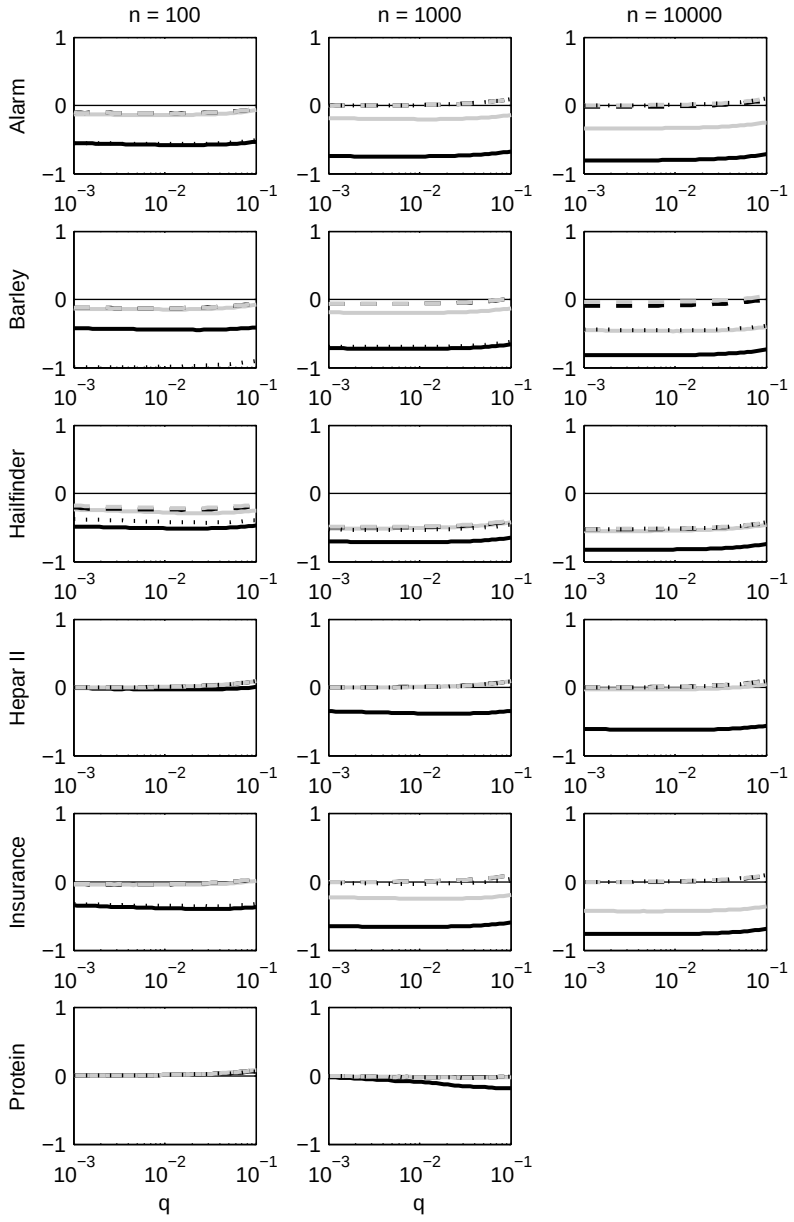



Figure 4.18: $q - \text{FDR}[\text{t}_\text{q}(\widehat{\text{FDR}})]$: bias of the False Discovery Rate (FDR) controlling procedure with FDR estimator $\widehat{\text{FDR}}_{\text{BY}}$ (Benjamini and Yekutieli [7]) with each FDR threshold $q$ for each network (rows), sample size $n$ (columns) and upper limit *max-k* on conditioning set cardinality. X-axes are in logarithmic-10 scale. The legend is the same as in Fig. 4.15. *Heuristic* denotes the heuristic power rule. As *max-k* increases from 0 to 3, $q - \text{FDR}[\text{t}_\text{q}(\widehat{\text{FDR}})]$ increases at a decreasing rate. For *max-k* $= 0$, bias is negative for $n = 1000$ on Protein and for $n \in \{1000, 10000\}$ on Alarm, Hepar II and Insurance; for *max-k* $= 1$, bias is negative for $n \in \{1000, 10000\}$ on Alarm and Insurance; bias with the heuristic power rule is (almost) non-negative in all of the aforementioned cases.

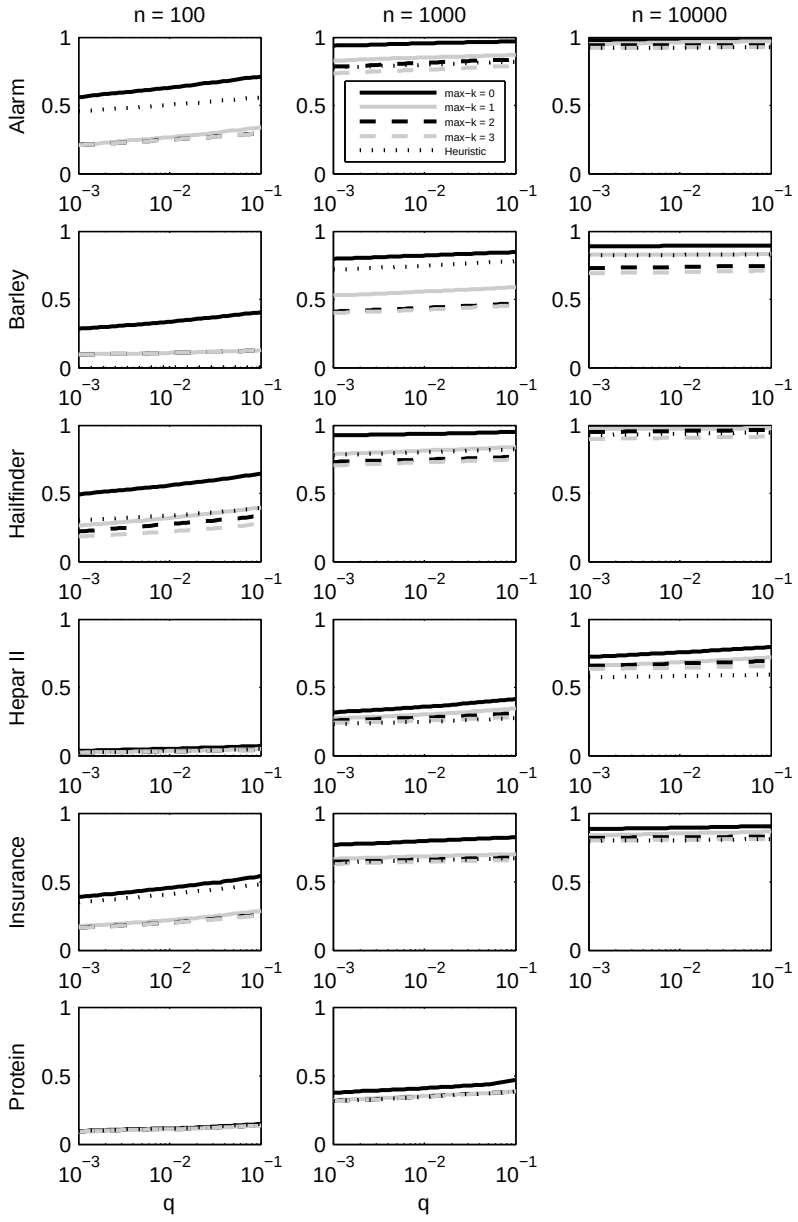



Figure 4.19: $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$: power after applying the False Discovery Rate (FDR) controlling procedure with FDR estimator $\widehat{\mathrm{FDR}} = \widehat{\mathrm{FDR}}_{\mathrm{BY}}$ (Benjamini and Yekutieli [7]) with each FDR threshold $q$ for each network (rows), sample size $n$ (columns) and upper limit *max-k* on conditioning set cardinality. X-axes are in logarithmic-10 scale. *Heuristic* denotes the heuristic power rule. As *max-k* increases from 0 to 3, $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ decreases at a decreasing rate. Compared to the heuristic power rule: *max-k* $= 0$ increases $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ in all cases; *max-k* $= 1$ noticeably increases $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ for $n = 100$ on Barley, for $n = 1000$ on Hepar II and for $n = 10000$ on Hepar II and Insurance, but also noticeably decreases $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ for $n = 100$ on Alarm and Insurance and for $n = 1000$ on Barley; *max-k* $\in \{2, 3\}$ noticeably increases $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ for $n = 100$ on Barley and $n = 10000$ on Hepar II, but also noticeably decreases $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ for $n = 100$ on Alarm, Hailfinder and Insurance, for $n = 1000$ on Barley and Hailfinder and for $n = 10000$ on Barley.

### 4.2.4 Varying the significance level

Fast [12] demonstrates that increasing the significance level $\alpha$ decreases the false negatives (i.e. increases the power) but also increases false positives at a much faster rate, resulting in more errors overall. These effects are more pronounced for the smallest sample size used, $n = 500$.

We applied MMPC-skeleton on each of the 100 samples of the Alarm, Barley, Hailfinder, Hepar II, Insurance and Protein networks, using the first $n \in \{100, 1000, 10000\}$ observations ($n \in \{100, 1000\}$ for Protein) and significance level $\alpha \in \{0.001, 0.01, 0.05, 0.1, 0.2\}$. For the rest parameters we used the same values as in Section 3.4. For each $\alpha$, we only report quantities for $t \leq \alpha$ (see Section 3.3.3). In the figures however, we fix quantities at their value at $\alpha$ for $t > \alpha$, in order to enhance readability.

- We confirm the finding of Fast [12] regarding false positives (Fig. 4.20) in most cases. The effect is more pronounced as $t$ increases or $n$ decreases. $\text{FDR}(t)$ does not necessarily increase at a faster rate than $\pi(t)$ (Fig. 4.21) does when $\alpha$ increases. Compared to $\alpha = 0.05$, $\alpha = 0.2$ increases while $\alpha \in \{0.001, 0.01\}$ decreases the FDR of large $t$ in most cases.

- We also confirm the finding of Fast [12] regarding false negatives (Fig. 4.21) in most cases. The effect are more pronounced as $t$ increases or $n$ decreases. Compared to $\alpha = 0.05$, $\alpha = 0.2$ increases while $\alpha \in \{0.001, 0.01\}$ decreases the power of large $t$ in many cases.

- As $\alpha$ increases, $\widehat{\text{FDR}}_{\text{BY}}$ bias (Fig. 4.22) increases; the effect is more pronounced as $t$ increases. For $\alpha \in \{0.001, 0.01\}$, $\widehat{\text{FDR}}_{\text{BY}}$ is conservative in some cases and not too large as for $\alpha = 0.05$. However, for $\alpha \in \{0.001, 0.01\}$, $\widehat{\text{FDR}}_{\text{BY}}$ is not conservative in some cases where it is for $\alpha = 0.05$.

- The differences in power of the FDR controlling procedure (Fig. 4.23) are small. The differences in bias are neglectable so the corresponding figure is not shown. This is because $\text{E}[\text{t}_\text{q}(\widehat{\text{FDR}}_{\text{BY}})]$ (Fig. 4.24) is about the same for all $\alpha$ and belongs to the lower part of the range of $t$ in Fig. 4.21 and Fig. 4.20, where the difference in $\pi(t)$ is small and $\widehat{\text{FDR}}_{\text{BY}}$ bias is about the same, respectively.

The results are summarized as follows:

- When $\alpha$ increases both FDR and power increase.

- An $\alpha < 0.05$ results in less conservative (but not necessarily conservative) FDR estimation.

- The effect of varying $\alpha \in \{0.001, 0.01, 0.05, 0.1, 0.2\}$ to FDR control is neglectable.

Therefore, we would not recommend using $\alpha \in \{0.001, 0.01, 0.1, 0.2\}$ over $\alpha = 0.05$ when estimating or controlling the FDR.

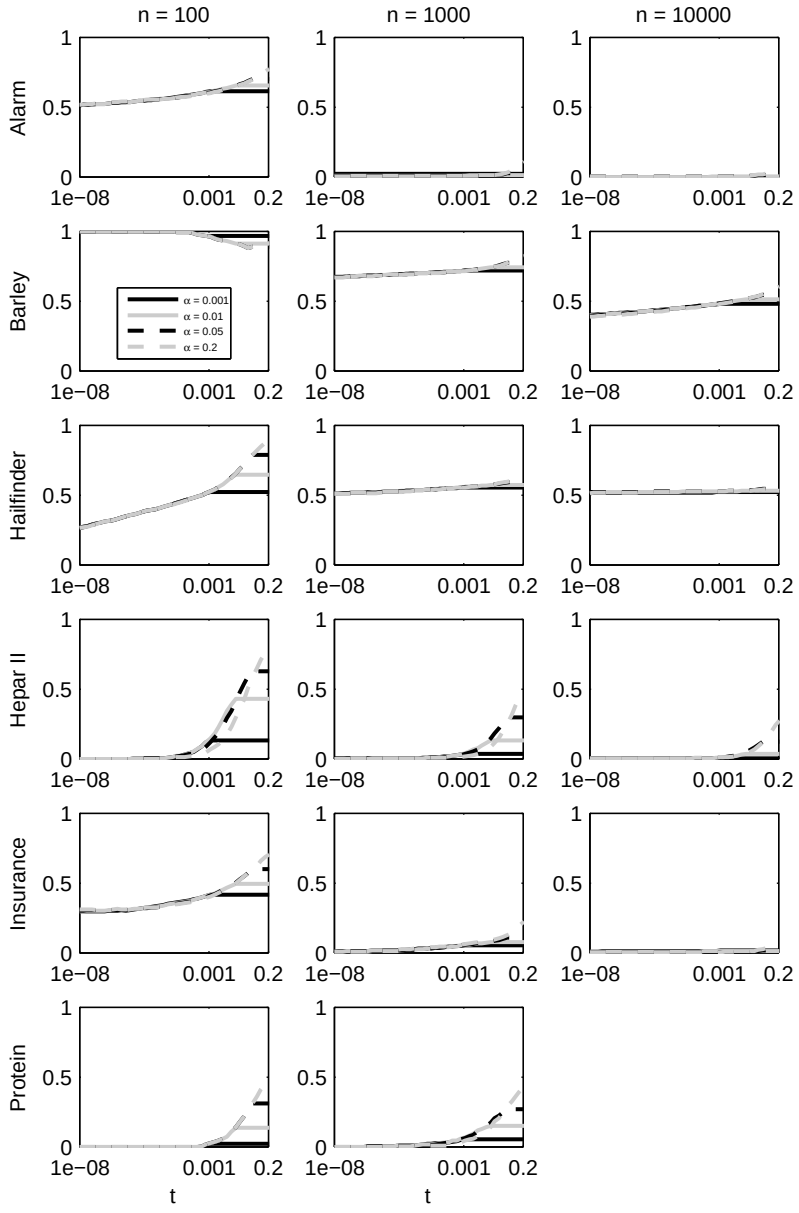



Figure 4.20: FDR($t$): False Discovery Rate (FDR) of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and significance level $\alpha$. X-axes are in logarithmic-10 scale. For each $\alpha$, FDR($t$) is fixed at FDR($\alpha$) for $t > \alpha$ to enhance readability. As $\alpha$ increases, FDR($t$) increases except for $n = 100$ on Barley and for some $t$ on Hepar II; the effect is more pronounced as $t$ increases or $n$ decreases. Compared to $\alpha = 0.05$, $\alpha = 0.2$ noticeably increases the FDR of large $t$ for $n = 100$ on all networks except Barley and for $n = 1000$ on Hepar II and Protein while $\alpha \in \{0.001, 0.01\}$ noticeably decreases the FDR of large $t$ for $n = 100$ on all networks except Barley, for $n = 1000$ on Barley, Hepar II, Insurance and Protein and for $n = 10000$ on Barley and Hepar II.

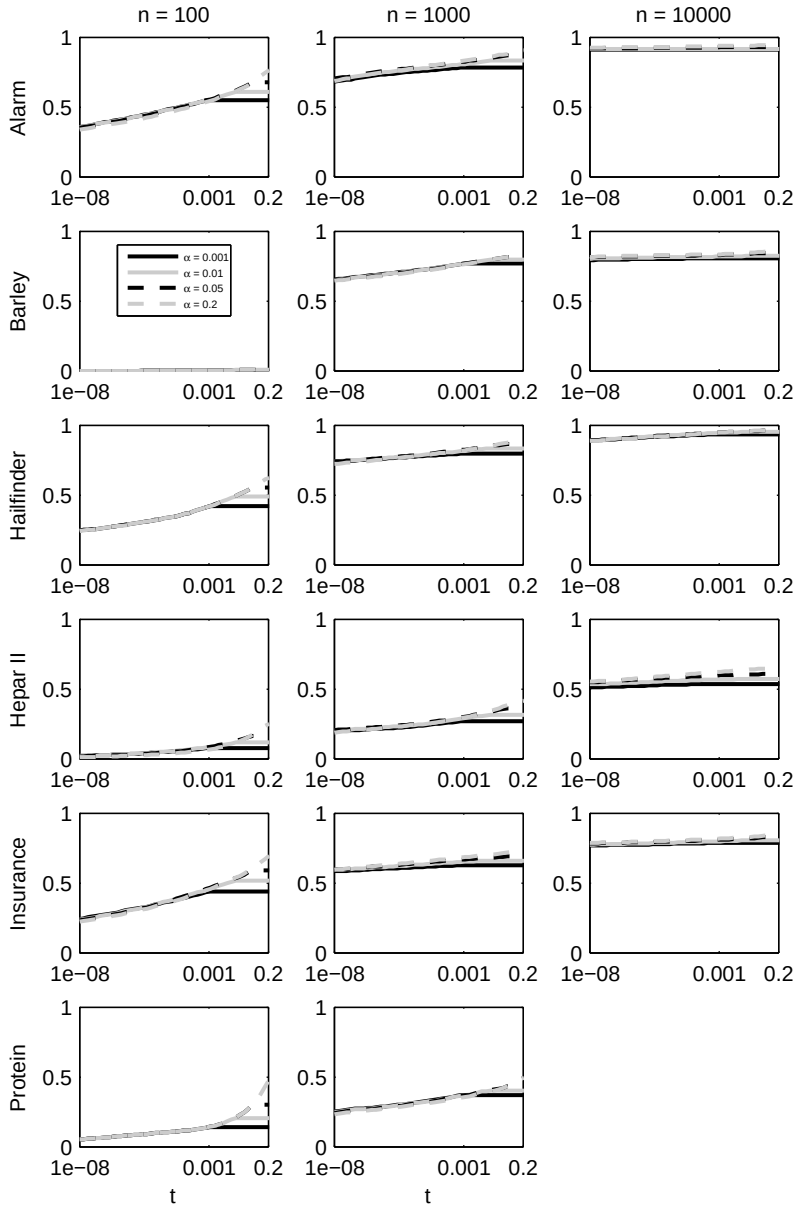



Figure 4.21: $\pi(t)$: power of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and significance level $\alpha$. X-axes are in logarithmic-10 scale. For each $\alpha$, $\pi(t)$ is fixed at $\pi(\alpha)$ for $t > \alpha$ to enhance readability. As $\alpha$ increases $\pi(t)$ increases; the effect is more pronounced as $t$ increases or $n$ decreases. Compared to $\alpha = 0.05$, $\alpha = 0.2$ noticeably increases the power of large $t$ for $n = 100$ on Alarm, Hailfinder, Insurance and Protein, for $n = 1000$ on Insurance and for $n = 10000$ on Hepar II while $\alpha \in \{0.001, 0.01\}$ noticeably decreases the power of large $t$ for $n = 100$ on all networks except Barley, for $n = 1000$ on all networks and for $n = 10000$ on Hepar II.

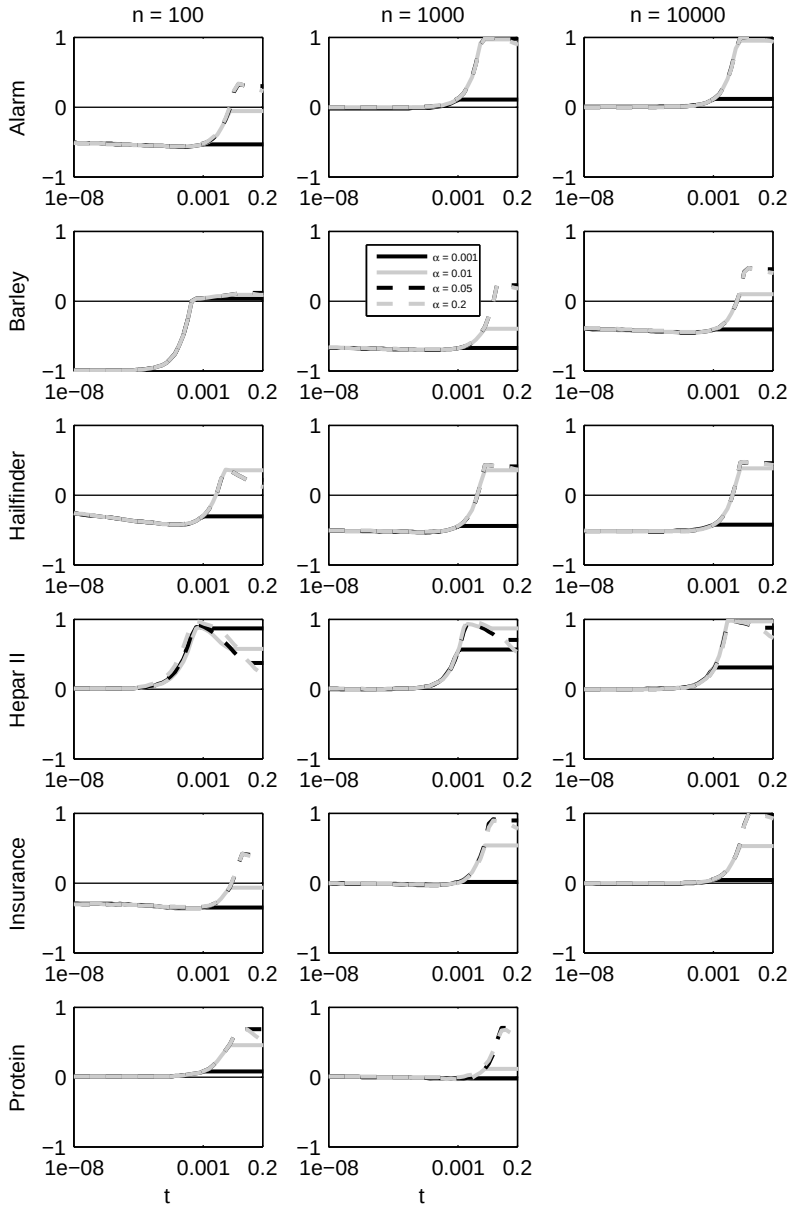



Figure 4.22: $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)] - \mathrm{FDR}(t)$: bias of the False Discovery Rate (FDR) estimator $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ (Benjamini and Yekutieli [7]) at each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and significance level $\alpha$. X-axes are in logarithmic-10 scale. For each $\alpha$, $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)] - \mathrm{FDR}(t)$ is fixed at $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BY}}(\alpha)] - \mathrm{FDR}(\alpha)$ for $t \geq \alpha$ to enhance readability. $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)] - \mathrm{FDR}(t)$ is the skeleton identification bias. As $\alpha$ increases, $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ bias increases except for large $t$ on Hepar II; the effect is more pronounced as $t$ increases. Bias for $\alpha = 0.001$ is positive for $n = 100$ on Protein, for $n = 1000$ on Alarm, Insurance and Protein and for $n = 10000$ on Alarm, Hepar II and Insurance and not too large as for $\alpha = 0.05$. However, bias for $\alpha = 0.001$ is negative for $n = 100$ on Alarm, Hailfinder and Insurance, for $n = 1000$ on Barley and Hailfinder and for $n = 10000$ on Barley and Hailfinder while it is not for $\alpha = 0.05$. Bias for $\alpha = 0.01$ is positive for $n \in 1000$ on Protein and for $n = 10000$ on Barley and Insurance and not too large as for $\alpha = 0.05$. However, bias for $\alpha = 0.01$ is noticeably negative for $n = 1000$ on Barley while it is positive for $\alpha = 0.05$.

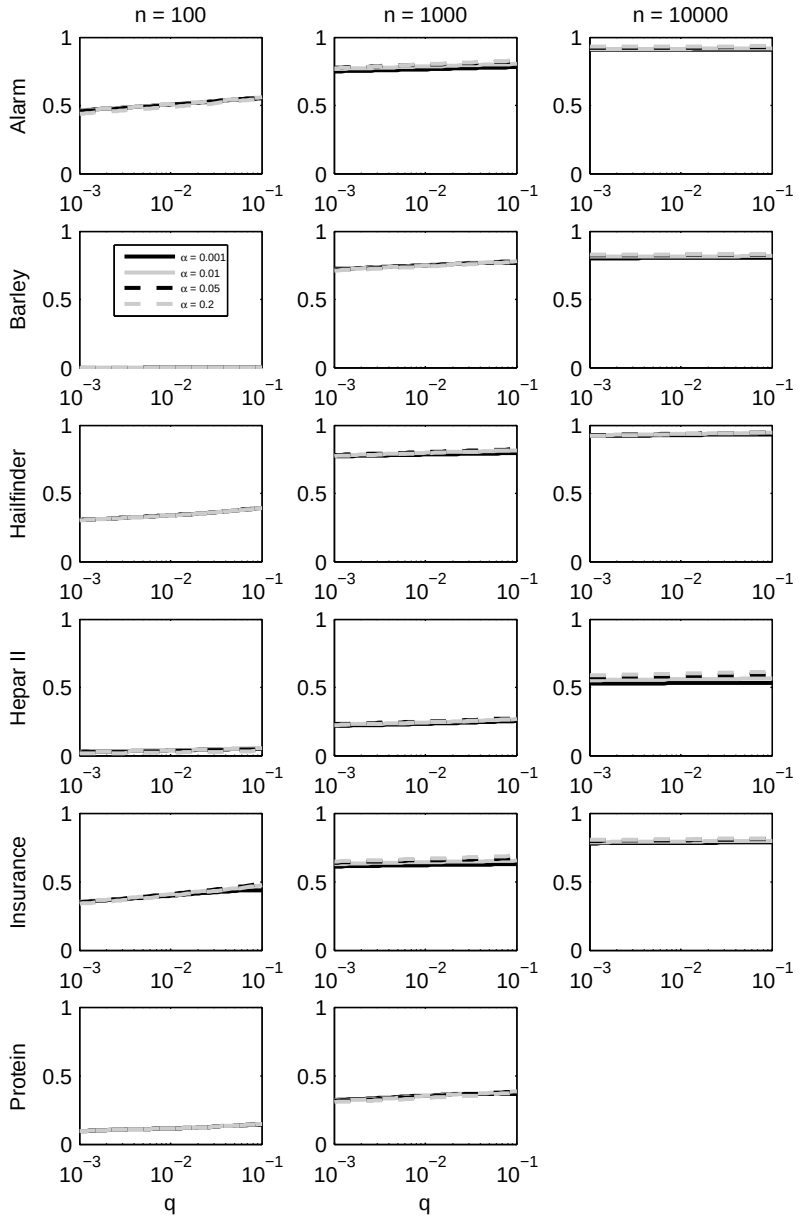



Figure 4.23: $\pi[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$: power of the Benjamini-Hochberg [7] False Discovery Rate (FDR) controlling procedure with each FDR threshold $q$ for each network (rows), sample size $n$ (columns) and significance level $\alpha$. X-axes are in logarithmic-10 scale. Differences are small because the expected p-value threshold returned by the procedure (Fig. 4.24) is about the same and belongs to the lower part of the range of $t$ in Fig. 4.21 where the difference in $\pi(t)$ is small.

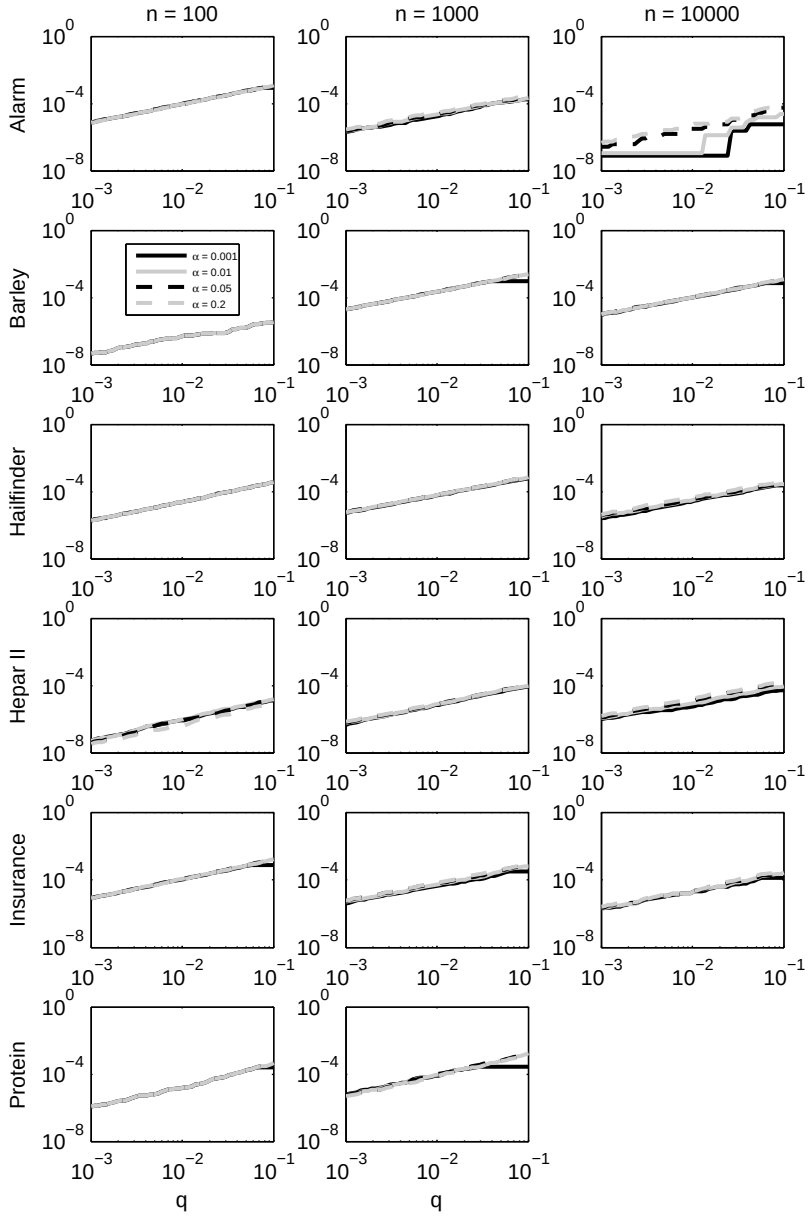



Figure 4.24: $\mathrm{E}[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$: expected p-value threshold returned by the Benjamini-Hochberg [7] False Discovery Rate (FDR) controlling procedure with each FDR threshold $q$ for each network (rows), sample size $n$ (columns), and significance level $\alpha$. All axes are in logarithmic-10 scale. In most cases, for all $\alpha$ $\mathrm{E}[\mathrm{t_q}(\widehat{\mathrm{FDR}}_{\mathrm{BY}}))]$ is about the same and belongs to the lower part of the range of $t$ in Fig. 4.21 and Fig. 4.20, where the difference in $\pi(t)$ is small and $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ bias is about the same, respectively.

### 4.2.5 Varying the test statistic

Fast [12] evaluates the Cochran-Mantel-Haenszel (CMH) test of independence as a means to increase power compared to the usual G test. He demonstrates that the CMH test actually decreases power and increases the FPR on all three networks (Alarm, Insurance and Win95pts) used in the evaluation. Therefore, we do not consider the CMH test.

The p-value of the G test is calculated assuming that the test statistic follows the $\chi^2$ distribution. However, the latter is only asymptotically true. Thus, the calculated p-value is only asymptotically correct. Tsamardinos and Borboudakis [44] demonstrate that the G test with degrees of freedom calculated according to Steck and Jaakkola [36] underestimates the p-value when the sample size is small. Tsamardinos and Borboudakis [44] devise a permutation test that is well-calibrated, i.e., its FPR matches the significance level $\alpha$. We do not consider this test as a means to increase the power of the tests either, because not underestimating the p-value would only decrease the power of the tests.

# Chapter 5

# Relaxing the definition of false discovery

The results of the previous chapter indicate that the goal of accurately estimating and controlling the FDR in all cases using the common FDR estimators may be unrealistic. Thus, we pursue a more realistic goal, that of accurately estimating and controlling the FDR according to a *relaxed* definition of false discovery whose rate is *guaranteed* to be conservatively estimated using an appropriate estimator.

## 5.1 A relaxed definition of false discovery

Recall from Section 4.1.2 that, for the estimation of the FDR to be conservative, only false discoveries need to have an upper-bounded p-value. If we ignore false discoveries *without* a p-value upper bound theoretically guaranteed by Lemma 3.2, then $\widehat{\text{FDR}}_{\text{BY}}$ is a conservative estimator of the proportion of false discoveries *with* such bound among all discoveries, as we prove below. This leads us to our relaxed definition of false discovery:

**Definition 5.1.** *(relaxed false discovery) A relaxed false discovery is a falsely discovered link for which the tests of conditional independence of its ends given the sets of parents of either of them are reliable according to the employed reliability criterion.*

**Theorem 5.1.** *Suppose a Bayesian network $(\mathbb{G}, P)$. Further suppose that an algorithm instantiating Algorithm Template 2.2 is applied on a sample from $P$ and $X - Y$ is a relaxed false discovery. If*

1. *$\mathbb{G}$ and $P$ are faithful to each other,*
2. *all tests attempted by the algorithm are completed and*
3. *completed tests do not yield a false negative result,*

*then $p_{\neg\, Adj(X,Y)}$ is upper-bounded by the maximal among $p_{I_P(\{X\},\{Y\}|\mathbf{Z})}$ for $\mathbf{Z} \in \mathbf{C}_{XY}$:*

$$p_{\neg\, Adj(X,Y)} \leq \max_{\mathbf{Z} \in \mathbf{C}_{XY}} p_{I_P(\{X\},\{Y\}|\mathbf{Z})}$$

*Proof.* The proof follows immediately from Lemma 3.2 and Definition 5.1. □

The corresponding FDR definition is the following:

**Definition 5.2.** *(relaxed False Discovery Rate)*

$$FDR_{relax} \triangleq E\left[\frac{V_{relax}}{R \vee 1}\right] = E\left[\frac{V_{relax}}{R}\middle| R > 0\right] Pr(R > 0)$$

*where $V_{relax}$ is the number of relaxed false discoveries.*

The relaxed FDR is conservatively estimated by $\widehat{\text{FDR}}_{\text{BY}}$ regardless of the dependence among the p-values:

**Theorem 5.2.** *$\widehat{FDR}_{BY}$ is a conservative estimator of the relaxed FDR:*

$$E[\widehat{FDR}_{BY}] \geq FDR_{relax}$$

*Proof.* If we consider only hypotheses with theoretically guaranteed upper-bounded p-values, then $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ is a conservative estimator of the FDR:

$$E\left[\frac{[m-\mathrm{V}_{\overline{\mathrm{relax}}}(t)]\cdot t\cdot\left(\sum_{i=1}^{m-\mathrm{V}_{\overline{\mathrm{relax}}}(t)}\frac{1}{i}\right)}{[\mathrm{R}(t)-\mathrm{V}_{\overline{\mathrm{relax}}}(t)]\vee 1}\right]\geq \mathrm{E}\left[\frac{\mathrm{V}_{\mathrm{relax}}(t)}{[\mathrm{R}(t)-\mathrm{V}_{\overline{\mathrm{relax}}}(t)]\vee 1}\right]$$

where $\mathrm{V}_{\overline{\mathrm{relax}}}(t)$ is defined as:

$$\mathrm{V}_{\overline{\mathrm{relax}}}(t)\triangleq \mathrm{V}(t)-\mathrm{V}_{\mathrm{relax}}(t)$$

By multiplying both sides of the inequality with $\{[\mathrm{R}(t)-\mathrm{V}_{\overline{\mathrm{relax}}}(t)]\vee 1\}/[\mathrm{R}(t)\vee 1]$,

$$E\left[\frac{[m-\mathrm{V}_{\overline{\mathrm{relax}}}(t)]\cdot t\cdot\left(\sum_{i=1}^{m-\mathrm{V}_{\overline{\mathrm{relax}}}(t)}\frac{1}{i}\right)}{\mathrm{R}(t)\vee 1}\right]\geq \mathrm{E}\left[\frac{\mathrm{V}_{\mathrm{relax}}(t)}{\mathrm{R}(t)\vee 1}\right] \tag{5.1}$$

For $\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)$ it holds that

$$\begin{aligned}\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)&=\mathrm{E}\left[\frac{m\cdot t\cdot\left(\sum_{i=1}^{m}\frac{1}{i}\right)}{\mathrm{R}(t)\vee 1}\right]\\&\geq \mathrm{E}\left[\frac{[m-\mathrm{V}_{\overline{\mathrm{relax}}}(t)]\cdot t\cdot\left(\sum_{i=1}^{m-\mathrm{V}_{\overline{\mathrm{relax}}}(t)}\frac{1}{i}\right)}{\mathrm{R}(t)\vee 1}\right]\end{aligned} \tag{5.2}$$

From inequalities 5.1 and 5.2 above,

$$\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)\geq \mathrm{E}\left[\frac{\mathrm{V}_{\mathrm{relax}}(t)}{\mathrm{R}(t)\vee 1}\right]=\mathrm{FDR}_{\mathrm{relax}}(t)$$

□

$\widehat{\mathrm{FDR}}_{\mathrm{BH}}$ can be proved to be a conservative estimator of the relaxed FDR in a similar fashion, assuming independent null p-values or positive regression dependence of the p-values on the null p-values.

A similar relaxed definition of false discovery is also introduced by Tsamardinos and Brown [45], although no theoretical proof regarding the conservative estimation of its rate is given there.

## 5.2 Experimental results

We computed the relaxed FDR for the runs in Section 3.4 and used the computations of $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ in Section 4.1.1 to evaluate $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ in estimation and strong control of the relaxed FDR this time. We calculated $V_{\mathrm{relax}}$ by comparing the output skeleton not with the true one as we did for $V$, but with the one that has the same links as the true skeleton plus every non-link for which a test of conditional independence of its ends given the sets of parents of either of them is unreliable according to the employed reliability criterion.

- As $n$ increases, $\mathrm{FDR}(t)$ and $\mathrm{FDR}_{\mathrm{relax}}(t)$ get closer (Fig. 5.1). For this reason, $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ bias curves when estimating and controlling the relaxed FDR also get closer to each other (Fig. 5.2 and 5.3).

- For $n = 100$ on four networks and for $n \in \{1000, 10000\}$ on one of them, the relaxed FDR is conservatively estimated and strongly controlled with $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$, while the FDR is not.

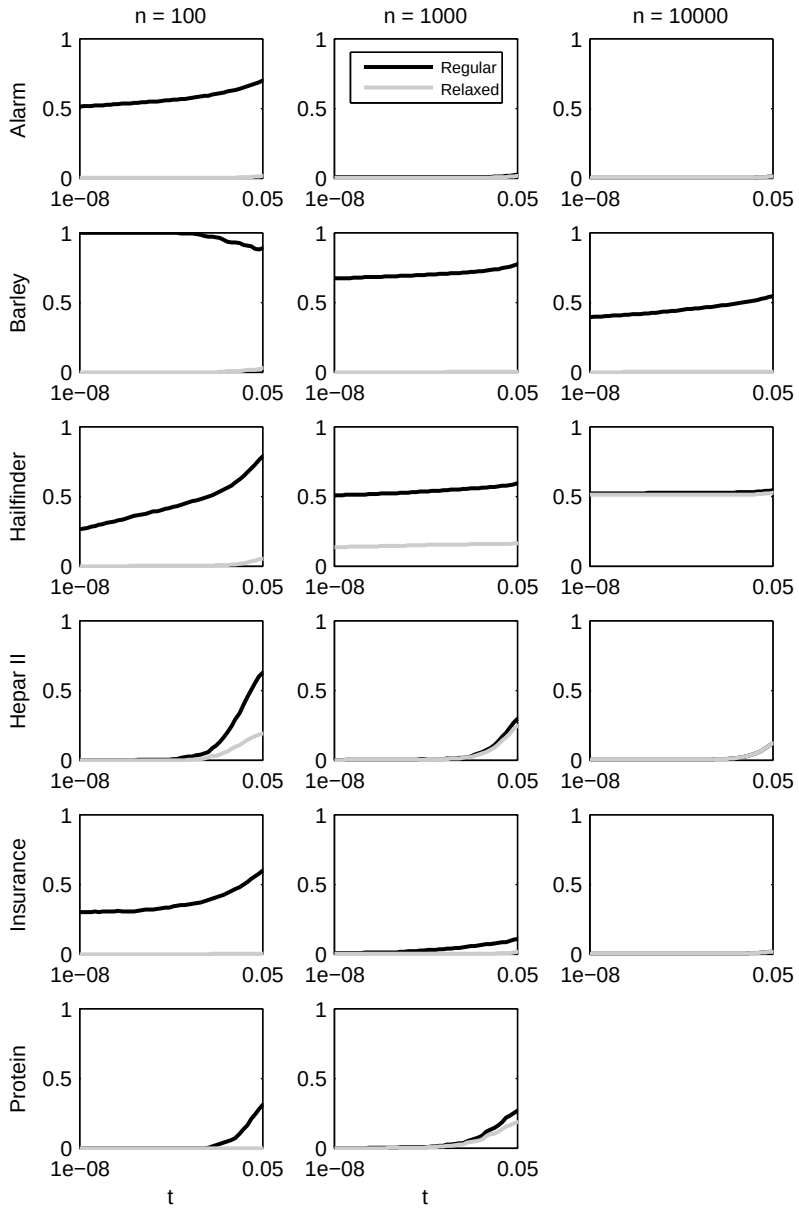



Figure 5.1: FDR($t$): False Discovery Rate (FDR) of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and definition of false discovery. X-axes are in logarithmic-10 scale. $\mathrm{FDR}_{\mathrm{relax}}$ denotes the rate of relaxed false discoveries (see text for the definition). As $n$ increases, $\mathrm{FDR}_{\mathrm{relax}}(t)$ and FDR($t$) get closer.

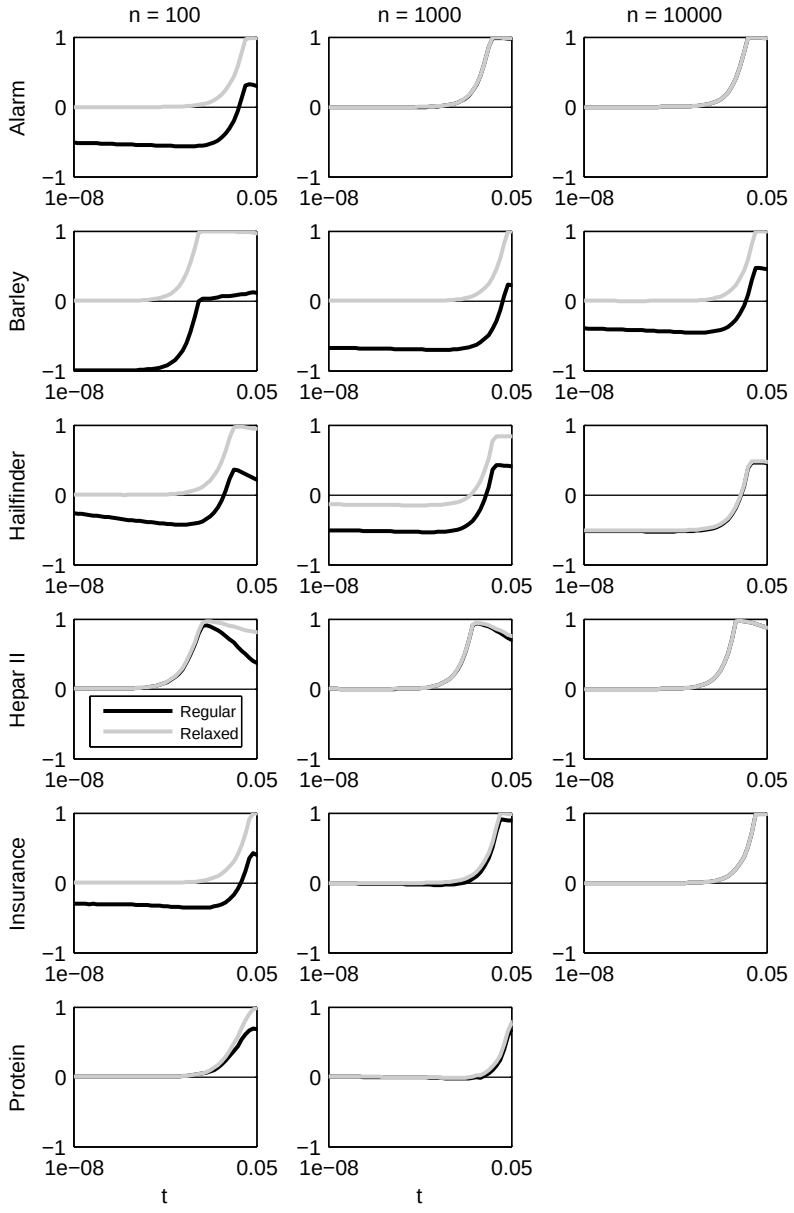



Figure 5.2: $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)] - \mathrm{FDR}(t)$: bias of the False Discovery Rate (FDR) estimator $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ (Benjamini and Yekutieli [7]) of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and definition of false discovery. X-axes are in logarithmic-10 scale. $\mathrm{FDR}_{\mathrm{relax}}$ denotes the rate of relaxed false discoveries (see text for the definition). As $n$ increases, $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)] - \mathrm{FDR}(t)$ and $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)] - \mathrm{FDR}_{\mathrm{relax}}(t)$ get closer. For $n = 100$ on Alarm, Barley, Hailfinder and Insurance and for $n \in \{1000, 10000\}$ on Barley, $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)] - \mathrm{FDR}_{\mathrm{relax}}(t)$ is non-negative (except for some small invisible minor discrepancies) while $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)] - \mathrm{FDR}(t)$ is.

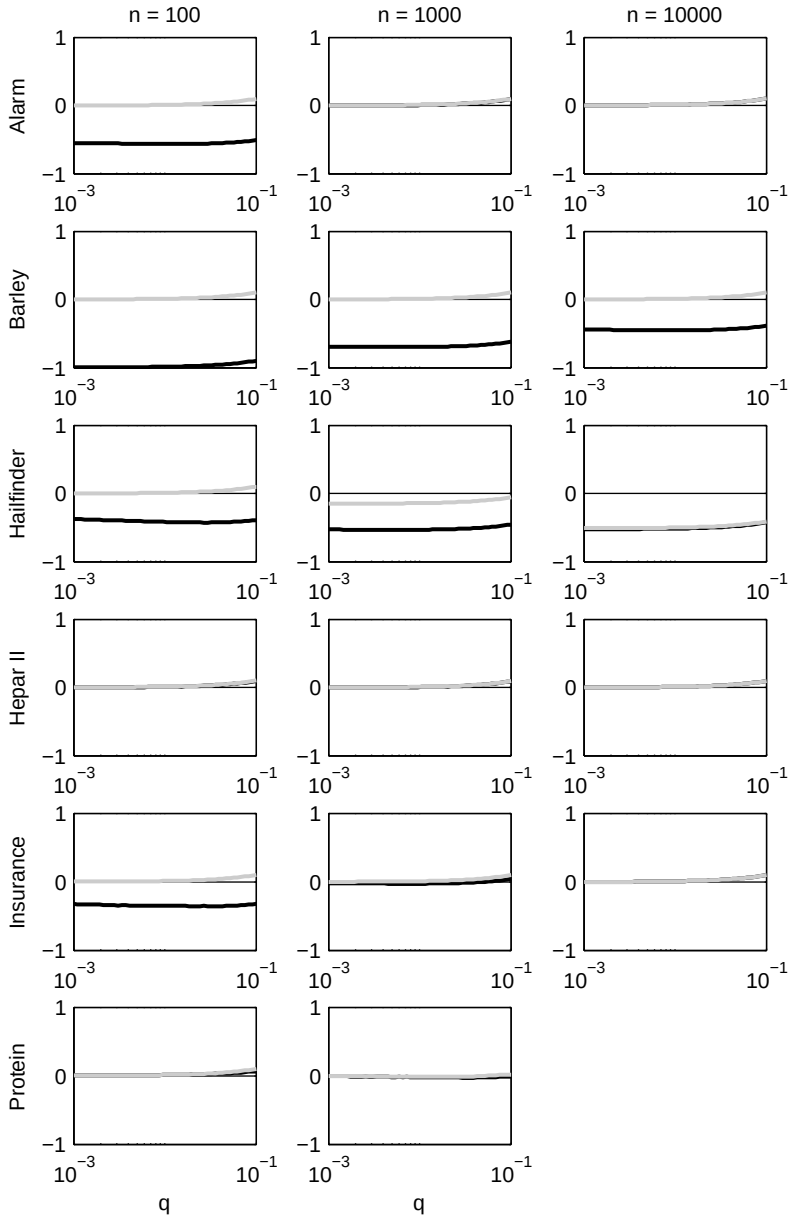



Figure 5.3: $q - FDR[t_q(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$: bias of the Benjamini and Yekutieli [7] False Discovery Rate (FDR) controlling procedure with each FDR threshold $q$ for each network (rows), sample size $n$ (columns) and definition of false discovery. X-axes are in logarithmic-10 scale. The legend is the same as in Fig. 5.2. $\mathrm{FDR}_{\mathrm{relax}}$ denotes the rate of relaxed false discoveries (see text for the definition). For $n = 100$ on Alarm, Barley, Hailfinder and Insurance and for $n \in \{1000, 10000\}$ on Barley, $q - \mathrm{FDR}_{\mathrm{relax}}[t_q(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ is non-negative (except for some small invisible minor discrepancies) while $q - FDR[t_q(\widehat{\mathrm{FDR}}_{\mathrm{BY}})]$ is.

## 5.3 Varying the upper limit on conditioning set cardinality

The relaxed definition of false discovery is intuitive only if the reliability criterion is intuitive. Such a criterion is the conditioning cardinality rule. When employing this rule, a relaxed false discovery is a falsely discovered link whose ends have *max-k* parents at most.

We computed the relaxed FDR for the runs of Section 3.4 and used the computed values of $\widehat{\text{FDR}}_{\text{BY}}$ to evaluate $\widehat{\text{FDR}}_{\text{BY}}$ in estimation and control of the relaxed FDR.

- As *max-k* increases from 0 to 3, $\text{FDR}_{\text{relax}}(t)$ (Fig. 5.1) increases at a decreasing rate. Compared to the heuristic power rule: *max-k* $= 0$ decreases $\text{FDR}_{\text{relax}}(t)$ in some cases; *max-k* $= 1$ decreases $\text{FDR}_{\text{relax}}(t)$ in two cases but increases it in some others; *max-k* $\in \{2, 3\}$ also increases $\text{FDR}_{\text{relax}}(t)$ in the previous cases.

- As *max-k* increases from 0 to 3, the evolution of $\widehat{\text{FDR}}_{\text{BY}}$ bias (Fig. 5.2 and 5.3) is inconsistent among the cases. For *max-k* $= 0$, bias is non-negative or slightly negative in two cases, while it is noticeably negative for the heuristic power rule. For *max-k* $= 1$, bias is noticeably negative in one case, while it is non-negative or slightly negative for the heuristic power rule. For *max-k* $\in \{2, 3\}$, bias is noticeably negative in two cases, while it is non-negative or slightly negative for the heuristic power rule.

The above results, together with the results of Section 4.2.3, are summarized as follows:

- When using the conditioning cardinality rule, as *max-k* increases from 0 to 3:
  - FDR and power decrease at a decreasing rate.
  - Relaxed FDR increases at a decreasing rate.
  - FDR estimation and control are getting more conservative at a decreasing rate.

- Compared to the heuristic power rule, the conditioning cardinality rule with *max-k* $= 0$:
  - Increases FDR in most cases.
  - Decreases the relaxed FDR in some cases.
  - Increases power in all cases.
  - Worsens FDR estimation and control in some cases.
  - Improves relaxed FDR estimation and control in two cases.

- Compared to the heuristic power rule, the conditioning cardinality rule with *max-k* $= 1$:
  - Decreases FDR in some cases but increases it in some others.

  - Decreases the relaxed FDR in two cases but increases it in some others.
  - Increases power in some cases but decreases it in some others.
  - Worsens FDR estimation and control in some cases.
  - Worsens relaxed FDR estimation and control in one case.

- Compared to the heuristic power rule, the conditioning cardinality rule with $max\text{-}k \in \{2, 3\}$:
  - Decreases FDR in some cases.
  - Increases the relaxed FDR in some cases.
  - Increases power in some cases but decreases it in some others.
  - Worsens relaxed FDR estimation and control in two cases.

Thus, we would not recommend using the conditioning cardinality rule with $max\text{-}k \in \{0, 1, 2, 3\}$ over the default reliability criterion when utilizing the relaxed definition of the FDR, even if the former criterion is more intuitive than the heuristic power rule.

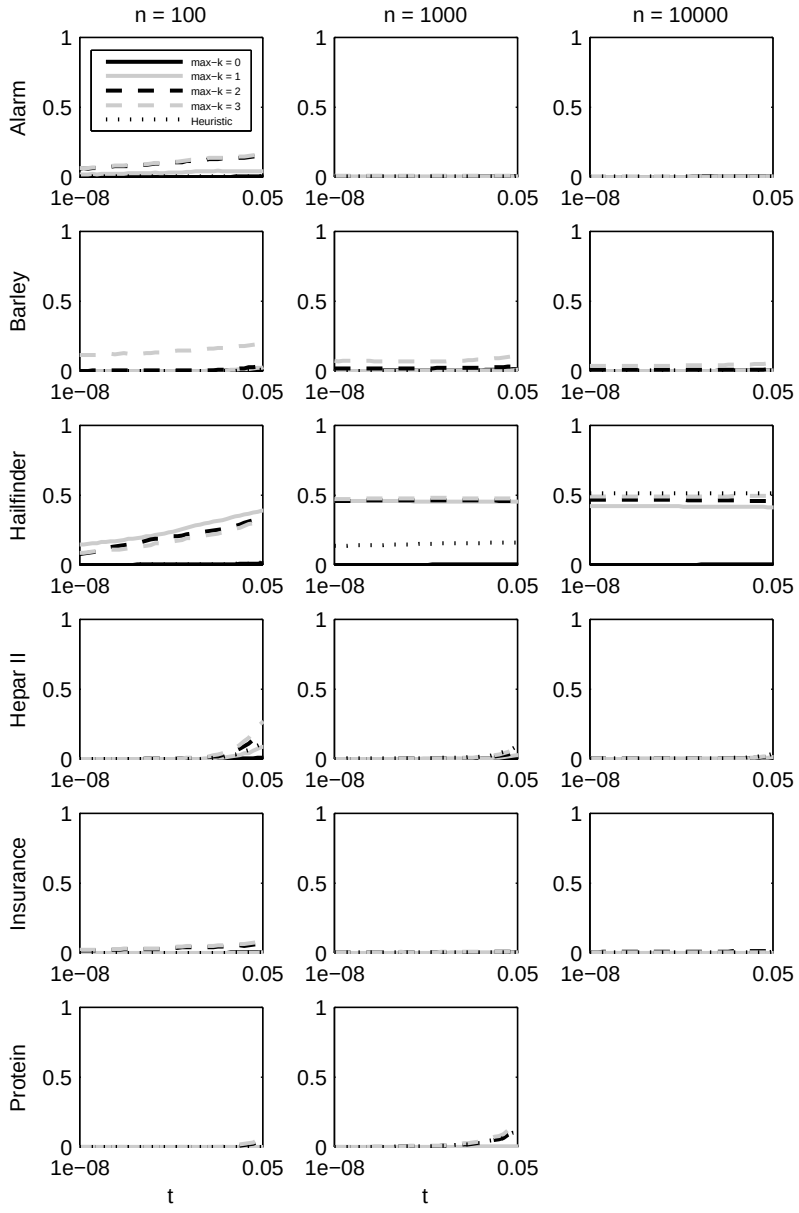



Figure 5.4: $\mathrm{FDR}_{\mathrm{relax}}(t)$: relaxed False Discovery Rate (FDR) of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and upper limit *max-k* on conditioning set cardinality. X-axes are in logarithmic-10 scale. *Heuristic* denotes the heuristic power rule. As *max-k* increases from 0 to 3, $\mathrm{FDR}_{\mathrm{relax}}(t)$ increases at a decreasing rate, except on Hailfinder for $n = 100$. Compared to the heuristic power rule: *max-k* $= 0$ decreases $\mathrm{FDR}_{\mathrm{relax}}(t)$ for $n = 100$ on Hepar II, for $n = 1000$ on Hailfinder, Hepar II and Protein and for $n = 10000$ on Hailfinder; *max-k* $= 1$ decreases $\mathrm{FDR}_{\mathrm{relax}}(t)$ for $n = 1000$ on Protein and for $n = 10000$ on Hailfinder but increases it for $n = 100$ on Alarm and Hailfinder and for $n = 1000$ on Hailfinder; *max-k* $\in \{2, 3\}$ also increases $\mathrm{FDR}_{\mathrm{relax}}(t)$ in the previous cases.

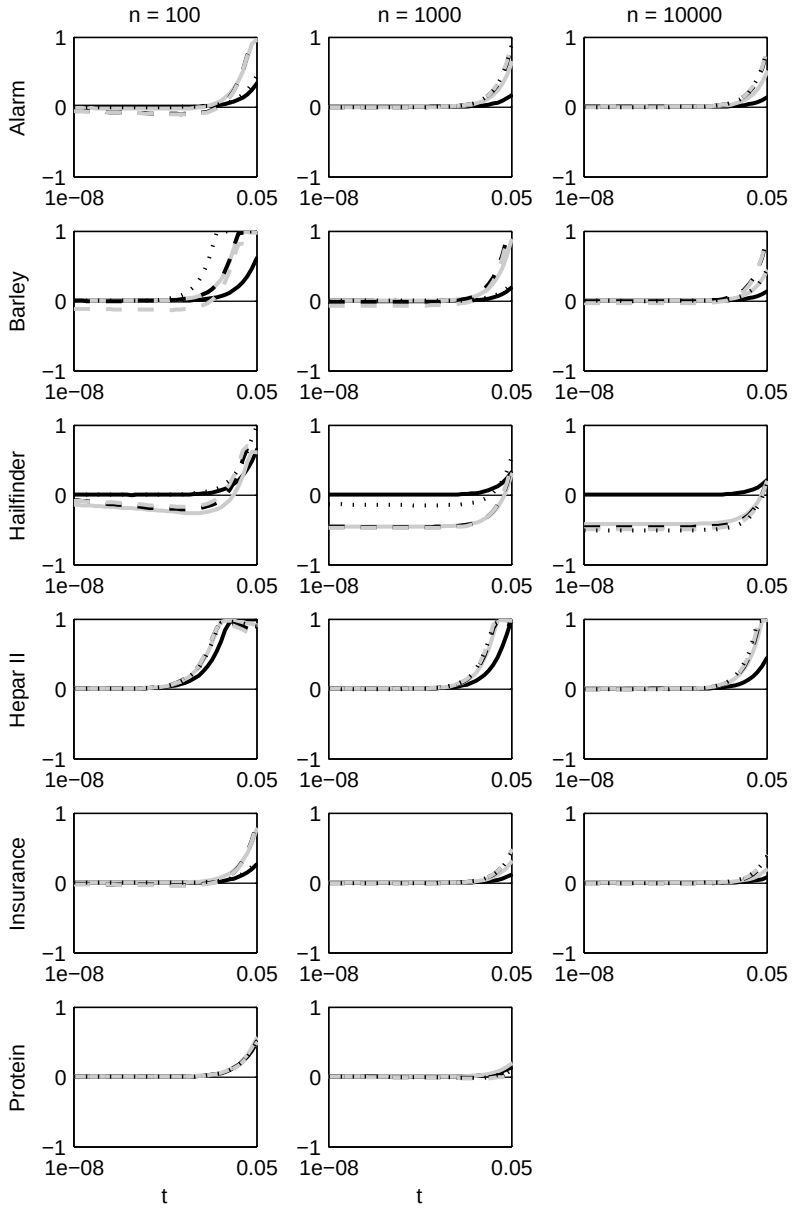



Figure 5.5: $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)] - \mathrm{FDR}_{\mathrm{relax}}(t)$: bias of the False Discovery Rate (FDR) estimator $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ (Benjamini and Yekutieli [7]) when estimating the relaxed FDR ($\mathrm{FDR}_{\mathrm{relax}}$, see text for definition) of each p-value threshold $t$ for each network (rows), sample size $n$ (columns) and upper limit *max-k* on conditioning set cardinality. X-axes are in logarithmic-10 scale. *Heuristic* denotes the heuristic power rule. The legend is the same as in Fig. 5.4. As *max-k* increases from 0 to 3, the evolution of bias is inconsistent among the cases and depends on the relative evolution of $\mathrm{E}[\widehat{\mathrm{FDR}}_{\mathrm{BY}}(t)]$ (figure not included) and $\mathrm{FDR}_{\mathrm{relax}}(t)$ (Fig. 5.4). For *max-k* $= 0$, bias is non-negative or slightly negative for $n \in \{1000, 10000\}$ on Hailfinder, while it is noticeably negative for the heuristic power rule. For *max-k* $= 1$, the bias of small $t$ is noticeably negative for $n = 100$ on Hailfinder, while it is non-negative or slightly negative for the heuristic power rule. For *max-k* $\in \{2, 3\}$, the bias of small $t$ is noticeably negative for $n = 100$ on Alarm and Hailfinder, while it is non-negative or slightly negative for the heuristic power rule.

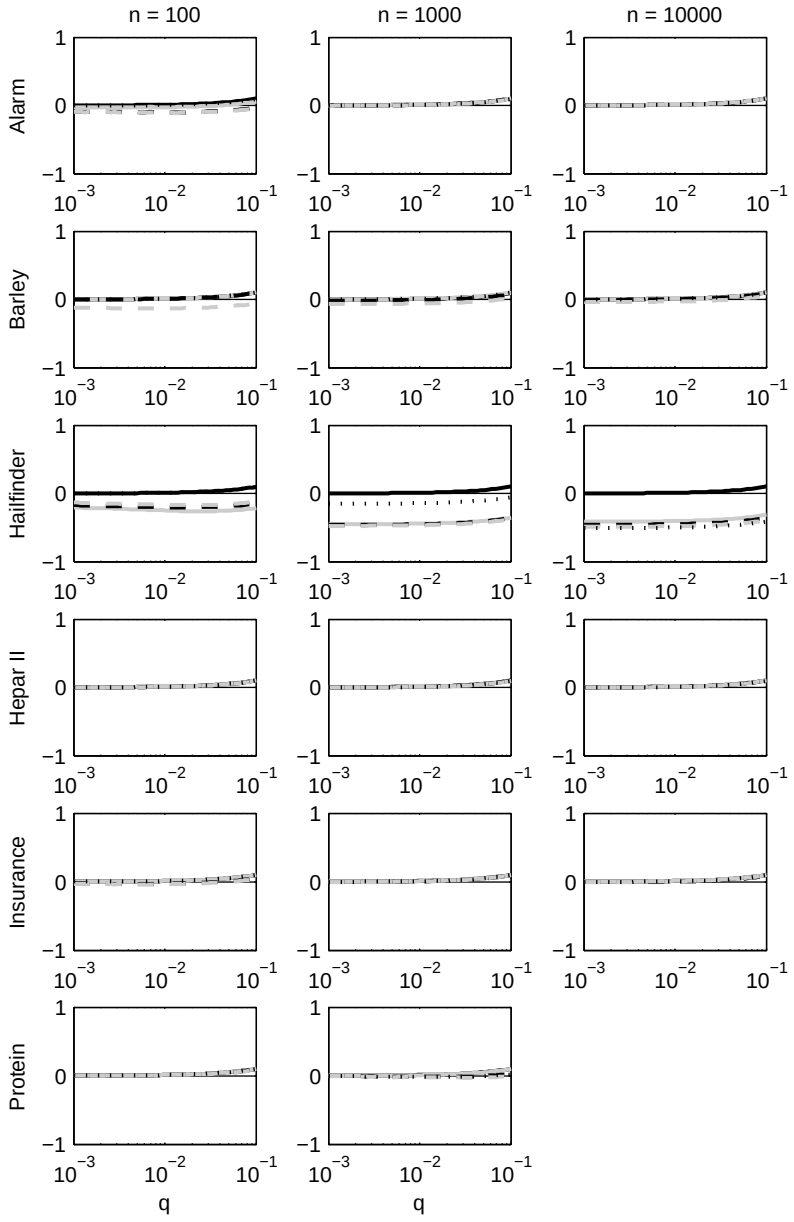



Figure 5.6: $q - \mathrm{FDR}_{\mathrm{relax}}(q)$: bias of the Benjamini and Yekutieli [7] False Discovery Rate (FDR) controlling procedure when controlling the relaxed FDR ($\mathrm{FDR}_{\mathrm{relax}}$, see text for definition) at each FDR threshold $q$ for each network (rows), sample size $n$ (columns) and upper limit *max-k* on conditioning set cardinality. X-axes are in logarithmic-10 scale. *Heuristic* denotes the heuristic power rule. The legend is the same as in Fig. 5.4. As *max-k* increases from 0 to 3, the evolution of the procedure's bias is not consistent among the networks because the evolution of $\widehat{\mathrm{FDR}}_{\mathrm{BY}}$ bias is not consistent (Fig. 5.5). For *max-k* $= 0$, bias is non-negative or slightly negative for $n \in \{1000, 10000\}$ on Hailfinder, while it is noticeably negative for the heuristic power rule. For *max-k* $= 1$, the bias is noticeably negative for $n = 100$ on Hailfinder, while it is non-negative or slightly negative for the heuristic power rule. For *max-k* $\in \{2, 3\}$, the is noticeably negative for $n = 100$ on Alarm and Hailfinder, while it is non-negative or slightly negative for the heuristic power rule.

# Chapter 6

# Summary and future work

## 6.1 Summary

The focus of this thesis is the problem of estimating and controlling the False Discovery Rate (FDR) in learning the skeleton (set of edges without regard of direction) of a network. The main contributions as summarized as follows:

- A unified approach to estimation and control in Bayesian network skeleton identification
- An experimental evaluation of the most common FDR estimator in both tasks over several networks and sample sizes, employing both simulated data as well as real biological data
- A discussion about the relationship of utilizing FDR in skeleton identification with other approaches to assessing confidence in structure learning
- An experimental evaluation of several approaches to improving estimation and control of the FDR
- A relaxed definition of false discovery and its experimental evaluation

## 6.2 Future work

Our work opens new directions in the utilization of FDR in learning Bayesian network structure and in estimating structural uncertainty in general:

- The permutation approach to estimating the FDR introduced in Section 4.1.2 is yet to be experimentally evaluated.
- Even in cases where the FDR is not accurately estimated and controlled, we have demonstrated that the calculated link absence p-values are good relative scores in the classification of links as present or not present. An experimental comparison of skeleton identification to Bayesian model averaging and the bootstrap in classification of links is yet to be conducted.
- In our evaluations we used relatively small networks. However, in typical biological applications such as learning Bayesian network structure from gene expression data the number of variables is large compared to the available sample size. Evaluating the performance of FDR estimators in low sample with large Bayesian networks with the supposed characteristics of real gene networks would be of great interest.
- In this work we have focused on global structure learning, i.e., learning the structure of the whole network. However, the FDR in local learning, i.e., learning the neighbors of a single node, may be of interest. For example, one would like to find the genes that share a link with a target gene or experimental condition and then estimate and/or control the proportion of false links among the discovered ones. Utilization of the FDR in local learning has some special issues [47] and has to be studied separately.

# Appendix

## Data sources

Summary statistics of the Bayesian networks used throughout this thesis are given in Table 1. Protein is the assumed Bayesian network the observational flow cytometry measurements made by Sachs et. al. [32] are generated from (see Section 3.4). The Win95pts network was developed at Microsoft Research and contributed to the community by Jack Breese. All networks except Protein were downloaded from the following two online repositories:

**Bayesian Network Repository (BNR)**
`http://www.cs.huji.ac.il/site//labs/compbio/Repository/`

**GeNIe & SMILE Network Repository (GS)**
`http://genie.sis.pitt.edu/networks.html`

The discretized data from Sachs et. al. [32] were obtained from the *BDAGL* (Bayesian DAG learning) software package by Daniel Eaton and Kevin Murphy. BDAGL is downloadable at:
`http://www.cs.ubc.ca/~murphyk/Software/BDAGL/`

Table 1: Summary statistics of the Bayesian networks used throughout this thesis. $|\mathbf{V}|$ denotes the number of nodes, $|\mathbf{E}|$ the number of edges, $\overline{|\mathbf{PA}_X|}$ the mean number of parents and $\overline{|\mathbf{D}_X|}$ the mean number of levels for categorical networks.

| Name | $\|\mathbf{V}\|$ | $\|\mathbf{E}\|$ | $\overline{\|\mathbf{PA}_X\|}$ | $\overline{\|\mathbf{D}_X\|}$ | Repository | Reference |
|---|---|---|---|---|---|---|
| Alarm | 37 | 46 | 1.24 | 2.84 | BNR | [5] |
| Barley | 48 | 84 | 1.75 | 8.77 | GS | [21] |
| Hailfinder | 56 | 66 | 1.18 | 3.98 | GS | [1] |
| Hepar II | 70 | 66 | 1.76 | 2.31 | GS | [27] |
| Insurance | 27 | 52 | 1.93 | 3.30 | BNR | [8] |
| Protein | 11 | 20 | 1.82 | | | [32] |
| Win95pts | 76 | 112 | 1.47 | 2.00 | GS | |

## Software

- The *Probabilistic Graphical Model Toolbox* (PGM Toolbox) for MATLAB® is developed by the author and was used in the experiments of this thesis. PGM Toolbox is available for download at:
  `http://sourceforge.net/projects/pgm-toolbox/`
- The POWER correction is available within the *PowerBayes* software package by Andrew Fast. PowerBayes can be downloaded from:
  `http://kdl.cs.umass.edu/powerbayes/`
- The MATLAB code of the experiments is available for download at the website of Bio-Informatics Laboratory:
  `http://www.ics.forth.gr/bil/`

# Bibliography


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

[2] A. Agresti. Categorical data analysis.

[3] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos. Local causal and markov blanket induction for causal discovery and feature selection for classification part i: Algorithms and empirical evaluation. *Journal of Machine Learning Research, Special Topic on Causality*, 11:171–234, 2010.

[4] C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. 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, Special Topic on Causality*, 11:235–284, 2010.

[5] IA Beinlich, HJ Suermondt, RM Chavez, and GF Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. *AI in Medicine in Europe*.

[6] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):pp. 289–300, 1995.

[7] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[8] J. Binder, D. Koller, S. Russell, and K. Kanazawa. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29(2):213–244, 1997.

[9] M.H. DeGroot, M.J. Schervish, X. Fang, L. Lu, and D. Li. *Probability and statistics*, volume 298. Addison-Wesley Boston, MA, USA;, 1986.

[10] D. Eaton and K. Murphy. Bayesian structure learning using dynamic programming and mcmc. *Proceedings of the Twenty-Third Confererence on Uncertainty in Artificial Intelligence (UAI 2007)*, 2007.

[11] A. Fast, M. Hay, and D. Jensen. Improving Accuracy of Constraint-Based Structure Learning. 2009.

[12] A.S. Fast. *Learning the Structure of Bayesian Networks with Constraint Satisfaction*. PhD thesis, University of Massachusetts Amherst, 2010.

[13] T. Fawcett. ROC graphs: Notes and practical considerations for data mining researchers. *HP Laboratories technical report*, 2003.

[14] S.E. Fienberg. The analysis of cross-classified categorical data. 1977.

[15] N. Friedman, M. Goldszmidt, and A. Wyner. On the application of the bootstrap for computing confidence measures on features of induced Bayesian networks. *AI&STAT VII*, 1999.

[16] N. Friedman and D. Koller. Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks. *Machine learning*, 50(1):95–125, 2003.

[17] M. Grzegorczyk and D. Husmeier. Improving the structure mcmc sampler for bayesian networks by introducing a new edge reversal move. *Machine Learning*, 71(2):265–305, 2008.

[18] D. Heckerman, C. Meek, and G. Cooper. A Bayesian Approach to Causal Discovery. *Computation, causation, and discovery*, page 141, 1999.

[19] M. Koivisto. Advances in exact bayesian structure discovery in bayesian networks. In *Proc. of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 241–248. Citeseer.

[20] M. Koivisto and K. Sood. Exact Bayesian structure discovery in Bayesian networks. *The Journal of Machine Learning Research*, 5:549–573, 2004.

[21] K. Kristensen and I.A. Rasmussen. The use of a Bayesian network in the design of a decision support system for growing malting barley without use of pesticides. *Computers and Electronics in Agriculture*, 33(3):197–217, 2002.

[22] J. Li and Z. J. Wang. Controlling the false discovery rate of the association/causality structure learned with the pc algorithm. *J. Mach. Learn. Res.*, 10:475–514, 2009.

[23] J. Listgarten and D. Heckerman. Determining the number of non-spurious arcs in a learned dag model: Investigation of a bayesian and a frequentist approach. In *23rd Conference on Uncertainty in Artificial Intelligence*. Citeseer, 2007.

[24] D. Madigan, J. York, and D. Allard. Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, pages 215–232, 1995.

[25] C. Meek. Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the eleventh international conference on uncertainty in artificial intelligence*, 1995.

[26] R.E. Neapolitan. *Learning bayesian networks*. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[27] A. Onisko. *Probabilistic Causal Models in Medicine: Application to Diagnosis of Liver Disorders*. PhD thesis, Ph. D. Dissertation, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Science, Warsaw, 2003.

[28] P. Parviainen and M. Koivisto. Exact structure discovery in bayesian networks with less space. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 436–443. AUAI Press, 2009.

[29] P. Parviainen and M. Koivisto. Bayesian structure discovery in bayesian networks with less space. pages 589–596, 2010.

[30] R. Robinson. Counting unlabeled acyclic digraphs. *Combinatorial mathematics V*, pages 28–43, 1977.

[31] S.J. Russell, P. Norvig, J.F. Canny, J.M. Malik, and D.D. Edwards. *Artificial intelligence: a modern approach*, volume 74. Prentice hall Englewood Cliffs, NJ, 1995.

[32] K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger, and G.P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523, 2005.

[33] J. Schäfer and K. Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754, 2005.

[34] J.P. Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46, 1995.

[35] P. Spirtes, C.N. Glymour, and R. Scheines. *Causation, prediction, and search*. 2000.

[36] H. Steck and T. Jaakkola. On the Dirichlet Prior and Bayesian Regularization. In *NIPS*, pages 697–704, 2002.

[37] B. Steinsky. Enumeration of labelled chain graphs and labelled essential directed acyclic graphs. *Discrete mathematics*, 270(1-3):267–278, 2003.

[38] J. D. Storey. False Discovery Rates. *International Encyclopedia of Statistical Science, Lovric M (editor)*, 2010.

[39] J.D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[40] J.D. Storey. The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003.

[41] J.D. Storey, J.E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

[42] J.D. Storey and R. Tibshirani. *Estimating the positive false discovery rate under dependence, with applications to DNA microarrays*. 2001.

[43] J. Tian and R. He. Computing posterior probabilities of structural features in bayesian networks. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 538–547. AUAI Press, 2009.

[44] I. Tsamardinos and G. Borboudakis. Permutation Testing Improves Bayesian Network Learning. *Machine Learning and Knowledge Discovery in Databases*, pages 322–337, 2010.

[45] I. Tsamardinos and L. E. Brown. Bounding the false discovery rate in local bayesian network learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, volume 2, pages 1100–1105. AAAI Press, 2008.

[46] I. Tsamardinos, L. E. Brown, and C. F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning Journal*, 65:31–78, 2006.

[47] I. Tsamardinos, L.E. Brown, and S. Triantafylloy. Controlling the False Discovery Rate in Bayesian Network Structure Learning. 2008.