ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

A Unified Approach to Estimation and Control of the False Discovery Rate in Bayesian Network Skeleton Identification

Angelos P. Armen and Ioannis Tsamardinos

University of Crete - Computer Science Department
Heraklion, Crete - Greece

Foundation of Research and Technology, Hellas - Institute of Computer Science
Heraklion, Crete - Greece

Abstract. Constraint-based Bayesian network (BN) structure learning algorithms typically control the False Positive Rate (FPR) of their skeleton identification phase. The False Discovery Rate (FDR), however, may be of greater interest and methods for its utilization by these algorithms have been recently devised. We present a unified approach to BN skeleton identification FDR estimation and control and experimentally evaluate the performance of FDR estimators in both tasks over several networks. We demonstrate that estimation is too conservative for most networks and strong control at common FDR thresholds is not achieved with some networks; finally, we identify the possible causes of this situation.

1 Introduction

Consider the following example of BN skeleton identification from related work [2]: suppose a network of 100 genes, each one sharing a *link* (i.e., edge without regard of direction) with 3 other genes on average, i.e., there are 150 out of $100 \cdot (100-1)/2 = 4950$ possible links. An algorithm that learns the set of links (i.e., the *skeleton*) with FPR $= 5\%$ and power $= 90\%$ discovers a network with $150 \cdot 90\% = 135$ true links and $(4950-150) \cdot 5\% = 240$ false links, on average. Then the expected proportion of false links, i.e., the *False Discovery Rate* is $240/(240+135) = 64\%$. When we are interested in having mostly true positives among our discoveries, FDR, and not FPR, is the error rate of choice.

2 Constraint-based Bayesian network structure learning

The pair (G, P) of a directed acyclic graph (DAG) $G = (V, E)$ and a probability distribution P of the variables in some set $\mathbf{V}$ is a *Bayesian network* if it satisfies the *Markov condition*: every variable (equivalently *node*) is independent of any subset of its non-descendants conditioned on its parents [4]. Based on this condition, G *entails* (i.e., implies) even more conditional independences; when these are *all and only* in P, we say that G and P are *faithful* to each other [4]. Under faithfulness, two nodes are linked if and only if there is no subset of the rest nodes that renders the two nodes conditionally independent. When there is a DAG faithful to a distribution, there is usually more than one; however, since

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

they all entail the same conditional independences, they all share the same set of links, i.e., the same skeleton.

The goal of BN *structure learning* is to find a DAG G faithful to a distribution P given a sample of P [4]. *Constraint-based* structure learning algorithms work in two phases: first the conditional independences in P are identified and then they are used as *constraints* in generating G [4]. The first phase is called *constraint* or *skeleton identification* because it corresponds to learning the skeleton of G.

Typical constraint-based algorithms exploit a theorem that states that, under faithfulness, if two nodes are not linked, there exists a subset of the parents of one of them that renders the nodes conditionally independent [4]. For each pair of nodes, these algorithms search for a subset of *supersets* of the parents (the parents are, of course, unknown to the algorithms) of each of the two nodes that renders these nodes conditionally independent. If such subset is found, the pair is no longer considered; otherwise, a link between these nodes is discovered.

Conditional independences are identified by performing hypothesis tests at a given significance level α. A test, however, is only attempted if there is sufficient power according to a reliability criterion; otherwise it is ignored [5]. Even if a test is attempted, the computation of its p-value may not be possible [5]; again, the test is ignored. If there exist a DAG faithful to P and all statistical decisions made are correct, constraint-based algorithms are proven to find such a DAG.

3 False Discovery Rate

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

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

where $\mathbf{V}$ is the number of rejected true null hypotheses and 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 utilize FDR. The first one, *control*, is to set an FDR threshold q and find a p-value threshold t such that *strong control* of FDR under q is achieved, i.e., $FDR(t) \leq q$, where $FDR(t)$ is the FDR resulting from rejecting all hypotheses with p-value $\leq t$. The procedure below is proven to achieve strong control, assuming independent p-values [6] or *positive regression dependence* of the p-values on each of the null p-values [7]:

Algorithm 1 Benjamini-Hochberg (BH) FDR control procedure

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

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

The second approach to FDR utilization is *estimation*: set a p-value threshold t and estimate $FDR(t)$ in a conservative manner, i.e., $E[\widehat{FDR}(t)] \geq FDR(t)$. It was introduced by Storey [8], along with a family of estimators[1] proven to be conservative when the p-values are independent:

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

$\widehat{FDR}(t)$ can be used to define valid FDR control procedures [3]: taking the largest t s.t. $\widehat{FDR}(t) \leq q$ corresponds to applying the BH procedure.

4 False Discovery Rate of skeleton identification

4.1 Skeleton identification as a multiple testing procedure

In order to utilize FDR, skeleton identification is viewed as a multiple testing procedure, the null hypotheses being the absence of links. To test the hypothesis $H_{\neg X-Y}$ of absence of a link between nodes X and Y, constraint-based algorithms complete the tests of the hypotheses $H_{I(X,Y|\mathbf{Z})}$ and obtain test statistics $g_{I(X,Y|\mathbf{Z})}$ and p-values $p_{I(X,Y|\mathbf{Z})}$ for some set $\mathbf{S}_{X-Y}$ of subsets $\mathbf{Z}$ of $\mathbf{V}$. The p-value $p_{\neg X-Y}$ of $H_{\neg X-Y}$ is the probability, when $H_{\neg X-Y}$ is true, that the statistics $G_{I(X,Y|\mathbf{Z})}$ are as extreme or more extreme than the obtained statistics $g_{I(X,Y|\mathbf{Z})}$:

$$p_{\neg X-Y} = Pr\left(\bigcap_{\mathbf{Z} \in \mathbf{S}_{X-Y}} \left\{ |G_{I(X,Y|\mathbf{Z})}| \geq |g_{I(X,Y|\mathbf{Z})}| \right\} \middle| \neg X-Y \right)$$

Unfortunately, $p_{\neg X-Y}$ is unavailable. However, $p_{I(X,Y|\mathbf{Z})}$ can be used to upper-bound $p_{\neg X-Y}$ thanks to the following theorem:

Theorem 1. *If (1) there is a DAG faithful to the probability distribution P, (2) all conditional independence tests considered by the algorithm are completed (i.e., return a p-value), and (3) the* realized *power of the output skeleton is 1, i.e., all true links are discovered, then the p-value $p_{\neg X-Y}$ of the link absence (LA) hypothesis $H_{\neg X-Y}$ is upper-bounded by the maximal $\max_{\mathbf{Z} \in \mathbf{S}_{X-Y}} p_{I(X,Y|\mathbf{Z})}$ among the p-values $p_{I(X,Y|\mathbf{Z})}$ of the conditional independence (CI) hypotheses $H_{I(X,Y|\mathbf{Z})}$ tested by the algorithm:*

$$p_{\neg X-Y} \leq \max_{\mathbf{Z} \in \mathbf{S}_{X-Y}} p_{I(X,Y|\mathbf{Z})}$$

Proof. First consider the hypothesis $H_{\exists \mathbf{Z} \in \mathbf{S}_{X-Y} : I(X,Y|\mathbf{Z})}$ that there is a set $\mathbf{Z}$ in $\mathbf{S}_{X-Y}$ that renders X and Y conditionally independent. Tsamardinos and Brown [1] prove that the p-value $p_{\exists \mathbf{Z} \in \mathbf{S} : I(X,Y|\mathbf{Z})}$ of $H_{\exists \mathbf{Z} \in \mathbf{S}_{X-Y} : I(X,Y|\mathbf{Z})}$ is upper-bounded by the maximal $\max_{\mathbf{Z} \in \mathbf{S}_{X-Y}} p_{I(X,Y|\mathbf{Z})}$ among the p-values $p_{I(X,Y|\mathbf{Z})}$ of the CI hypotheses $H_{I(X,Y|\mathbf{Z})}$. When (1), (2) and (3) hold, by design of constraint-based

[1] Storey's estimators also include 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 this context we use $\hat{\pi}_0(\lambda) = 1$ instead.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

algorithms $\mathbf{S}_{X-Y}$ contains all subsets of $Pa(X)$ and all subsets of $Pa(Y)$. Then the hypotheses $H_{\exists \mathbf{Z} \in \mathbf{S}: I(X,Y|\mathbf{Z})}$ and $H_{\neg X-Y}$ are equivalent. Hence, $p_{\neg X-Y} = p_{\exists \mathbf{Z} \in \mathbf{S}: I(X,Y|\mathbf{Z})} \leq \max_{\mathbf{Z} \in \mathbf{S}} p_{I(X,Y|\mathbf{Z})}$. □

The assumptions above are also stated in [1] but not formally included in a theorem. No formal proof for the upper bounds themselves is given in [2] too. If we assume independent maximal CI p-values, constraint-based algorithms implicitly control the FPR at the α level: an LA hypothesis is accepted if a CI p-value exceeds α, or equivalently, if the maximal CI p-value exceeds α [2].

4.2 Estimation and control of skeleton identification FDR

Tsamardinos and Brown [1] follow the *estimation* approach in *local* BN learning; the latter is concerned with learning the set $PC(X)$ of parents and children of a *single* node X. First, $PC(X)$ is learned by performing tests of conditional independence at the α level, thus implicitly controlling the local learning FPR at the α level (assuming independence of the maximal CI p-values). Then the local learning FDR is estimated by $\widehat{FDR}(\alpha)$.

Li and Wang [2] follow the *control* approach in skeleton identification (*global* BN learning). They modify the skeleton identification phase of the *PC* algorithm and come up with *PC_{FDR}–skeleton*, a skeleton identification algorithm with *embedded* FDR control. PC_{FDR}–skeleton is proven to strongly control the FDR at a given level q under assumptions similar to those of Theorem 1. PC_{FDR}–skeleton does not accept a CI hypothesis (and subsequently, an LA hypothesis) when its p-value exceeds some FPR threshold α but instead applies the BH procedure to the up-to-date maximal CI p-values with an FDR threshold q after an up-to-date maximal CI p-value is updated. Given that, if an FDR control procedure strongly controls FDR given some p-values, it also strongly controls FDR given upper bounds on those p-values, PC_{FDR}–skeleton strongly controls FDR when it terminates.

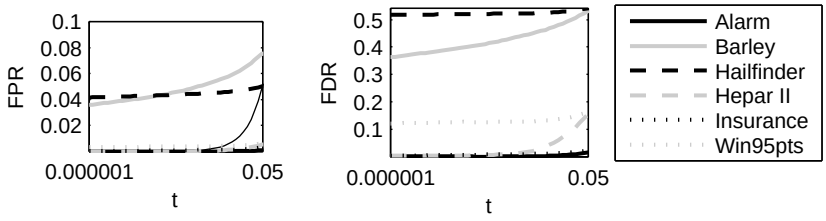
In this work we adapt the method of Tsamardinos and Brown [1] to skeleton identification, whose FDR is estimated by $\widehat{FDR}(\alpha)$. Moreover, instead of only computing $\widehat{FDR}(\alpha)$, we compute $\widehat{FDR}$ at *all* maximal CI p-values; doing so allows for subsequent FDR control at *any* level q, instead of having to fix q in advance as with PC_{FDR}–skeleton. On the other hand, PC_{FDR}–skeleton does not require an FPR threshold α to be specified.

5 Experimental results

We generated 100 random samples of size 5000 from the Alarm, Barley, Hailfinder, Hepar II, Insurance and Win95pts networks from online repositories and applied the skeleton identification phase of the *MMHC* algorithm [5], with $\alpha = 0.05$, on each sample. Expectations are estimated by the respective means.

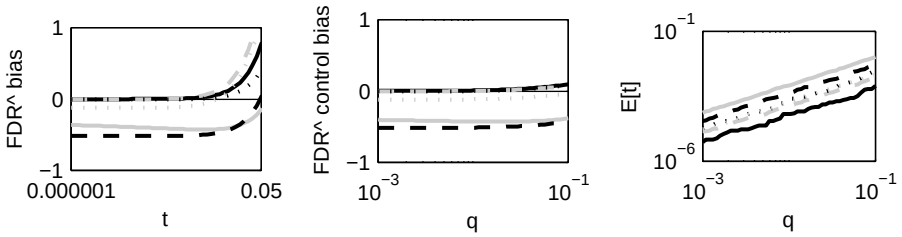
For FDR estimation, we computed $\widehat{FDR}(t)$ at the same 50 logarithmically spaced in $[10^{-8}10^{-1}]$ p-value thresholds t for all samples of each network, because

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.



(a) (b)

Fig. 1: FPR (a) and FDR (b) as functions of the p-value (FPR) threshold t. X-axes are in logarithmic-10 scale. Both quantities vary greatly among networks and increase as t increases. For $t = \alpha = 0.05$ they are equal to the skeleton identification FPR and FDR respectively. The thin black curve in (a) is $FPR(t) = t$. FPR is not controlled at any level t for Barley, Hailfinder and Win95pts.



(a) $E[\widehat{FDR}(t)] - FDR(t)$ ($\widehat{FDR}$ bias at p-value threshold t) increases as t increases and is mostly negative, (i.e., $\widehat{FDR}(t)$ is not conservative) for Barley, Hailfinder and Win95pts. The bias at $t = \alpha = 0.05$ is the skeleton identification FDR bias and is too positive for all networks except Barley and Hailfinder.

(b) $q - FDR(q)$ ($\widehat{FDR}$ "control bias" at FDR threshold q). FDR is tightly controlled (i.e., control bias is positive and small) at all q with Alarm, at $q > 10^{-2}$ for Hepar II and Insurance but at no q for the rest networks.

(c) $E[t(q)]$, the expected p-value threshold t returned by the BH procedure with FDR threshold q. The y-axis is in logarithmic-10 scale. For q s.t. $FDR(q)$ is tightly controlled (Fig. 2b), the corresponding $E[t(q)]$ belongs to the lower part of the range of t, where $\widehat{FDR}(t)$ bias is positive and small (Fig. 2a).

Fig. 2: FDR estimation and control bias. X-axes are in logarithmic-10 scale.

different sets of links (skeletons) with different maximal CI p-values are learned from different samples. $FPR(t)$ and $FDR(t)$ (Fig. 1) vary greatly among networks; [2] $\widehat{FDR}$ is slightly not conservative (i.e., its bias, $E[\widehat{FDR}(t)] - FDR(t)$, is slightly negative) at the smaller t for Alarm, Hepar II and Insurance and absolutely not conservative for the rest networks except for t close to α (Fig. 2a). The skeleton identification FDR estimation is too conservative because it corresponds to $\widehat{FDR}(\alpha)$, which is too conservative for most networks.

[2] We also considered an alternative definition of the FDR, called the *positive* FDR (pFDR) [8]. We found $FDR(t) = pFDR(t)$ for all networks so we did not consider pFDR any further.

ESANN 2011 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2011, i6doc.com publ., ISBN 978-2-87419-044-5. Available from http://www.i6doc.com/en/livre/?GCOI=28001100817300.

For FDR control, we applied the BH procedure to the maximal CI p-values from each sample of each network to control the FDR of each network at 50 logarithmically spaced in $[10^{-3}10^{-1}]$ FDR levels q. Whether strong control is achieved is an immediate result of the estimators being conservative: $\widehat{FDR}$ achieves strong control at any threshold q for Alarm, slightly fails for Hepar II and Insurance for $q < 10^{-2}$ and fails miserably for the rest networks (Fig. 2b).

There are two possible causes for the lack of accuracy of FDR estimation and control: Either (1) the dependence of the maximal CI p-values is not supported by the estimators (which assume independence or positive regression dependence for the p-values) or (2) there are maximal CI p-values that are not upper bounds on the LA p-values, or both. These issues are not inherent to our approach: the assumptions of [2] are similar to ours and the estimators used there are the same as here. PC_{FDR}–skeleton is evaluated with custom networks and at $q = 0.05$ only; we plan to compare both approaches at various q using repository networks.

6 Conclusions and future work

We presented a unified approach to BN skeleton identification FDR estimation and control and demonstrated that estimation is too conservative for most networks and strong control at common FDR thresholds is not achieved with some of the networks we used. We identified the two possible causes of this lack of accuracy and we are currently working on eliminating them.

Acknowledgement: This research was partially funded by ICS-FORTH.

References

[1] 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.

[2] 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.

[3] 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.

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

[5] 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.

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

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