# Bayesian network structure learning with causal effects in the presence of latent variables

**Kiattikun Chobtham** K.CHOBTHAM@QMUL.AC.UK

**Anthony C. Constantinou** A.CONSTANTINOU@QMUL.AC.UK

*Bayesian Artificial Intelligence research lab, Risk and Information Management Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London.*

## Abstract

Latent variables may lead to spurious relationships that can be misinterpreted as causal relationships. In Bayesian Networks (BNs), this challenge is known as learning under causal insufficiency. Structure learning algorithms that assume causal insufficiency tend to reconstruct the ancestral graph of a BN, where bi-directed edges represent confounding and directed edges represent direct or ancestral relationships. This paper describes a hybrid structure learning algorithm, called CCHM, which combines the constraint-based part of cFCI with hill-climbing score-based learning. The score-based process incorporates Pearl's do-calculus to measure causal effects, which are used to orientate edges that would otherwise remain undirected, under the assumption the BN is a linear Structure Equation Model where data follow a multivariate Gaussian distribution. Experiments based on both randomised and well-known networks show that CCHM improves the state-of-the-art in terms of reconstructing the true ancestral graph.

**Keywords:** ancestral graphs; causal discovery; causal insufficiency; probabilistic graphical models.

## 1. Introduction and related works

A Bayesian Network (BN) is a type of a probabilistic graphical model that can be viewed as a Directed Acyclic Graph (DAG), where nodes represent uncertain variables and arcs represent dependency or causal relationship between variables. The structure of a BN can be learned from data based on three main classes of structure learning: constraint-based, score-based and hybrid learning. The first type relies on conditional independence tests to construct the skeleton and orientate edges, whereas the second type searches over the space of possible graphs and returns the graph that maximises a fitting score. Hybrid learning refers to algorithms that combine both constraint-based and score-based classes of learning.

A common problem when learning BNs from data is that of causal insufficiency, where data fail to capture all the relevant variables. Variables not captured by data are often referred to as latent variables (also known as unobserved, unmeasured, or hidden variables). In the real world, latent variables are impossible to avoid either because data may not be available or simply because we are unaware of their existence. A special case of a latent variable, referred to as a latent confounder, is an unobserved common cause of two or more observed variables in a BN. While known latent variables pose less of a problem in knowledge-based BNs, where methods exist that enable users to model latent variables not present in the data under the assumption the statistical outcomes are already influenced by the causes an expert might identify as variables missing from the dataset (Constantinou et al., 2016), they can be a problem in structure learning. This is because child nodes that share an unobserved common cause will be found to be directly related, even when they are

not, and this is a widely known problem that gives rise to spurious correlations in the presence of confounding.

The traditional DAG has proven to be unsuitable when structure learning is performed under the assumption that some variables are latent. This is because a DAG assumes causal sufficiency and hence, it is not designed to capture latent variables. Ancestral graphs have been proposed as a solution to this problem, and represent an extension of DAGs that capture hidden variables. Specifically, the Maximal Ancestral Graph (MAG) (Richardson and Spirtes, 2000) is a special case of a DAG where arcs indicate direct or ancestral relationships, and bi-directed edges represent confounding. Moreover, a Partial Ancestral Graph (PAG) represents a set of Markov equivalent MAGs (Spirtes et al., 2001), in the same way a Complete Partial Directed Acyclic Graph (CPDAG) represents a set of Markov equivalent DAGs. Fig 1 illustrates an example of a DAG with latent variables $L_1$ and $L_2$, along with its corresponding Markov equivalent MAGs and the PAG of Markov equivalent MAGs. Both types of ancestral graph, MAGs and PAGs, can be used to represent causally insufficient systems.

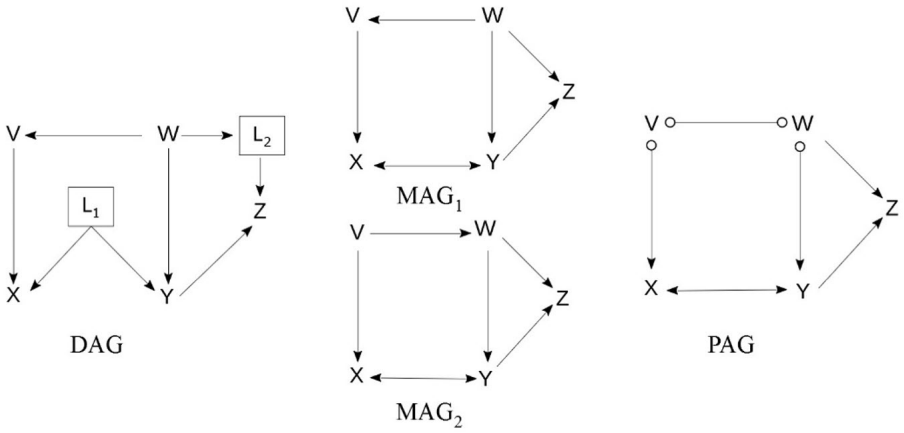


**Figure 1** A causal DAG with observed variables $\{V, W, X, Y, Z\}$ ∪ latent variables $\{L_1, L_2\}$, with two corresponding Markov equivalent MAGs, and the Markov equivalent PAG of MAGs, where o-o and o→ represent edges in which the orientation is uncertain.

The most popular BN structure learning algorithm for causally insufficient systems is the constraint-based FCI, which is based on the PC algorithm (Spirtes et al., 2001). Modified versions of FCI have been published in the literature and include the augmented FCI which improves the orientation phase by extending the orientation rules of FCI from four to ten (Zhang, 2008), the conservative FCI (cFCI) which uses additional conditional independence tests to restrict unambiguous orientations and improve the identification of definite colliders (Ramsey et al., 2012), and the RFCI which skips some of the orientation rules in FCI and performs fewer conditional independence tests that make the algorithm faster and more suitable to problems that involve thousands of variables, in exchange for a minor reduction in the accuracy of the learned graph (Colombo et al., 2011). These constraint-based algorithms assume the joint probability distribution is a perfect map that is faithful to the true graph, although this will often not be true when working with real data. Moreover, the orientation phase depends on the accuracy of the skeleton and hence, any errors from the first phase are propagated to the orientation phase. GFCI relaxes these issues by incorporating the score-based approach of FGS (Ramsey, 2015), which is an enhanced version of Greedy Equivalence Search

(GES), thereby producing a hybrid learning algorithm that outperforms the constraint-based versions of FCI (Ogarrio et al., 2016).

In addition to the FCI variants, other algorithms have been proposed that are based on different approaches to structure learning. These include the GSPo, the M$^3$HC and the GSMAG algorithms. Specifically, the GSPo is an order-based search algorithm that performs greedy search over the space of independence maps (IMAPs) to determine the minimal IMAP (Bernstein et al., 2019). This is achieved by defining a partial ordered set (poset) that is linked to the IMAP, expressed as a discrete optimisation problem. However, GSPo uses a random starting point for a poset, and this makes the algorithm non-deterministic since each run is likely to produce a different result. On the other hand, the M$^3$HC is a hybrid learning algorithm (Tsirlis et al., 2018) that adds a constraint-based learning phase to the greedy search of the GSMAG algorithm (Triantafillou and Tsamardinos, 2016). Both M$^3$HC and GSMAG assume the data are continuous and normally distributed, and Tsirlis et al. (Tsirlis et al., 2018) showed that hybrid algorithms such as M$^3$HC and GFCI demonstrate better performance over the other relevant constraint-based algorithms.

This paper builds on recent developments in BN structure learning under causal insufficiency and describes a novel structure learning algorithm that combines constraint-based and score-based learning with causal effects to learn Gaussian BNs (GBNs). The paper is organised as follows: Section 2 describes the CCHM algorithm, Section 3 describes the evaluation process, Section 4 presents the results, and we provide our concluding remarks and a discussion for future work in Section 5.

## 2. Conservative rule and Causal effect Hill-climbing for MAG (CCHM)

CCHM is a hybrid structure learning algorithm defined as a Structural Equation Model (SEM), under the assumption the data are continuous and follow a multivariate Gaussian distribution. The process of CCHM can be divided into two phases. The first phase adopts the conditional independence steps of cFCI to construct the skeleton of the graph and to further classify definite colliders as whitelist and definite non-colliders as blacklist. The second phase involves score-based learning that uses the Bayesian Information Criterion (BIC) as the objective function, adjusted for MAGs, where edge orientation is augmented with causal effect measures. These steps are described in more detail in the subsections that follow.

### 2.1 Definite colliders (whitelist) and definite non-colliders (blacklist)

Conditional independence tests are used to determine the edges between variables and to produce the skeleton graph. A p-value associates with each statistical test result, which is used to sort conditional independencies in ascending order. An alpha hyperparameter is then used as the cut-off threshold in establishing independence. For each conditional independency $A \perp\!\!\!\perp B | Z$, $Z$ is recorded as the separation set (Sepset) of variables $A$ and $B$. The orientation of edges is determined by a method inherited from cFCI, where extra conditional independence tests over all unshielded triples determine the classification for each of those triples as either a definite collider or a definite non-collider:

- Given unshielded triple $A$-$C$-$B$, perform conditional independence tests on $A$ and $B$ over all neighbours of $A$ and $B$.
- If $C$ is **NOT** in all Sepsets of $A$ and $B$, add $A$-$C$-$B$ to the whitelist as a definite collider.

- If $C$ is in **ALL** Sepsets of $A$ and $B$, add $A$-$C$-$B$ to the blacklist as a definite non-collider.

### 2.2 Bayesian Information Criterion (BIC) for MAG

The score-based learning part of CCHM involves hill-climbing greedy search that minimises the BIC score, which is a function for goodness-of-fit based on Occam's razor principle (Darwiche, 2009). The BIC score balances the Log-Likelihood (LL) fitting against a penalty term for model dimensionality. CCHM adopts the BIC function used in the M$^3$HC and GSMAG algorithms which is adjusted for MAGs (Tsirlis et al., 2018). Formally, given a dataset over vertices $V$ with a distribution $\mathcal{N}(0, \Sigma)$ where $\Sigma$ is a covariance matrix calculated from the dataset, a unique solution $Y$ is found where $\hat{\Sigma} = (I - \mathcal{B})^{-1} \Omega (I - \mathcal{B})^{-t}$. MAG $\mathcal{G}$ is constructed from linear equations $Y = \mathcal{B} \cdot Y + \epsilon$, where $Y = \{Y_i | i \in V\}$, $\mathcal{B}$ is a $V \times V$ coefficient matrix for the directed edge $j$ to $i$ $\{\beta_{ij}\}$, I is an identity matrix, $\epsilon$ is a positive random error vector for the bidirected edge $j$ to $i$ $\{\omega_{ij}\}$, and the error covariance matrix $\Omega = Cov(\epsilon) = \{\omega_{ii}\}$. The BIC score is then calculated as follows (Richardson and Spirtes, 2000):

$$BIC\left(\hat{\sum}\middle|\mathcal{G}\right) = -2\ln\left(l_{\mathcal{G}}\left(\hat{\sum}\middle|\mathcal{G}\right)\right) + \ln(N)\left(2|V| + |E|\right) \qquad (1)$$

where $l_{\mathcal{G}}$ is likelihood function, $|V|$ and $|E|$ are the size of nodes and edges that are part of the complexity penalty term, and $N$ is the sample size. Similar to the factorisation property of DAGs, the score $l_{\mathcal{G}}\left(\hat{\sum}\middle|\mathcal{G}\right)$ can be decomposed into c-components $(S_k)$ of $\mathcal{G}$ which refer to the connected components that are partitioned by removing all directed edges (Nowzohour et al., 2015):

$$l_{\mathcal{G}}\left(\hat{\sum}\middle|\mathcal{G}\right) = -\frac{N}{2}\sum_k S_k \qquad (2)$$

$$\text{where } S_k = |C_k| \cdot \ln(2\pi) + ln\left(\frac{|\hat{\Sigma}_{\mathcal{G}_k}|}{\prod_{j \in \mathrm{Pa}_{\mathcal{G}_k}} \sigma^2_{kj}}\right) + \frac{N-1}{N} \cdot tr\left[\hat{\Sigma}^{-1}_{\mathcal{G}_k} S_{\mathcal{G}_k} - |\mathrm{Pa}_{\mathcal{G}}(C_k) \setminus \{C_k\}|\right]$$

and where $C_k$ denotes the set of nodes for each c-component $k$, $\mathcal{G}_k$ is the marginalisation from $C_k$, with all their parent nodes defined as $\mathrm{Pa}_{\mathcal{G}}(C_k)$ in $C_k$, $\sigma^2_{kj}$ represents the diagonal $\hat{\Sigma}_{\mathcal{G}_k}$ of the parent node k. The likelihood $\hat{\Sigma}$ is determined by the RICF algorithm (Drton et al., 2006).

### 2.3 Direct causal effect criteria

Because the BIC is a Markov equivalent score, it is incapable of orientating all edges from statistical observations. Optimising for BIC under causal insufficiency returns a PAG, or one of the MAGs that are part of the equivalence class of the optimal PAG. In this paper, we are interested in orientating all edges and discovering a MAG. We achieve this using Pearl's do-calculus (Pearl, 2000) to measure the direct causal effect on edges that the BIC score fails to orientate. The direct causal effect is estimated by intervention that renders the intervening variable independent of its parents.

## Theorem: Single-door criterion for direct effect

Single-Door Criterion for direct effect (Pearl, 2000): Given $X \to Y$, path coefficient is identifiable and equal to the regression coefficient if

- There existed a set of variable $Z$ such that $Z$ contains no descendant of $Y$
- $Z$ is d-separated set of $X$ and $Y$ in subgraph removing $X \rightarrow Y$

The interpretation of the path coefficient $(\beta)$ in the regression of the single-door criterion theorem can be expressed as the direct causal effect determined by the rate of change of $E[Y]$ given intervention $X$ (Maathuis et al., 2009) as follows.

$$\beta = \frac{\partial}{\partial x} E[Y|\, do(x)] = E[Y|do(X\ = x+1)] - E[Y|do(X = x)] \text{ for any value of } x$$

This assumes that all casual effect parameters are identifiable, and that the path coefficient or the direct causal effect is the regression coefficient estimated from the likelihood function. Let $A \rightarrow B$ be the edge in the ground truth graph, the SEM $B = \beta_A A + \epsilon_B$, if we assume that we have $A \sim \mathcal{N}\left(\mu_A, \sigma_A^2\right), \epsilon_B \sim \mathcal{N}(0, \sigma_{\epsilon_B}^2)$, and $\epsilon_B$ and $A$ are independent. Thus, $E[B] = \beta_A E[A]\,,\ {\sigma^2}_B = {\beta_A}^2 {\sigma^2}_A + \sigma_{\epsilon_B}^2$. For every pair $A$ and $B$ in the learned graph, two causal graphs where $A \rightarrow B$ and $A \leftarrow B$ need to be constructed to measure the direct causal effects. Specifically,

- For graphs $A \rightarrow B$, do the intervention on A; i.e., $do(a)$ (Pearl, 2000) (page 161)

$$\beta_A \ = \ \frac{E[BA]}{E[A^2]} \qquad (3)$$

- For graphs $B \rightarrow A$, do the intervention on B; i.e., $do(b)$.

$$\beta_B \ = \ \frac{E[AB]}{E[B^2]} \qquad (4)$$

From (3) and (4);

$$\frac{\beta_A}{\beta_B} = \ \frac{E\left[B^2\right]}{E[A^2]} \ = \ \frac{E[B]^2 + {\sigma^2}_B}{E[A]^2 + {\sigma^2}_A}$$

Substitute $E[B] = \beta_A E[A]\,,\ {\sigma^2}_B = {\beta_A}^2 {\sigma^2}_A + \sigma_{\epsilon_B}^2$ from the graph,

$$= \ \frac{\beta_A^2 E[A]^2 + {\beta_A}^2 {\sigma^2}_A + \sigma_{\epsilon_B}^2}{E[A]^2 + {\sigma^2}_A} \ = \ \beta_A^2 + \frac{\sigma_{\epsilon_B}^2}{E[A]^2 + {\sigma^2}_A} \qquad (5)$$

If $E[A]\ = \mu_A = 0, {\sigma^2}_A = 1$ and ${\sigma^2}_{\epsilon_B} = 1$ in (5)

$$\frac{\beta_A}{\beta_B} = \ {\beta_A}^2 + 1; \text{ we have the probability } (|\beta_A| > |\beta_B|) = 1$$

Algorithm 1 describes the steps of CCHM in detail.

Algorithm 1: CCHM (Conservative rule and Causal effect Hill-climbing for MAG)

Input: significance threshold alpha, maximum Sepset size $n$
Output: MAG
*// Search for a skeleton (Step 1 and 2 are the first and second steps of the cFCI Algorithm)*
Step 1 Set up a complete undirected graph and initialise Sepset $Z$ with size =0
**Repeat**
remove edges between each pair of nodes $A$ and $B$ that become independent conditional on Sepset $Z$, as determined by the significance level alpha
**Until** all Sepset $Z$ size $= n$ have been tested
Step 2 Given unshielded triple $A - C - B$, perform conditional independence tests on $A$ and $B$ given all neighbours of $A$ and $B$ as determined by the significance level alpha
a. If $C$ is **NOT** in all Sepsets of $A$ and $B$, add $A - C - B$ to the whitelist as a definite collider
b. If $C$ is in **ALL** Sepsets of $A$ and $B$, add $A - C - B$ to the blacklist as a definite non-collider
Step 3 Orientate as many edges as possible in the skeleton graph given the whitelist, and retrieve the BIC score of the resulting graph using the equation (1)
*// Score-based learning with do-calculus*
Step 5 **Repeat**
**For each** pair($A$,$B$), in ascending order by p-value
Use equation (1) to calculate the BIC scores for each edge $A \rightarrow B$, $A \leftarrow B$ and $A \leftrightarrow B$
**If** i) BIC decreases, ii) the result graph remains acyclic, and iii) the result triple is not in blacklist
**If** edges $A \rightarrow B$, $A \leftarrow B$ and $A \leftrightarrow B$ produce unequal BIC scores
Add the edge $A \rightarrow B$, $A \leftarrow B$ or $A \leftrightarrow B$ that minimises the BIC score
**Else**
Calculate the direct causal effect $\beta$ for edges $A \rightarrow B$ and $A \leftarrow B$

$$\beta_A = \mathrm{E}(\mathrm{B}|\mathrm{do}(\mathrm{A} = \mathrm{a} + 1)) - \mathrm{E}(\mathrm{B}|\mathrm{do}(\mathrm{A} = \mathrm{a}))$$
$$\beta_B = \mathrm{E}(\mathrm{A}|\mathrm{do}(\mathrm{B} = \mathrm{b} + 1)) - \mathrm{E}(\mathrm{A}|\mathrm{do}(\mathrm{B} = \mathrm{b}))$$

Orientate $A \rightarrow B$ or $A \leftarrow B$ that maximises the absolute causal effect
**End If**
**End If**
**End**
**Until** no undirected edges remain

## 3. Evaluation

The graphs produced by the CCHM algorithm are compared to the outputs of the M$^3$HC, GSPo, GFCI, RFCI, FCI, and cFCI algorithms, when applied to the same data. The M$^3$HC algorithm was tested using the MATLAB implementation by (Triantafillou et al., 2019), the GFCI and RFCI algorithms were tested using the Tetrad-based rcausal package in R (Wongchokprasitti, 2019), and the GSPo algorithm was tested using the causaldag Python package by Squires (Squires, 2018). The computational time of CCHM is compared to the M$^3$HC, FCI and cFCI, which are based on the same MATLAB package.

All experiments are based on synthetic data. However, we divide them into experiments based on data generated from BNs which had their structure and dependencies randomised, and data generated from real-world BNs. Randomised BNs were generated using Triantafillou's (Triantafillou et al., 2019) MATLAB package. We created a total of 600 random Gaussian DAGs that varied in variable size, max in-degree, and sample size. Specifically, 50 DAGs were generated for each combination of variables $V$ and max in-degree settings $\mathcal{D}$ where $V = \{10, 20, 50, 70, 100, 200\}$ and $\mathcal{D}=\{3, 5\}$. Each of those 600 graphs was then used to generate two datasets of sample sizes 1,000 and 10,000, for a total of 1,200 datasets. Data were generated assuming linear Gaussian parameters $\mu$=0 and $\sigma^2$=1 and uniformly random coefficients $\pm$[0.1,0.9] for each parent set to avoid very weak

or very strong edges. Approximately 10% of the variables in the data are made latent in each of the 600 datasets.

In addition to the randomised networks, we made use of four real-world Gaussian BNs taken from the bnlearn repository (Scutari, 2019). These are the a) *MAGIC-NIAB* (44 nodes) which captures genetic effects and phenotypic interactions for Multiparent Advanced Generation Inter-Cross (MAGIC) winter wheat population, b) *MAGIC-IRRI* (64 nodes) which captures genetic effects and phenotypic interactions for MAGIC indica rice population, c) *ECOLI70* (46 nodes) which captures the protein-coding genes of *E. coli*, and d) *ARTH150* (107 nodes) which captures the gene expressions and proteomics data of *Arabidopsis Thaliana*. Each of these four BNs was used to generate data, with the sample size set to 10,000. For each of the four datasets, we introduced four different rates of latent variable: 0%, 10%, 20% and 50%. This made the total number of real-world datasets 16; four datasets per BN.

The following hyperparameter settings are used for all algorithms: a) alpha=0.01 for the fisher's z hypothesis test for datasets sampled from the randomised [1] BNs, b) alpha=0.05, 0.01, 0.001 (all cases tested) for datasets generated by the real-world BNs, and c) the max Sepset size of the conditioning set is set to '4' so that runtime is maintained at reasonable levels. The maximum length of discriminating paths is also set to '4' for the four FCI-based algorithms (this is the same as the max Sepset size). For GSPo, the depth of depth-first search is set to '4' and the randomised points of posets to '5' (these are the default settings). Because GSPo is a non-deterministic algorithm that generates a different output each time it is executed, we report the average scores obtained over five runs. Lastly, all algorithms were restricted to a four-hour runtime limit.

Further, because the algorithms will output either a PAG or a MAG, we convert all MAG outputs into the corresponding PAGs. The accuracy of the learned graphs is then assessed with respect to the true PAG. The results are evaluated using the traditional measures of Precision and Recall, the Structural Hamming Distance (SHD) which represents the difference in the number of edges and edge orientations between the learned and the true graphs, and the Balance Scoring Function (BSF) that balances the score proportional to the number of edges, and no edges, in the true graph by taking into consideration all four confusion matrix parameters as follows (Constantinou, 2020). Specifically, $BSF = 0.5\left(\frac{TP}{a} + \frac{TN}{i} - \frac{FP}{i} - \frac{FN}{a}\right)$ where $a$ is the number of edges in the true graph, $i$ is the number of direct independences in the true graph, and $i = \frac{n(n-1)}{2} - a$. The BSF score ranges from -1 to 1, where 1 refers to the most accurate graph (i.e., matches the true graph), 0 refers to a baseline performance equal to that of a fully connected or an empty graph, and -1 refers to the worst possible graph (i.e., the reverse of the true graph).

## 4. Results

### 4.1 Random Gaussian Bayesian Networks

Fig 2 presents the Precision and Recall scores achieved by each of the algorithms achieve on the datasets generated by the randomised BNs. The scores are averaged across the different settings of variable size and max in-degree. Note that because there was no noteworthy difference between the overall results obtained from the two different data sample sizes, we only report the results based on

1. Because of the large number of experiments based on the randomised graphs (i.e., 600), we had to restrict the alpha parameter to 0.01 for all algorithms in those experiments.

sample size 10,000. Therefore, the results and conclusions based on the datasets with sample size 10,000 also hold for the datasets with sample size 1,000.

Overall, the results show that the CCHM outperforms all other algorithms in terms of both Precision and Recall, and across all settings excluding Recall under max in-degree 5 where GSPo ranks highest (Fig 2b). While GSPo appears to perform best when the number of variables is lowest, its performance decreases sharply with the number of variables, and fails to produce a result within the 4-hour time limit when the number of variables is highest.

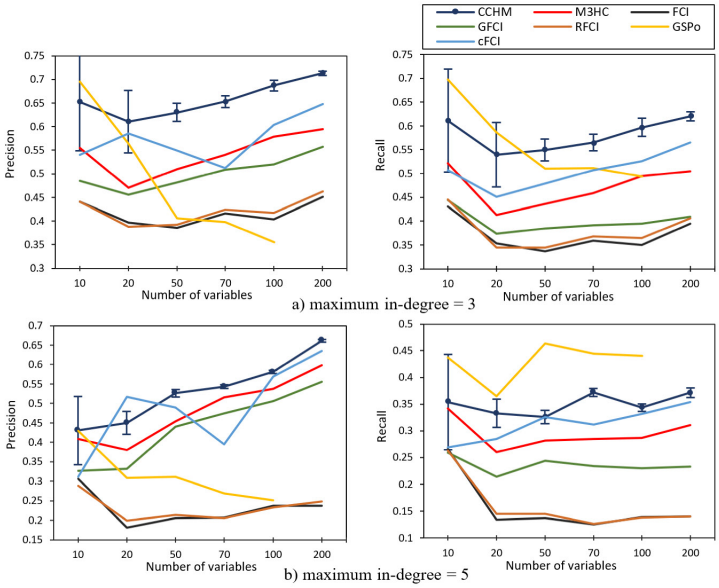


**Figure 2.** Average Precision and Recall scores of the algorithms (including score variances for CCHM) for each combination of variable size and max in-degree settings (50 graphs per combination). The results are based on synthetic data with sample size 10,000 and assume that 10% of the variables are latent.

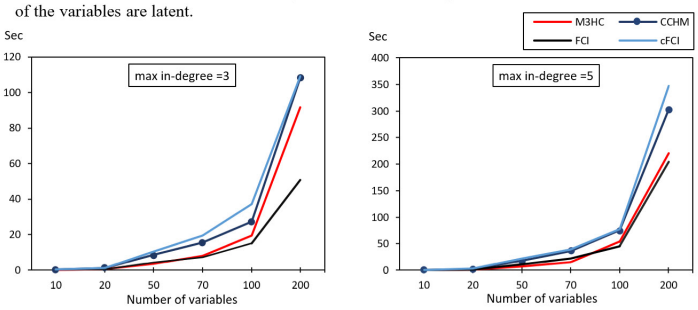


**Figure 3.** Average computation time of the algorithms for each combination of variable size and max in-degree settings (50 graphs per combination). The results are based on synthetic data with sample size 10,000 and assume that 10% of the variables are latent.

The results show no noticeable difference between FCI and its variant RFCI, whereas the cFCI and GFCI show strong improvements over FCI, with cFCI outperforming all the other FCI-based algorithms. Moreover, the performance of cFCI is on par with that of M$^3$HC. Note that while CCHM employs the BIC objective function of M$^3$HC, CCHM outperforms M$^3$HC in both sparse (Fig 2a) and dense (Fig 2b) graphs. This result provides empirical evidence that the conservative rules used in the constraint-based phase of CCHM and the do-calculus used in the score-based phase of CCHM have indeed improved structure learning performance.

Fig 3 compares the average runtime of CCHM to the runtimes of the other algorithms. The runtime comparison is restricted to algorithms that are based on the same MATLAB implementation on which CCHM is based. The results show that CCHM is marginally faster than cFCI and slower than the other algorithms, with the worst case scenario observed when the number of variables is highest, where CCHM is approximately two times slower than FCI.

Fig 4 presents the SHD and BSF scores, along with the corresponding numbers of edges generated by each algorithm. Both the SHD and BSF metrics rank CCHM highest, and these results are consistent with the Precision and Recall results previously depicted in Fig 2. The number of edges produced by CCHM is in line with the number of edges produced by the other algorithms, and this observation provides confidence that CCHM achieves the highest scores due to accuracy rather than due to the number of edges, which may sometimes bias the result of a metric (Constantinou, 2020). One inconsistency between the SHD and other metrics involves the GFCI algorithm, where SHD ranks lower than all the other FCI-based algorithms, something which contradicts the results of Precision, Recall, and BSF. Interestingly, while GSPo produces the highest BSF scores for graphs that incorporate just 10 variables, its performance diminishes drastically with the number of variables and quickly becomes the worst performer (refer to the BFS scores in Fig 4a); an observation that is largely consistent with the results in Fig 2.

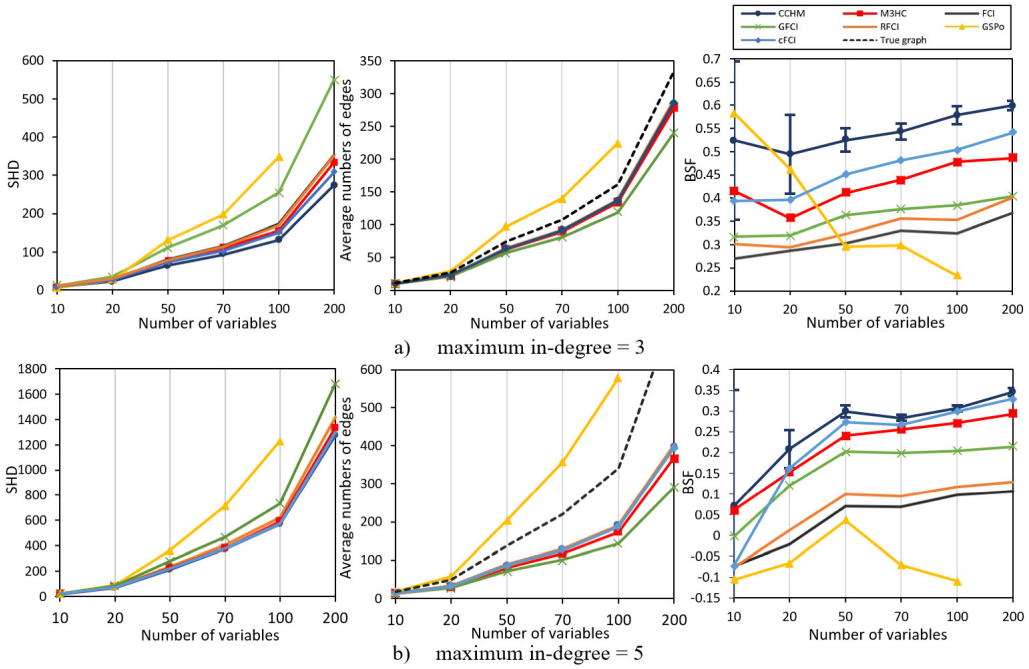


**Figure 4.** Average number of edges, SHD and BSF scores of the algorithms (including BSF variances for CCHM) for each combination of variable size and max in-degree settings (50 graphs per combination). The results are based on synthetic data with sample size 10,000 and assume that 10% of the variables are

## 4.2 Real-world Gaussian Bayesian Networks

The reduced number of experiments that associate with the real-world GBNs (i.e., 16 instead of 600 randomised experiments) enabled us to also test the sensitivity of the algorithms on the alpha hyperparameter, which reflects the significance cut-off point in establishing independence. Fig 5

presents the SHD scores for each of the four real-world GBNs, and over different rates of latent variables. The results are restricted to the top three algorithms for each case study, and this is because we report three different results for each of the top three algorithms based on the three different hyperparameter inputs alpha specified in Fig 5.

Only four algorithms (CCHM, M$^3$HC, cFCI and GSPo) achieved a top-three performance in any of the four networks, and this suggests that the relative performance between algorithms is rather consistent across the different case studies. While there is no clear relationship between the rate of latent variables and SHD score, the results do suggest that the accuracy of the algorithms decreases with the rate of latent variables in the data. This is because while we would expect the SHD score to decrease with less variables in the data, since less variables lead to potentially fewer differences between the learned and the true graphs (refer to Fig 4), the results in Fig 5 reveal a weak increasing trend in SHD score with the rate of latent variables in the data.

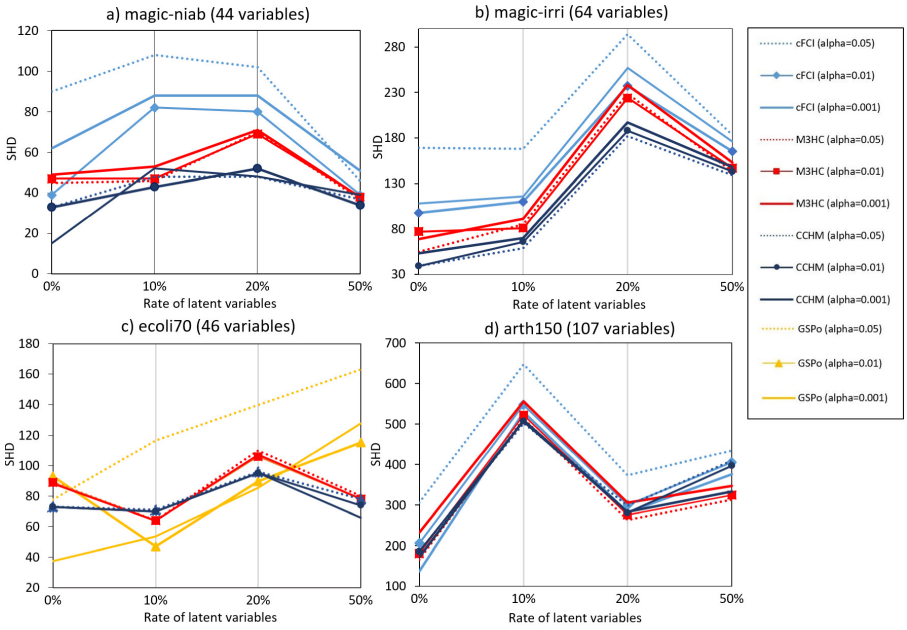


**Figure 5.** The SHD scores of the top three algorithms in each of the four Gaussian BNs, over three different input settings for hyperparameter alpha. The results are based on synthetic data with sample size 10,000.

Overall, the CCHM algorithm was part of the top three algorithms in all the four case studies. Specifically, CCHM generated the lowest SHD error in networks (a) and (b). The results in network (c) were less consistent, with GSPo ranked 1$^{st}$ at latent variable rates of 10% and 20%, and CCHM ranked 1$^{st}$ at latent variable rates of 0% and 50%. In contrast, the results based on network (d) show no noteworthy differences in the performance between the three top algorithms. Overall, the results suggest that cFCI and GSPo are much more sensitive to the alpha hyperparameter compared to the CCHM and M$^3$HC algorithms, and that CCHM generally performs best when alpha=0.01.

## 5. Discussion and future work

This paper builds on recent developments in BN structure learning under causal insufficiency and describes a novel structure learning algorithm, called CCHM, that combines constraint-based and score-based learning with causal effects to learn GBNs. The constraint-based part of CCHM adopts

features from the state-of-the-art cFCI algorithm, whereas the score-based part is based on traditional hill-climbing greedy search that minimises the BIC score. CCHM applies Pearl's do-calculus as a method to orientate the edges that both constraint-based and score-based learning fail to do so from observational data. The results show that CCHM outperforms the state-of-the-art algorithms in the majority of the experiments, which include both randomised and real-world GBNs.

A limitation of this work is that the algorithm assumes linear GBNs and that the data are continuous. Future work will extend this approach to discrete BNs, where causal insufficiency remains an important open problem (Jabbari et al., 2017). Other directions include investigating different strategies in the way the do-calculus effect is applied to the process of structure learning; e.g., it can be applied directly to the calculation of the BIC score during score-based learning, or computed as the total causal effect of the graph using do-calculus rules or via back-door adjustment with graph surgery. Lastly, causal insufficiency represents just one type of data noise that exist in real-world datasets, and future work will also investigate the effects of causal insufficiency when combined with other types of noise in the data.

## Acknowledgments

This research was supported by the ERSRC Fellowship project EP/S001646/1 on *Bayesian Artificial Intelligence for Decision Making under Uncertainty* (Constantinou, 2018), by The Alan Turing Institute in the UK under the EPSRC grant EP/N510129/1, and by the Royal Thai Government Scholarship offered by Thailand's Office of Civil Service Commission (OCSC).

## References

D. I. Bernstein, B. Saeed, C. Squires, and C. Uhler. Ordering-based causal structure learning in the presence of latent variables. *ArXiv*, abs/1910.09014, 2019.

D. Colombo, M. Maathuis, M. Kalisch, and T. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics - ANN STATIST*, 40, 04 2011. doi: 10.1214/11-AOS940.

A. Constantinou, N. Fenton, and M. Neil. Integrating expert knowledge with data in bayesian networks: Preserving data-driven expectations when the expert variables remain unobserved. *Expert Systems with Applications*, 56, 03 2016. doi: 10.1016/j.eswa.2016.02.050.

A. C. Constantinou. Bayesian artificial intelligence for decision making under uncertainty. *Engineering and Physical Sciences Research Council (EPSRC)*, EP/S001646/1, 2018.

A. C. Constantinou. Evaluating structure learning algorithms with a balanced scoring function. arXiv: 1905.12666 [cs.LG], 2020.

A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. 01 2009. ISBN 978-0-521-88438-9. doi: 10.1017/CBO9780511811357.

M. Drton, M. Eichler, and T. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10, 01 2006. doi: 10.1145/1577069.1755864.

F. Jabbari, J. Ramsey, P. Spirtes, and G. F. Cooper. Discovery of causal models that contain latent variables through bayesian scoring of independence constraints. *Machine learning and knowledge discovery in databases : European Conference, ECML PKDD*, 2017:142–157, 2017.

M. H. Maathuis, M. Kalisch, and P. Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, Dec 2009. ISSN 0090-5364. doi: 10.1214/09-aos685.

C. Nowzohour, M. Maathuis, and P. Bühlmann. Structure learning with bow-free acyclic path diagrams. *arXiv*, 2015.

J. M. Ogarrio, P. Spirtes, and J. Ramsey. A hybrid causal search algorithm for latent variable models. In A. Antonucci, G. Corani, and C. P. Campos, editors, *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*, pages 368–379, 2016.

J. Pearl. Causality: Models, reasoning, and inference, second edition. *Causality*, 29, 01 2000. doi: 10.1017/CBO9780511803161.

J. Ramsey, J. Zhang, and P. Spirtes. Adjacency-faithfulness and conservative causal inference. *CoRR*, abs/1206.6843, 2012.

J. D. Ramsey. Scaling up greedy equivalence search for continuous variables. *CoRR*, abs/1507.07749, 2015.

T. Richardson and P. Spirtes. Ancestral graph markov models. *Annals of Statistics*, 30, 11 2000. doi: 10.1214/aos/1031689015.

M. Scutari. *Bnlearn dataset repository*, 2019. URL `https://www.bnlearn.com/bnrepository`.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search, 2nd Edition*, volume 1 of *MIT Press Books*. The MIT Press, August 2001. ISBN ARRAY(0x479b6ad8).

C. Squires. *causaldag Python library*, 2018. URL `https://github.com/uhlerlab/causaldag`.

S. Triantafillou and I. Tsamardinos. Score-based vs constraint-based causal learning in the presence of confounders. In *CFA@UAI*, 2016.

S. Triantafillou, K. Tsirlis, V. Lagani, and I. Tsamardinos. *MATLAB library*, 2019. URL `https://github.com/mensxmachina/M3HC`.

K. Tsirlis, V. Lagani, S. Triantafillou, and I. Tsamardinos. On scoring maximal ancestral graphs with the max–min hill climbing algorithm. *International Journal of Approximate Reasoning*, 102, 08 2018. doi: 10.1016/j.ijar.2018.08.002.

C. Wongchokprasitti. *R-causal R Wrapper for Tetrad Library, v1.1.1*, 2019. URL `https://github.com/bd2kccd/r-causal`.

J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172, 11 2008. doi: 10.1016/j.artint.2008.08.001.