Empirical Bayesian Approaches for Robust Constraint-based Causal Discovery under Insufficient Data

Zijun Cui[1] , **Naiyu Yin**[1] , **Yuru Wang**[*2] and **Qiang Ji**[1]
[1]Rensselaer Polytechnic Institute
[2]Northeast Normal University
cuiz3@rpi.edu, yinn2@rpi.edu, wangyr915@nenu.edu.cn, jiq@rpi.edu

Abstract

Causal discovery is to learn cause-effect relationships among variables given observational data and is important for many applications. Existing causal discovery methods assume data sufficiency, which may not be the case in many real world datasets. As a result, many existing causal discovery methods can fail under limited data. In this work, we propose Bayesian-augmented frequentist independence tests to improve the performance of constraint-based causal discovery methods under insufficient data: 1) We firstly introduce a Bayesian method to estimate mutual information (MI), based on which we propose a robust MI based independence test; 2) Secondly, we consider the Bayesian estimation of hypothesis likelihood and incorporate it into a well-defined statistical test, resulting in a robust statistical testing based independence test. We apply proposed independence tests to constraint-based causal discovery methods and evaluate the performance on benchmark datasets with insufficient samples. Experiments show significant performance improvement in terms of both accuracy and efficiency over SOTA methods.

1 Introduction

Learning causal relations has been a fundamental and widely-investigated topic. The causal relations are captured by a directed acyclic graph (DAG), and a directed link in DAG captures cause-effect relation between two variables connected by the link. Specifically, a directed link from node X to node Y indicates the cause-effect relation between cause variable X and effect variable Y. Causal discovery aims at learning a DAG capturing causal-effect relationships among a set of random variables from observational data. Existing causal discovery methods focus on learning a DAG with high confidence from sufficient data samples. Not much attention, however, has been paid to performance improvement of causal discovery under limited data. Such work is important, as even in the era of big data, there are still domains in which the availability of data is very limited. For example, in biological or clinical disciplines, data can be severely insufficient either because of high cost or lack of cases from which data is collected [Mukherjee and Speed, 2007]. Furthermore, even for applications with a vast amount of data, the data may not adequately cover all possible states of the nodes, leading to insufficient data for certain states. For example, the observed data under the absence of earthquake is adequate, while the observed data under the occurrence of earthquake is limited, due to the fact that earthquake rarely happens in nature.

Constraint-based causal discovery methods apply independence tests to determine a DAG from observational data and can be performed globally or locally. Global approaches aim at learning cause-effect relationships among all random variables, such as PC-stable [Colombo and Maathuis, 2014], and Sepset consistent PC (SC-PC) [Li *et al.*, 2019]. Global causal discovery methods discussed above learn DAGs that are in the same markov equivalent class of ground truth DAG. Further tests under certain assumptions about the graph or data distribution are needed to resolve the causal ambiguity [Glymour *et al.*, 2019]. In this paper, we focus on learning markov equivalent DAGs. In contrast to global approaches, local approaches identify the direct causes and effects of a target variable, represented by a causal Markov Blanket [Gao and Ji, 2015; Yang *et al.*, 2021]. A causal Markov Blanket captures local relationships of a target variable by identifying its parents, children, and spouses. For both global and local approaches, the main challenge of constraint-based causal discovery methods is that their performance highly depends on the accuracy of the independence test. Independence test error, even one mistake in independence decision, can propagate throughout the graph, causing a sequence of errors and resulting in an erroneous DAG with incorrect orientations [Spirtes, 2010]. Hence, to perform a robust constraint-based causal discovery, it is crucial to improve the robustness of the independence test.

To improve the causal discovery performance under insufficient data, we propose to introduce Bayesian approaches to independence tests for accurate and efficient constraint-based causal discovery. Specifically, two Bayesian-augmented frequentist independence tests are proposed, whereby we use Bayesian approach to reliably estimate, under low data regime, independence test statistics used by frequentist independence tests. For both MI estimation (Sec.3.1) and hy-

*Corresponding author

pothesis likelihood estimation (Sec.3.2), we employ Bayesian inference to calculate statistics by considering the entire parameter space instead of using a point estimate one. Given the estimated Bayesian statistics, we follow the standard frequentist framework to perform independence test. The proposed Bayesian-augmented independence tests are then applied to improve the constraint-based causal structure learning. We evaluate both local and global causal discovery performance with proposed independence tests on benchmark datasets and compare them to state-of-the-art methods. We empirically demonstrate the effectiveness of the proposed Bayesian approaches in improving both the accuracy and efficiency of the local and global causal discovery under insufficient data.

2 Related Work

To handle causal discovery under insufficient data, some methods downsize the problem domain to sub-domains. Rohekar *et al.*, [2020] approximated the structure by performing independence tests with a small fixed size of the condition set. The structure was then refined by iteratively increasing the condition set. A similar idea was explored in the Recursive Autonomy Identification (RAI) method [Yehezkel and Lerner, 2009]. Related works along this line always assume that there exist sufficient data for sub-domains. Besides, Claassen and Heskes [2012] estimated the posterior distribution of the independence hypothesis between two variables, based on which reliability was quantified. The causal discovery was then processed in decreasing order of reliability. Rohekar *et al.*, [2018] estimated posterior distribution of DAG through bootstrap samples. The negative effect from independence tests error was minimized through model averaging.

Some causal discovery methods address the limited data issue by directly improving the independence test [Marx and Vreeken, 2018]. A Bayesian-augmented frequntist independence test based on Bayes Factor (BF) was proposed [Natori *et al.*, 2017] whereby Bayesian parameter estimate is employed in computing BF while the value of BF is then applied to a frequentist independence test. The proposed independence test is incorporated into RAI, achieving competitive DAG learning performance. However, a threshold is required in [Natori *et al.*, 2017] and the selection of threshold can be heuristic. Instead, we propose to formulate the Bayes Factor into a well-defined statistical independence test without requiring threshold tuning.

In addition, different approaches have been proposed for robust independence tests under insufficient data. These methods, however, are not aimed at improving the causal discovery performance. Seok and Seon Kang [2015] improved the estimation of mutual information (MI) by partitioning the whole sample space into sub-regions. For better MI estimation under limited data, Bayesian approaches have been widely considered [Hutter, 2002; Archer *et al.*, 2013]. Another category of recent independence test techniques are focused on developing non-parametric methods to improve efficiency, such as CCIT [Sen *et al.*, 2017] and RCIT [Strobl *et al.*, 2019]. These works assume the availability of sufficient data and are mainly focused on continuous variables, while we are focused on discrete ones.

3 Proposed Methods

We consider two types of independence test: MI based and statistical testing based independence tests. We introduce Bayesian approaches to improve both types of independence tests through a Bayesian-augmented frequentist framework. Particularly, for MI based approach, we employ empirical Bayesian approach for better MI estimation under limited data. For statistical testing based approach, we consider the empirical Bayesian estimation of hypothesis likelihood and formulate it into χ^2 statistical independence test, providing an accurate p-value under limited data.

3.1 Bayesian Approach for Mutual Information Based Independence Test

The mutual information (MI) of two discrete random variables X and Y is defined as $MI(X;Y) = \sum_{i=1}^{K_x}\sum_{j=1}^{K_y} P(x_i,y_j)\log\frac{P(x_i,y_j)}{P(x_i)P(y_j)}$, where K_x and K_y denote the total number of possible states of X and Y respectively. $P(x_i,y_j)$, $P(x_i)$, and $P(y_j)$ represent the joint probability of (X,Y), and the marginal probabilities of X and Y respectively. By definition, $MI(X;Y) = 0$ if and only if X and Y are independent. In practice, the true MI is unknown, and the estimated MI is always larger than zero. In the following, we denote the probability distribution parameters as $\boldsymbol{\theta}$, i.e., $P(x_i) = \theta_i, P(y_j) = \theta_j$ and $P(x_i,y_j) = \theta_{ij}$. Conventionally, MLE is employed to estimate $\boldsymbol{\theta}$ from data as $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})$, where $P(D|\boldsymbol{\theta})$ is the likelihood of parameter $\boldsymbol{\theta}$ given the data D. MI is then estimated as $MI = MI(X;Y|\hat{\boldsymbol{\theta}})$. When data is insufficient, MLE is not reliable [Geweke and Singleton, 1980] and MI tends to be overestimated. Instead, the full Bayesian MI is estimated from data over the entire parameter and hyper-parameter space, i.e.,

$$\begin{aligned} \hat{MI}^{fB} &= MI(X;Y|D) \\ &= \int\int MI(X;Y|\boldsymbol{\theta},\alpha)p(\boldsymbol{\theta},\alpha|D)d\boldsymbol{\theta}d\alpha \\ &= \int\int MI(X;Y|\boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha,D)p(\alpha|D)d\boldsymbol{\theta}d\alpha \end{aligned} \tag{1}$$

where α is the hyper-parameter for symmetric Dirichlet prior of $\boldsymbol{\theta}$. The full Bayesian MI is the expected MI over the joint posterior distribution of the parameters and hyper-parameter, i.e., $p(\boldsymbol{\theta},\alpha|D)$. The integration over hyper-parameter α can be computationally challenging [Archer *et al.*, 2013]. Instead of marginalizing out α, we propose to maximize it out. Particularly, we approximate the integration over the hyper-parameter space by its mode α^* that maximizes a posterior (MAP) of α, i.e., $\alpha^* = \arg\max_\alpha p(\alpha|D)$. By assuming uniform distribution of $p(\alpha)$, we have $\alpha^* = \arg\max_\alpha p(D|\alpha)$. The likelihood $p(D|\alpha)$ can be computed as,

$$p(D|\alpha) = N!\frac{\Gamma(K\alpha)}{\Gamma(K\alpha+N)}\prod_{i=1}^{K}\frac{\Gamma(\alpha+n_i)}{\Gamma(\alpha)n_i!} \tag{2}$$

where K is the number of states for the random variable, n_i is the number of samples for state i, and $N = \sum_i^K n_i$. $P(D|\alpha)$

follows Polya distribution and $\Gamma(x)$ is the gamma function. We solve for α^* with a fixed-point update [Minka, 2000]. Given α^*, the full Bayesian method is converted to the empirical Bayesian method, and we have the proposed empirical Bayesian MI $\hat{MI}^{eB}$ defined as,

$$\hat{MI}^{eB} = \int MI(X;Y|\boldsymbol{\theta})p(\boldsymbol{\theta}|D,\alpha^*)d\boldsymbol{\theta} \tag{3}$$

with a closed-form solution:

$$\begin{aligned}\hat{MI}^{eB} &= \psi(N+\alpha^*K+1) \\ &- \sum_{ij}\frac{n_{ij}+\alpha^*}{N+\alpha^*K}[\psi(n_i+\alpha^*K_y+1) \\ &+ \psi(n_j+\alpha^*K_x+1) - \psi(n_{ij}+\alpha^*+1)\end{aligned} \tag{4}$$

where $\psi(x)$ is the digamma function. n_i and n_j are the number of samples for $X=i$ and $Y=j$ respectively, and n_{ij} is the number of samples for $(X,Y)=(i,j)$. Given the estimated MI, we compare it against a pre-defined threshold for independence test. If MI is smaller than the threshold, two random variables will be declared to be independent, and dependent otherwise.

3.2 Bayesian Approach for Statistical Testing Based Independence Test

We now introduce our proposed Bayesian approach to improve the statistical testing based independence test. We firstly consider a standard independence test, G test [McDonald, 2009], which is a likelihood ratio test with null hypothesis assuming two random variables are independent. G test is a widely used statistical test. As the same with other statistical tests, G test doesn't require threshold tuning and the significance level is set to be 5% by default. The formula for the statistic G reads as $G = -2\sum_{i=1}^{K_x}\sum_{j=1}^{K_y} n_{ij}\ln\frac{\hat{\theta}_i\hat{\theta}_j}{\hat{\theta}_{ij}}$ where $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} P(D|\boldsymbol{\theta})$. Samples $D=\{D_n\}_{n=1}^N$ are $i.i.d$ given parameter $\hat{\boldsymbol{\theta}}$ and the statistic G follows asymptotic $\chi^2_{df=(K_x-1)(K_y-1)}$ distribution, based on which a statistical test can be performed. As MLE parameter estimates are not reliable under insufficient data, leading to inaccurate estimation of the likelihood of hypothesis, we instead consider the empirical Bayesian estimation. Specifically, we employ the Bayes Factor (BF) [Kass and Raftery, 1995] which defines the ratio of expected likelihoods of null hypothesis(H_0) and that of the alternative hypothesis(H_1) over all possible parameter settings with the posterior distributions of parameters under null and alternate hypothesis respectively,

$$BF = \frac{P(D|H_0,\alpha^0)}{P(D|H_1,\alpha^1)} = \frac{\int P(D|\boldsymbol{\theta})P(\boldsymbol{\theta}|H_0,\alpha^0)d\boldsymbol{\theta}}{\int P(D|\boldsymbol{\theta})P(\boldsymbol{\theta}|H_1,\alpha^1)d\boldsymbol{\theta}} \tag{5}$$

where α^0 and α^1 are the hyper-parameters for the symmetric Dirichlet prior under null and alternate hypothesis respectively. Both hypothesis likelihoods $P(D|H_0,\alpha^0)$ and $P(D|H_1,\alpha^1)$ can be analytically solved, and BF can be computed. However, BF can't be directly applied to a statistical test because samples $D=\{D_n\}_{n=1}^N$ are not $i.i.d$ given hyper-parameter α and BF no longer follows the χ^2 distribution under the null hypothesis. Instead, we propose to approximate $P(D|\alpha)$ by a multinomial distribution and calculate modified parameters of multinomial distribution with α taken into account, as both capture the distributions for integer random variables, i.e.,

$$P(D|\alpha) \approx P(D|\tilde{\boldsymbol{\theta}}) = \frac{N!}{\prod_{i=1}^K n_i!}\prod_{i=1}^K \tilde{\theta}_i^{\,n_i} \tag{6}$$

where K is the total number of states, and $N=\sum_{i=1}^K n_i$ is the total number samples with n_i being the number of samples for state i. $\tilde{\theta}_i$ are the modified parameters of the multinomial distribution. $\tilde{\theta}_i = \frac{g(n_i,\alpha)}{g(N,K\alpha)}$ with $g(n_i,\alpha) = an_i + b\alpha$ where $\Lambda = \begin{pmatrix} a \\ b \end{pmatrix}$ are unknown coefficients. By plugging the $P(D|\alpha)$ (defined in Eq. 2) into Eq. 6, it is clear that to satisfy Eq. 6, we must have $n_i \ln g(n_i,\alpha) = \ln\Gamma(n_i+\alpha) - \ln\Gamma(\alpha)$. Given $\{n_i\}_{i=1}^K$ and α, we can construct a system of K such equations through which we can solve for Λ^*, i.e.,

$$\Lambda^* = \arg\min_{\Lambda} ||M\Lambda - T||_2^2 = (M^tM)^{-1}M^tT \tag{7}$$

with $M = \begin{pmatrix} n_1,\alpha \\ n_2,\alpha \\ \dots \\ n_K,\alpha \end{pmatrix}$, $T = \begin{pmatrix} t(n_1,\alpha) \\ t(n_2,\alpha) \\ \dots \\ t(n_K,\alpha) \end{pmatrix}$, and $t(n_i,\alpha) = \exp(\frac{1}{n_i}(\ln\Gamma(n_i+\alpha) - \ln\Gamma(\alpha)))$. Given Λ^*, we have $\tilde{\theta}_i$ as $\tilde{\theta}_i = \frac{g(n_i,\alpha)}{g(N,K\alpha)} = \frac{a^*n_i+b^*\alpha}{a^*N+b^*K\alpha}$. $P(D|\tilde{\boldsymbol{\theta}})$ can well approximate $P(D|\alpha)$. Our proposed approximation is different from the method provided in [Minka, 2000] where the Polya distribution $P(D|\alpha)$ is interpreted as a multinomial distribution with modified counts $\tilde{n}_i$. In addition, our proposed estimation can better approximate the Polya distribution given the symmetric Dirichlet prior compared to [Minka, 2000]. We then approximate the hypothesis likelihood under null and alternative hypothesis respectively and obtain a modified Bayes Factor $\tilde{BF}$

$$\tilde{BF} = \frac{P(D|\tilde{\boldsymbol{\theta}},H_0)}{P(D|\tilde{\boldsymbol{\theta}},H_1)} = \frac{\prod_{i=1}^{K_x}\tilde{\theta}_i^{\,n_i}\prod_{j=1}^{K_y}\tilde{\theta}_j^{\,n_j}}{\prod_{i=1,j=1}^{K_xK_y}\tilde{\theta_{ij}}^{\,n_{ij}}} \tag{8}$$

We obtain the statistic BF_{chi2} for the statistical test as,

$$BF_{chi2} = -2\ln\tilde{BF} = -2\sum_{i=1}^{K_x}\sum_{j=1}^{K_y} n_{ij}\ln\frac{\tilde{\theta}_i\tilde{\theta}_j}{\tilde{\theta_{ij}}} \tag{9}$$

The statistic BF_{chi2} asymptomatically follows the $\chi^2_{df=(K_x-1)(K_y-1)}$ distribution. If p-value is smaller than the significance level, we reject the null hypothesis and accept the alternative hypothesis. It is worth noting that BF can be directly applied for a frequentist independence test where a pre-defined threshold η is required [Natori *et al.*, 2017]. The value of the threshold is unconstrained [Kass and Raftery, 1995] making it hard to be properly selected. Instead, our approach only requires a significant level for independence test which is usually set to be 5% by default.

Dataset	Size	SHD			#Independence Test		
		cI^{eB}	cBF_{chi2}	CMB	cI^{eB}	cBF_{chi2}	CMB
CHILD	100	2.90±0.28	2.65±0.40	5.94±0.65	1008	1154	16869
	300	2.61±0.26	2.64±0.59	6.95±0.63	1709	1926	14578
	500	2.29±0.31	2.24±0.84	4.52±0.58	2524	4751	13873
	MEAN	2.60	**2.51**	5.80	**1747**	2610	15107
INSURANCE	100	3.89±0.34	3.98±0.39	7.18±0.66	1261	1363	22168
	300	3.47±0.21	3.24±0.12	7.59±0.57	1541	2977	18043
	500	3.11±0.21	2.98±0.13	7.20±0.67	1477	3949	14881
	MEAN	3.49	**3.40**	7.32	**1426**	2763	18364
ALARM	100	2.69±0.07	2.39±0.19	5.20±0.71	1424	1109	27492
	300	2.50±0.19	2.27±0.15	4.36±0.83	2398	3885	14900
	500	2.40±0.11	2.26±0.19	3.53±0.62	2807	4766	11328
	MEAN	2.53	**2.31**	4.36	**2210**	3253	17907
HAILFINDER	500	3.33±0.02	4.22±0.04	7.90±0.11	676	1923	183350
	800	3.56±0.01	4.49±0.13	7.12±0.09	1098	2145	169705
	1000	3.56±0.09	4.45±0.08	7.10±0.11	1924	2621	119815
	MEAN	**3.48**	4.39	7.37	**1233**	2229	157620
CHILD3	500	2.46±0.23	2.53±0.18	4.72±0.28	7168	7417	14789
	800	3.01±0.13	2.67±0.11	3.57±0.21	6720	7802	9765
	1000	2.90±0.07	2.57±0.23	3.09±0.19	8424	8285	9516
	MEAN	2.79	**2.59**	3.79	**7437**	7835	11357
CHILD5	500	2.87±0.05	2.62±0.19	5.00±0.15	5234	11126	16819
	800	2.66±0.21	3.02±0.13	5.75±0.32	8236	11424	51967
	1000	2.82±0.23	2.99±0.07	4.34±0.19	13384	9956	36888
	MEAN	**2.78**	2.88	5.03	**8951**	10835	26322

Table 1: Local causal discovery performance under insufficient data

4 Experiments

We evaluate both the local and global constraint-based causal discovery performance on benchmark datasets. Through exhaustive experiments, we show that our approaches can significantly improve causal discovery performance in terms of both accuracy and efficiency over state-of-the-art methods. Besides, we compare proposed independence tests to state-of-the-art independence tests to further show the effectiveness of the proposed methods.

Experiment Settings. We employ six benchmark datasets[1] that are widely used for causal discovery evaluation: CHILD, INSURANCE, ALARM, HAILFINDER, CHILD3 and CHILD5. The causal discovery performance is evaluated in terms of both accuracy and efficiency. For accuracy, we employ the structural hamming distance (SHD) [Tsamardinos *et al.*, 2006a]. SHD computes the number of extra and incorrect (missing and reverse) edges in the learned causal structure compared to the ground truth one. For efficiency, we consider the number of conducted independence test. We perform evaluation on a number of small sized datasets. These small sample sizes are chosen to mimic insufficient data scenario through significantly small number of samples per configuration. For each sample size, we repeat 10 runs and report the averaged performance over 10 runs. In addition, we report standard derivation of SHD. Experiments are performed on a laptop with a 8-Core Intel Core i9 processor with CPU only.

[1]https://www.bnlearn.com/bnrepository/.

4.1 Local Constraint-based Causal Discovery

For the local causal discovery, we employ Causal Markov Blanket (CMB) [Gao and Ji, 2015], which is the state-of-the-art method. CMB employs constraint-based approach and performs conditional independence test using MI to identify the CMB of a target node. We incorporate the proposed independence tests into CMB and compare to the original CMB. We denote cI^{eB} as the CMB with empirical Bayesian MI estimation and cBF_{chi2} as CMB with BF_{chi2} independence test. SHD is 0 if learned CMB is identical to the ground truth CMB.

From Table 1, we can see that both cI^{eB} and cBF_{chi2} outperform the CMB on all datasets in terms of both accuracy and efficiency under insufficient data. The number of performed independence test reduces dramatically. On ALARM dataset, cI_{eB} only performs 2210 independence tests on average, while CMB requires 17907 tests on average. The proposed methods improve the accuracy significantly. On INSURANCE dataset, cBF_{chi2} improves the averaged SHD by 3.92 compared to CMB. From the results we can see that, by introducing Bayesian approaches, both the accuracy and the efficiency can be improved. Comparing the performance between the two proposed methods, cBF_{chi2} achieves overall better accuracy, and cI^{eB} is more efficient with the fewest number of independence test on all datasets.

It is worth noting that the number of independence test increases with reduced samples in CMB, but decreases with the proposed methods. The reason is that under insufficient data,

Dataset	Size	SHD				#Independence Test			
		rI^{eB}	rBF_{chi2}	RAI-BF	PC-Stable	rI^{eB}	rBF_{chi2}	RAI-BF	PC-Stable
CHILD	100	21.6±2.1	24.2±2.3	30.4±3.7	23.8±1.7	283	314	893	559
	300	19.9±2.7	17.7±1.8	23.5±4.4	22.6±1.9	342	546	997	986
	500	17.6±1.7	16.0±2.9	22.6±2.4	24.4±2.2	424	754	975	1317
	MEAN	19.7	**19.3**	25.5	23.6	**350**	538	955	954
INSURANCE	100	48.9±1.3	50.1±2.9	54.9±3.6	52.0 ±1.5	486	604	905	1217
	300	47.3±0.8	44.5±2.0	46.6±3.2	50.2±3.1	576	986	1011	1250
	500	49.5±1.8	39.4±3.0	47.1±2.2	50.7±2.5	662	1200	1120	2326
	MEAN	48.6	**44.7**	49.5	51.0	**575**	930	1012	1598
ALARM	100	44.5±2.2	42.7±2.3	48.4±5.8	45.8 ±4.9	891	958	1591	2215
	300	40.7±3.0	36.1±4.5	35.3±5.4	34.6±2.7	1158	1752	1881	3398
	500	40.0±3.1	29.8±5.1	29.8±5.2	36.5 ±5.7	1433	2018	2098	3992
	MEAN	41.7	**36.2**	37.8	39.0	**1161**	1576	1857	3202
HAILFINDER	500	88.0±2.0	98.3 ±1.5	118.0±1.0	91.6±1.0	2024	2587	6171	3267
	800	85.0±1.7	106.3 ±2.1	124.7 ±6.7	99.7±1.2	1983	3726	7847	3423
	1000	92.3±4.5	108.3 ±2.3	131.3 ±3.2	101.8±2.2	2638	3073	16618	3603
	MEAN	**88.4**	104.3	124.7	97.7	**2215**	3129	10212	3431
CHILD3	500	67.6±3.2	54.3±2.6	79.6±4.9	81.2±2.8	2693	3796	5422	4963
	800	65.8±2.5	52.9±2.8	74.0±3.7	79.9±2.4	3941	4587	5106	6026
	1000	61.5±3.8	52.3±3.9	71.0±6.5	81.4±2.7	4723	5170	5980	6846
	MEAN	65.0	**53.2**	74.9	80.8	**3786**	4518	5503	5945
CHILD5	500	122.0±2.6	109.3±5.1	134.0 ±2.6	113.9±2.4	6966	8646	10038	10253
	800	121.7±3.8	105.3±4.0	132.3±6.7	120.1±2.9	10249	10431	9337	10708
	1000	116.3±2.9	105.7±2.5	126.3±7.0	123.4±1.7	10375	10494	11174	11070
	MEAN	120.0	**106.8**	126.3	119.1	**9197**	9857	11174	10677

Table 2: Global causal discovery performance under insufficient data

MLE will lead to an overestimated MI. Hence, conventional MI based independence test is likely to declare dependence when data size is small, resulting in a large number of independence test. As the sample size increases, the incorrect dependency declarations will be corrected and the number of independence tests will decrease. On the other hand, our methods are more accurate and show a preference of independence under insufficient data, resulting a small number of performed independence test.

4.2 Global Constraint-based Causal Discovery

Majority of global causal discovery algorithms are under causal sufficiency assumption, whereby all random variables are observed in data and there is no latent variable. However, causal sufficiency assumption can be violated since the real data may fail to capture the values for all the variables, leaving some variables to be latent. To address this issue, several recent causal discovery methods [Ramsey *et al.*, 2012; Colombo *et al.*, 2012] have been developed to identify latent common confounders of the observed variables. In our evaluations, we mainly focus on standard algorithms that are under causal sufficiency assumptions. We firstly employ RAI [Yehezkel and Lerner, 2009] as our baseline and compare to two state-of-the-art methods. Then, to demonstrate that our proposed methods can consistently improve causal discovery performance, we consider well-known DAG learning algorithms: PC [Spirtes *et al.*, 2000] and MMHC [Tsamardinos *et al.*, 2006b] as two additional baselines. In the end, we consider the algorithms without causal sufficiency assumption to demonstrate that our proposed methods can be applied to different causal discovery methods, independent of the existence of latent confounders.

Global causal discovery with causal sufficiency assumption. We employ RAI as our baseline algorithm and incorporate the proposed independence tests. We denote rI^{eB} as the RAI with empirical Bayesian MI estimation, and rBF_{chi2} as RAI with BF_{chi2} independence test. We compare our approaches to two state-of-the-art methods: RAI-BF method [Natori *et al.*, 2017] and PC-stable [Colombo and Maathuis, 2014]. SC-PC can't be performed under insufficient data smoothly, and thus we exclude this method for comparison. SHD is 0 if the learned DAG and the ground truth DAG belong to the same equivalence class.

From Table 2, we can see that rBF_{chi2} outperforms RAI-BF and PC-stable on almost all datasets in terms of both accuracy and efficiency. rI^{eB} also achieves overall better accuracy and significantly improves efficiency. For example, on CHILD3, rBF_{chi2} improves the SHD by 21.7 and 27.6 compared to RAI-BF and PC-stable. In terms of efficiency, on HAILFINDER, rI^{eB} only performs 2215 independence tests in average, while RAI-BF requires 10212 tests in average. Comparing between the two proposed methods, rBF_{chi2} achieves better accuracy and rI^{eB} achieves better efficiency. With the proposed methods, the number of independence tests decreases due to the reduced samples for all datasets, which is consistent with the conclusion we have from the local causal discovery. In addition, both RAI-BF and PC-stable show a preference of independence under insufficient data, leading to the decreased number of independence tests with reduced number of samples.

Since BF_{chi2} essentially is only an approximate of original BF, BF with the optimal threshold should outperform BF_{chi2}

in principle. However, selecting the optimal threshold for BF can be challenging and incorrect thresholds can lead to inferior causal discovery performance. Instead of fixing the threshold of RAI-BF with its default value, we consider the optimal performance of RAI-BF with tuned thresholds for comparison. According to the experimental results, RAI-BF with the optimally tuned threshold at best achieves comparable performance compared to rBF_{chi2} in terms of both accuracy and efficiency, which is expected. While rBF_{chi2} only requires a fixed significance level (5% by default) without additional tuning process.

To further show that our proposed methods can consistently improve the causal discovery performance, we consider another two widely used DAG learning algorithms: PC [Spirtes *et al.*, 2000] and MMHC [Tsamardinos *et al.*, 2006b]. We incorporate the proposed methods into PC and MMHC for evaluation. According to the experimental results, our proposed methods can consistently improve the DAG learning performance, particularly with PC. For example, on ALARM, PC with BF_{chi2} achieves averaged SHD 40.5, while PC only achieves averaged SHD 58.2. Overall, BF_{chi2} achieves better accuracy and I^{eB} achieves better efficiency with both PC and MMHC on different datasets.

Global causal discovery without causal sufficiency assumption. To demonstrate that our robust independent tests can also be applied to causal discovery without causal sufficiency assumption, we employ the conservative FCI (cFCI) method [Ramsey *et al.*, 2012] as our baseline. cFCI is considered as the state-of-the-art causal discovery algorithm that identifies latent confounders. We denote cI^{eB} as the cFCI

Dataset (**MEAN**)	SHD			#Independence Test		
	cI^{eB}	cBF_{chi2}	cFCI	cI^{eB}	cBF_{chi2}	cFCI
CHILD	49.1	**35.1**	50.4	**109**	417	2289
INSURANCE	118.7	**94.9**	121.3	**147**	593	4836
ALARM	105.3	**78.2**	94.7	**397**	902	7361
HAILFINDER	**153.2**	220.2	339.0	**368**	2024	82683
CHILD3	204.3	135.2	**103.6**	**692**	2858	4009
CHILD5	250.7	**159.0**	178.8	**1145**	5161	7068

Table 3: Global causal discovery performance (with latent confounder) under insufficient data

with empirical Bayesian MI estimation, and cBF_{chi2} as cFCI with BF_{chi2} independence test. We compare our approaches to cFCI with default g^2 statistical based independence test[2]. As we can see from Table 3, cI^{eB} achieves best efficiency by performing the smallest number of independence tests. In terms of accuracy, cBF_{chi2} achieves overall better performance. The consistent performance improvement further demonstrates that the proposed independence test can improve the causal discovery performance under insufficient data, independent of the existence of latent confounders.

4.3 Bayesian Approaches for Independence Tests

To compare the proposed independence tests to state-of-the-art methods, we firstly perform a direct evaluation of proposed independence tests on synthetic data, and we then compare to state-of-the-art methods in terms of causal discovery performance on benchmark datasets. On synthetic data, we compare to three state-of-the-art independence tests: adaptive partition [Seok and Seon Kang, 2015], empirical Bayesian with fixed α [Hutter, 2002] and full Bayesian method [Archer *et al.*, 2013]. We evaluate the performance in terms of both accuracy and efficiency. Experimental results show that the proposed methods achieve better accuracy with significantly improved efficiency. More importantly, we compare proposed independence tests to two state-of-the-art methods: adaptive partition and empirical Bayesian with fixed α methods in terms of causal discovery performance on benchmark datasets. Because the full Bayesian method is of high computational complexity, making it impractical to be applied to constraint-based causal discovery, we exclude the comparison to this method. We incorporate the adaptive partition method and the empirical Bayesian with fixed α method to RAI (denoted as rI^{AdP} and rI^{eBFix} respectively). As

Dataset (**MEAN**)	SHD				
	rI^{AdP}	rI^{eBFix}	rI^{eB}	rBF_{chi2}	RAI-BF
CHILD	26.5	23.9	19.7	**19.3**	25.5
INSURANCE	53.2	49.1	48.6	**44.7**	49.5
ALARM	46.9	40.9	41.7	**36.2**	37.8
HAILFINDER	**70.8**	91.2	88.4	104.3	124.7
CHILD3	81.6	66.3	65.0	**53.2**	74.9
CHILD5	129.9	121.6	120.0	**106.8**	126.3

Table 4: Accuracy comparison to SoTA independence tests

shown in Table 4, our methods achieve overall better accuracy than rI^{AdP} and rI^{eBFix} on different datasets. For example, on CHILD3, rBF_{chi2} achieves averaged SHD 53.2, significantly better than rI^{AdP} which achieves averaged SHD 81.6. In terms of efficiency evaluation, rI^{eB} also achieves competitive efficiency. Overall, our proposed methods outperform other SoTA independence tests in terms of causal discovery performance. On HAILFINDER, because rI^{AdP} tends to declare independence, the learned DAG contains fewer false positive edges compared to other methods and thus its averaged SHD is the best.

5 Conclusion

In this paper, we introduce Bayesian methods for robust constraint-based causal discovery under insufficient data. Two Bayesian-augmented frequentist independence tests are proposed for reliable statistic estimation under a frequentist independence test framework. Specifically, we propose: 1) an effective empirical Bayesian method for accurate estimation of mutual information under limited data; 2) a Bayesian statistical testing method for independence test by formulating the Bayes Factor into the well-defined χ^2 statistical test. We apply the proposed methods to both local and global causal discovery algorithms and evaluate their performance against state-of-the-art methods on different benchmark datasets. The experiments show that, by introducing Bayesian approaches, the proposed methods not only outperform the competing methods in terms of accuracy, but also improve efficiency significantly.

[2]https://github.com/striantafillou/causal-graphs.

Acknowledgments

This work is supported in part by a DARPA grant FA8750-17-2-0132, and in part by the Rensselaer-IBM AI Research Collaboration (http://airc.rpi.edu), part of the IBM AI Horizons Network (http://ibm.biz/AIHorizons).

References

[Archer *et al.*, 2013] Evan Archer, Il Park, and Jonathan Pillow. Bayesian and quasi-bayesian estimators for mutual information from discrete data. *Entropy*, 15(5):1738–1755, 2013.

[Claassen and Heskes, 2012] Tom Claassen and Tom Heskes. A bayesian approach to constraint based causal inference. *arXiv preprint arXiv:1210.4866*, 2012.

[Colombo and Maathuis, 2014] Diego Colombo and Marloes H Maathuis. Order-independent constraint-based causal structure learning. *The Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

[Colombo *et al.*, 2012] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.

[Gao and Ji, 2015] Tian Gao and Qiang Ji. Local causal discovery of direct causes and effects. *Advances in Neural Information Processing Systems*, 28:2512–2520, 2015.

[Geweke and Singleton, 1980] John F Geweke and Kenneth J Singleton. Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association*, 75(369):133–137, 1980.

[Glymour *et al.*, 2019] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

[Hutter, 2002] Marcus Hutter. Distribution of mutual information. In *Advances in neural information processing systems*, pages 399–406, 2002.

[Kass and Raftery, 1995] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

[Li *et al.*, 2019] Honghao Li, Vincent Cabeli, Nadir Sella, and Hervé Isambert. Constraint-based causal structure learning with consistent separating sets. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 2019.

[Marx and Vreeken, 2018] Alexander Marx and Jilles Vreeken. Stochastic complexity for testing conditional independence on discrete data. In *NeurIPS 2018 Workshop on Causal Learning*, 2018.

[McDonald, 2009] John H McDonald. *Handbook of biological statistics*, volume 2. sparky house publishing Baltimore, MD, 2009.

[Minka, 2000] Thomas Minka. Estimating a dirichlet distribution. *Technical report, MIT*, 2000.

[Mukherjee and Speed, 2007] Sach Mukherjee and Terence P Speed. Markov chain monte carlo for structural inference with prior information. *University of California; Berkeley*, 2007.

[Natori *et al.*, 2017] Kazuki Natori, Masaki Uto, and Maomi Ueno. Consistent learning bayesian networks with thousands of variables. In *Advanced Methodologies for Bayesian Networks*, pages 57–68, 2017.

[Ramsey *et al.*, 2012] Joseph Ramsey, Jiji Zhang, and Peter L Spirtes. Adjacency-faithfulness and conservative causal inference. *arXiv preprint arXiv:1206.6843*, 2012.

[Rohekar *et al.*, 2018] Raanan Y Rohekar, Yaniv Gurwicz, Shami Nisimov, Guy Koren, and Gal Novik. Bayesian structure learning by recursive bootstrap. In *Advances in Neural Information Processing Systems*, pages 10525–10535, 2018.

[Rohekar *et al.*, 2020] Raanan Y Rohekar, Yaniv Gurwicz, Shami Nisimov, and Gal Novik. A single iterative step for anytime causal discovery. *arXiv preprint arXiv:2012.07513*, 2020.

[Sen *et al.*, 2017] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. *arXiv preprint arXiv:1709.06138*, 2017.

[Seok and Seon Kang, 2015] Junhee Seok and Yeong Seon Kang. Mutual information between discrete variables with many categories using recursive adaptive partitioning. *Scientific Reports*, 5, 06 2015.

[Spirtes *et al.*, 2000] Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

[Spirtes, 2010] Peter Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11(5), 2010.

[Strobl *et al.*, 2019] Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1), 2019.

[Tsamardinos *et al.*, 2006a] Ioannis Tsamardinos, Laura E. Brown, and Constantin F. Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, Oct 2006.

[Tsamardinos *et al.*, 2006b] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

[Yang *et al.*, 2021] Shuai Yang, Hao Wang, Kui Yu, Fuyuan Cao, and Xindong Wu. Towards efficient local causal structure learning, 2021.

[Yehezkel and Lerner, 2009] Raanan Yehezkel and Boaz Lerner. Bayesian network structure learning by recursive autonomy identification. *Journal of Machine Learning Research*, 10(Jul):1527–1570, 2009.