Causal Discovery with Hidden Confounders using the Algorithmic Markov Condition (Supplementary Material)

David Kaltenpoth[1] **Jilles Vreeken**[1]

[1]CISPA Helmholtz Center for Information Security, Saarbrücken

A APPENDIX

A.1 PROOFS

Proof of Proposition 1. For any two $i, j \in S$ we know that, since they are direct descendants of Z, $X_i \not\perp\!\!\!\perp X_j \mid U$ for any $U \subset \{X_1, \dots, X_m\} \setminus \{X_i, X_j\}$. Hence all edges $\{X_i, X_j\}$ are in G so that S is a clique in G. □

Proof of Theorem 2. fix We prove this statement in two steps. First, we show that all b_{ij} are identifiable. Let $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, l\}$. Then, by assumption (A2) there exists a distinct quadruple (X_i, X_u, X_v, X_w) of nodes that are conditionally independent given Z_j. In order to make every quadruple (X_i, X_u, X_v, X_w) be dependent conditional on Z_j, it would have to have either an edge between them or a common predecessor, which would require at least $|S_j| - 3$ incoming edges to S_j from sources that are not Z_j.

Therefore, for any two variables (X_λ, X_μ) in our quadruple we know that $\sigma_{\lambda\mu} = \operatorname{cov}(X_\lambda, X_\mu) = b_\lambda b_\mu$ and in particular

$$\sigma_{iu}\sigma_{vw} = b_i b_u b_v b_w = \sigma_{iv}\sigma_{uw}$$

We can therefore write

$$b_i^2 = \sigma_{iu}\sigma_{iv}/\sigma_{uv}.$$

Furthermore, no quadruple $(X_i, X_{u'}, X_{v'}, X_{w'})$ this is not conditionally independent given Z can satisfy the constraint $\sigma_{iu}\sigma_{vw} = \sigma_{iv}\sigma_{uw}$ by assumption (A3). Hence, all b_{ij}^2 are identifiable, and since we know the sign of at least one b_{ij} for the given j, we therefore know the sign of each b_{ij} for fixed j. However, since j was arbitrary, B is identifiable in its entirety.

Now, knowing the values of B we can determine the distribution $P(X \mid Z)$, which depends only on A and σ_ϵ^2. Since $X \mid Z$ is now a linear Gaussian SEM with equal variances, the identifiability of A and σ_ϵ^2 follows from the work of Peters and Bühlmann [2012] on the identifiability of the equal variance model. □

Proof of Theorem 3. To prove the first statement, let Z be jointly independent and let there be no edges $X \to Z$. Pick P such that $P(Z = 0) = 1$. Then Z contains no information about X so that $K(P(X, Z)) \leq K(P(X)) + K(P(Z)) = K(X) + O(1)$, with constant $K(P(Z)) = O(1)$ independent of $P(X)$.

For the second statement, consider the case where the true generating mechanism for X does not include any latent variables for any subset X_S. Then as noted in the AMC and the discussion preceding it, *all* information needed to compress $P(X)$ is already present in the graph G_X^* giving the optimal factorization of $P(X)$. Hence $K(P(X, Z)) \geq K(P(X)) + K(P(Z|X)) > K(P(X))$. □

Proof of Theorem 4. Assume that Eq. (4) holds in the limit $n \to \infty$. Then there is a model class $\mathcal{M} \in \mathfrak{M}$ such that $L(x \mid \mathcal{M}) < L(x \mid \mathcal{M}_0)$. Further, as we use a refined MDL score that means there is an optimal $P^* \in \mathcal{M}$ such that

$L(x \mid \mathcal{M}) = L(x \mid P^*) + O(n^{-1})$. Due to the consistency [Grünwald, 2007] of refined MDL scores this means that almost surely X has generating distribution $P^*(X)$. As P^* is a joint distribution over (X, Z) this means that x is the observed part of a joint sample (x, z) from $P^*(X, Z)$. □

Proof of Theorem 5. As we've seen in the proof of Theorem 2, for each X_i there exists a distinct quadruple (X_i, X_u, X_v, X_w) conditionally independent given Z_j by (A2). Hence, all correlations between these four variables can be explained by the parameters in B. Furthermore, by Proposition 1, no pair of variables X_μ, X_λ can be d-separated in any DAG over X, so that by setting $b_{ij} = 0$ we would require *at least* four additional entries of A to be non-zero, instead of only one in B.

Hence, since in the limit we have $\widehat{b}_{ij}\widehat{b}_{vj} - \sigma_{iv} \to 0$, the matrix $\widehat{B}$ converges towards B. Furthermore, given a good approximation of $P(X)$ and of B, we obtain a good approximation of A by the results of Van de Geer and Bühlmann [2013]. □

Proof of Theorem 6. This follows directly from Theorem 5 and the fact that for $n \to \infty$, our MDL score is equivalent to BIC [Grünwald, 2007]. □

Proof of Proposition 7. By Proposition 1, we know that S_j^* forms a clique in the graph G^* inferred by a consistent $\mathcal{A}$. This clique is maximal due to no node being in the Markov Blanket of all $s \in S_j^*$. Further, since x^n is a sample from Eq. (6) we know from the MDL principle for selecting nested model classes [Grünwald, 2007] that in the limit no other set can be compressed better by introducing a confounder than S_j^* itself. □

A.2 IMPLEMENTATION DETAILS

We implemented CDHC in Python using PyMC3 for posterior inference using ADVI with default parameters. All code is available for research purposes. Experiments were run single-threaded on a standard commodity laptop and each experiment finished within minutes.

A.3 COMPUTING CONFIDENCES

For CDHC we measure its confidence as the relative gain in compression due to addition of confounders, $C_{\mathcal{A}} = (L_{\mathcal{A}} - L_{\text{CDHC-}\mathcal{A}})/L_{\mathcal{A}} \geq 0$ and for NOTEARS we use the normalized difference between the initial (empty network) and final (discovered network) score obtained from optimization. None of GFCI, 3OFF2 or DCD come with readily computable confidence scores so we treat them in *the way most favorable to them* by assuming that their best performances are also their most confident.

A.4 ADDITIONAL RESULTS FOR GGSL AND GES

We now provide additional details on the results of GGSL and GES in Table 1. We compute all confidences as described in the previous section. For comparison we also include the results of CDHC. Since they are most interpretable, we include only the F_1^{net} score here. For other metrics too, however, both GGSL and GES perform similarly to NOTEARS so that CDHC significantly outperforms them. In particular, neither of them can find any confounders.

	Data evaluated		
Method	1%	50%	100%
CDHC	**0.92**	**0.85**	**0.53**
GGSL	0.68	0.42	0.24
GES	0.64	0.35	0.21

Table 1: Comparison of CDHC, GGSL and GES in terms of F_1^{net} scores. The performance of each method is shown for each of 1%, 50% and 100% of the data evaluated (corresponding to leftmost, median and right-most points in a decision-rate plot). We see that CDHC clearly outperforms both competitors by a large margin and further that both GGSL and GES perform similarly to NOTEARS.

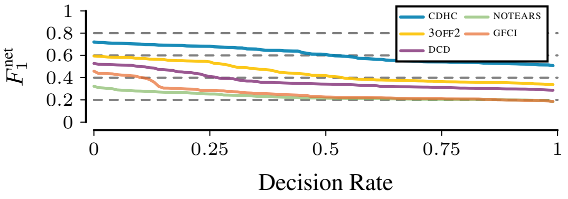


Figure 1: [Higher is better] Decision rate for CDHC and its competitors on the REGED dataset. Overall, CDHC outperforms all other methods both for points where they are confident as well as those where they are not.

A.5 ADDITIONAL DETAILS ON SIGNIFICANCE TESTING

To verify that CDHC significantly outperforms its competitors we use the Bayesian signed rank test [Benavoli et al., 2014]. It explicitly models the probability that one model is significantly better than the other *in practice* by introducing a *region of practical equivalence* (rope) specified by parameter r. Two methods are considered to perform equally well if the difference of scores for the methods lies in $[-r, r]$. We pick $r = 0.05$ [Benavoli et al., 2014] but the conclusion remains the same for values $r \in (0, 0.15]$. Since the test was designed for two competing methods, for each dataset we compare CDHC with the best-performing of its competitors, which we refer to as OPT.

To compare the two methods over all samples, we aggregate the F_1 scores for both CDHC and OPT and take their differences $z_i = F_{1,i}^{\text{OPT}} - F_{1,i}^{\text{CDHC}}, i \in \{1, \ldots, q\}$. To include the prior assumption that both methods are equally good, we include a pseudo-observation $z_0 = 0$, i.e. that both methods are precisely equally good. We take weights $w = (w_0, \ldots, w_q) \sim \text{Dirichlet}(s, 1, \ldots, 1)$ where s corresponds to the number of times we obtained z_0. This is commonly set to be $s = 0.5$, but due to our large number of experiments its influence on the posterior is minor. The posterior probabilities are computed as

$$\theta_{\text{OPT}} = \sum_{i,j=0}^{q} w_i w_j I_{(2r,\infty)}(z_i + z_j)$$

$$\theta_{\text{rope}} = \sum_{i,j=0}^{q} w_i w_j I_{[-2r,2r]}(z_i + z_j)$$

$$\theta_{\text{CDHC}} = \sum_{i,j=0}^{q} w_i w_j I_{(-\infty,-2r)}(z_i + z_j)$$

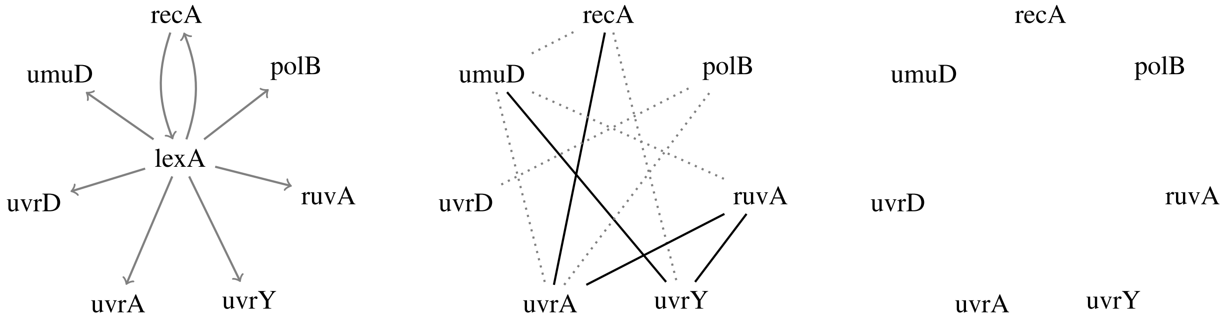
where $\theta_{\text{OPT}}, \theta_{\text{CDHC}}$ are the posterior probabilities that OPT, respectively CDHC are better by at least a margin r, while θ_{rope} is the posterior probability that they perform equally well up to said margin. The distribution of θ is not analytically tractable, but we can evaluate it empirically by sampling values for w. Such a sample is precisely what is depicted in Fig. 5 in barycentric coordinates.

A.6 REALISTIC DATA: REGED

Next, we consider realistic synthetic data from REGED [Guyon et al., 2008], which is based on human lung-cancer microarray gene expression data. Since the available samples are non-i.i.d., the causal relationships are nonlinear, and the ground truth is known from gene intervention studies, it provides a good benchmark for CDHC.

To make CDHC applicable to the REGED dataset we consider the following setup. For each node X_i in the ground truth graph G^* with $k \geq 5$ children, the set of which we denote by $C = C_i$, we select a random subset $R = R_i$ also consisting of k nodes of G^* which are not contained in the Markov boundary of any X_i in G^* and do not have a common parent. We then consider the induced subgraph G_i over the nodes $C_i \cup R_i \cup \{i\}$. However, the data given to each method is only over the variables $X_{C \cup R}$ from which we compute the results in Fig. 1.

We show the results for F_1^{net} for different methods in a DR plot in Fig. 1. Even though the data violates our assumptions, CDHC outperforms its competitors by a large margin. Moreover, even for those sets of variables where CDHC is only moderately confident, it still performs better than its competitors at their *most* confident. This suggests that CDHC works reliably even when the true model deviates from our assumptions.



(a) Ground Truth

(b) 3OFF2

(c) NOTEARS

Figure 2: 3OFF2 and NOTEARS on the SOS dataset. As before, only (potentially) confounded pairs are drawn in the figures. We see that 3OFF2, like GFCI and DCD cannot determine all nodes to be jointly confounded. Meanwhile NOTEARS assumes causal sufficiency and therefore finds no indication of confounding.

A.7 OTHER METHODS ON THE SOS NETWORK

In Fig. 2 we show the results of 3OFF2 and NOTEARS on the SOS dataset. Like GFCI, 3OFF2 can only give indications about which pairs might be confounded, and for the majority of pairs it is not confident. Meanwhile, by its very design NOTEARS has no notion that confounders might be involved.

References

A. Benavoli, F. Mangili, G. Corani, M. Zaffalon, and F. Ruggeri. A bayesian wilcoxon signed-rank test based on the dirichlet process. In *ICML*, pages 1026–1034. JMLR, 2014.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

Isabelle Guyon, Constantin Aliferis, Greg Cooper, André Elisseeff, Jean-Philippe Pellet, Peter Spirtes, and Alexander Statnikov. Datasets of the causation and prediction challenge. Technical report, 2008.

Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *arXiv:1205.2536*, 2012.

Sara Van de Geer and Peter Bühlmann. ℓ_0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics*, 41(2):536–567, 2013.