Inferring Cause and Effect in the Presence of Heteroscedastic Noise

Sascha Xu [1] **Osman Mian** [1] **Alexander Marx** [◦ 2] **Jilles Vreeken** [◦ 1]

Abstract

We study the problem of identifying cause and effect over two univariate continuous variables X and Y from a sample of their joint distribution. Our focus lies on the setting where the variance of the noise may be dependent on the cause. We propose to partition the domain of the cause into multiple segments when the noise indeed is dependent. To this end, we minimize a scale-invariant, penalized regression score, finding the optimal partitioning using dynamic programming. We show under which conditions this allows us to identify the causal direction for the linear setting with heteroscedastic noise, for the non-linear setting with homoscedastic noise, as well as empirically confirm that these results generalize to the non-linear and heteroscedastic case. Altogether, the ability to model heteroscedasticity translates into an improved performance in telling cause from effect on a wide range of synthetic and real-world datasets.

1. Introduction

Causal discovery based on conditional independence tests can identify causal graphs up to Markov equivalence. To learn fully directed graphs, we need to disambiguate between Markov equivalent graphs, which entails inferring the causal direction between pairs of statistically dependent variables. This problem is known as bi-variate causal inference. Pearl (2000) showed that it is impossible to tell cause from effect from observational data without making additional assumptions about the data generating process.

What assumptions to make is the central question in causal inference. While these should be strong enough as to permit formal guarantees on identifiability of the model, they should at the same time be as likely as possible to hold in practice. The perhaps most common assumption in bivariate causal inference is the additive noise model, where noise is independent from the cause (Bühlmann et al., 2014; Peters et al., 2014; Shimizu et al., 2006; Hoyer et al., 2009). As would be expected, methods based on this assumption are successful when it holds, but as Tagasovska et al. (2020) show, these indeed fail when noise does depend on the cause, as for example, with location-scaled noise.

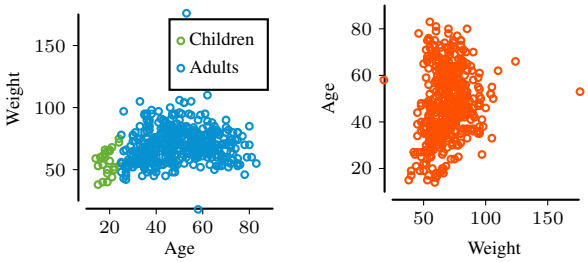

Figure 1. Age vs. Weight from the Tübingen cause-effect benchmark dataset (Mooij et al., 2016). Different noise variances for adolescents/young adults and adults overlap in the anti-causal direction and create the asymmetry to distinguish cause from effect.

In this work, we focus on the bi-variate setting and propose a causal model that includes heteroscedastic noise distributions. That is, rather than wishing it away, we explicitly permit the variance of the noise to depend on the cause.

We provide conditions under which this causal model is identifiable by explicitly modelling the scale of the noise with a BIC-regularized likelihood model. In particular, we give proofs for the linear setting with heteroscedastic noise, and for the non-linear setting with homoscedastic noise. Further, we empirically validate that our method can identify the causal direction for non-linear models with heteroscedastic noise, as well.

As an example, consider Fig. 1, in which we depict the Age vs. Weight pair from the Tübingen cause-effect benchmark dataset. In the causal direction (left) it is possible to find two regions with different noise, while in the anti-causal direction (right) it is not. Under our assumptions this translates into a higher negative log-likelihood for the causal direction, even if we rescale the variables between zero and one (Reisach et al., 2021).

[◦]Joint senior author [1]CISPA Helmholtz Center for Information Security, Saarbrücken, Germany [2]ETH Zürich & ETH AI Center, Zürich, Switzerland. Correspondence to: Sascha Xu <sascha.xu@cispa.de>, Jilles Vreeken <jv@cispa.de>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

To perform causal inference in practice, we propose a fitting process that automatically divides the domain of the presumed cause into segments of different noise variances by optimizing a BIC-based score. We propose an efficient dynamic-programming-based algorithm, HECI, that can determine the optimal scoring model in quadratic time despite an exponential search space. We can then by simply comparing the scores we so obtain for X to Y, and vice versa, determine cause from effect.

We provide a thorough empirical evaluation on synthetic and real world data, comparing to a wide range of methods for bi-variate causal inference. These results show that our method, HECI, performs as well as the strongest competitor whenever noise is homoscedasctic, and is the strongest whenever noise does depend on the cause.

The remainder of this paper is structured as usual. In Sec. 2 we formally specify our causal model, and show under which conditions it is identifiable. We present HECI, an effective algorithm for heteroscedastic-noise based causal inference in Sec. 3. We discuss related work in Sec. 4 and empirically evaluate HECI in Sec. 5. We wrap up with discussion and conclusions in Sec. 6. We postpone all proofs to the Supplementary.

2. Theory

We consider causal inference from independent and identically distributed (iid) measurements of two continuous random variables X and Y. Further, we assume that there is no selection bias and that X and Y are not affected by an unobserved common cause. Under those assumptions, our task reduces to deciding between the two Markov equivalent DAGs $X \to Y$ and $X \leftarrow Y$. To tackle this problem, we need to impose assumptions on the underlying causal model (Pearl, 2000; Peters et al., 2017), which we define below.

2.1. Heteroscedastic Noise Models

In the most general setting, we assume that cause and effect can be generated from the following structural equation model (SEM)

$$\begin{aligned} X &:= N_X \\ Y &:= f(X) + s_\alpha(X) \cdot N_Y \,, \end{aligned} \tag{1}$$

where $N_X \perp\!\!\!\perp N_Y$, N_X, N_Y have zero mean and variances σ_X^2, resp. σ_Y^2, f is a smooth function, and $s_\alpha(X)$ a function controlling the amplitude of the variance of noise added to Y. The function $s_\alpha(X) : \mathbb{R} \mapsto \mathbb{R}^+$ may be dependent on X and a scale factor α. For now, we leave the above causal model in its general form and first provide some intuition about $s_\alpha(X)$. Subsequently, we specify for which functions f, s_α, and noise distributions N_X, N_Y we can identify the correct causal model.

As specified above, the causal model in Eq. (1) can express various noise settings. In the simplest setting, in which we define $s_\alpha(X) = c$ as a constant function, Eq. (1) reduces to the classical additive noise model (Peters et al., 2017). More interesting to us, however, is if the noise scales with respect to X. For example, there could be a threshold t s.t. $s_\alpha(x)$ for $x < t$ is smaller than for $x \geq t$ (Fig. 1), or the noise could just fan out scaled by location (see Fig. 2).

In the following, we first show that both such cases can be identified in the linear case and then extend our analysis to the non-linear setting.

2.2. Linear Functions

First, we discuss linear functions, for which $f(X) = \beta_0 + \beta_1 X$. It has been shown for $s_\alpha(X) = 1$ that the linear SEM is identifiable from the L_2-loss for Gaussian SEMs with equal error variances (Peters & Bühlmann, 2014) and for non-Gaussian SEMs (Loh & Bühlmann, 2014). These results have encouraged a line of optimization-based approaches which minimize the global L_2-loss to learn the causal DAG (Zheng et al., 2018; 2020; Lachapelle et al., 2020; Ng et al., 2020; Yu et al., 2019), an advancement that recently became feasible by proposing a continous constraint to enforce a DAG structure (Zheng et al., 2018). Another line of research even proved that cause and effect are identifiable for linear SEMs with heterogeneous noise (Park, 2020), by showing that

$$\mathrm{E}\left[\mathrm{Var}(X \mid Y)\right] = \sigma_X^2 - \frac{\beta_1^2 \sigma_X^4}{\sigma_Y^2 + \beta_1^2 \sigma_X^2}$$

is smaller than $\mathrm{E}\left[\mathrm{Var}(Y \mid X)\right] = \sigma_Y^2$ if $\frac{\sigma_Y^2}{\sigma_X^2} > (1 - \beta_1^2)$.

Despite the recent success of these approaches, Reisach et al. (2021) note that these scores are not robust w.r.t. to scaling. In particular, they showed that standardizing the data significantly reduces the performance of approaches based on the above principle.

In Lemma 1, we generalize this statement to linear models, where X and Y might be rescaled arbitrarily, and for which $s_\alpha(X) = \alpha$.

Lemma 1 *Let X be a random variable with variance σ_X^2 and let $Y = \beta_0 + \beta_1 X + \alpha N_Y$, where N_Y has variance σ_Y^2 and $\alpha > 0$ is a scaling parameter. Further, let $X' = a + bX$ and $Y' = c + dY$ be the rescaled versions of X and Y, then*

$$\frac{E[\mathrm{Var}(X' \mid Y')]}{E[\mathrm{Var}(Y' \mid X')]} = \left(\frac{b}{d}\right)^2 \frac{\sigma_X^2}{\alpha^2 \sigma_Y^2 + \beta_1^2 \sigma_X^2} \,,$$

where $\sigma_X^2 = \mathrm{Var}(X)$ and $\alpha^2 \sigma_Y^2 + \beta_1^2 \sigma_X^2 = \mathrm{Var}(Y)$.

The proof for Lemma 1 is provided in Supplementary Material A.1. To see how scaling affects the identifiability result, consider that we standardize X and Y by subtracting their means and dividing by the corresponding standard deviation. From Lemma 1, we can immediately see that the expected variance for both directions will be equal, which renders the problem non-identifiable. Even if we additionally add both marginal variances to the score, i.e. compare $\mathrm{Var}(X')+\mathrm{E}\left[\mathrm{Var}(Y' \mid X')\right]$ to $\mathrm{Var}(Y')+\mathrm{E}\left[\mathrm{Var}(X' \mid Y')\right]$ we will arrive at an equality since $\mathrm{Var}(X') = \mathrm{Var}(Y') = 1$ after standartization. Further, if the support of X and N_Y is bounded, we can derive the same result for normalization (rescaling X and Y between zero and one) if $\alpha \to 0$.

In the following, we show how we can break this symmetry. Consider the age vs. weight example shown in Fig. 1 where the variance in weight is lower in children and young adults than in adults. Mathematically, this could be expressed as the case where s_α is a stepfunction, i.e.

$$s_\alpha(x) = \begin{cases} \alpha c_1, \text{ if } x < a \\ \alpha c_2, \text{ otherwise} \end{cases}$$

where $0 < a < 1$ is some cut-off, at which the scale of the variance changes. Our approach models the local noise variance by its log-likelihood, which we will show in section 2.6 to be the weighted mean of the logarithmic variance. In the causal direction we are able to separate the variance into the step function $s_\alpha(x)$ such that

$$\begin{aligned} &P(X < a) \cdot \log((\alpha c_1)^2) + P(X \geq a) \cdot \log((\alpha c_2)^2) \\ &= \log\left((\alpha c_1)^{2 \cdot P(X<a)} \cdot (\alpha c_2)^{2 \cdot P(X \geq a)}\right). \end{aligned}$$

Here the log-likelihood corresponds to the logarithm of the geometric mean. In the anti-causal direction the noise overlaps s.t. we cannot separate the two noise distributions and instead need to model them with a single variance. According to Lemma 1, in the limit of $\alpha \to 0$ we obtain the same expected variance in both directions (after normalization), hence the variance over this single interval is the arithmetic mean of both independent variance terms, i.e.

$$\begin{aligned} &\log\left(\mathrm{Var}(X|Y)\right) \\ &= \log\left(P(X < a) \cdot (\alpha c_1)^2 + P(X \geq a) \cdot (\alpha c_2)^2\right). \end{aligned}$$

Since the geometric mean is smaller or equal than the arithmetic mean, with equality if and only if $c_1 = c_2$, the negative log-likelihood we calculated for the causal direction is smaller than that for the anti-causal direction.

Based on this intuition, we generalize this scheme onto the continuous case, where we compare the integral of logarithmic noise variances in both directions. We define N_Y to have unit variance, such that the variance of the scaled noise term is fully determined by the square of the amplitude function $s_\alpha(x)^2$. We let the noise be arbitrarily scaled as $s_\alpha(x) = \alpha \cdot s(x)$, with the additional constraint that the scaling function s is strictly positive.

Theorem 1 *Given a causal model as specified in Eq.* (1)*, assume that*

(1) N_X, N_Y have finite support, and X and Y are normalized to obtain values within $[0, 1]$

(2) f is a linear function with $f(X) = \beta_0 + \beta_1 X$ and g is its inverse

(3) N_Y is unbiased with unit variance and strictly positive scale function $s_\alpha(x) = \alpha s(x)$, with $s_\alpha(x) \to 0$ if $\alpha \to 0$.

In that case it holds that in the limit of $\alpha \to 0$,

$$\begin{aligned} &\int_0^1 p_Y(y) \cdot \log\left(\mathrm{Var}(X|Y = y)\right) dy \\ &\geq \int_0^1 p_X(x) \cdot \log\left(\mathrm{Var}(Y|X = x)\right) dx\,, \end{aligned}$$

with equality, if and only if the conditional variance of the noise scaling $\mathrm{Var}\left(s(X)|Y\right) = 0$, i.e. there is no overlap of noise with different amplitude $s(x)$ in the domain $\mathcal{Y}$.

The proof to Theorem 1 is provided in Supplementary Material A.2. In the following, we sketch out the main idea of the proof. For the causal direction we know that $\mathrm{Var}(Y|X = x) = s_\alpha(x)$ according to our causal model. In the anti-causal direction, however, the noise variance in the linear case is obtained by the weighted integral of noise amplitudes $s_\alpha(x)$ over all x which are mapped to $y = f(x) + s_\alpha(x)N$. Consequently we may express the first term of the Theorem 1 as

$$\int_0^1 p_Y(y) \cdot \log\left(\int_0^1 p_{X|Y=y}(x) \cdot s_\alpha(x)^2 dx\right) dy\,.$$

Through Jensens inequality it follows that

$$\geq \int_0^1 p_Y(y) \cdot \left(\int_0^1 p_{X|Y=y}(x) \cdot \log\left(s_\alpha(x)^2\right) dx\right) dy\,.$$

This relationship holds with equality if and only if $s(X|y)$ is constant for all $y \in \mathcal{Y}$. Formally, this condition is met if the conditional variance of the noise scaling is zero, i.e. $\mathrm{Var}(s(X)|Y) = 0$. For example this is fulfilled in the homoscedastic setting, where $s(x) = c$. Now, if we integrate over y we get

$$\begin{aligned} &\int_0^1 p_X(x) \cdot \log\left(s_\alpha(x)^2\right) dx \\ &= \int_0^1 p_X(x) \cdot \log\left(\mathrm{Var}(Y|X = x)\right) dx\,, \end{aligned}$$

which resembles the second term in Theorem 1.

Theorem 1 shows that we can identify the causal direction under heteroscedastic noise for linear SEMs, even after normalizing the data. The limit of $\alpha \to 0$ is necessary to make a definite statement over the regression error in the linear case. Empirically we observe that we can distinguish cause and effect even for high noise levels.

2.3. Non-Linear Functions

Next, we consider the case in which f can be a non-linear function and start again with the homoscedastic noise setting in which $s_\alpha(X) = \alpha$. In this setting, we can build upon well known identifiability results. In particular, Blöbaum et al. (2018) prove that cause and effect can be identified from the expected variances, i.e.

$$\lim_{\alpha \to 0} \frac{\mathrm{E}\left[\mathrm{Var}(X \mid Y)\right]}{\mathrm{E}\left[\mathrm{Var}(Y \mid X)\right]} > 1$$

under similar assumptions as above. That is, X has compact support in $[0, 1]$, Y is rescaled to $[0, 1]$ with compact support, f is an invertible non-linear function and the noise variable N_Y is unbiased with variance equal to one. In addition, Blöbaum et al. (2018) note that for non-invertible functions, the statement trivially holds due to the information loss in the anti-causal direction. Similar statemets also exist for additive noise models (Zhang & Hyvärinen, 2009).

Thus the only case we did not cover, yet are non-linear invertible functions with heteroscedastic noise. Based on Theorem 1 and the identifiability results for non-linear additive noise functions Blöbaum et al. (2018) (when $\alpha \to 0$), we conjecture that the inequality also holds for cases where f is non-linear, invertible and the noise is heteroscedastic as in Theorem 1, with $s_\alpha(X) \cdot N_Y \to 0$ when $\alpha \to 0$. In our empirical evaluation, we can validate this conjecture and show that our method perfoms well, even in cases where there is a lot of difference in the noise amplitude, i.e. heteroscedasticity.

2.4. Multivariate Causal Models

To use HECI beyond the bi-variate setting, we recommend to first apply a causal discovery algorithm that identifies the Markov equivalence class of the true DAG, e.g. the PC algorithm (Spirtes et al., 2000), the GES algorithm (Chickering, 2002) or their extensions. After that, one can use HECI to infer the remaining undirected edges.

2.5. Inference

To infer the causal direction between X and Y, we instantiate the discrete counterpart of the integrals in Theorem 1. The score below is used to infer the causal direction. For a sample $\{x_i, y_i\}_{i=1}^n$ each point has the weight $\frac{1}{n}$, consequently the score of the causal direction is

$$Score_{X \to Y} = \sum_{i=1}^{n} \frac{1}{n} \log(s_\alpha(x_i)^2) \,.$$

We say that X causes Y if $Score_{X \to Y} < Score_{Y \to X}$, that Y causes X if $Score_{X \to Y} > Score_{Y \to X}$ and do not decide if both quantities are equal. In practice, we instantiate the above score using a penalized negative log-likelihood, such as the Bayesian Information Criterion, which allows to regularize the model fitting process for f, s_α, and avoid overfitting. Next, we link the above score to the empirical negative log-likelihood of the regression errors.

2.6. Empirical Log-Likelihood

Given a sample $\{x_i.y_i\}_{i=1}^n$ drawn iid from the joint distribution of X and Y, the residuals $r_i = y_i - f(x_i)$ have zero mean and variance determined by $s_\alpha(x_i)^2$ from our causal model. The empirical negative log-likelihood of the residuals, assuming a normal distribution, is

$$\begin{aligned} -\log\left[L_{X \to Y}(s_\alpha^2, f)\right] &= -\log\left[\prod_{i=1}^{n} p\left(r_i | x_i; s_\alpha^2\right)\right] \\ &= \frac{1}{2}\sum_{i=1}^{n} \log\left[s_\alpha(x_i)^2\right] + \frac{1}{2}\sum_{i=1}^{n} \frac{r_i^2}{s_\alpha(x_i)^2} + \frac{n}{2}\log(2\pi) \,. \end{aligned}$$

Since f and s_α^2 are not known in advance, they have to be estimated ($\hat{f}$ and $\hat{s}^2$). We will show that $-\log(L_{X \to Y}) \propto \frac{2}{n} \cdot Score_{X \to Y}$, which means that we can identify cause and effect by minimizing the negative log-likelihood.

The local variance could be estimated pointwise through the residuals $\hat{s}(x)^2 = \mathrm{mean}(\{\hat{r}_i^2 | x_i = x\})$, which is asymptotically consistent. Practically, there is limited data, thus we model the variance as locally constant inside an interval with a step function. If the domain of X can be partitioned in m non-overlapping bins s.t. within each bin_j the empirical variance has the constant value $\hat{\sigma}_j^2$, we can write the empirical negative log-likelihood w.r.t. this partitioning $\hat{\mathcal{P}}$ as

$$-\log\left[L_{X \to Y}(\hat{\sigma}^2, \hat{f}, \hat{\mathcal{P}})\right] = \sum_{j=1}^{m} \frac{n_j}{2} \cdot \log(\hat{\sigma}_j^2) \,, \qquad (2)$$

where n_j relates to the number of data points falling within bin_j. The full derivation is provided in Supplementary Material B. Since $\hat{\sigma}_j^2$ is the local estimate for the noise variance $s_\alpha(x_i)$, we get that the negative log-likelihood approximates our score above as

$$Score_{X \to Y} \propto -\frac{2}{n} \cdot \log\left[L_{X \to Y}(\hat{\sigma}^2, \hat{f}, \hat{\mathcal{P}})\right] \,.$$

For the inverse direction, we can derive the corresponding negative log-likelihood similarly. Thus, if we optimize this

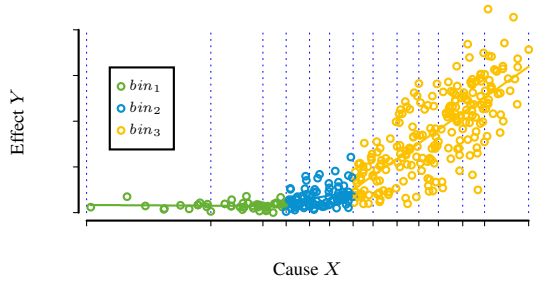

Figure 2. Fitted causal model with heteroscedastic noise. The dashed blue lines indicate the initial binning. The green, blue and yellow segments indicate the optimal partitioning.

log-likelihood, the identifiability guarantees of Theorem 1 and the results for non-linear functions by Blöbaum et al. (2018) hold asymptotically. In the next section, we explain how we compute $\hat{\mathcal{P}}$ and $\hat{f}$ to minimize the corresponding negative log-likelihood via dynamic programming.

3. Algorithm

In the previous section, we established the heteroscedastic noise model and the identifiability conditions based on the expected value of the noise variance logarithm. It uses the residuals of the fitted function $\hat{f}$ under a partition $\hat{\mathcal{P}}$. However, ordinary least squares and other methods estimate $\hat{f}(x)$ assuming *homoscedastic noise*.

In view of this, we present the HECI algorithm for **He**teroscedastic-noise based **C**ausal **I**nference. The regressor domain is divided up into segments, where least squares based regression models are fitted. This way we implicitly estimate $s^2(x)$ as a step function with which we can compute the negative log-likelihood from Eq. (2).

To find the optimal partition and function, three components are required: the binning scheme, that defines the feasible partitions, the regularizing scoring function and the optimization algorithm itself.

3.1. Binning

We initiate the binning algorithm with b equal-width bins that partition the domain of X. A local function is fitted inside a single bin or over multiple, neighboring bins. Each bin is defined as the interval $bin_j : [min_j, max_j)$. In Fig. 2, these are marked by the blue dashed lines. The bins are adjacent with $max_j = min_{j+1}$ for $j \in [1, b-1]$ and have $min_1 = 0$ and $max_b = 1$.

The initial equal-width bins are defined such that $max_j = min_j + \Delta$. The initial bin width Δ must be chosen carefully, especially in cases with limited data. We therefore require a min support of 10 unique data points per bin. In our experiments, we set $\Delta = 0.05$, with which the best performance was achieved. An analysis of the impact of the Δ parameter on performance is provided in the Supplementary Material C. From the set of initial bins $\{bin_j\}_{i=1}^{b}$, the task is to find a partition $\hat{\mathcal{P}}$ of neighboring bins and the underlying function $\hat{f}$, which minimizes the negative log-likelihood, as described below.

3.2. Scoring Models

The combined model of partition $\hat{\mathcal{P}}$ and function $\hat{f}$ is scored based on the empirical log-likelihood and a parameter penalty. The cardinality of the partition is denoted as $|\hat{\mathcal{P}}|$. We use the Bayesian Information Criterion (BIC) to regularize the size of the partition and the complexity of the fitted functions, i.e.

$$BIC\left(\hat{f}, \mathcal{P}, \hat{\hat{\sigma}}\right) := -2 \cdot \log\left[L\left(\hat{\sigma}^2, \hat{f}, \hat{\mathcal{P}}\right)\right] + \log(n) \cdot \left(||\beta_{\hat{f}}||_0 + |\hat{\mathcal{P}}|\right) .$$

With BIC, we may now approach the task of finding the combination of local functions which minimize it. To improve readability we omit $\hat{\sigma}^2$ as it is constant per element of the partition $\hat{\mathcal{P}}$. As we saw in the previous section, the data likelihood is decomposable into independent, additive components. In particular, the BIC score of a given model partitioned at bin_{a-1} is additive, i.e.

$$BIC(f, \bigcup_{j=1}^{b} bin_j) = BIC(f_1, \bigcup_{j=1}^{a-1} bin_j) + BIC(f_2, \bigcup_{k=a}^{b} bin_k) .$$

We make use of this fact for our proposed algorithm to find the optimal model within our binned search space.

3.3. HECI: Dynamic Programming Optimization

The binning provides b possible points, where the domain may be partitioned, and thus 2^b possible partitions in total. The problem is structured however, and allows to find the optimal model in b^2 fits via dynamic programming.

For a single bin_j, the best model $\tilde{f}_{j,j}$ is determined by the best scored polynomial $f_{j,j}$ (linear to cubic). For groups of neighboring bins, which we will call segments from now on, there are two possibilities for the optimal model $\tilde{f}_{p,q}$:

- Fitting a local function $f_{p,q}$ for the segment from bin_p to bin_q, or
- Combining two optimal functions $\tilde{f}_{p,a}$ and $\tilde{f}_{a+1,q}$ for smaller segments, where $p \leq a < q$.

Note, that the optimal functions $\tilde{f}_{p,a}$ and $\tilde{f}_{a+1,q}$ for the smaller segments may in turn be a combination as well. The algorithm to compute the optimal model $\tilde{f}_{1,b}$ over the

entire domain is as follows. First, for all segments from bin_p to bin_q ($p, q \in [1, b], p \leq q$), the local polynomial functions $f_{p,q}$ are fitted. To choose the polynomial degree, we use BIC and minimize

$$f_{p,q} = \arg\min_f BIC\left(f, \bigcup_{j=p}^{q} bin_j\right) .$$

The optimal model for the entire domain is attained in a bottom-up approach. The single bin optimal models $\tilde{f}_{j,j}$ are initialized with the local functions $f_{j,j}$. To compute the optimal models $\tilde{f}_{p,q}$ for segments consisting of $m = q-p+1$ bins, all combinations of functions with splitpoint $a \in [p, q-1]$ are checked. This requires to have the optimal models for all segments of size $m-1$ and smaller available. The best of the combined functions or the local function is chosen based on the BIC.

$$\tilde{f}_{p,q} = \begin{cases} f_{p,q}, & \text{if } BIC(f_{p,q}, \bigcup_{j=p}^{q} bin_j) \leq \\ & \quad BIC(\tilde{f}_{p,a}, \bigcup_{j=p}^{a} bin_j) + \\ & \quad BIC(\tilde{f}_{a+1,q}, \bigcup_{k=a+1}^{a} bin_k) \\ \tilde{f}_{p,a} \cup \tilde{f}_{a,q} & \text{otherwise} \end{cases}$$

Once all optimal models of size m have been determined, the segment size is incremented by one and the process is repeated, until $m = b$. At this point, we have attained the optimal model for the entire domain according to the BIC score. The model defines a partition $\hat{\mathcal{P}}$, defined through the selected split-points a_j and the function $\hat{f}$ defined by the the locally fitted polynomials in the partition.

One such fitted model can be seen in Fig. 2. From the initial b bins, we find the optimal partition and local functions, which are marked as blue, orange and green, using the described bottom-up approach. Like our causal model, the variance is modelled as locally constant, but different between each segment.

3.4. Complexity

The complexity of our algorithm is as follows. There are $\frac{b^2+b}{2}$ permutations of $p, q \in [1, b], p \leq q$. A local polynomial function $f_{p,q}$ is fitted with ordinary least squares in linear time $\mathcal{O}(n)$. The process to find an optimal model $\tilde{f}_{p,q}$ needs to compare at most b scores and is in $\mathcal{O}(b)$. Since the number of bins b is smaller than the number of samples n, the overall computational complexity of HECI is $\mathcal{O}(b^2 \cdot n)$.

4. Related Work

To infer cause an effect from observational data, we need to impose assumptions about the generating mechanism,

Algorithm 1: HECI(X, Y, Δ)

```
Normalize X and Y to [0, 1];
b ← 1/Δ;
for j = 1 : b do
    bin_j ← {x_i, y_i | x_i ∈ [(j − 1) · Δ, j · Δ)}
for j = 1 : b, k = j : b do
    bin_{j−k} ← ⋃_{l=j,...,k} bin_l;
    score[j, k] ← min BIC(f, ⋃_{l=j}^{k} bin_l);
for size = 2 : b, start = 1 : b − size do
    end ← start + size;
    for split = start + 1 : end − 1 do
        left ← score[start, split];
        right ← score[split, end];
        if left + right < score[start, end] then
            score[start, end] ← left + right; ;
Score_{X→Y} ← score[1, b];
Compute Score_{Y→X} in the same way;
Predict causal direction by lowest score, undecided
  when Score_{X→Y} = Score_{Y→X};
```

since the to possible DAGs $X \to Y$ and $X \leftarrow Y$ are Markov equivalent (Verma & Pearl, 1990; Pearl, 2000).

Most well known are additive noise models (ANMs) (Peters et al., 2017). In essence, ANMs assume that the effect is generated as a deterministic function of the cause X and an additive noise term N_Y. For a broad range of function classes and distributions (Shimizu et al., 2006; Hoyer et al., 2009; Peters et al., 2011; Hu et al., 2018; Zhang & Hyvärinen, 2009), it has been shown that such an ANM does not exist in the inverse direction–i.e. the noise N_X will not be independent of Y. One of the most prominent examples is the linear non-Gaussian additive noise model, LiNGAM (Shimizu et al., 2006). A recent proposal based on ANMs is NNCL (Wang & Zhou, 2021). It partitions the domain of the cause into two bins, fits a linear models for each bin, and then checks whether the ANM holds for the partitioned model. Similarly, CDCI (Duong & Nguyen, 2021) discretizes the cause domain via rounding and compares the conditional divergence for both directions. CDCI does, however, not come with strong identifiability guarantees. Different to those methods, we consider a more general class of partitions, non-linear functions, heterosedastic noise and we base our score on the L_2-loss.

Another large class of approaches is based on the principle of independent mechanisms (Janzing et al., 2012; Sgouritsa et al., 2015), or its information-theoretic variant: the algorithmic independence of conditionals (Budhathoki & Vreeken, 2016; Marx & Vreeken, 2017; Stegle et al., 2010; Tagasovska et al., 2020; Mian et al., 2021). Both pos-

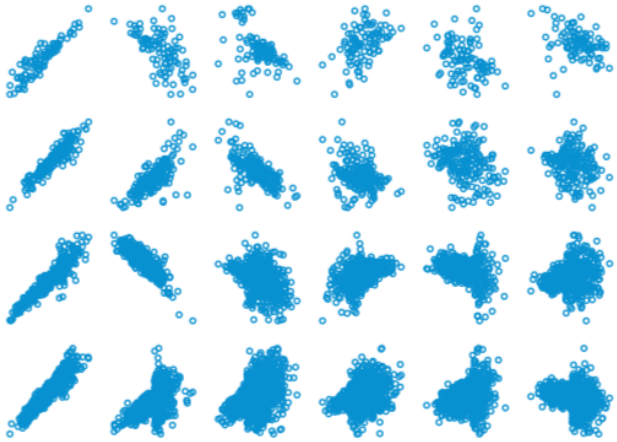

Figure 3. Generated linear cause-effect pairs. The variance of the Gaussian noise changes at a random cutoff near the mean. The noise variance increases from left to right. Pairs of 100 datapoints (top row) up to 1000 datapoints (bottom row) are sampled.

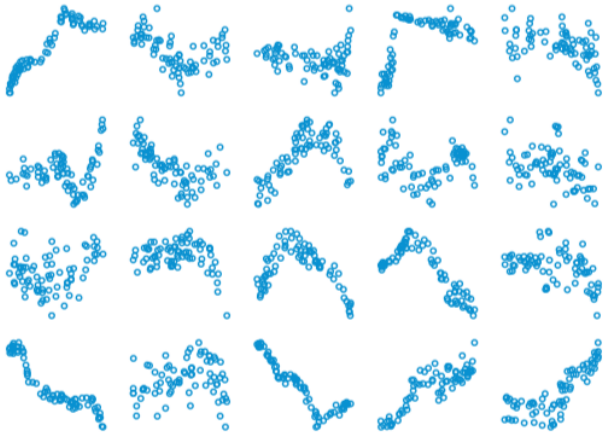

Figure 4. Generated non-linear cause effect pairs. The noise variance changes in a stepwise manner. The step parameter increases from left to right. On the left data with constant noise variance (homoscedastic) is found, while on the right side the most heteroscedastic data is located.

tulates base their inference on the assumption that $P(X)$ is (algorithmically) independent of $P(Y \mid X)$, while the same does not hold for the factorization of the anti-causal direction, i.e. $P(Y)$ is not (algorithmically) independent of $P(X \mid Y)$ (Peters et al., 2017; Janzing & Schölkopf, 2010). Janzing et al. (2012) define the approach IGCI which relies on the principle of independent mechanisms and considers the setting where the effect is a deterministic function of the cause. In practice, they derive a score based on differential entropy. SLOPE (Marx & Vreeken, 2017) and QCCD (Tagasovska et al., 2020) are two recent proposals that aim to approximate the algorithmic Markov condition. Although they empirically perform well, both do not have identifiability guarantees. A more detailed overview is provided by Marx & Vreeken (2022).

Closely related methods to our work are the ones that base their inference rules on regression error. Two such approaches for purely bi-variate pairs are RECI (Blöbaum et al., 2018), which compares the expected regression error, and SLOPPY (Marx & Vreeken, 2019), which considers L_0-penalized regression errors. CAM (Bühlmann et al., 2014) is designed to find a general causal graph, but can decide causal direction for the bivariate case using regularized log-likelihood by building upon identifiability results for additive noise models. Further, identifiability results based on the L_2 have been proven for linear SEMs with equal error variances (Peters & Bühlmann, 2014), non-Gaussian SEMs (Loh & Bühlmann, 2014) and linear SEMs with heterogeneous noise (Park, 2020). Building upon these identifiability results, Zheng et al. (2018) introduced a continuous DAG constraint to speed up the DAG search via continuous optimization based on the L_2-loss. This idea has been extended in various ways, by allowing for non-linear functions (Zheng et al., 2020; Lachapelle et al., 2020), employing different constraints to enforce a DAG structure (Ng et al., 2020) or using different architectures to minimize the objective, e.g. graph neural networks (Yu et al., 2019). For a broader overview, we refer to Vowels et al. (2021).

In this paper, we focus purely on the bi-variate setting, in which the above approaches are comparable to CAM. Different to CAM and RECI, we additionally privide identifiability results for linear SEMs with heteroscedastic noise.

5. Experiments

In this section, we empirically evaluate HECI on both synthetic data and the real-world Tübingen cause and effect pairs (Mooij et al., 2016) benchmark. We will compare it to a wide range of state-of-the-art bivariate causal inference methods. As identifiable approaches that assume an additive noise model, we compare to CAM (Bühlmann et al., 2014) as a representative for regression-based log-likelihood approaches and to RESIT (Peters et al., 2014) a state-of-the-art ANM-based method. Further, we compare to QCCD (Tagasovska et al., 2020) as an approach that explicitly models heteroscedastic noise, SLOPPY (Marx & Vreeken, 2019) and IGCI (Janzing et al., 2012) as the state-of-the-art information theoretic approaches. Finally, we also compare to NNCL (Wang & Zhou, 2021) and CDCI (Duong & Nguyen, 2021) as the bivariate causal inference approaches that discretize the domain of the cause, where NNCL uses piecewise/non-invertible functions and CDCI utilizes conditional divergence. HECI is implemented in Python and we provide the source code as well as

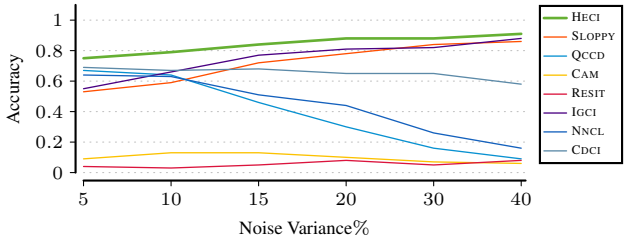

Figure 5. [Higher is better] Accuracy in determining cause from effect for linear functions with 2 noise variances under increasing noise levels.

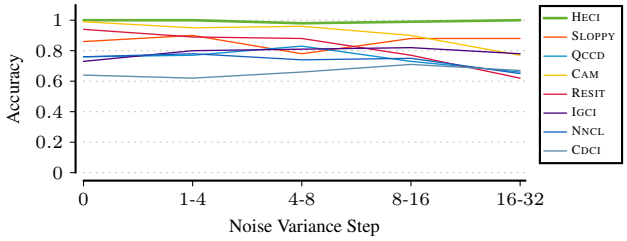

Figure 6. [Higher is better] Accuracy in determining cause from effect for increasing heteroscedasticity, measured by the step height of the variance step function.

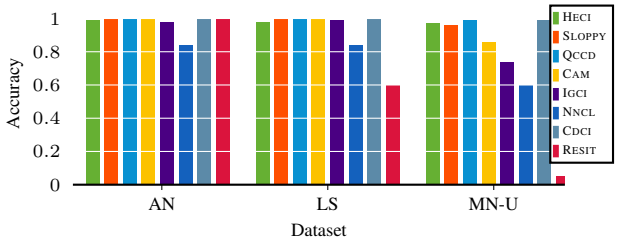

Figure 7. [Higher is better] Accuracy over benchmark synthetic data with Additive Noise (AN), Location Scaling (LS) and Multiplicative Noise (MN-U).

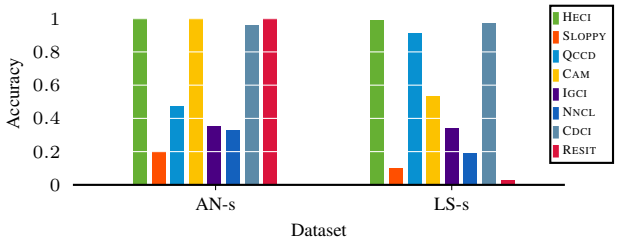

Figure 8. [Higher is better] Accuracy over benchmark synthetic data of sigmoidal functions with Additive Noise (AN-s) and Location Scaling Noise (LS-s).

the synthetic data for research purposes.[1] All experiments were executed on a 4-core Intel i7 machine with 16 GB RAM, running Windows 10. For each instance, HECI was able to decide the causal direction in less than 5 seconds.

5.1. Synthetic Data

We test HECI on two different settings. First, we generate synthetic data according to our assumed causal model in Eq. (1). Next, we use the synthetic data of Gaussian processes provided by Tagasovska et al. (2020) over different noise settings.

Linear Functions We start by generating cause effect pairs with linear functions and known ground truth, displayed in Fig. 3. The cause X is sampled from a normal distribution and is linked to the effect with $Y = \beta_0 + \beta_1 X + s(X)N_Y$. We sample 100 pairs of 100, 200, 500 and 1000 datapoints for all noise settings, which will be explained now. The noise is heteroscedastic Gaussian, where the variance function is a step function scaled in relation to the support S_Y of Y.

$$s(x) = \begin{cases} [0.05, 0.4] \cdot S_Y & \text{if } x < [0.3, 0.7] \\ [1.5, 2.5] \cdot [0.05, 0.4] \cdot S_Y & \text{otherwise} \end{cases}$$

We run all methods, and plot their average accuracies in Fig. 5. The results show that we can identify cause and effect for linear functions with overlapping noise, and give empirical backing for Theorem 1. High noise levels beyond the guarantees are solved especially well. We attribute this to the increased overlap of noise distributions that comes with larger noise variances. Overall, our method is ahead of its competition on all noise levels. The performance of RESIT and CAM also highlights the advantage of our causal model over plain ANMs. Because the independence of noise and cause is violated, the inference criterion of RESIT identifies the anti-causal direction. QCCD models heteroscedastic noise as well, but does not provide explicit guarantees for the linear case and cannot handle it. In contrast, our causal model accomodates both homoscedastic and heteroscedastic noise.

Non-Linear Functions Next, we consider non-linear functions. We do so by relating cause to effect via a non-linear cubic spline function. For each causal pair, we first randomly choose the noise to be either Gaussian or uniform. In accordance with the premise of this paper, we also vary the noise variance for each segment of the spline. This level of heteroscedasticity is controlled through a step parameter which determines how much the noise variance changes between the segments. An example for low and high heteroscedasticity is shown in Fig. 4. The step parameter is sampled uniformly from five different settings which we show in Fig. 6. Setting the step to 0 implies constant noise variance i.e. homoscedasticity. We generate a total of 100 pairs for each setting with a starting noise level of $20 - 30\%$ and 3 segments of 25-50 points.

[1] https://eda.mmci.uni-saarland.de/heci/

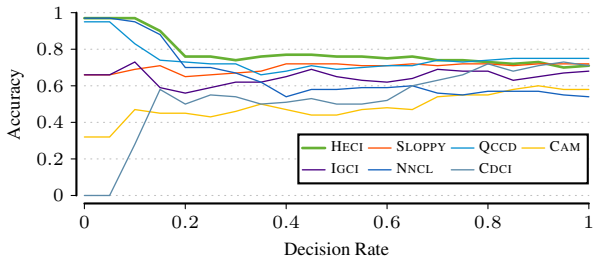

Figure 9. [Higher is better] Accuracy (weighted) over the Tübingen cause-effect pairs, ordered by decreasing heteroscedasticity ($\sigma^2_{max}/\sigma^2_{min}$).

We show the results of this experiments in Fig. 6. We see that HECI is able to robustly infer cause from effect under homogeneous noise and increasingly heteroscedastic noise conditions. The other approaches either only work well for homoscedastic noise, degrading rapidly as the noise variance increases across the dataset (RESIT, CAM), have a high variance in accuracy (QCCD and SLOPPY), or obtain lower accuracy than HECI throughout all settings (IGCI and CDCI). These results support our conjecture that the identifiability of the linear heteroscedastic and homoscedastic non-linear case extends to the heteroscedastic non-linear case.

Location Scaled and Multiplicative Noise After confirming that HECI is able to identify the correct causal directions on data that follows, or comes close to our assumptions, we next evaluate HECI on five synthetic benchmark datasets where our assumptions do not necessarily hold. For this we consider the data proposed by Tagasovska et al. (2020). These datasets consist of three different noise models, namely additive (AN), location scaled (LS) and multiplicative (MN-U), with as the underlying data generating process either a Gaussian process or an invertible sigmoidal function (AN-s, LS-s). We report the accuracy over the first three in Fig. 7. We see that HECI is robust to each of the three different noise settings. HECI is also robust when the underlying process consists of invertible sigmoidal functions, as shown in Fig. 8. In contrast, all all other methods except for CDCI deteriorate significantly in at least one of these two settings. These experiments support the conclusion that our method is able to discover and model heteroscedasticity, and consequently tell cause from effect, even when the generating mechanisms are outside of our causal model.

5.2. Tübingen Cause-Effect pairs

Last, we benchmark on the real-world benchmark Tübingen Cause-Effect pairs dataset. Since the main proposal of this paper is the inclusion of heteroscedasticity into the causal model, we are less interested in the overall accuracy, but are particularly concerned with how well methods do for those pairs that exhibit non-stationary noise. To this end we sort the cause–effect pairs by heteroscedasticity, measured by the proportion $\sigma^2_{max}/\sigma^2_{min}$ (maximum/minimum variance fitted by HECI in the causal direction), and report accuracy over the top-k pairs.

We show these results in Fig. 9. We see that among its competitors, HECI obtains the overall highest accuracy when we force it to decide over the most heteroscedastic half of the dataset. Overall, it achieves an average accuracy of **0.71**, which is on par with the next closest competitors QCCD and SLOPPY. These results corroborate that our causal model and the HECI algorithm are effective in dealing with non-constant noise encountered in real-world data.

6. Conclusion

In this paper we propose a causal model that sets itself apart from existing work by explicitly modelling local noise; by which it is particularly well-suited for a wide range of real-world applications. We show that we can identify the true causal model for linear functions with heteroscedastic noise, and non-linear functions homoscedastic noise. Through empirical evaluation, we show that our method can even identify the correct causal direction for non-linear funcions with heteroscedastic noise. On an observational sample we can compute our solution efficiently via dynamic programming and regularized with BIC. Through extensive experiments, we show that our method, HECI, indeed performs well on a wide range of benchmarks—especially in the target scenarios with high heteroscedacity. This advantage also shows on the real world Tübingen Cause-Effect pairs, in particular for those with a wide difference in variance of noise, and points towards the regularity and importance of heteroscedastic noise conditions.

As a continuation of this work, we aim to adapt the causal model and algorithm to introduce smoothness and outlier resistance to the fitted functions. Furthermore, we would like to expand local functions from polynomials to include more powerful models such as splines.

References

Blöbaum, P., Janzing, D., Washio, T., Shimizu, S., and Schölkopf, B. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics*, pp. 900–909. PMLR, 2018.

Budhathoki, K. and Vreeken, J. Causal inference by compression. In *Proceedings of the 16th IEEE International Conference on Data Mining (ICDM), Barcelona, Spain*, pp. 41–50. IEEE, 2016.

Bühlmann, P., Peters, J., Ernest, J., et al. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014.

Chickering, D. M. Optimal Structure Identification With Greedy Search. *Journal of Machine Learning Research*, 3(Nov):507–554, 2002.

Duong, B. and Nguyen, T. Bivariate causal discovery via conditional divergence. In *First Conference on Causal Learning and Reasoning*, 2021.

Hoyer, P., Janzing, D., Mooij, J., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 689–696, 2009.

Hu, S., Chen, Z., Partovi Nia, V., CHAN, L., and Geng, Y. Causal inference and mechanism clustering of a mixture of additive noise models. In *Advances in Neural Information Processing Systems*, pp. 5212–5222. PMLR, 2018.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Technology*, 56(10):5168–5194, 2010.

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B., and Schölkopf, B. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

Lachapelle, S., Brouillard, P., Deleu, T., and Lacoste-Julien, S. Gradient-based neural dag learning. In *International Conference on Learning Representations*, 2020.

Loh, P.-L. and Bühlmann, P. High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research*, 15(1): 3065–3105, 2014.

Marx, A. and Vreeken, J. Telling cause from effect using mdl-based local and global regression. In *2017 IEEE international conference on data mining (ICDM)*, pp. 307–316. IEEE, 2017.

Marx, A. and Vreeken, J. Identifiability of cause and effect using regularized regression. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 852–861, 2019.

Marx, A. and Vreeken, J. Formally justifying mdl-based inference of cause and effect. *Proceedings of the AAAI Workshop on Information Theoretic Causal Inference and Discovery*, 2022.

Mian, O., Marx, A., and Vreeken, J. Discovering fully oriented causal networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

Ng, I., Ghassami, A., and Zhang, K. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33, 2020.

Park, G. Identifiability of additive noise models using conditional variances. *Journal of Machine Learning Research*, 21(75):1–34, 2020.

Pearl, J. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000.

Peters, J. and Bühlmann, P. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Identifiability of causal graphs using functional models. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI*, pp. 589–598. AUAI Press, 2011.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Reisach, A., Seiler, C., and Weichwald, S. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34, 2021.

Sgouritsa, E., Janzing, D., Hennig, P., and Schölkopf, B. Inference of Cause and Effect with Unsupervised Inverse Regression. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 38: 847–855, 2015.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7, 2006.

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., and Richardson, T. *Causation, prediction, and search*. MIT press, 2000.

Stegle, O., Janzing, D., Zhang, K., Mooij, J. M., and Schölkopf, B. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems*, pp. 1687–1695. PMLR, 2010.

Tagasovska, N., Chavez-Demoulin, V., and Vatter, T. Distinguishing cause from effect using quantiles: Bivariate quantile causal discovery. In *International Conference on Machine Learning*, pp. 9311–9323. PMLR, 2020.

Verma, T. and Pearl, J. Equivalence and synthesis of causal models. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI*. AUAU Press, 1990.

Vowels, M. J., Camgoz, N. C., and Bowden, R. D'ya like dags? a survey on structure learning and causal discovery. *arXiv preprint arXiv:2103.02582*, 2021.

Wang, B. and Zhou, Q. Causal network learning with non-invertible functional relationships. *Computational Statistics & Data Analysis*, 156:107141, 2021.

Yu, Y., Chen, J., Gao, T., and Yu, M. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pp. 7154–7163. PMLR, 2019.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence (UAI*, pp. 647–655. AUAU Press, 2009.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Zheng, X., Dan, C., Aragam, B., Ravikumar, P., and Xing, E. Learning sparse nonparametric dags. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3414–3425. PMLR, 2020.

A. Identifiability

A.1. Error Scaling

Lemma 1 *Let X be a random variable with variance σ_X^2 and let $Y = \beta_0 + \beta_1 X + \alpha N_Y$, where N_Y has variance σ_Y^2 and $\alpha > 0$ is a scaling parameter. Further, let $X' = a + bX$ and $Y' = c + dY$ be the rescaled versions of X and Y, then*

$$\frac{E[\mathrm{Var}(X' \mid Y')]}{E[\mathrm{Var}(Y' \mid X')]} = \left(\frac{b}{d}\right)^2 \frac{\sigma_X^2}{\alpha^2\sigma_Y^2 + \beta_1^2\sigma_X^2} \,,$$

where $\sigma_X^2 = \mathrm{Var}(X)$ and $\alpha^2\sigma_Y^2 + \beta_1^2\sigma_X^2 = \mathrm{Var}(Y)$.

Proof: *By the law of total variation, we can derive that*

$$\mathrm{Var}(Y) = \mathrm{E}\left[\mathrm{Var}(Y \mid X)\right] + \mathrm{Var}(\mathrm{E}\left[Y \mid X\right]) = \alpha^2\sigma_Y^2 + \beta_1^2\sigma_X^2$$

and similarly that

$$\mathrm{E}\left[\mathrm{Var}(X \mid Y)\right] = \mathrm{Var}(X) - \mathrm{Var}(\mathrm{E}\left[X \mid Y\right]) = \sigma_X^2 - \frac{\beta_1^2\sigma_X^4}{\alpha^2\sigma_Y^2 + \beta_1^2\sigma_X^2} \,.$$

Thus, we can write

$$\begin{aligned}
\frac{E[\mathrm{Var}(X' \mid Y')]}{E[\mathrm{Var}(Y' \mid X')]} &= \left(\frac{b}{d}\right)^2 \frac{E[\mathrm{Var}(X \mid Y)]}{E[\mathrm{Var}(Y \mid X)]} \\
&= \left(\frac{b}{d}\right)^2 \frac{\sigma_X^2 - \frac{\beta_1^2\sigma_X^4}{\alpha^2\sigma_Y^2 + \beta_1^2\sigma_X^2}}{\alpha^2\sigma_Y^2} \\
&= \left(\frac{b}{d}\right)^2 \frac{\sigma_X^2}{\alpha^2\sigma_Y^2 + \beta_1^2\sigma_X^2} = \left(\frac{b}{d}\right)^2 \frac{\mathrm{Var}(X)}{\mathrm{Var}(Y)} \,.
\end{aligned}$$

□

A.2. Linear Functions

Theorem 1 *Given a causal model as specified in Eq.* (1)*, assume that*

(1) N_X, N_Y have finite support, and X and Y are normalized to obtain values within $[0, 1]$

(2) f is a linear function with $f(X) = \beta_0 + \beta_1 X$ and g is its inverse

(3) N_Y is unbiased with unit variance and strictly positive scale function $s_\alpha(x) = \alpha s(x)$, with $s_\alpha(x) \to 0$ if $\alpha \to 0$.

In that case it holds that in the limit of $\alpha \to 0$,

$$\begin{aligned}
&\int_0^1 p_Y(y) \cdot \log\left(\mathrm{Var}(X|Y = y)\right) dy \\
\geq &\int_0^1 p_X(x) \cdot \log\left(\mathrm{Var}(Y|X = x)\right) dx \,,
\end{aligned}$$

with equality, if and only if the conditional variance of the noise scaling $\mathrm{Var}\left(s(X)|Y\right) = 0$, i.e. there is no overlap of noise with different amplitude $s(x)$ in the domain $\mathcal{Y}$.

Proof: *Let the cause X have support S_X (before rescaling it to $[0, 1]$) and the minimum value of 0, which can be achieved by shifting X and adjusting β_0. By assumption, $N_Y \perp\!\!\!\perp N_X$ and N_Y has unit variance. Then, the support of Y before rescaling is $\beta_1 S_X + \epsilon_{\alpha,\sigma_Y}$, where $\epsilon_{\alpha,\sigma_Y}$ amounts for the increased support of Y by adding the scaled noise term*

αN_Y, which by Assumption (1) has finite support. Thus, $\epsilon_{\alpha,\sigma_Y}$ is bounded and $\epsilon_{\alpha,\sigma_Y} \to 0$ if $\alpha \to 0$. We normalize X, Y and $s_\alpha(X)$ as

$$\tilde{X} = \frac{X}{S_X},\ \tilde{Y} = \frac{Y - \beta_0}{\beta_1 S_X + \epsilon_{\alpha,\sigma_Y}},\ \tilde{s}_\alpha(X) = \frac{s_\alpha(X)}{\beta_1 S_X + \epsilon_{\alpha,\sigma_Y}}$$

To continue, note that the conditional variance $\mathrm{Var}(Y|X)$ is equivalent to the expected least squares error $\mathrm{E}\left[(Y - f(X))^2\right]$ where $f(X) = \mathrm{E}[Y|X]$. The rescaled variables are shown to be linked with the identity function $f(\tilde{X}) = \tilde{X}$ and $g(\tilde{Y}) = \tilde{Y}$ in the limit of $\alpha \to 0$. We begin by transforming the initial SEM to use the rescaled variables $\tilde{X}$ and $\tilde{Y}$.

$$\begin{aligned}
& Y = f(X) + s_\alpha(X) \cdot N_Y \\
\iff & Y - \beta_0 = \beta_1 X + s_\alpha(X) \cdot N_Y \\
\iff & \tilde{Y} = \frac{\beta_1}{\beta_1 S_X + \epsilon_{\alpha,\sigma_Y}} X + \frac{s_\alpha(X)}{\beta_1 S_X + \epsilon_{\alpha,\sigma_Y}} \cdot N_Y \\
\iff & \tilde{Y} = \frac{\beta_1}{\beta_1 S_X + \epsilon_{\alpha,\sigma_Y}} S_X \tilde{X} + \tilde{s}_\alpha(X) \cdot N_Y
\end{aligned}$$

We may express the conditional expectation of the normalized $\tilde{Y}$ given $\tilde{X}$ as

$$\begin{aligned}
\mathrm{E}\left[\tilde{Y}|\tilde{X}\right] &= \mathrm{E}\left[\frac{\beta_1}{\beta_1 S_X + \epsilon_{\alpha,\sigma_Y}} S_X \tilde{X} + \tilde{s}_\alpha(X) \cdot N_Y | \tilde{X}\right] \\
&= \mathrm{E}\left[\frac{\beta_1}{\beta_1 S_X + \epsilon_{\alpha,\sigma_Y}} S_X \tilde{X} | \tilde{X}\right] + \mathrm{E}\left[\tilde{s}_\alpha(X) \cdot N_Y | \tilde{X}\right]
\end{aligned}$$

The righthand term disappears as the noise N_Y is unbiased and is independent of the scaling $\tilde{s}_\alpha(X)$ and $\tilde{X}$ since by assumption $N_X \perp\!\!\!\perp N_Y$.

In the limit $\alpha \to 0$, the noise support $\epsilon_{\alpha,\sigma_Y}$ trends to zero. Then, the normalization factors of $\tilde{X}$ cancel each other out and it remains that $\mathrm{E}\left[\tilde{Y}|\tilde{X}\right] = \mathrm{E}\left[\tilde{X}|\tilde{X}\right] = \tilde{X}$. Similarly it can be shown, that $\mathrm{E}\left[\tilde{X}|\tilde{Y}\right] = \tilde{Y}$. We see that for the normalized variables in the limit of $\alpha \to 0$, the functional relationships $\mathrm{E}\left[\tilde{Y}|\tilde{X}\right] = f\left(\tilde{X}\right) = \tilde{X}$ in the causal direction and $\mathrm{E}\left[\tilde{X}|\tilde{Y}\right] = g\left(\tilde{Y}\right) = \tilde{Y}$ in the anti-causal direction both are the identity function.

Continuing, we assume the variables X and Y to be normalized in the scheme described above and drop the $\tilde{X}$ notation. In the causal direction we can express the conditional variance of Y given X as

$$\begin{aligned}
\mathrm{Var}(Y|X) &= \mathrm{E}\left[(Y - f(X))^2\right] = \mathrm{E}\left[(f(X) + s_\alpha(X) \cdot N_Y - f(X))^2\right] = \mathrm{E}\left[s_\alpha(X)^2\right] \\
\mathrm{Var}(Y|X = x) &= \mathrm{E}\left[s_\alpha(X)^2 | x\right] = \int_0^1 p_{X|X=x}(x') \cdot s_\alpha(x')^2 dx' = s_\alpha(x)^2
\end{aligned}$$

Thus we can give the continuous score of the causal direction as

$$\begin{aligned}
& \int_0^1 p_X(x) \cdot \log\left(\mathrm{Var}(Y|X = x)\right) dx \\
= & \int_0^1 p_X(x) \cdot \log\left(s_\alpha(x)^2\right) dx\,.
\end{aligned}$$

In the anti-causal direction with $g(Y) = Y$ for $\alpha \to 0$, the conditional variance is

$$\begin{aligned}
\mathrm{Var}(X|Y) &= \mathrm{E}\left[(X - g(Y))^2\right] = \mathrm{E}\left[(X - g(f(X) + s_\alpha(X) \cdot N))^2\right] = \mathrm{E}\left[(X - X - s_\alpha(X) \cdot N)^2\right] = \mathrm{E}\left[s_\alpha(X)^2\right] \\
\mathrm{Var}(X|Y = y) &= \mathrm{E}\left[s_\alpha(X)^2 | Y = y\right] = \int_0^1 p_{X|Y=y}(x) s_\alpha(x)^2 dx\,.
\end{aligned}$$

This renders the continuous score of the anti-causal direction

$$\begin{aligned}&\int_0^1 p_Y(y) \cdot \log\left(\mathrm{Var}(X|Y=y)\right) dy \\ =&\int_0^1 p_Y(y) \cdot \log\left(\int_0^1 p_{X|Y=y}(x) \cdot s_\alpha(x)^2 dx\right) dy\,.\end{aligned}$$

The logarithm is a concave function and we use Jensens inequality to show

$$\begin{aligned}&\geq \int_0^1 p_Y(y) \cdot \left(\int_0^1 p_{X|Y=y}(x) \cdot \log\left(s_\alpha(x)^2\right) dx\right) dy \\ &= \int_0^1 \left(\int_0^1 p_Y(y) \cdot p_{X|Y=y}(x) \cdot \log\left(s_\alpha(x)^2\right) dy\right) dx \\ &= \int_0^1 p_X(x) \cdot \log\left(s_\alpha(x)^2\right) dx \\ &= \int_0^1 p_X(x) \cdot \log\left(\mathrm{Var}(Y|X=x)\right) dx\,.\end{aligned}$$

To prove under which conditions both terms are equal, we note that the term on the left side of the inequality

$$\log\left(\int_0^1 p_{X|Y=y}(x) \cdot s_\alpha(x)^2 dx\right) = \log\left(\mathrm{E}\left[s_\alpha(X)^2|Y=y\right]\right),$$

is equal to the logarithm of the expected value for the squared noise scaling function of X at $Y=y$. On the other hand, if the order of the expectation and the logarithm is switched up, we obtain

$$\int_0^1 p_{X|Y=y}(x) \cdot \log\left(s_\alpha(x)^2\right) = \mathrm{E}\left[\log\left(s_\alpha(X)^2\right)|Y=y\right].$$

This corresponds to the expected value of the logarithmic squared noise scaling of X at $Y=y$. As the logarithm is strictly concave, Jensen's inequality lends the conclusion that

$$\log\left(\mathrm{E}\left[s_\alpha(X)|Y=y\right]\right) \geq \mathrm{E}\left[\log\left(s_\alpha(X)^2\right)|Y=y\right].$$

Equality holds for a given $y \in \mathcal{Y}$ if and only if $s_\alpha(X|Y=y)$ is constant. Throughout the whole domain this is enforced by $\mathrm{Var}(s(X)|Y)=0$. This relationship holds most prominently with equality for homoscedastic noise with $s(x)=c$. □

B. Empirical Log-Likelihood

Given a sample $\{x_i.y_i\}_{i=1}^n$ drawn iid from the joint distribution of X and Y the negative log-likelihood for the $X \to Y$ direction with normal distributed residuals $r_i = y_i - f(x_i)$ can be expressed as

$$
\begin{aligned}
p(r|x_i; s_\alpha) &= \frac{1}{\sqrt{2\pi s_\alpha(x_i)^2}} \exp\left(-\frac{r^2}{2s_\alpha(x_i)^2}\right) \\
-\log\left[L_{X\to Y}(s_\alpha^2, \hat{f})\right] &= -\log\left[\prod_{i=1}^{n} p(r_i|x_i; s_\alpha^2)\right] \\
&= -\sum_{i=1}^{n} \log(p(r_i|x_i; s_\alpha^2)) \\
&= -\sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi s_\alpha(x_i)^2}} \exp\left(-\frac{r^2}{2s_\alpha(x_i)^2}\right)\right) \\
&= -\sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi s_\alpha(x_i)^2}}\right) + \log\left(\exp\left(-\frac{r^2}{2s_\alpha(x_i)^2}\right)\right) \\
&= -\sum_{i=1}^{n} -\frac{1}{2}\log\left(2\pi s_\alpha(x_i)^2\right) - \frac{r^2}{2s_\alpha(x_i)^2} \\
&= \left(\frac{1}{2}\sum_{i=1}^{n} \log\left[s_\alpha(x_i)^2\right]\right) + \left(\sum_{i=1}^{n} \frac{1}{2}\frac{r_i^2}{s_\alpha(x_i)^2}\right) + \left(\sum_{i=1}^{n} \frac{1}{2}\log(2\pi)\right) \\
&= \frac{1}{2}\sum_{i=1}^{n} \log\left[s_\alpha(x_i)^2\right] + \frac{1}{2}\sum_{i=1}^{n} \frac{r_i^2}{s_\alpha(x_i)^2} + \frac{n}{2}\log(2\pi)\ .
\end{aligned}
$$

B.1. Heteroscedastic Noise

For heteroscedastic noise, the negative log-likelihood can be derived in a similar fashion. Let the domain of X be partitioned in m non-overlapping bins s.t. within each $\text{bin}_j \subset \{x_i, y_i\}_{i=1}^n$ there are n_j points, then the variance $\hat{\sigma}_j^2$ is estimated constant as $\hat{\sigma}_j^2 = \frac{1}{n_j}\sum_{x_i \in \text{bin}_j} r_i^2$.

The empirical negative log-likelihood w.r.t. a partitioning $\hat{\mathcal{P}}$ with m non-overlapping bins can be expressed as

$$
\begin{aligned}
-\log\left[L_{X\to Y}(\hat{\sigma}^2, f, \hat{\mathcal{P}})\right] &= \frac{1}{2}\sum_{i=1}^{n} \log\left[\hat{\sigma}^2(x_i)\right] + \frac{1}{2}\sum_{i=1}^{n} \frac{r_i^2}{\hat{\sigma}^2(x_i)} \\
&= \frac{1}{2}\sum_{j=1}^{m} n_j \log(\hat{\sigma}_j^2) + \frac{1}{2}\sum_{j=1}^{m} \sum_{x_k \in \text{bin}_j} \frac{r_k^2}{\hat{\sigma}_j^2} \\
&= \frac{1}{2}\sum_{j=1}^{m} n_j \log(\hat{\sigma}_j^2) + \frac{1}{2}\sum_{j=1}^{m} \frac{1}{\hat{\sigma}_j^2} \sum_{x_k \in \text{bin}_j} r_k^2 \\
&= \frac{1}{2}\sum_{j=1}^{m} n_j \log(\hat{\sigma}_j^2) + \frac{1}{2}\sum_{j=1}^{m} \frac{1}{\hat{\sigma}_j^2} n_j \hat{\sigma}_j^2 \\
&= \frac{1}{2}\sum_{j=1}^{m} n_j \log(\hat{\sigma}_j^2) + \frac{1}{2}\sum_{j=1}^{m} n_j \\
&= \frac{1}{2}\sum_{j=1}^{m} n_j \log(\hat{\sigma}_j^2) + \frac{n}{2}\ ,
\end{aligned}
$$

so that the last term only depends on n and is thus equal in both directions. After dropping we obtain

$$
-\log\left[L_{X\to Y}(\hat{\sigma}^2, \hat{f}, \hat{\mathcal{P}})\right] = \sum_{j=1}^{m} \frac{n_j}{2} \cdot \log(\hat{\sigma}_j^2).
$$

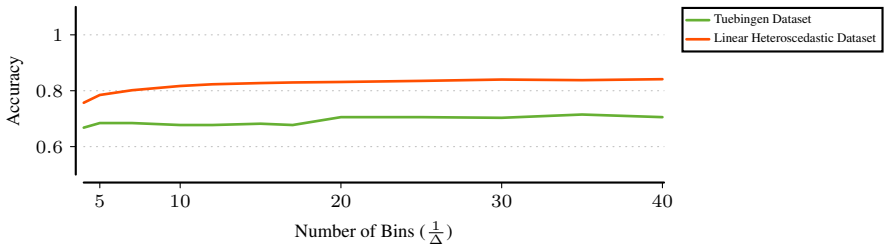

Figure 10. [Higher is better] Accuracy in determining cause from effect for an increasing number of bins on the simulated linear heteroscedastic (see Sec. 5.1) and the Tuebingen dataset

C. Setting the Binning Parameter Δ

The parameter Δ determines the width and therewith number of initial bins. In the limit $\Delta \to 0$, with the number of initial bins growing as a sub-linear function of n, we can recover the true noise variance function. In practice, however, each bin must contain sufficient points for variance estimation. Thus, on a finite sample, setting Δ is a trade-off between running time, which depends quadratically on the number of initial bins, and accuracy through the more accurate approximation of the borders between regions in which the noise changes.

To show that Δ has a small influence in practice, we re-run our experiments on the linear heteroscedastic and Tuebingen datasets for different values of Δ, and plot the results in Fig. 10. We can see that Δ has barely any influence on the accuracy of HECI, and plateaus beyond 20 initial bins (corresponding to $\Delta = 0.05$, which we use in the main body of paper).

In general, starting with a more fine grained grid allows for better estimation of the noise variance, since the binning algorithm finds the optimal solution given the starting bins. Each initial bin though has to at least contain multiple data points to allow for an initial noise estimate and regression step. Smaller values of Δ may thus come at a cost of running time while providing marginal gains in accuracy.