Causal Inference with Heteroscedastic Noise Models

Sascha Xu,[1,2] **Alexander Marx,** [3] **Osman Mian,** [1] **Jilles Vreeken,** [1]

[1] CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
[2] Saarland University, Saarbrücken, Germany
[3] ETH AI Center, Zürich, Switzerland
s8xgxuu@stud.uni-saarland.de, alexander.marx@ai.ethz.ch, osman.mian@cispa.de, jv@cispa.de

Abstract

We study the problem of identifying the cause and the effect between two univariate continuous variables X and Y. The examined data is purely observational, hence it is required to make assumptions about the underlying model. Often, the independence of the noise from the cause is assumed, which is not always the case for real world data. In view of this, we present a new method, which explicitly models heteroscedastic noise. With our HEC algorithm, we can find the optimal model regularized, by an information theoretic score. In thorough experiments we show, that our ability to model heteroscedastic noise translates into a superior performance on a wide range of synthetic and real-world datasets.

Introduction

Causal discovery algorithms based on conditional independence test are unable to discover fully oriented causal graphs. To disambiguate between Markov equivalent graphs, the causal direction between two variables must be inferred, a problem known as bivariate causal inference. Pearl (2000) showed that it is impossible to tell cause from effect from observational data without additionally making assumptions about the data generating process. Causal methods must therefore put lots of care into their modelling assumptions, such that it is both possible to guarantee that cause and effect can be identified under these assumptions, as well as that those assumptions are as likely to hold in practice as possible. Many methods build upon the assumption that noise is completely independent from the cause (Bühlmann et al. 2014; Peters et al. 2014; Shimizu et al. 2006; Hoyer et al. 2009). Tagasovska, Chavez-Demoulin, and Vatter (2020) show that methods, that do so, fail when the data generating process includes, for example, location-scaled noise.

In this work, we propose a method that sets itself apart from the state of the art by explicitly modelling heteroscedastic noise. Heteroscedacity describes the phenomenon of a different noise variance within the domain of the regressor. Rather than wishing it away, we propose a causal model that builds upon the additive noise model, but explitictly permits heteroscedacity. The cornerstone of our approach is a fitting process that automatically divides up the domain into segments of noise with different variances. We show under which assumptions we can idenfity the true causal direction using the Bayesian information criterion. We propose an efficient dynamic-programming based algorithm, HEC, that can determine the optimal scoring model in quadratic time despite an exponential search space. Together, the addition to the additive noise model and the optimal algorithm allow us to demonstrate that HEC outperforms a wide range of state of the art methods on synthetic and real world benchmarks that exhibit non-stationary noise.

Theory

We consider the problem of inferring cause and effect between two dependent, continuous random variables X and Y, under causal sufficiency assumption. That is, we assume that there exists no unobserved confounding variable Z, which causes both X and Y. Consequently, 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, Janzing, and Schölkopf 2017), which we define below. Before that, we introduce the notation used throughout this paper.

Notation

We refer to a sample of size n drawn from the distribution P of a random variable X as $\{x_i\}_{i=1}^{n}$. Lowercase letters x denote values from the domain $\mathcal{X}$ of X. Further, we follow the convention of denoting the parameters of a function f as β_f, where $||\beta_f||_0$ is the L_0 norm of the parameter vector.

Causal Model

Unlike most state-of-the-art approaches, we do not assume an independence between cause and noise, but instead allow for heteroscadastic noise, which may depend on the cause.

Assumption 1 *(Causal Model). The effect Y is generated from the cause X and noise variable N as*

$$Y = f(X) + s(X) \cdot N \ ,$$

where f is a non-linear function and s is a scaling function, which we specify further in Assumption 2.

Assumption 2 *(Heteroscedastic Noise). The scaled noise (which may depend on X) is constructed from a standard Gaussian variable N and a strictly positive scaling function $s : \mathcal{X} \to \mathbb{R}^+$—i.e. the variance of the scaled noise variable $s(X) \cdot N$ is equal to $s^2(X)$.*

Assumption 3 *(Compact Supports). The distribution of X and the distribution of Y has compact support, so that X and Y attain values within 0 and 1 (similar to the assumption made by Blöbaum et al. (2018)).*

One of the main advantages of the above causal model is that it can express various noise settings. In particular, if $s(x) = c$ is just a mapping to a constant c, the above model reduces to an additive noise model. More interesting to us, however, is noise that may fan out scaled by location, which can be expressed with $s(x) = ax + b$. We provide an example for such a generative mechanism in Fig. 1. As shown, we approximate the mechanism by modeling the variance $s^2(x)$ as a piecewise constant function. That is, we assume that we can construct a partitioning $\mathcal{P}$ of the domain of X s.t. $s^2(x)$ is constant within a bin of the partition, but may vary between bins.

Assuming the model above, we will now explain how to infer the causal direction between X and Y.

Inference

To infer the causal direction between X and Y, we follow a recent line of research suggesting that it suffices to compare the expected error (i.e. the residuals) when fitting a non-linear function for the causal and anti-causal direction (Blöbaum et al. 2018; Marx and Vreeken 2019).

In particular, for the low-noise setting Blöbaum et al. (2018) prove that

$$\mathbb{E}[(Y - f(X))^2] \leq \mathbb{E}[(X - g(Y))^2] \ ,$$

where f is the function minimizing the expected error when fitting a regression function from X to Y and g is the corresponding function minimizing the expected error in the anti-causal direction. Although the assumptions of the original approach—i.e. asserting low-noise, compact supports (Assumption 3) and additive noise are quite restrictive, our empirical evaluation suggests that it is applicable to a much more general setting.

To approximate the above inference criterium, we do not directly compare the residual errors, but instead compare the negative log-likelihoods w.r.t. the residuals under. That is, we refer to the negative log-likelihood of residuals when fitting a model from X to Y as $-\log L_{X\to Y} \approx n \log \hat{\sigma}^2$, which is an increasing function of the empirical error. Similarly, we denote the negative log-likelihood of the inverse model by $-\log L_{Y\to X}$. Thus, we say that X causes Y if $-\log L_{X\to Y} < -\log L_{Y\to X}$, that Y causes X if $-\log L_{X\to Y} > -\log L_{Y\to X}$ and do not decide if both quantities are equal. For our assumed causal model, i.e. under Assumptions 1–3, we can express the negative log-likelihood as follows.

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

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

Notice that last term depends only on n and can thus be dropped, as it is the same for the inverse direction. Further, if the variance is estimated homogeneously over the entire domain, the maximum likelihood estimator is $\hat{\sigma}^2_{global} = \frac{1}{n}\sum_{i=1}^n r_i^2$. In our causal model, this corresponds to the empirical variance function $\hat{s}^2(x) = \hat{\sigma}^2_{global}$. Hence, we can reformulate the second term as

$$\sum_{i=1}^{n} \frac{r_i^2}{\hat{s}^2(x_i)} = \frac{1}{\hat{\sigma}^2_{global}}\left(\sum_{i=1}^{n} r_i^2\right) = \frac{n\hat{\sigma}^2_{global}}{\hat{\sigma}^2_{global}} \ ,$$

which only depends on n and may be dropped. Thus, for constant additive noise, the empirical negative log-likelihood can be expressed as $\frac{n}{2} \cdot \log(\hat{\sigma}^2_{global})$.

For heteroscedastic noise, the negative log-likelihood can be derived in a similar fashion. If the domain of X can be partitioned in m non-overlapping bins s.t. within each bin_j the variance $\hat{\sigma}_j^2$ is constant, the empirical negative log-likelihood w.r.t. a partitioning $\hat{\mathcal{P}}$ with m non-overlapping bins can be expressed as

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

where n_j relates to the number of data points falling within bin_j. For the inverse direction, we can derive the corresponding negative log-likelihood similarly.

In the next section, we will explain how we compute $\hat{\mathcal{P}}$ and $\hat{f}$ to minimize the corresponding negative log-likelihood via dynamic programming.

Algorithm

In the previous section, we established the heteroscedastic noise model and a log-likelihood based approach to identify the causal direction. It uses the residuals of the fitted function $\hat{f}$ under a partition $\hat{\mathcal{P}}$. However, ordinary least squares regression and other methods estimate $\hat{f}(x)$ under the assumption of homoscedastic noise.

In view of this, we present the HEC algorithm for **he**teroscadastic noise **c**ausal models. The regressor domain is divided up into segments, where least squares based regression models are fitted. This way we implicitely estimate $s^2(x)$ as locally constant, but globally different, heteroscedastic. 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.

[1] In practice, we use the logarithm with base 2 to refer to bits.

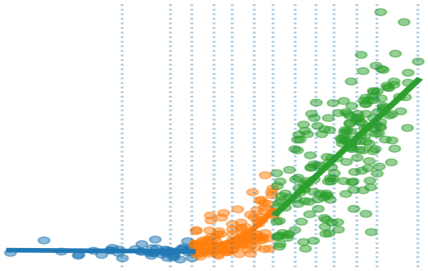

Figure 1: Fitted causal model with heteroscedastic noise. The blue, orange and green segments show the partioned domain with locally constant variance. The dashed blue lines show the initial bins for the partition.

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 $\text{bin}_j : [\min_j, \max_j)$. In Fig. 1, 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 compact supports $\min_1 = 0$ and $max_b = 1$. In practice, this is achieved through normalization of X.

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. From the set of initial bins $\{\text{bin}_j\}_{i=1}^{b}$, the task is to find a partition $\hat{\mathcal{P}}$ of neighboring bins with the underlying function $\hat{f}$, which minimizes the negative log likelihood, as described below.

In theory, we could approximate any noise variance $s^2(x)$ under the assumption that $n \to \infty$, where at the same time the maximal bin width goes to zero. Further, it must be ensured that the number of bins grows sub-linearly w.r.t. n, such that enough data is available for each bin.

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 $|\hat{\mathcal{P}}|$. The Akaike and Bayesian Information Criterion trade off the complexity of the fitted function with the predictive error. They offer a practical way to guide the model search.
AIC.

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

BIC.

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

For the scoring criterion, we opt for the stronger regularization of the BIC score. Intuitively, the stronger regularization helps us to avoid overfitting in cases where the true model is outside of our causal model and for large datasets, where gains of 2 bits are achieved easily.

With BIC, the task is to find the combination of local functions, which minimize it. As we saw in the previous section, the data likelihood is decomposable into independent, additive components. In particular, BIC of a model, which is partitioned at bin_{a-1} is additive, i.e.

$$\text{BIC}(f, \bigcup_{j=1}^{b} \text{bin}_j) = \text{BIC}(f_1, \bigcup_{j=1}^{a-1} \text{bin}_j) + \text{BIC}(f_2, \bigcup_{k=a}^{b} \text{bin}_k) \ .$$

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

HEC: 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. A similar algorithm for subgroup discovery is described in full detail by Nguyen and Vreeken (2016), or for histogram density estimation by Kontkanen and Myllymäki (2007).

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 multiple, neighboring bins, which we will call segments from now on, there exist two possibilities for the optimal model $\tilde{f}_{p,q}$:

- A local function $f_{p,q}$ for the segment from bin_p to bin_q
- A combination of 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

$$f_{p,q} = \arg\min_f \text{BIC}(f, \bigcup_{j=p}^{q} \text{bin}_j) \ .$$

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 and saved.

$$\tilde{f}_{p,q} = \begin{cases} f_{p,q}, & \text{if BIC}(f_{p,q}, \bigcup_{j=p}^{q} \text{bin}_j) \text{ is min} \\ \tilde{f}_{p,a} \cup \tilde{f}_{a,q} & \text{if BIC}(\tilde{f}_{p,a}, \bigcup_{j=p}^{a} \text{bin}_j) \\ & + \text{BIC}(\tilde{f}_{a+1,q}, \bigcup_{k=a+1}^{a} \text{bin}_k) \text{ is min} \end{cases}$$

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. 1. From the initial b bins, HEC uses the described bottom-up approach to find the optimal partition and local functions, which are marked as blue, orange and green. Like our causal model, the variance is modelled as locally constant, but different between each segment.

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 HEC is $\mathcal{O}(b^2 \cdot n)$. It means, that HEC finds the BIC-optimal partitioning and function in only a quadratic amount fits for the given bins.

Inference with HEC

With all described components we now predict the causal direction. First, X and Y are normalized to attain values between 0 and 1. In both directions, we fit the causal models with the described HEC algorithm. For the $X \to Y$ as well as the $Y \to X$ directions, we attain the empirical negative log-likelihood and predict the causal direction as the one corresponding to the lower negative log-likelihood as described in the theory section. Additionally, to take the complexity of the fitted function and partitioning into account, we use the regularized BIC scores to conduct the comparison and infer the causal direction.

Related Work

Causal inference from observational data is an important problem in science, and in recent years has received a lot of attention (Mian, Marx, and Vreeken 2021; Glymour, Zhang, and Spirtes 2019; Tagasovska, Chavez-Demoulin, and Vatter 2020; Wang and Zhou 2021). Constraint-based approaches that use conditional independence tests (Colombo and Maathuis 2014; Spirtes, Meek, and Richardson 1999) can identify causal models up to Markov equivalance, i.e. they cannot distinguish between the two Markov equivalent DAGs $X \to Y$ and $X \leftarrow Y$ (Verma, Pearl et al. 1991; Pearl 2000). To identify the causal direction between a pair of variates it is hence necessary to make additional assumptions about the generating mechanism.

The most common such assumption is the additive noise model (Peters, Janzing, and Schölkopf 2017), which has been exploited in various settings. In essence, additive noise models 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 and Hyvärinen 2009), it has been shown that such an additive noise model cannot (or, is extremely unlikely to) hold 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 extension of the additive noise model is NNCL (Wang and Zhou 2021), which similar to our approach partitions the domain of the cause into two, fits linear models for each bin, and then checks whether the additive noise assumption holds for the partitioned model. Different to NNCL, we consider a more general class of partitions, non-linear functions and follow the line of research based on comparing regression errors.

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 and Vreeken 2016; Marx and Vreeken 2017; Stegle et al. 2010; Tagasovska, Chavez-Demoulin, and Vatter 2020; Mian, Marx, and Vreeken 2021). Both postulates 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, Janzing, and Schölkopf 2017; Janzing and 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 and Vreeken 2017) and QCCD (Tagasovska, Chavez-Demoulin, and Vatter 2020) are two recent proposals that aim to approximate the algorithmic Markov condition. Although they empirically perform well, both do not have identifiability guarantees.

Closely related methods to our work are the ones that base their inference rules on regression error. Two such approaches are RECI (Blöbaum et al. 2018), which compares the expected regression error, and SLOPPY (Marx and 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. In contrast to our approach, none of these approaches are tailored towards heterogenous noise.

Experiments

In this section we empirically evaluate HEC on both synthetic data and the real-world Tübingen cause and effect pairs (Mooij et al. 2016) dataset. We will compare it to a wide range of state-of-the-art bivariate causal inference methods. As representative approaches that assume an additive noise model, we compare to CAM (Bühlmann et al. 2014) and RESIT (Peters et al. 2014). Further, we compare to SLOPPY (Marx and Vreeken 2019), SLOPE (Marx and Vreeken 2017), IGCI (Janzing et al. 2012) and QCCD (Tagasovska, Chavez-Demoulin, and Vatter 2020) as the state-of-the-art information theoretic approaches, and finally also to NNCL (Wang and Zhou 2021) as the bivariate causal inference approach for piecewise/non-invertible functions.

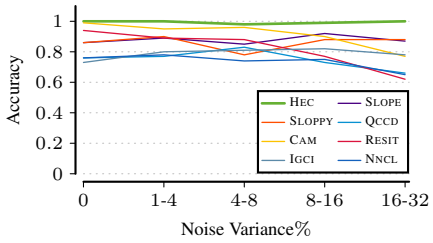

Figure 2: [Higher is better] Accuracy in determining cause from effect for increasing heteroscedasticity.

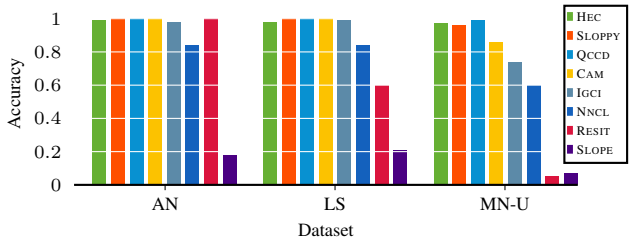

Figure 3: [Higher is better] Accuracy over benchmark synthetic data with Additive Noise (AN), Location scaling (LS) and Multiplicative noise (MN-U).

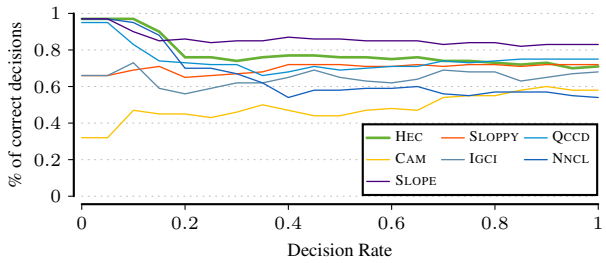

Figure 4: [Higher is better] Accuracy (weighted) over the Tübingen cause-effect pairs, ordered by decreasing heteroscedasticity.

HEC is implemented in Python and we provide the source code as well as the synthetic data for research purposes.[2] All experiments are executed on a 4-core Intel i7 machine with 16 GB RAM, running Windows 10. For each instance, HEC was able to decide the causal direction in less than 5 seconds.

Synthetic Data

We test HEC on two different settings. First, we generate synthetic data according to our assumed causal model (see Assumptions 1–3). Next, we use the synthetic data provided by Tagasovska, Chavez-Demoulin, and Vatter (2020) over different noise settings.

Heteroscedastic Noise We start by generating datasets with known ground truth. 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. We then introduce heteroscedasticity by dividing the domain of the causal variable into three continuous, but disjoint sections of 25 samples each, with each section having a different noise variance. The level of heteroscedasticity is controlled in each experiment through a step parameter which determines how much the noise variance changes between the segments. We sample the step parameter for each pair uniformly from five different settings which we show in Fig 2. Setting the step parameter to 0 implies constant noise variance i.e. homoscedasticity. We generate a total of 100 pairs for each setting.

We run all methods, and plot their average accuracies in Fig. 2. We see that HEC achieves a near-perfect accuracy in all of the settings, whilst also having the smallest variance between results. Other approaches either work well for homoscedastic noise, but degrade rapidly as the noise variance increases across the dataset (RESIT), have a high variance in accuracy (QCCD, SLOPE and SLOPPY) or are stable but have a lower accuracy than HEC throughout all settings (IGCI).

Location Scaled and Multiplicative Noise After confirming that HEC is able to identify the correct causal directions inside our causal model, we next evaluate HEC on three synthetic benchmark datasets proposed by Tagasovska, Chavez-Demoulin, and Vatter (2020), where our assumptions are unlikely to hold exactly. These datasets consist of three different variants of noise, namely additive (AN), location scaled (LS) and multiplicative (MN-U). We report the accuracy over each of these data sets in Fig 3. We see that HEC is robust to each of the three different noise settings alongside SLOPPY and QCCD, with the latter outperforming HEC by a slightest of margins. For other approaches we see that they can only handle additive noise (RESIT) or deteriorate notably for multiplicative noise settings (IGCI and NNCL).

Tübingen Cause-Effect pairs

Finally, as the real-world benchmark datasets we compare to the Tübingen Cause-Effect pairs. Overall, HEC achieves an accuracy of **0.71**, significantly beaten only by SLOPE and on par with the next closest competitors QCCD and SLOPPY.

Since the main objective of this paper is the inclusion of heteroscedasticity into the causal model, we examine this aspect further. That is, we sort the cause effect pairs by heteroscedasticity, which is measured by the proportion $\sigma^2_{max}/\sigma^2_{min}$ (maximum/minimum variance fitted by HEC in the causal direction).

Fig. 4 shows the accuracy in relation to the decided proportion of the pairs ordered by heteroscedasticity. Apart from SLOPE, HEC is superior to all other approaches if we decide the most heteroscedastic half of the dataset. Even QCCD, which does quantile regression with non-constant noise assumptions, is outperformed in this segment. It shows, that the causal model and the HEC algorithm are effective in dealing with highly variable noise. In addition, we achieve a generally strong performance on synthetic benchmarks, where methods like SLOPE fail. Overall, the results show the potential of causal inference in the presence of het-

[2] https://eda.mmci.uni-saarland.de/prj/hec/

eroscedastic noise.

Conclusion

In this paper we presented work in progress. We propose a causal model that sets itself apart from existing work by explicitely modelling heteroscedastic 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 using a broad range of information theoretic criteria, including AIC and BIC, as well as how to efficiently do so from observational data via dynamic programming. Through the experiments, we show that our method, HEC, 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. Finally, an investigation into identifiability of our causal model is to be conducted, with the goal of providing guarantees under certain conditions.

References

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

Budhathoki, K.; and Vreeken, J. 2016. Causal Inference by Compression. 41–50. IEEE.

Bühlmann, P.; Peters, J.; Ernest, J.; et al. 2014. CAM: Causal additive models, high-dimensional order search and penalized regression. 42(6): 2526–2556.

Colombo, D.; and Maathuis, M. H. 2014. Order-independent constraint-based causal structure learning. 15(1): 3741–3782.

Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*.

Hoyer, P.; Janzing, D.; Mooij, J.; Peters, J.; and Schölkopf, B. 2009. Nonlinear causal discovery with additive noise models. 689–696.

Hu, S.; Chen, Z.; Partovi Nia, V.; CHAN, L.; and Geng, Y. 2018. Causal Inference and Mechanism Clustering of A Mixture of Additive Noise Models. 5212–5222.

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

Janzing, D.; and Schölkopf, B. 2010. Causal Inference Using the Algorithmic Markov Condition. 56(10): 5168–5194.

Kontkanen, P.; and Myllymäki, P. 2007. MDL histogram density estimation. In *AISTATS*, 219–226.

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

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

Mian, O.; Marx, A.; and Vreeken, J. 2021. Discovering fully oriented causal networks.

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

Nguyen, H.-V.; and Vreeken, J. 2016. Flexibly mining better subgroups. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, 585–593. SIAM.

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

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

Peters, J.; Mooij, J. M.; Janzing, D.; and Schölkopf, B. 2011. Identifiability of Causal Graphs Using Functional Models. 589–598. AUAI Press.

Peters, J.; Mooij, J. M.; Janzing, D.; and Schölkopf, B. 2014. Causal discovery with continuous additive noise models.

Sgouritsa, E.; Janzing, D.; Hennig, P.; and Schölkopf, B. 2015. Inference of Cause and Effect with Unsupervised Inverse Regression. 38: 847–855.

Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; and Kerminen, A. 2006. A Linear Non-Gaussian Acyclic Model for Causal Discovery. 7.

Spirtes, P.; Meek, C.; and Richardson, T. 1999. An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, causation, and discovery*, 21: 1–252.

Stegle, O.; Janzing, D.; Zhang, K.; Mooij, J. M.; and Schölkopf, B. 2010. Probabilistic latent variable models for distinguishing between cause and effect. (26): 1687–1695.

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

Verma, T.; Pearl, J.; et al. 1991. Equivalence and synthesis of causal models.

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

Zhang, K.; and Hyvärinen, A. 2009. On the Identifiability of the Post-nonlinear Causal Model. 647–655. AUAU Press.