Quantile Causal Discovery

Anonymous Authors[1]

Abstract

Causal inference using observational data is challenging, especially in the bivariate case. Through the minimum description length principle, we link the postulate of independence between the generating mechanisms of the cause and of the effect given the cause to quantile regression. Based on this theory, we develop Quantile Causal Discovery (QCD), a new method to uncover causal relationships. Because it uses multiple quantile levels instead of the conditional mean only, QCD is adaptive not only to additive, but also to multiplicative or even location-scale generating mechanisms. To illustrate the effectiveness of our approach, we perform an extensive empirical comparison on both synthetic and real datasets. This study shows that QCD is robust across different implementations of the method (i.e., the quantile regression), computationally efficient, and compares favorably to state-of-the-art methods.

1 Introduction

Driven by the usefulness of causal inference in most scientific fields, an increasing body of research has contributed towards understanding the generative processes behind data. The aim is elevation of learning models towards more powerful interpretations: from correlations and dependencies towards *causation* (Pearl, 2009; Spirtes et al., 2000b; Dawid et al., 2007; Pearl et al., 2016; Schölkopf, 2019).

While the golden standard for causal discovery is randomized control trials (Fisher, 1936), experiments or interventions in a system are often prohibitively expensive, unethical, or, in many cases, impossible. In this context, an alternative is to use observational data to infer causal relationships (Spirtes et al., 2000b; Maathuis & Nandy, 2016). This challenging task has been tackled by many, often relying on testing conditional independence and backed up by heuristics (Maathuis & Nandy, 2016; Spirtes & Zhang, 2016; Peters et al., 2017).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

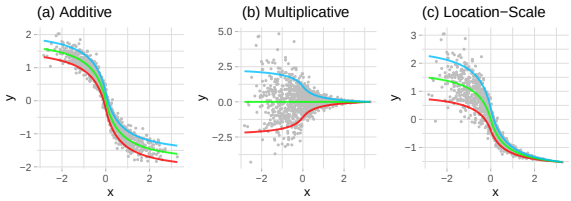

Figure 1. Diverse cause-effect ($X \to Y$) setups. Green - median, blue - 0.9^{th} quantile, red - 0.1^{th} quantile. Focusing on the mean only is not enough in the location-scale and multiplicative cases.

Borrowing from structural equations and graphical models, structural causal models (SCMs, Pearl et al., 2016; Peters et al., 2017) represent the causal structure of variables $X_1, \cdots, X_d$ using equations such as

$$X_c = f_c(X_{PA(c),\mathcal{G}}, N_c),\ c \in \{1, \ldots, d\},$$

where f_c is a causal mechanism linking the child/effect X_c to its parents/direct causes $X_{PA(c),\mathcal{G}}$, N_c is another variable independent of $X_{PA(c),\mathcal{G}}$, and $\mathcal{G}$ is the directed graph obtained from drawing arrows from parents to their children. Further complications arise when observing only two variables. In this case, one cannot distinguish between latent confounding ($X \leftarrow Z \to Y$) and direct causation ($X \to Y$ or $X \leftarrow Y$) without additional assumptions (Shimizu et al., 2006; Janzing et al., 2012; Peters et al., 2014; Lopez-Paz et al., 2015). A possible solution to this open question is to impose certain model restrictions. For example, (non-)linear additive noise models, with $Y = f(X) + N_Y$, provide a foundation for establishing *identifiability* (Shimizu et al., 2006; Hoyer et al., 2009; Peters et al., 2011). An extension is the post nonlinear model (Zhang & Hyvärinen, 2009), $Y = g(f(X) + N_Y)$, with g being an invertible function.

In Figure 1, we illustrate a limitation of causal discovery methods based on the conditional mean and refer to Section 2.2 for more details on the underlying generative mechanisms. From Figure 1(a), it is clear that the variance of the effect variable is independent of the cause. As a result, the independence between the cause and the effect's noise can be measured in various ways by substracting an estimate of the conditional mean. In Figure 1(b) and (c) however, the mechanisms are $Y = g(X)N_Y$ and $Y = f(X) + g(X)N_Y$ respectively. In other words, the variance of the effect also depends on the cause, i.e. it exhibits heteroskedasticity. Assume that $E(N_Y) = 0$, then in the multiplicative case, we have that $E(Y|X = x) = 0$. And it is not sensible to use the conditional mean to identify the causal direction. But since $\text{Var}(Y|X = x) = x^2$, the variance is more infor-

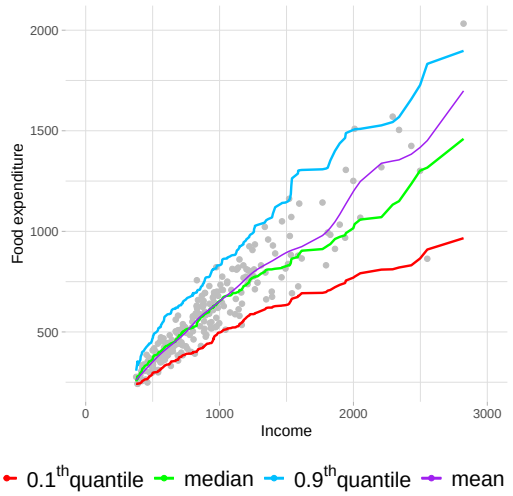

Figure 2. Heteroskedasticity in food expenditure as a function of income for Belgian working class households. Our method uses multiple quantiles to find the correct causal direction.

mative. Similarly, features of the conditional distributions (e.g., conditional spread) different from the location might help. In such cases, relying on mutliple (conditional) quantiles rather than on the mean only can help. As we show in Section 3, it allows us to correctly determine the causal direction across various benchmarks where additive-noise competitors fail when their assumptions are not met.

A classic example of real world data displaying such features is the impact of income on expenditure on meals: as an individual's income increases, so does its food expenditure and food expenditure variability. An explanation could be as follows: while a poorer person will spend a rather constant amount on inexpensive food, wealthier individuals occasionally buy inexpensive food and sometimes eat expensive meals. Or, that wealthier individuals have more leeway when deciding which fraction of their income to allocate to food expenditure, whereas poorer ones are constrained by necessity. This heteroskedastic effect can be observed in Figure 2, where our method use multiple quantiles to find the correct causal direction.

Another line of work avoids functional restrictions by relying on the *independence of cause and mechanism* postulate (Schölkopf et al., 2012; Peters et al., 2017):

Postulate 1 (Sgouritsa et al. 2015)**.** *The marginal distribution of the cause and the conditional distribution of the effect given the cause, corresponding to independent mechanisms of nature, are independent (i.e., they contain no information about each other).*

Information Geometric Causal Inference (IGCI) (Janzing et al., 2012) uses the postulate directly for causal discovery. Alternatively (Mooij et al., 2010) and (Janzing & Schölkopf, 2010) reformulate the postulate through asymmetries in Kolmogorov complexities (Kolmogorov, 1963) between marginal and conditionals distributions. However, the halting problem (Turing, 1937; 1938) implies that the Kolmogorov complexity is not computable, and approximations or proxies have to be derived to make the concept practical. In this context, (Mooij et al., 2010) proposes an approximation based on the minimum message length principle using Bayesian priors, while other metods are based on reproducing kernel Hilbert space embedding such as EMD (Chen et al., 2014), FT (Liu & Chan, 2017) and KCDC by Mitrovic et al. (2018). A related line of work suggests using the minimum description length (MDL, Rissanen, 1978) principle as a proxy for Kolmogorov complexity: (Budhathoki & Vreeken, 2017) uses MDL for causal discovery on binary data, and Slope (Marx & Vreeken, 2017; 2019) implements (local and) global functional relations using MDL based regression and is suitable for continuous data.

In this work, we build on a similar idea, using *quantile scoring* as a proxy for the Kolmogorov complexity through the MDL principle. To the best of our knowledge, quantiles have only been mentioned in a somewhat related context by (Heinze-Deml et al., 2018), where quantile predictions are used to exploit the invariance of causal models across different environments. As opposed to (Heinze-Deml et al., 2018), our method uses an asymmetry directly derived from the postulate, and therefore it does not require an additional variable for the environment.

To avoid the restrictive assumptions imposed by standard quantile regression techniques (e.g., linearity of the quantiles or additive relationships), we suggest fully nonparametric QCD implementations using three different approaches: *copulas, quantile forests or quantile neural networks*. We also show that QCD is robust to the choice regression approach. To the best of our knowledge, we are the first to explore the idea of using conditional quantiles to distinguish cause from effect in bivariate observational data. Our main contributions are:

- a new method based on quantile scoring to determine the causal direction without any assumptions on the class of causal mechanisms along with a theoretical analysis justifying its usage (Section 2),
- *quantile causal discovery* (QCD), an efficient implementation robust to the choice of underlying regressor (Section 3.1),
- a new benchmark set of synthetic cause-effect pairs from additive, location-scale and multiplicative noise models (Section 3.2),
- a comparative study to benchmark QCD against state-of-the-art alternatives (Section 3).

2 Causal Discovery using Quantiles

In this section, we develop our quantile-based method for distinguishing between cause and effect from continuous and discrete observational data.

Problem setting We restrict ourselves to bivariate cases by considering pairs of univariate random variables. We further simplify the problem by assuming the existence of a causal relationship but the absence of confounding, selection bias, and feedback.

2.1 Kolmogorov Complexity and Quantile Scoring

Let X and Y be two random variables with some joint distribution F, and where F_X, F_Y and $F_{Y|X}, F_{X|Y}$ are respectively the marginal and conditional distributions. Denote by $K(F)$ the Kolmogorov complexity of a distribution F, namely the length of the shortest computer program producing F as an output (see Janzing & Schölkopf, 2010, and references therein). The Kolmogorov complexity can be leveraged to for causal discovery through the following theorem:

Theorem 1 (Mooij et al. 2010)**.** *If Postulate 1 holds and X causes Y, then $K(F_X)+K(F_{Y|X}) \leq K(F_Y)+K(F_{X|Y})$.*

Stated differently, the causal direction between X and Y can be recognized as the least complex, that is the decomposition of the joint distribution leading to the lowest value of the Kolmogorov complexity. This is because, in this context, the direct translation of Postulate 1 is that the mutual information between F_X and $F_{Y|X}$ equals zero while that between F_Y and $F_{X|Y}$ does not. In other words, the number of bits saved when compressing X and Y jointly rather than compressing them independently is smaller when using the causal factorization of the joint distribution. Importantly, because we assumed the existence of a causal link, the asymmetry in Theorem 1 is not only necessary but also sufficient.

Remark 1. *Theorem 1 holds up to an additive constant, as the asymetry does not depend on the strings involved, but may depend on the Turing machines they refer to (see e.g., Janzing & Schölkopf, 2010; Hernández-Orallo & Dowe, 2010). When using the same Turing machine to compute all complexities, carrying this constant over is not required.*

Since the Kolmogorov complexity is not computable, we use the MDL principle (Rissanen, 1978) as a proxy: the less complex direction becomes the one allowing a better compression of the data. In other words, the correct causal direction allows to store information using the shortest description length, or code length (CL).

Two-step MDL encoding To construct a coding scheme satisfying the MDL principle, we use a two-stage approach (see e.g., Hansen & Yu, 2001, Section 3.1): first encode a model and then the data using that model. Assume that the goal is to compress $X = \{X_i\}_{i=1}^n$ with $X_i \in \mathcal{X}$ i.i.d. according to some distribution $F_X \in \mathcal{F}$, where the model class is known to be $\mathcal{F} = \{f_\theta(\cdot) \mid \theta \in \Theta\}$ and θ is an indexing parameter. When θ is known, Shanon's source coding theorem implies that an encoding based on f_θ is "optimal": on average, it achieves the lower bound on any lossless compression, that is the differential entropy $-n\int_{\mathcal{X}} f_\theta(z)\log f_\theta(z)dz$. And the code length of a dataset encoded using the known f_θ is given by minus its log-likelihood $\mathrm{CL}(X \mid \theta) = -\sum_{i=1}^n \log f_\theta(X_i)$. This optimal scheme thus amounts at transmitting the true parameter along with the data encoded using its known distribution:

$$\mathrm{CL}_\theta(X) = \mathrm{CL}(\theta) + \mathrm{CL}(X \mid \theta), \tag{1}$$

where the two terms on the right-hand side result from transmitting the model and the data encoded using the model (see e.g., Hansen & Yu, 2001, Section 3.1).

Encoding marginals via quantiles Since knowledge of the model class and true parameter is seldom achieved, consider compression using partial information about the data-generating process. Assuming that we only know a τ-quantile $q_{X,\tau} = \arg\inf_q \{q \mid F_X(q) = \tau\}$, we can compress the data by transmitting τ, $q_{X,\tau}$, and the "residuals" $E_X = \{X_i - q_{X,\tau}\}_{i=1}^n$. This encoding results in

$$\mathrm{CL}_\tau(X) = \mathrm{CL}(\tau) + \mathrm{CL}(q_{X,\tau}) + \mathrm{CL}(E_X \mid q_{X,\tau}, \tau). \tag{2}$$

To encode E_X, it is then natural to use the asymmetric Laplace (AL) distribution (see Aue et al., 2014; Geraci & Bottai, 2007; Yu et al., 2003) with density $f(z; q, \tau) = \tau(1-\tau)\exp(-S_\tau(q, z))$, where $S_\tau(\cdot,\cdot)$ is the quantile scoring (QS) function

$$S_\tau(x_1, x_2) = (\mathbb{I}\{x_1 \geq x_2\} - \tau)(x_1 - x_2).$$

Given that the distribution of E_X is generally unknown, using the AL is optimal in the sense that the population quantile is exactly the minimizer of the QS's expected value, namely $q_{X,\tau} = \mathrm{argmin}_q \mathbb{E}\left[S_\tau(q, X)\right]$. We revisit the link between QS and the AL in Section 2.2.

Using this encoding, the last term in the right-hand side of (2) is finally given by the negative of the AL log-likelihood (Rissanen, 1986), that is

$$\mathrm{CL}(E_X \mid q_{X,\tau}, \tau) = \sum_{i=1}^n S_\tau(q_{X,\tau}, X_i) - a_n(\tau), \tag{3}$$

where $a_n(\tau) = n\log(\tau(1-\tau))$.

Encoding conditionals via quantiles Next, consider the problem of compressing $\{X_i, Y_i\}_{i=1}^n$ with (X_i, Y_i) i.i.d. according to some joint distribution F. Because F can be decomposed using the marginal and the conditional distributions, one can proceed as above with F_X to compress X, and then use $F_{Y|X}$ to compress Y given X. Assume similarly that the only information about the conditional data-generating process is the conditional τ-quantile $q_{Y|X=x,\tau} = \arg\inf_q \left\{q \mid F_{Y|X=x}(q) = \tau\right\}$ of Y given $X = x$. Encoding a conditional distribution using (2) and (3) thus results in

$$\begin{aligned}\mathrm{CL}_\tau(Y \mid X) = \mathrm{CL}(\tau) + \mathrm{CL}(q_{Y|X,\tau}) \\ + \sum_{i=1}^n S_\tau(q_{Y|X=X_i,\tau}, Y_i) - a_n(\tau).\end{aligned} \tag{4}$$

In other words, the data is compressed by transmitting τ, $q_{Y|X,\tau}$, and the "residuals" $E_{Y|X} = \{Y_i - q_{Y|X=X_i,\tau}\}_{i=1}^n$. Note that, for a given fixed quantile level τ, while $q_{X,\tau}$ is a real number, $q_{Y|X,\tau}$ is generally a function of the conditioning variable.

Causal identification via quantiles The same idea can be applied to compress the data with the decomposition of the joint distribution F using the marginal F_Y and the conditional $F_{X|Y}$ distributions. But according to Theorem 1 and using CLs as proxies for Kolmogorov complexities, if X is

a cause of Y, one expects that for all $\tau \in (0,1)$

$$\mathrm{CL}_\tau(X) + \mathrm{CL}_\tau(Y \mid X) \leq \mathrm{CL}_\tau(Y) + \mathrm{CL}_\tau(X \mid Y)$$

with high probability as the sample size increases. We thus make the following "identifying" assumption for X to be a cause of Y:

Assumption 1. $P\left(\frac{CL_\tau(X)+CL_\tau(Y|X)}{CL_\tau(Y)+CL_\tau(X|Y)} \leq 1\right) \underset{n\to\infty}{\longrightarrow} 1$

This assumption is called identifying because, if the ratio of CLs equals 1 for all $\tau \in (0,1)$, quantile-based CLs cannot be leveraged for causal discovery. For any given sample size and quantile level τ, cases where $\mathrm{CL}(\tau) = \infty$, $\mathrm{CL}(q_{X,\tau}) = \infty$ or $\mathrm{CL}(q_{Y|X,\tau}) = \infty$ are problematic. The issue can be resolved by assuming for instance that all population quantities are computable and can be transmitted using a finite albeit potentially increasing precision.

Using O/o for the usual asymptotic notations, we draw a first link between CLs and QSs at the population level.

Theorem 2. *Assume that $CL(l) = o(n)$ for $l \in \{\tau, q_{X,\tau}, q_{Y,\tau}, q_{Y|X,\tau}, q_{X|Y,\tau})\}$, then Assumption 1 holds if and only if*

$$\frac{\mathbb{E}\left[S_\tau(q_{X,\tau}, X_i)\right] + \mathbb{E}\left[S_\tau(q_{Y|X=X_i,\tau}, Y_i)\right]}{\mathbb{E}\left[S_\tau(q_{Y,\tau}, Y_i)\right] + \mathbb{E}\left[S_\tau(q_{X|Y=Y_i,\tau}, X_i)\right]} \leq 1.$$

Because the population (conditional) quantiles and the expectations involved in the ratio are seldom known in practice, Theorem 2 cannot be leveraged directly for causal discovery. However, it can be used to determine whether a given causal model satisfies Assumption 1, as exmplified below. The proof can be found in Section A.1.

Example 1. *Consider X being a cause of Y in the linear model defined by $X = \theta_1 N_X$ and $Y = \gamma X + \theta_2 N_Y$ with N_X and N_Y two independent sources of noise and $\theta_1, \theta_2, \gamma > 0$ such that var(X) = var(Y). Note that the condition on the variances simply ensures that the variables have the same scale.*

Letting $N_X, N_Y \sim N(0,1)$, $X \sim N(0, \theta_1^2)$ and $Y \sim N(0, \gamma^2\theta_1^2 + \theta_2^2)$. The variance equality condition can be satisfied by choosing $\gamma \in (0,1)$ and letting $\theta_2 = \theta_1\sqrt{1-\gamma^2}$, resulting in $X \mid Y = y \sim N(\gamma y, (1-\gamma^2)\theta_1^2)$ and $Y \mid X = x \sim N(\gamma x, (1-\gamma^2)\theta_1^2)$. Using the fact that, if $Z \sim N(\mu, \sigma)$, then $\mathbb{E}\left[S_\tau(q_{Z,\tau}, Z)\right] = \sigma c(\tau)$, where $c(\tau) = e^{-\Phi^{-1}(\tau)^2/2}/\sqrt{2\pi}$ with Φ the standard normal cumulative distribution, it is then straightforward to verify that the ratio of expectations in Theorem 2 is equal to one independently of τ. In other words, the linear Gaussian model is not identifiable (Shimizu et al., 2006; Hoyer et al., 2009; Peters et al., 2014).

Causal discovery via quantiles In order to leverage Assumption 1 for causal discovery, we let $\widehat{q}_{X,\tau}$, $\widehat{q}_{Y,\tau}$, $\widehat{q}_{X|Y,\tau}$, $\widehat{q}_{Y|X,\tau}$ be estimators of the respective population quantiles, and further make the following assumption:

Assumption 2. $\widehat{q}_{X,\tau}$, $\widehat{q}_{Y,\tau}$, $\widehat{q}_{X|Y,\tau}$, $\widehat{q}_{Y|X,\tau}$ *satisfy*

- $|\widehat{q}_{X,\tau} - q_{X,\tau}| = o_p(1)$ *and* $|\widehat{q}_{Y,\tau} - q_{Y,\tau}| = o_p(1)$,
- $\left|\widehat{q}_{Y|X=x,\tau} - q_{Y|X=x,\tau}\right| = o_p(1)$ *for every x and* $\left|\widehat{q}_{X|Y=y,\tau} - q_{X|Y=y,\tau}\right| = o_p(1)$ *for every y,*
- *$CL(\widehat{q}_{X,\tau}) = o(n)$, $CL(\widehat{q}_{Y|X,\tau}) = o(n)$, $CL(\widehat{q}_{Y,\tau}) = o(n)$, and $CL(\widehat{q}_{X|Y,\tau}) = o(n)$,*

using O_p/o_p for stochastic boundedness and convergence in probability.

The first two bullet points simply state that the unconditional and conditional quantile estimators are consistent without rate. As for the third bullet point, note that having the same growth rate for the CL of all models does not prevent a smaller number of parameters in the correct causal direction (see e.g., Blöbaum et al., 2018; Marx & Vreeken, 2019). However, it means that, if the population quantiles $q_{\cdot,\tau}$ are replaced by estimators $\widehat{q}_{\cdot,\tau}$ in (2) and (4), the CLs of the models are all asymptotically dominated by the CLs of the residuals, namely the $\sum_{i=1}^n S_\tau(\cdot,\cdot)$ terms.

The third bullet point of Assumption 2 includes the important case where all CLs are $O(\log n)$ (Rissanen, 1983; 1986): discretizing a compact parameter space with a $n^{-1/2}$ grid (i.e., the magnitude of the estimation error) and transmitting an estimated parameter using a uniform encoder with this precision is optimal for regular parametric families. Using this $n^{-1/2}$ precision, each parameter thus leads to a cost of $1/2 \log n$.

For nonparametric (i.e., with a non-Euclidean parameter space) models, a similar idea can be applied, as estimators typically converge at a rate slower than $n^{-1/2}$. For instance, it is well known that the optimal grid size for histogram estimators of continuous densities with bounded first derivative is proportional to $n^{-1/3}$ (Theorem 6.11, Wasserman, 2006). And if each of the histogram heights is encoded using $n^{-1/2}$ grid (i.e., smaller than the estimation error), the resulting estimator's CL is proportional to $n^{1/3}/2 \log n$, which satisfies the third bullet point of Assumption 2. More generally, the fastest possible rate for kernel density estimators of continuous densities with k bounded derivatives is $n^{k/(2k+1)}$ (Theorem 6.31, Wasserman, 2006). As a result, discretizing the support on a grid proportional to $n^{-k/(2k+1)}$ and encoding the kernel values using a $n^{-1/2}$ grid encures an estimator's CL proportional to $n^{k/(2k+1)}/2 \log n$.

We can then state the following theorem:

Theorem 3. *Under Assumption 2, Assumption 1 holds if and only if*

$$P\left(\frac{S_{X,\tau} + S_{Y|X,\tau}}{S_{Y,\tau} + S_{X|Y,\tau}} \leq 1\right) \underset{n\to\infty}{\longrightarrow} 1,$$

with the scores $S_{X,\tau} = \sum_{i=1}^n S_\tau(\widehat{q}_{X,\tau}, X_i)$, $S_{Y|X,\tau} = \sum_{i=1}^n S_\tau(\widehat{q}_{Y|X=X_i,\tau}, Y_i)$ and similarly for Y and $X \mid Y$.

Note that Theorem 3 does not state that CLs and QSs are equivalent, but rather that inequalities in CLs imply inequalities in QSs with respect to a specific statistical model, and conversely. In other words, Theorem 3 implies that there is an equivalence between minimizing code length and quantile score. Hence, because of the MDL principle, the causal direction can be inferred from the lowest quantile score. The proof can be found in the supplementary material (Section A.2), but the intuition is as follows. Because of the third

bullet point in Assumption 2, QSs, namely the CLs of the residuals, asymptotically dominate the CLs of the models. As a result, using QSs corresponding to consistent models is sufficient for causal discovery.

Thanks to the stability (or invariance) of the true causal model, we expect Assumption 1 to hold over different quantile levels. However, since a single quantile is generally not enough to characterize a distribution, we further consider

$$\widehat{S}_X = \int_{[0,1]} S_{X,\tau} d\tau, \tag{5}$$

and similarly for $X \mid Y$, Y and $X|Y$. By pooling results at different quantile levels, we aim at better describing the marginal and conditional distributions. Arguing that estimating high and low (conditional) quantiles is hard, we could use only quantiles close to the median, that is integrating between over $[0.4, 0.6]$ instead of $[0, 1]$). We empirically found that this was more error prone when the generative models have asymmetries or multiplicative noises. Finally, we use averaging through integration rather than the maximal QS difference over quantile levels because the scale of QS is not uniform (e.g., the closer to 0.5 the higher). Hence, using maximization would essentially mean basing the decision on the median only, whereas properly capturing the spread of the data is also important.

Decision rule 1 (Quantile Causal Discovery)**.** *Let $S_{X\to Y} = \widehat{S}_X + \widehat{S}_{Y|X}$ and $S_{Y\to X} = \widehat{S}_Y + \widehat{S}_{X|Y}$ If $S_{X\to Y} < S_{Y\to X}$, conclude that X causes Y. If $S_{X\to Y} > S_{Y\to X}$, conclude that Y causes X. Otherwise, do not decide.*

2.2 Intuition

Consider mean regression from the point of view of the MDL principle: one seeks a model that allows the best compression of the data, measured in code length (CL). Using a two-part scheme as in Equation (1), the total CL is the sum of the conditional mean's CL and that of the residuals. In the context of least-squares regression, encoding the residuals using the normal distribution is then natural. The reason is that minimizing the residual sum of squares is equivalent to maximizing the Gaussian log-likelihood: given that the CL of the residuals is given by the negative log-likelihood of the distribution used to encode them, such a scheme is thus optimal.

A similar reasoning can be applied to quantile regression, where encoding the residuals using an asymmetric Laplace (AL) distribution (Koenker & Machado, 1999) is optimal. The reason is that the quantile score (QS) is related to the AL likelihood, which is equals $a_n(\tau) - QS$. Hence, by minimizing the QS, quantile regression also maximizes the AL likelihood. Intuitively, the likelihood corresponding to (conditional) residuals in the causal direction is higher, that is the QS and CL are smaller: the shortest CL corresponds to the largest AL likelihood/smallest QS, which establishes a link between minimizing QS and the MDL principle.

Let us now illustrate the key ideas in the theoretical justification of QCD. In Figure 3, we revisit the toy examples

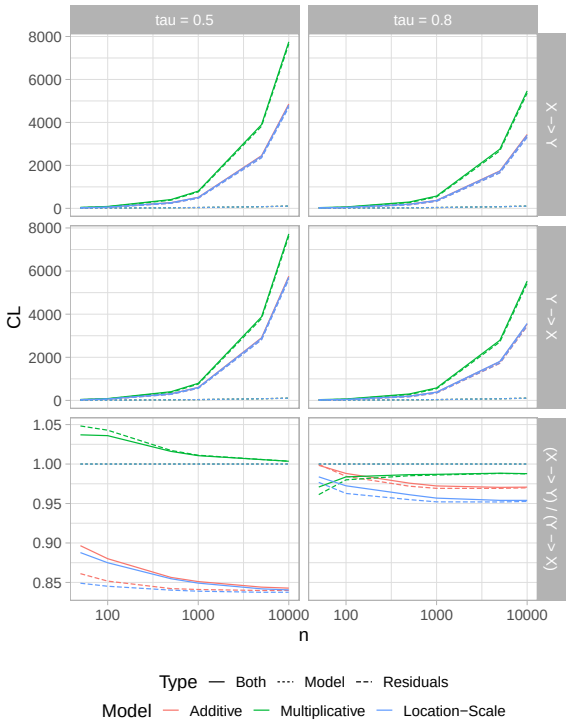

Figure 3. Verifying Assumption 1 for the setups of Figure 1. The CLs of the residuals asymptotically dominate the CLs of the marginal and conditional models. The multiplicative model is not identifiable using the median only. The correct causal direction has a lower CL.

from Figure 1, namely, the different setups: additive, multiplicative and location-scale causal pairs. We disentangle the CL's different components for both the causal $X \to Y$ and anti-causal $Y \to X$ direction, for two quantile levels $\tau \in \{0.5, 0.8\}$ and per generative model.

Curves are obtained by averaging over 100 repetitions.

As described in Section 2.1, causal discovery by MDL involves four summands for each possible direction. For instance, for $X \to Y$, we have

- two for the marginal and conditional model parameters, $\mathrm{CL}(\widehat{q}_{X,\tau})$ and $\mathrm{CL}_\tau(\widehat{q}_{Y|X,\tau})$,
- and two for the marginal and conditional residuals, $\mathrm{CL}(E_X \mid \widehat{q}_{X,\tau}, \tau)$ and $\mathrm{CL}(E_{X|Y} \mid \widehat{q}_{X|Y,\tau}, \tau)$.

In Figure 3, we use *Model* for the former and *Residuals* for latter and sum them by type. The sum of both model and residuals CLs is given by full lines. Unconditional and conditional quantiles are estimated respectively using the empirical quantile and kernel quantile regression. A uniform encoding is assumed in both cases, resulting in respective model CLs of $1/2\log(n)$ and $n^{1/3}/2\log(n)$.

In the top four panels of Figure 3, it is clear that the CLs of the residuals (i.e., the QSs) dominate the CLs of the marginal and conditional models. This illustrates the third bullet point in Assumption 2.

In the bottom two panels of Figure 3, we see that the ratio

of causal to anti-causal CLs generally[1] converges to some constant smaller than 1, which illustrates Assumption 1. It also means that Decision rule 1 would lead to the correct causal direction. The multiplicative noise mechanism has an interesting behaviour: considering the median only (i.e., $\tau = 0.5$), the ratio converges slowly to one, and therefore, the causal direction is not identifiable. However given that the ratio converges to some constant small than one for $\tau = 0.8$, pooling the decision over multiple quantiles as in Decision rule 1 would again lead to the correct causal direction.

3 Experiments

3.1 QCD implementation

Quantile regression Theorem 3 holds provided that the model is consistent and its complexity does not grow too fast. As such, Decision rule 1 can be implemented using any quantile regression approach that satisfies Assumption 2. In our experiments, we use three methods, namely nonparametric copulas (Geenens et al., 2017), quantile forests (Meinshausen, 2006), quantile neural networks (Cannon, 2018), and show that they yield qualitatively and quantitatively similar results. We refer to Appendix B for more details on the regression methods and their specific implementations. For estimating the overall quantile scores (aggregating multiple quantile levels), we use Legendre quadrature to approximate the integral over $[0, 1]$, as it is fast and precise for univariate functions. In other words, denoting by $\{w_j, \tau_j\}_{j=1}^m$ the m pairs of quadrature weights and nodes, we use $\int_0^1 g(\tau)d\tau \approx \sum_{j=1}^m w_j g(\tau_j)$, which when plugged into (2.1) yields $\widehat{S}_X = \sum_{j=1}^m w_j \widehat{S}_X(\tau_j)$, $\widehat{S}_Y = \sum_{j=1}^m w_j \widehat{S}_Y(\tau_j)$, $\widehat{S}_{X|Y} = \sum_{j=1}^m w_j \widehat{S}_{X|Y}(\tau_j)$, and $\widehat{S}_{Y|X} = \sum_{j=1}^m w_j \widehat{S}_{Y|X}(\tau_j)$. Summing over an equally spaced grid with uniform weights or using quadrature nodes and weights yields two valid approximations of an integral, and using one or the other should not matter. But the quadrature gives more importance to the center of the distribution (i.e., quantiles closer to 0.5 have a higher weight). Note that, to compute scores free of scale bias, the variables are transformed to the standard normal scale.

Computational complexity QCD scales linearly with the size of input data, $O(n)$ for copulas, $O(n_{tree} \times n_{attr} \times depth \times n)$ for random forests and $O(epochs \times n_{weights} \times n)$ corresponding to the choice of regressor. However, using multiple quantile levels does not increase the complexity significantly since all of the suggested implementations allow for simultaneous conditional quantile estimation, that is, we estimate a single model at each possible causal direction [2], from which we compute all of the requested quantiles. Roughly speaking, since $n >> m$, the overall complexity scales with $\mathcal{O}(n)$. As such, QCD compares favorably to nonparametric methods relying on computationally intensive procedures, for an instance based on kernels (Chen et al., 2014; Hernández-Lobato et al., 2016) or Gaussian processes (Hoyer et al., 2009; Mooij et al., 2010; Sgouritsa et al., 2015).

The parameter m can be used to control for the trade-off between the computational complexity and the precision of the estimation. We recommend the value $m = 3$ which, makes it possible to capture variability in both location and scale. Setting $m = 1$ is essentially equivalent to using only the conditional median for causal discovery, a setting that suitable for distributions with constant variance. An empirical analysis of the choice of m is provided in the following section. In what follows, we report results for QCD with $m = 3$ if not stated otherwise.

3.2 Datasets, baselines and metrics

Benchmarks For simulated data, we first rely on the following scenarios (Mooij et al., 2016): *SIM* (without confounder), *SIM-ln* (with low noise), *SIM-G* (with distributions close to Gaussian), and *SIM-c* (with latent confounder). There are 100 pairs of size $n = 1000$ in each of these datasets.

As a second benchmark, inspired by (Peters et al., 2014), we generate a diverse dataset of additive, location-scale and multiplicative causal pairs. We include nonlinear additive noise (*AN*) models of the form $Y = f(X) + E_Y$ for some deterministic function f with $E_Y \sim \mathcal{N}(0, \sigma)$, $X \sim \mathcal{N}(0, \sqrt{2})$, and $\sigma \sim \mathcal{U}[1/5, \sqrt{2/5}]$. In *AN*, f is an arbitrary nonlinear function simulated using Gaussian processes (GP, Rasmussen & Williams, 2006) with a Gaussian kernel of bandwidth one. Since the functions in *AN* are often non-injective, we include *AN-s* to explore the behavior of QCD in injective cases. In this setup, f are sigmoids as in (Bühlmann et al., 2014). The third experiment considers location-scale (*LS*) data generating processes with both the mean and variance of the effect being functions of the cause, that is $Y = f(X) + g(X)E_Y$, and E_Y and X are similar as for the additive noise models. *LS* and *LS-s* then correspond to the Gaussian processes and sigmoids described for *AN* and *AN-s*. Finally, the fourth experiment considers multiplicative models (*MN*) as $Y = f(X)E_Y$, with $f(X)$ sampled as sigmoid functions and $E_Y \sim \mathcal{U}(0, 1)$. In each of the second, third, and fourth experiments, we simulate 100 pairs of size $n = 1000$. All pairs have equal weights with variable ordering according to a coin flip, therefore resulting in balanced datasets. Example datasets for each of the simulated experiments are shown in the supplementary material (Appendix F).

For real data, we use the Tübingen CE benchmark (version Dec 2017), consisting of 108 pairs from 37 different domains, from which we consider only the 99 pairs that have univariate continuous or discrete cause and effect vari-

[1] Because we assumed the same uniform encoding for both directions, the ratio of CLs of models is always equal to one.

[2] Except for QCD, where the nature of copula models allows for joint estimation.

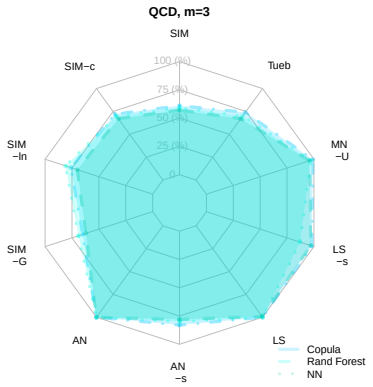

Figure 4. Quantile Causal Discovery achieves consistent results across all benchmarks regardless of the practical implementation.

ables. When evaluating the performance on this dataset we included the pairs' corresponding weights which accounts for potential bias in cases where pairs were selected from same multivariable dataset.

Baselines On simulated data, we compare QCD to state-of the-art approaches, namely RESIT (Peters et al., 2014), biCAM (Bühlmann et al., 2014), LinGaM (Shimizu et al., 2006), and GR-AN (Hernández-Lobato et al., 2016), which are ANM-based, and IGCI (Janzing & Schölkopf, 2010), EMD (Chen et al., 2014), RECI (Blöbaum et al., 2018), Slope (Marx & Vreeken, 2017) and Sloppy (Marx & Vreeken, 2019), based on the independence postulate.

We also consider other methods such as PNL-MLP (Zhang & Hyvärinen, 2009), GPI (Mooij et al., 2010), ANM (Hoyer et al., 2009), and CURE (Sgouritsa et al., 2015). For the real data benchmark, GR-AN was evaluated over 15 subsamples limited to 500 observations, and the results are then averaged. Implementation details and hyper parameters for all baselines are described in the supplementary material (Section C.1). Our code and datasets are available in the submitted supplementary package and http://coolplace.org.

Evaluation metrics As (Mooij et al., 2016), we use the *accuracy* for *forced decisions* and the *area under the receiver operating curve* (ROC) for *ranked decisions*. The former corresponds to forcing the compared methods to decide the causal direction. The later corresponds to using heuristic scores allowing to rank confidence in the directions along with ROC/AUC as performance measure. As a confidence heuristic for the ranked decisions for QCD, we use same score as (24) in (Mooij et al., 2016), that is $\hat{C} = -S_{X\to Y} + S_{Y\to X}$.

3.3 Results and discussion

QCD robustness wrt to its implementation From Figure 4 it is clear that the choice of implementation in estimating the conditional quantiles has no significant impact and QCD provides consistent results across all benchmarks. Further results that confirm the robustness to implementation for different values of m are shown in Appendix B, Figure 8 . In the remaining of the paper we will show the copula-based results.

Selection of m In an ablation study we explored the significance of the parameter m with regards to different sample sizes. Looking at the figures in Figure 5, we can clearly see in which cases multiple quantile levels can indeed increase the accuracy, namely for sigmoid (i.e., harder to detect) causal mechanisms. For confounded data, increasing m seems to help too, albeit faintlier. Additionally, we can notice that higher values of m have more pronounced effect as the sample size increases. This was used to improve our results on real data, namely by setting the parameter m to 1 when $n < 200$ and $m = 3$ for the rest.

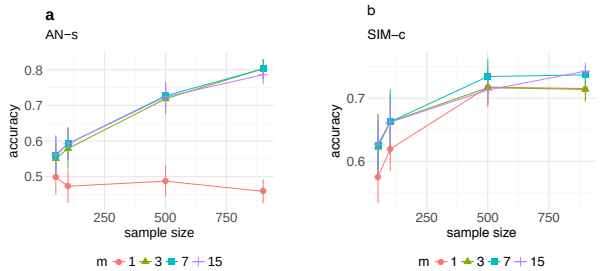

Figure 5. The accuracy increases with m and the sample size.

Comparison to baselines In Figure 6, we compare causal discovery algorithms across simulated datasets with regards to accuracy. Tabulated numbers are in Section D.1.

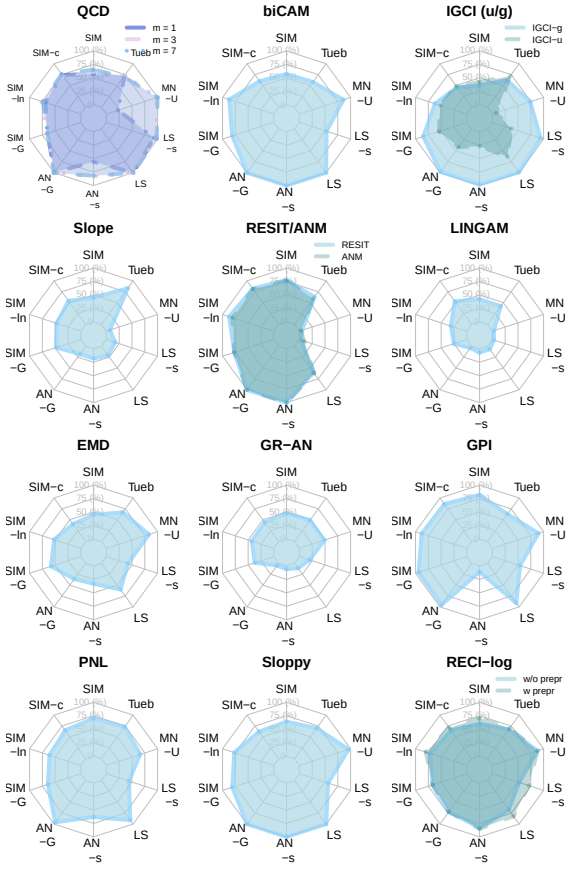

Figure 6. Accuracy of QCD and competitors.

There is no single baseline is an overall best performer, however, we notice that QCD has most consistent results across

all benchmarks. Starting with the SIM benchmarks, we notice that GPI achieves highest accuracy in all four scenarios, followed with similar results by RESIT/ANM [3]. On this benchmark, QCD behaves similarly to the rest of the baselines, while being more robust in the confounded scenario where others achieve results on the scale of random guess. Interestingly, higher values of the parameter m improve the results in such pairs.

The results are significantly different for the AN, LS and MN scenarios. biCAM, ANM and RESIT easily handle the AN pairs since their underlying assumptions are met, while we can notice some discrepancy in the LS and MN scenarios where there is an interaction between the noise and the cause. Similarly, LINGAM does not perform well on any of the datasets, which are all highly nonlinear, hence violating its assumptions. IGCI can handle any scenario with the gaussian reference measure, while this is not the case with the uniform measure[4]. In the LS generative models where not only the mean, but the variance of the effect changes with the cause only IGCI-g was on-par with QCD, but QCD is still better than IGCI-g on the SIM benchmark and real data pairs. On the other hand, more flexible methods such as PNL and EMD had difficulties in the non-injective cases. QCD has satisfactory ($> 75\%$ accuracy) for all different data generative mechanisms (AN, LS and MN). From the baselines presented, RECI's model allows for dependence between the noise and the cause. Although it shows good results in practice, it is important to note that the outcome depends significantly on the preprocessing step as well as the selection of the regressor, for which still, there is no clear guidance. For mote details on the variations of the proposed solution, we refer to the original paper (Blöbaum et al., 2018). On the contrary, QCD achieves the same results no matter the implementation, which makes it straightforward to use.

With real data pairs, Table 1 shows that QCD[5] is highly competitive in terms of weighted accuracy, with only Slope achieving better overall results. However, note that Slope does well only on this dataset and performs poorly on synthetic benchmarks, while QCD performs well under diverse setups. Additionally, we include accuracy decision-rate plot in Figure 7. Note that QCD provides statistically significant results (i.e., compared to a coin flip) and is second out of the 6 best performing algorithms with respect to weighted accuracy. In Appendix D, Figure 10, we further provide accuracy decision rate plot and ROC curves with all baselines. Moreover, the efficiency of our method is highlighted

[3]Because RESIT is an R version based on the MATLAB ANM, we overlay their results on a single chart.

[4]Selecting the reference measure is the method's most sensitive part and is as difficult as selecting the right kernel/bandwidth for a specific task (Janzing et al., 2012; Mooij et al., 2016).

[5]Results are averaged over 30 repetitions to account for the effect of the jittering in the discrete pairs.

Table 1. Results for the Tübingen Benchmark.

	QCD	IGCI-u/g	biCAM	Slope	LINGAM	RESIT	Sloppy
Acc	0.68	0.67/0.61	0.57	0.75	0.3	0.53	0.59
Weig Acc	0.75	0.72/0.62	0.58	0.83	0.36	0.63	0.7
AUC ROC	0.71	0.67	0.61	0.84	0.3	0.56	0.67
CPU	7 min.	2 sec.	10 sec.	25 min.	3.5 sec.	12 h	1.3 min
	EMD	**GRAN**	**GPI**	**PNL-MLP**	**ANM**	**CURE**	**RECI**
Acc	0.55	0.4	0.6	0.75	0.6	0.6	0.63
Weig Acc	0.6	0.5	0.63	0.73	0.6	0.54	0.7
AUC ROC	0.53	0.47	0.61	0.7	0.45	0.61	0.68
CPU	4.6 days	NA	30 days	8.3 h	3.2 days	NA	1.2h

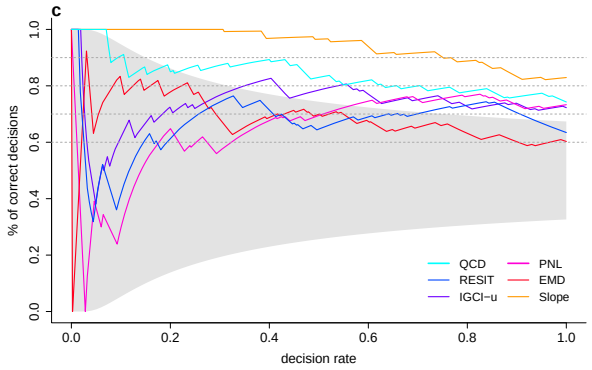

Figure 7. Accuracy decision-rates for the top baselines on the Tübingen Benchmark.

in the last row of Table 1, where QCD is able to go over the whole dataset in ~ 7 minutes. As for other nonparametric methods, only IGCI and Sloppy are faster, Slope is twice as slow, RESIT 55 times, PNL 71 times, and the others required days to go through the whole dataset of had to be averaged on subsamples due to slow execution (GRAN).

Overall, we can conclude that compared to baselines QCD performs well in both real and simulated scenarios therefore being more robust to different generative models while also having computational advantages.

4 Conclusion

In this work, we develop a causal discovery method based on conditional quantiles. We propose QCD, an effective implementation of our method based on nonparametric quantile regression, compares favorably to state-of-the-art methods on both simulated and real datasets. The new method should be preferred mainly because of three reasons. First, quantiles are less sensitive to outliers than the mean, thus QCD is robust to contamination or heavy-tails, hence, it does not require preprocesing ("cleaning") step. Second, we make no assumption on the parametric class considered, thus allowing for a wide range of mechanisms, whereas baselines perform worse when their assumptions are not satisfied. Third, our implementation is computationally more efficient than other nonparametric methods. There are currently three directions that we are exploring to extend this work. First, QCD can be extended to multivariate cause-effect pairs. Second, it is interesting to see how quantile scores can fit in the invariant causal prediction framework (Peters et al., 2016). Third, the computational efficiency of QCD is promising in the context of extensions to higher dimensional datasets. As such, ongoing research leverages existing graph discovery algorithms for hybrid learning, as suggested in the supplementary material.

References

An Introduction to Causal Inference. *The International Journal of Biostatistics*, 2010.

Graph estimation with joint additive models. *Biometrika*, 2014.

Structural Intervention Distance for Evaluating Causal Graphs. *Neural Computation*, 27(3):771–799, mar 2015.

Aas, K., Czado, C., Frigessi, A., and Bakken, H. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics*, 44(2):182–198, 2009.

Aue, A., Cheung, R. C. Y., Lee, T. C. M., and Zhong, M. Segmented model selection in quantile regression using the minimum description length principle. *Journal of the American Statistical Association*, 109(109):507–1241, 2014.

Bauer, A. and Czado, C. Pair-copula Bayesian networks. *Journal of Computational and Graphical Statistics*, 25 (4):1248–1271, 2016a.

Bauer, A. and Czado, C. Pair-copula Bayesian networks. *Journal of Computational and Graphical Statistics*, 25 (4):1248–1271, 2016b.

Bauer, A., Czado, C., and Klein, T. Pair-copula constructions for non-Gaussian DAG models. *The Canadian Journal of Statistics*, 40(1):86–109, 2012.

Bedford, T., Cooke, R. M., et al. Vines–a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031–1068, 2002.

Besag, J. Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

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, 2018.

Breiman, L. Random forests. *Machine learning*, 45(1): 5–32, 2001.

Budhathoki, K. and Vreeken, J. MDL for Causal Inference on Discrete Data.

Budhathoki, K. and Vreeken, J. Causal inference by compression. In *ICDM*, pp. 41–50, 2017.

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

Cannon, A. J. Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic Environmental Research and Risk Assessment*, pp. 3207–3225, 2018. ISSN 1436-3259.

Chalupka, K., Eberhardt, F., and Perona, P. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017.

Chang, Y., Li, Y., Ding, A., and Dy, J. A robust-equitable copula dependence measure for feature selection. *AISTATS*, 41:84–92, 2016.

Chen, Z., Zhang, K., Chan, L., and Schölkopf, B. Causal discovery via reproducing kernel hilbert space embeddings. *Neural Computation*, 26(7):1484–1517, 2014.

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

Cui, R., Groot, P., and Heskes, T. Copula PC algorithm for causal discovery from mixed data. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9852 LNAI, pp. 337–392. Springer, Cham, sep 2016.

Dawid, C. A. et al. Fundamentals of statistical causality. 2007.

Drton, M. and Maathuis, M. H. Structure Learning in Graphical Modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

Elidan, G. Copula Bayesian Networks. In *NIPS 23*, pp. 559–567, 2010.

Elidan, G. Copulas in machine learning. In *Copulae in mathematical and quantitative finance*, pp. 39–60. Springer, 2013.

Embrechts, P., Mcneil, E., and Straumann, D. Correlation: Pitfalls and alternatives. *Risk Magazine*, 1999.

Ernest, J. *Causal inference in semiparametric and nonparametric structural equation models ETH Library*. PhD thesis, 2016.

Fisher, R. A. Statistical Methods for Research Workers.

Fisher, R. A. Statistical methods for research workers. *Especially Section*, 21, 1936.

Flaxman, S. R., Neill, D. B., and Smola, A. J. Gaussian Processes for Independence Tests with Non-iid Data in Causal Inference. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):21–22, 2016.

Freedman, D. and Diaconis, P. On the histogram as a density estimator:l2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57(4):453–476, Dec 1981. ISSN 1432-2064. doi: 10.1007/BF01025868. URL https://doi.org/10.1007/BF01025868.

Geenens, G., Charpentier, A., and Paindaveine, D. Probit transformation for nonparametric kernel estimation of the copula density. *Bernoulli*, 23(3):1848–1873, 2017.

Geraci, M. and Bottai, M. Quantile regression for longitudinal data using the asymmetric laplace distribution. *Biostatistics*, 8(1):140–154, 2007.

Gneiting, T. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, jun 2011.

Goudet, O., Kalainathan, D., Caillou, P., Guyon, I., Lopez-Paz, D., and Sebag, M. Causal Generative Neural Networks. 2017a.

Goudet, O., Kalainathan, D., Caillou, P., Lopez-Paz, D., Guyon, I., Sebag, M., Tritas, A., and Tubaro, P. Learning functional causal models with generative neural networks. *preprint*, nov 2017b. arXiv: 1709.05321.

Guennebaud, G., Jacob, B., and Others. Eigen v3, 2010. URL http://eigen.tuxfamily.org.

Hall, P. and Hannan, E. J. On stochastic complexity and nonparametric density estimation. *Biometrika*, 75(4):705–714, 1988.

Hansen, M. H. and Yu, B. Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746–774, 2001.

Harrell Jr, F. E., with contributions from Charles Dupont, and others., M. Hmisc: Harrell Miscellaneous, 2017. URL https://cran.r-project.org/package=Hmisc.

Harris, N. and Drton, M. PC Algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14:3365–3383, 2013.

Heinze-Deml, C., Peters, J., and Meinshausen, N. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

Hernández-Lobato, D., Morales Mombiela, P., Lopez-Paz, D., and Suárez, A. Non-linear causal inference using Gaussianity measures. *Journal of Machine Learning Research*, 17(1):939–977, 2016.

Hernández-Orallo, J. and Dowe, D. L. Measuring universal intelligence: Towards an anytime intelligence test. *Artificial Intelligence*, 174(18):1508–1539, 2010.

Hoeffding, W. A Non-Parametric Test of Independence. *The Annals of Mathematical Statistics*, 19(4):546–557, 1948.

Hoyer, P. O., Janzing, D., Mooij, J., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *NIPS 22*, pp. 689–696, 2009.

Hyvärinen , A. and Smith, S. M. Pairwise Likelihood Ratios for Estimation of Non-Gaussian Structural Equation Models. *Journal of Machine Learning Research*, 14:111–152, 2013.

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

Janzing, D., Peters, J., Mooij, J., and Schölkopf, B. Identifying confounders using additive noise models.

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.

Joe, H. Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters. In *Distributions with fixed marginals and related topics*, pp. 120–141. Institute of Mathematical Statistics, 1996.

Karra, K. and Mili, L. Hybrid Copula Bayesian Networks. In *Conference on Probabilistic Graphical Models*, volume 52, pp. 240–251, 2016.

Koenker, R. and Machado, J. A. Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, 94(448):1296–1310, 1999.

Koenker, Roger . *Quantile regression*. Econometric Society monographs, 2005.

Kolmogorov, A. N. On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 369–376, 1963.

Kpotufe, S., Sgouritsa, E., Janzing, D., and Schölkopf, B. Consistency of causal inference under the additive noise model. In *ICML 31*, pp. 478–486, 2014.

Lichman, M. {UCI} Machine Learning Repository, 2013. URL http://archive.ics.uci.edu/ml.

Liu, F. and Chan, L.-W. Causal inference on multidimensional data using free probability theory. *IEEE Transactions on Neural Networks and Learning Systems*, 2017.

Liu, H., Lafferty, J., and Wasserman, L. The Nonparanormal: semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10:2295–2328, 2009.

Loader, C. *Local regression and likelihood*. Springer Science & Business Media, 2006.

Lopez-Paz, D. *From Dependence to Causation*. PhD thesis, University of Cambridge, 2016.

Lopez-Paz, D., Hernandez-Lobato, J. M., and Schölkopf, B. Semi-supervised domain adaptation with copulas. *NIPS 26*, pp. 674–682, 2013.

Lopez-Paz, D., Muandet, K., Schölkopf, B., and Tolstikhin, I. Towards a learning theory of cause-effect inference. In *ICML 32*, pp. 1452–1461, 2015.

Maathuis, M. H. and Nandy, P. A review of some recent advances in causal inference. In *Handbook of Big Data*. CRC Press, 2016.

Mandros, P., Boley, M., and Vreeken, J. Discovering reliable approximate functional dependencies. *KDD*, pp. 355–364, 2017.

Marx, A. and Vreeken, J. Telling cause from effect using MDL-based local and global regression. In *ICDM*, 2017.

Marx, A. and Vreeken, J. Identifiability of cause and effect using regularized regression. In *ACM SIGKDD*, 2019.

Meinshausen, N. Quantile regression forests. *JMLR*, 2006.

Mitrovic, J., Sejdinovic, D., and Teh, Y. W. Causal inference via kernel deviance measures. In *Advances in Neural Information Processing Systems*, pp. 6986–6994, 2018.

Mooij, J., Janzing, D., Peters, J., and Schölkopf, B. Regression by dependence minimization and its application to causal inference in additive noise models. In *ICML 26*, pp. 745–752, 2009.

Mooij, J. M., Stegle, O., Janzing, D., Zhang, K., and Schölkopf, B. Probabilistic latent variable models for distinguishing between cause and effect. In *NIPS 23*, pp. 1687–1695, 2010.

Mooij, J. M., Peters, J., Janzing, D., Zscheischler, J., and Schölkopf, B. Distinguishing cause from effect using observational data: methods and benchmarks. *Journal of Machine Learning Research*, 17:1–102, 2016.

Müller, D. and Czado, C. Selection of sparse vine copulas in high dimensions with the lasso. may 2017.

Nagler, T. and Vatter, T. vinecopulib: High Performance Algorithms for Vine Copula Modeling in C++. http://vinecopulib.org, 2017.

Nagler, T. and Vatter, T. rvinecopulib: high performance algorithms for vine copula modeling, 2018. URL https://cran.r-project.org/package=rvinecopulib.

Oates, C. J., Smith, J. Q., and Mukherjee, S. Estimating Causal Structure Using Conditional DAG Models. *Journal of Machine Learning Research*, 17:1–23, 2016a.

Oates, C. J., Smith, J. Q., and Mukherjee, S. Estimating Causal Structure Using Conditional DAG Models. *Journal of Machine Learning Research*, 17:1–23, 2016b.

Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Series in Representation and Reasoning, 1988.

Pearl, J. *Causality*. Cambridge University Press, 2009.

Pearl, J., Glymour, M., and Jewell, N. P. *Causal Inference in Statistics: A Primer*. John Wiley & Sons, 2016.

Peters, J. and Ernest, J. *CAM: Causal Additive Model (CAM)*, 2015. URL https://cran.r-project.org/package=CAM.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Identifiability of causal graphs using functional models. In *UAI 27*, pp. 589–598, 2011.

Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1):2009–2053, 2014.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press (available on-line), 2017.

Pircalabelu, E., Claeskens, G., and Gijbels, I. Copula directed acyclic graphs. *Statistics and Computing*, 27(1): 55–78, 2017.

R Core Team. R: A language and environment for statistical computing, 2017. URL https://www.r-project.org/.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*, volume 1. MIT Press Cambridge, 2006.

Rissanen, J. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Rissanen, J. A Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics*, 11(2):416–431, 1983.

Rissanen, J. Stochastic complexity and modeling. *The annals of statistics*, pp. 1080–1100, 1986.

Rissanen, J. Complexity and Information in Modeling. pp. 1–16, 2005.

Rissanen, J., Speed, T. P., and Yu, B. Density estimation by stochastic complexity. *IEEE Transactions on Information Theory*, 38(2):315–323, 1992.

Rojas-Carulla, M., Baroni, M., and Lopez-Paz, D. Causal Discovery Using Proxy Variables. 2017.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, 308 (5721):523–529, apr 2005.

Scaillet, O., Charpentier, A., and Fermanian, J.-D. The estimation of copulas: Theory and practice. Technical report, Ensae-Crest and Katholieke Universiteit Leuven, NP-Paribas and Crest; HEC Genve and Swiss Finance Institute, 2007.

Schäling, B. *The Boost C++ Libraries*. 2011.

Schölkopf, B. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. In *ICML 29*, pp. 1255–1262, 2012.

Schölkopf, B., Hogg, D. W., Wang, D., Foreman-Mackey, D., Janzing, D., Simon-Gabriel, C.-J., and Peters, J. Modeling confounding by half-sibling regression. *Proceedings of the National Academy of Sciences*, 2016.

Sgouritsa, E., Janzing, D., Hennig, P., and Schölkopf, B. Inference of cause and effect with unsupervised inverse regression. In *AISTATS 38*, pp. 847–855, 2015.

Shimizu, S., Hoyer, P., Hyvärinen , Aapo, and Antti, K. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

Sklar, A. Fonctions de répartition à n dimensions et leurs marges. *Publications de L'Institut de Statistique de L'Université de Paris*, 8:229–231, 1959.

Smyth, G. K. Numerical integration. *Encyclopedia of Biostatistics*, pp. 3088–3095, 2005.

Spirtes, P. and Glymour, C. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

Spirtes, P. and Zhang, K. Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, pp. 165–191, 2016.

Spirtes, P., Meek, C., and Richardson, T. Causal Inference in the Presence of Latent Variables and Selection Bias. In *UAI 11*, pp. 499—-506, 1995.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT press, 2000a.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT press, Causation2000, 2000b.

Tagasovska, N., Ackerer, D., and Vatter, T. Copulas as high-dimensional generative models: Vine copula autoencoders. In *Advances in Neural Information Processing Systems*, pp. 6525–6537, 2019.

Takeuchi, I., Bengio, Y., and Kanamori, T. Robust Regression with Asymmetric Heavy-Tail Noise Distributions. *Neural Computation*, 14(10):2469–2496, oct 2002.

Tran, D., Blei, D. M., and Airoldi, E. M. Copula variational inference. In *NIPS 28*, pp. 3564–3572, 2015.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, oct 2006.

Turing, A. M. On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, 2(1):230–265, 1937.

Turing, A. M. On computable numbers, with an application to the Entscheidungsproblem. A correction. *Proceedings of the London Mathematical Society*, 2(1):544–546, 1938.

Vatter, T. and Chavez-Demoulin, V. Generalized additive models for conditional dependence structures. *Journal of Multivariate Analysis*, 2015.

Vereshchagin, N. and Vitanyi, P. Kolmogorov's Structure Functions and Model Selection. *IEEE Transactions on Information Theory*, 50(12):3265–3290, dec 2004.

Vereshchagin, N. K. and Vitányi, P. M. B. Kolmogorov 's structure functions and model selection. *IEEE Transactions on Information Theory*, 50(12):3265–3290, 2004.

Vreeken, J. Causal Inference by Direction of Information. *SIAM*, 2015. URL http://eda.mmci.uni-saarland.de/pubs/2015/ergo-vreeken.pdf.

Wasserman, L. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

Wieczorek, A., Wieser, M., Murezzan, D., and Roth, V. Learning sparse latent representations with the deep copula information bottleneck. *ICLR*, 2018.

Yu, H. and Dauwels, J. Modeling Spatio-Temporal Extreme Events Using Graphical Models. *IEEE Transactions on Signal Processing*, 64(1), 2016.

Yu, K., Lu, Z., and Stander, J. Quantile regression: applications and current research areas. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3): 331–350, 2003.

Zhang, K. and Hyvärinen, A. On the identifiability of the post-nonlinear causal model. In *UAI 25*, pp. 647–655, 2009.

Zhang, K., Muandet, K., Wang, Z., and Others. Domain adaptation under target and conditional shift. *ICML 30*, pp. 819–827, 2013.

Zhang, K., Schölkopf, B., Spirtes, P., and Glymour, C. Special Topic: Machine Learning Learning Causality and Causality-Related Learning: Some Recent Progress Learning Causal Relations. *National Science Review*, pp. nwx137, 2017.

Zhang, Kun and Wang, Zhikun and Zhang, Jiji and Sch{\"o}lkopf, B. On Estimation of Functional Causal Models: General Results and Application to Post-Nonlinear Causal Model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):13, 2016.