Causal structure-based root cause analysis of outliers

Kailash Budhathoki [1] **Lenon Minorics** [1] **Patrick Blöbaum** [1] **Dominik Janzing** [1]

Abstract

Current techniques for explaining outliers cannot tell what *caused* the outliers. We present a formal method to identify "root causes" of outliers, amongst variables. The method requires a causal graph of the variables along with the functional causal model. It quantifies the contribution of each variable to the target outlier score, which explains to what extent each variable is a "root cause" of the target outlier. We study the empirical performance of the method through simulations and present a real-world case study identifying "root causes" of extreme river flows.

1. Introduction

Outlier detection has been studied extensively over the years (Aggarwal, 2013; Akoglu, 2021; Chandola et al., 2009; Blázquez-García et al., 2021). Besides the development in outlier detection, we have also made some progress in recent years towards explaining outliers (Knorr & Ng, 1999; Liu et al., 2018; Macha & Akoglu, 2018; Idé et al., 2021). When the purpose of explaining outliers is to take *actions* (e.g. fixing a cloud service that slowed down a website), explanations should have *causal* relations to the target outliers. Existing methods, however, do not provide *causal* explanations; they only describe observed correlations to target outliers. A formal way to define "root causes" of outliers seems to be missing.

Our problem setup is simple. We consider the scenario where the value x_n of a target variable X_n has been flagged as an outlier by an existing outlier detection algorithm. We jointly observed values $(x_1, \ldots, x_n) =: \mathbf{x}$ of variables $(X_1, \ldots, X_n) =: \mathbf{X}$. Our goal is to identify the "root causes" of the outlier x_n amongst variables $X_1, \ldots, X_n$.[1]

[1]Amazon Research Tübingen. Correspondence to: Kailash Budhathoki <kaibud@amazon.com>.

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

[1]We assume that variables $X_1, \ldots, X_n$ contain the root causes of outliers. If other variables explain the outliers, we need to include them in the analysis, search for root causes among them and their ancestors.

We make three key assumptions to this end.

A_1 The causal relationships between variables $X_1, \ldots, X_n$ is known in the form of a directed acyclic graph, also called causal graph (Pearl, 2009).

A_2 The causal graph comes with the functional causal model (FCM) (Pearl, 2009) that describes how each variable X_j is generated from its parents PA_j in the causal graph. In an FCM, each variable X_j is a function f_j of its parents PA_j in the causal graph and an unobserved noise term N_j, i.e.,

$$X_j := f_j(\mathrm{PA}_j, N_j), \tag{1}$$

where the noises $N_1, \ldots, N_n$ are statistically jointly independent (Pearl, 2009).

A_3 The FCM is invertible (Zhang et al., 2015). That is, we can recover the noise value n_j—of the noise term N_j—corresponding to the observed variable X_j from its observed value x_j and the values pa_j of its parents.

This paper presents a formal method to identify "root causes" of outlier x_n among variables $X_1, \ldots, X_n$ by quantifying the contribution of each noise N_j to the outlier score of x_n. This notion of contribution captures the contribution of the "causal mechanism" of event x_j (at node X_j) to the outlier score of x_n. We illustrate the outcome of the method on a hypothetical example in Figure 1. There are two key steps involved in formalising the method:

- As there exists a multitude of outlier detection algorithms (Aggarwal, 2013), our method should be applicable to most, if not all, of them. To this end, we introduce information-theoretic (IT) outlier scores, which probabilistically calibrate existing outlier scores (Section 2). The probabilistic calibration also renders IT outlier scores comparable across variables with different dimension, range, and scaling.
- We then present our method based on counterfactuals, which are the third rung on Pearl's ladder of causation (Pearl & Mackenzie, 2018) (Section 3). In particular,

(a) Service Dependency graph **(b)** Observed latencies

(c) Causal graph of services' latencies **(d)** Results of our method

Figure 1: (**a**) Dependencies between cloud services that empower a hypothetical retail website. $A \to B$ indicates A uses B to serve client requests. (**b**) For one client (marked by the dashed line), we observe an extremely high latency (i.e., time delay between request and response) in the web service. What *caused* the outlier? Other services also have high latencies for that client. High latencies in services upstream in the dependency graph can result from high latencies in services downstream or issues in the upstream services themselves. (**c**) By inverting the dependency graph, we obtain the causal graph of latencies of services. From training samples of observed latencies, we estimate the associated functional causal models (FCMs). (**d**) Our method uses the FCMs to identify the web service itself and the database service as potential *causes* of the extremely high latency in the web service as their contributions to the outlier are high. Experts can use this information to diagnose issues in web service and database service for that client.

> we measure the contribution of each noise term N_j to the IT score of x_n in terms of logarithmic decrease of the likelihood of x_n had the causal mechanisms at X_j been "normal" (Section 3.2). The contributions are symmetrized using the concept of Shapley values (Shapley, 1953) from game theory (Theorem 3.1).

We compare and contrast related work in Section 4. In the experiments (Section 5), we first study the performance of our method on synthetic data. Then we present a case study on identifying the root causes of extreme river flows. Finally, we conclude in Section 6. The implementation is available from the `gcm` module (Blöbaum et al., 2022) in DoWhy. We provide the scripts for the experiments as supplementary material. All proofs are in the appendix.

2. Information-theoretic outlier scores

We start by introducing "information-theoretic" outlier score that probabilistically calibrates existing outlier scores, which is key to developing our *general* method for identifying root causes of outliers.

It is commonly agreed upon that outliers are "rare" events that differ significantly from the "majority" of data objects (Hawkins, 1980; Aggarwal, 2013; Akoglu, 2021). From an information-theoretic viewpoint, the rarer an event, the more information it carries (Cover & Thomas, 2006). The notion of *information-theoretic outlier scores* formalises this insight, which we define below.

Definition 1 (Information-theoretic outlier scores)**.** *Let X be a random variable (r.v.) with values in $\mathcal{X}$ and distribution P_X. Let $\tau : \mathcal{X} \to \mathbb{R}$ be a measurable "feature map" that maps elements of $\mathcal{X}$ to the real line $\mathbb{R}$. The information theoretic (IT) outlier score $S_X^\tau : \mathcal{X} \to \mathbb{R}_0^+$, corresponding to the transformation τ, of an event $x \in \mathcal{X}$ is given by*

$$S_X^\tau(x) \coloneqq -\log P\{\tau(X) \geq \tau(x)\}. \tag{2}$$

Using an arbitrary feature map τ in the definition allows us to calibrate existing outlier scores, i.e., τ can be an existing outlier score, e.g., isolation forest (Liu et al., 2008).

As X is a random variable and τ is measurable, $\tau(X)$ is also a random variable. Assign $Y \coloneqq \tau(X)$ and $y \coloneqq \tau(x)$. Then $P\{Y \geq y\}$ is the probability of events of Y that are extreme than y. As such, $P\{Y \geq y\}$ measures the extremeness of the event $\tau(x)$ in feature space $\tau(\mathcal{X})$. From information theory, we know that $-\log P(Y = y)$ measures the information content of an event y (Cover & Thomas, 2006). Therefore, $-\log P\{Y \geq y\}$ measures the information content of an event y in terms of its extremeness.

Note that IT outlier score considers the distribution over feature space $\tau(\mathcal{X})$, instead of input space $\mathcal{X}$. Therefore, what an extreme event is, according to an IT outlier score, not only depends on the distribution of X, but also the feature map τ. This way, one can easily define outliers also for multi-variate X or other domains. It can also assign high score to low density regions between clusters in multimodal distributions—by choosing $\tau(x) \coloneqq -\log p(x)$.

In the example below, we show how **z**-score, a commonly used outlier score, can be calibrated into an IT outlier score. We provide more examples in the appendix.

Example 1 (**z**-score)**.** **z**-*score measures the normalised absolute distance from the mean, i.e.* $\mathbf{z}(x) \coloneqq |x - \mu_X|/\sigma_X$*, where μ_X is the mean and σ_X is the standard deviation of X. By setting $\tau(x) \coloneqq \mathbf{z}(x)$, we obtain the IT outlier score*

$$S_X^{\mathbf{z}}(x) = -\log P\left\{|X - \mu_X| \geq |x - \mu_X|\right\},$$

where σ_X is ignored as $\sigma_X \geq 0$ and it scales both sides.

2.1. Properties of IT outlier scores

Besides probabilistically calibrating existing outlier scores, IT scores also possess other properties that are desirable for any outlier score, which we formalise in the lemmas below.

Lemma 2.1 (Tail probability of IT outlier scores)**.** *Let X be a random variable (r.v.) with values in $\mathcal{X}$ and distribution P_X. Every information-theoretic outlier score satisfies the following properties:*

$$P\{S_X^\tau(X) \geq c\} \leq e^{-c} \quad \forall c \in \mathbb{R}_0^+ \tag{3}$$

$$P\{S_X^\tau(X) \geq S_X^\tau(x)\} = e^{-S_X^\tau(x)} \quad \text{for } P_X\text{-almost all } x \in \mathcal{X}. \tag{4}$$

where P_X denotes the distribution of X. Conversely, if $S_X^\tau : \mathcal{X} \to \mathbb{R}_0^+$ is measurable and satisfies (3) *and* (4)*, then there exists a measurable function $\tau : \mathcal{X} \to \mathbb{R}$ such that*

$$S_X^\tau(x) = -\log P\{\tau(X) \geq \tau(x)\}, \quad x \in \mathcal{X}.$$

If S_X^τ is surjective then equality holds in (3)*.*

Remark. In words, IT outlier scores reflect the heuristic idea that extreme outliers should happen rarely since the probability that $S_X^\tau(X)$ is large decreases exponentially.

The probabilistic calibration of IT scores has advantages beyond the fact that they are comparable across variables with different dimension, range, and scaling. This is because they obey simple rules that they inherit from basic probability theory, like the following. For any two events A and B, we have $P(A \cap B)/P(A) \leq 1 \implies P(A \mid B)/P(A) \leq 1/P(B)$. Therefore, if event A is a-priori very unlikely, then conditioning on an event B can render A more likely—relative to the likelihood of A—only by the factor $1/P(B)$ at most.

Likewise, observing large outlier scores of one variable can render large outliers of other variables more likely, but those whose score are *much* larger than the observed one are still unlikely. This is formalized by the following simple lemma:

Lemma 2.2 (Relations between outlier scores)**.** *For any $\delta \in \mathbb{R}_0^*$, for almost all $c \in S_X^\tau(\mathcal{X})$, we have*

$$P\{S_Y^\tau(Y) \geq c + \delta \,|S_X^\tau(X) \geq c\} \leq e^{-\delta}.$$

Suppose that we observe event $X = x$ has the outlier score 100 in some datasets. If we now select all datasets for which $X = x$ has score 100 *or more*, it is not unexpected that a second variable Y also has events with outlier scores up to 100 or slightly above, if Y is strongly coupled to X. But Y showing scores significantly above 100 should still be rare.[2] Note that this conclusion holds regardless of how X and Y are causally related.

[2]Note the following subtlety: Lemma 2.2 holds only for those c that really occur as outlier score. If $S_X^\tau(X)$ is a *non-surjective* outlier score which never attains values between 10 and 1000; conditioning on $S_X^\tau(X) \geq 100$ implicitly conditions on $S_X^\tau(X) \geq 1000$, which can render scores around 1000 quite likely.

We can also re-interpret Lemma 2.2 in a causal way as follows. Assume X is an unconfounded cause of Y and hence $P^{do(X:=x)}(Y) = P(Y \mid X = x)$, where $P^{do(X:=x)}(Y)$ denotes the distribution of Y after the atomic intervention (Pearl, 2009, Chap. 3) of setting X to x, keeping everything else in the system fixed. Further let us, for some $c \in S_X^\tau(\mathcal{X})$, define the randomized intervention $do(S_X^\tau(X) \geq c)$ by randomizing X according to the conditional distribution $P(X \mid S_X^\tau(X) \geq c)$. That is, we generate outliers having scores at least c according to their natural relative likelihood. Then Lemma 2.2 implies

$$P^{do(S_X^\tau(X) \geq c)}\{S_Y^\tau(Y) \geq c + \delta\} \leq e^{-\delta}. \tag{5}$$

In this sense, outliers of the causes most likely only cause outliers in the effect whose scores are not significantly above the ones they were driven by. From here onwards, we will drop X and τ in S_X^τ whenever it is clear from the context to which variable and feature map we refer to.

3. Contribution-based root cause analysis

We first present the intuition behind our method (Section 3.1). Then we formalize counterfactuals (a key concept) in our language (Section 3.2). Finally we present our method for identifying "root causes" of a target outlier (Section 3.3), and then address key points when applying the method in practice (Section 3.4).

3.1. Intuition behind our method

Three key points guide our method.

First, to qualify an upstream node as the "root cause" of an outlier event x_n, we ask the counterfactual question, "Would the event x_n not have been an outlier had we assigned rather "normal" causal mechanisms at the node instead of the existing mechanism associated with the outlier x_n?" By focusing on causal mechanisms, we can separate the contribution of the node itself from that inherited from its parents. This notion of "root cause" is in the spirit of the notion of "actual cause" of an outcome defined by Halpern & Pearl (2005). Here our focus is on the *outlierness* of the event x_n, rather than the actual value x_n itself. To get counterfactuals, we require a functional causal model (FCM) (Pearl, 2009), in addition to the causal graph.

Second, we consider the *degree* to which a node is a root cause. This is in the spirit of "graded causation" of Halpern & Hitchcock (2013), who argue that the degree to which something is a cause also a question of whether the instantiations of the other variables are "normal". There, "normality" is not necessarily defined in the sense of statistical regularity, but possibly also in the sense of an ethical norm. Here we define normality w.r.t. statistical regularity.

Third, a node X_j's contribution to the extremeness of x_n

boils down to the contribution of its noise term N_j in a statistical sense because statistical properties of observed variables are derived from the noise terms in an FCM. To see this, assume, without loss of generality, that X_n is a sink node, i.e. it has no descendants in the causal graph. Its structural equation is $X_n \coloneqq f_n(\mathrm{PA}_n, N_n)$. By recursively resolving the parents in terms of their parents (applying Eq. 1 recursively) until we have reached root nodes, we can write X_n as a function of all noise variables $\mathbf{N} \coloneqq (N_1, \ldots, N_n)$:

$$X_n \coloneqq f(N_1, \ldots, N_n) = f(\mathbf{N}), \tag{6}$$

where $P_{\mathbf{N}} = P_{N_1} \times \cdots \times P_{N_n}$. Using this representation, we see that any observed upstream nodes X_j's contribution to the distribution of X_n, thereby the extremeness of x_n, comes from the distribution P_{N_j} of its noise N_j.

In summary, counterfactuals are key to our formal method for identifying "root causes" of outliers.

3.2. Counterfactuals for root cause analysis

Next, we briefly explain counterfactuals in our formal language. The notion of counterfactuals we use in our method concerns "causal mechanisms", which differs slightly from the usual treatment of counterfactuals in the graphical causal model literature (Pearl, 2009, Ch. 7).

To explain our notion of counterfactuals, we consider the canonical representation of FCMs (Peters et al., 2017, Ch. 3), also referred to as the response function framework (Balke & Pearl, 1994). We illustrate the idea for a bivariate DAG $X \to Y$. But it generalises to more then two variables. The FCM at Y is a stochastic function, given by

$$Y \coloneqq f(X, N), \tag{7}$$

which reduces to a deterministic function of X for a fixed value n (by slightly abusing the notation) of the noise N:

$$Y \coloneqq f(X, n)$$

That is, if X and Y take values in $\mathcal{X}$ and $\mathcal{Y}$ respectively, then noise N acts as a random switch that selects different functions from $\mathcal{X}$ to $\mathcal{Y}$. Without loss of generality, we can therefore assume that N takes values in the set of functions from $\mathcal{X}$ to $\mathcal{Y}$, denoted by $\mathcal{Y}^{\mathcal{X}}$. Then we can rewrite the structural equation of Y (Eq 7) as

$$Y \coloneqq N(X). \tag{8}$$

The FCM $Y \coloneqq f(X, N)$ has now turned into a probability distribution P_N on the set of deterministic functions $\mathcal{Y}^{\mathcal{X}}$. The representation in Eq. 8 with the distribution P_N on $\mathcal{Y}^{\mathcal{X}}$ is the canonical representation of the FCM $Y \coloneqq f(X, N)$.

For example, if X and Y are binary, i.e., $\mathcal{X} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1\}$, then there are four possible functions from $\mathcal{X}$ to $\mathcal{Y}$, i.e., $\mathcal{Y}^{\mathcal{X}} = \{\mathbf{0}, \mathbf{1}, \mathtt{ID}, \mathtt{NOT}\}$, where $\mathbf{0}$ and $\mathbf{1}$ denote constant functions that always map to 0 and 1 respectively, and $\mathtt{ID}$ and $\mathtt{NOT}$ denote identity and negation respectively.

Suppose $X \to Y$ is the causal graph of variables X and Y, and we jointly observe their values (x, y) with a deterministic function $h \in \mathcal{Y}^{\mathcal{X}}$ identified by the value n of the noise term N. An alternative value $\tilde{n}$ of N would identify a different deterministic function $\tilde{h}$ from $\mathcal{Y}^{\mathcal{X}}$. The function $\tilde{h}$ is then a *counterfactual* causal mechanism at node Y as $\tilde{h}$ is not associated with the factual values (x, y) we observed.

From here, we can now formalize our notion of counterfactuals to more than two variables. Suppose we jointly observed values $\mathbf{x} \coloneqq (x_1, \ldots, x_n)$ of variables $\mathbf{X} \coloneqq (X_1, \ldots, X_n)$. The deterministic functions $h_1, \ldots, h_n$ at nodes $\mathbf{X}$ associated with the factuals $\mathbf{x}$ are identified by the values $\mathbf{n} \coloneqq (n_1, \ldots, n_n)$ of the corresponding noise terms $\mathbf{N} \coloneqq (N_1, \ldots, N_n)$, where $h_j \in \mathcal{X}_j^{\mathcal{PA}_j}$ with $\mathcal{X}_j$ and $\mathcal{PA}_j$ being the support of X_j and PA_j respectively.

Let $\mathcal{U} \coloneqq \{1, \ldots, n\}$ denote the index set and $\mathcal{I} \subset \mathcal{U}$ be its subset. We use $rd(\mathbf{N}_{\mathcal{I}})$ to denote the operation of randomizing some noises $\mathbf{N}_{\mathcal{I}}$ according to some joint distribution $\tilde{P}_{\mathbf{N}_{\mathcal{I}}}$ (not necessarily their true joint distribution $P_{\mathbf{N}_{\mathcal{I}}}$), whilst the remaining noises are kept fixed, i.e., $\mathbf{N}_{\bar{\mathcal{I}}} \coloneqq \mathbf{n}_{\bar{\mathcal{I}}}$, with $\bar{\mathcal{I}} = \mathcal{U} \setminus \mathcal{I}$. The operation $rd(\mathbf{N}_{\mathcal{I}})$ corresponds to assigning "normal" causal mechanisms to nodes $\mathbf{X}_{\mathcal{I}}$, whilst keeping the causal mechanisms of other nodes $\mathbf{X}_{\bar{\mathcal{I}}}$ fixed (to the causal mechanisms associated with the factuals $\mathbf{x}_{\bar{\mathcal{I}}}$, which are identified by noises $\mathbf{n}_{\bar{\mathcal{I}}}$). The action $rd(\mathbf{N}_{\mathcal{I}})$ yields a counterfactual distribution $P^{rd(\mathbf{N}_{\mathcal{I}})}(\mathbf{X})$ of the variables $\mathbf{X}$, where counterfactuals are w.r.t. alternative causal mechanisms. This counterfactual distribution is the key ingredient of our method.

Although the operation $rd(\mathbf{N}_{\mathcal{I}})$ suggests intervention on the noise terms $\mathbf{N}_{\bar{\mathcal{I}}} \coloneqq \mathbf{n}_{\bar{\mathcal{I}}}$ (which is infeasible if we think of the exogenous noise of something that is not under our control, and even worse, not even observable), we can interpret it as an intervention on observed variables $\mathbf{X}_{\bar{\mathcal{I}}}$ instead: for each $X_j \in \mathbf{X}_{\bar{\mathcal{I}}}$, just simulate an iid copy $\tilde{N}_j$ of the noise N_j and set $X_j \coloneqq f_j(\mathrm{pa}_j, \tilde{n}_j)$ if value $\tilde{n}_j$ was obtained.

Although unobserved, in practice, noise values can be recovered from the samples drawn from the observed joint distribution $P_{\mathbf{X}}$ subject to appropriate assumptions, e.g., additive noise (Peters et al., 2017, Ch. 4).

3.3. Root cause analysis quantifying contributions

Next, using the intuition and the counterfactuals introduced earlier, we formalise our method for identifying root causes of an outlier x_n amongst variables $X_1, \ldots, X_n$. In particular, we quantify the contribution of each unobserved noise term N_j (corresponding to X_j) to the IT outlier score of x_n.

To this end, we first rewrite the IT score of x_n, i.e., $S(x_n)$, in terms of noises, i.e.,

$$\begin{aligned} S(x_n) &\coloneqq -\log P\{\tau(X_n) \geq \tau(x_n)\} \\ &\coloneqq -\log P\{\tau(f(\mathbf{N})) \geq \tau(f(\mathbf{n}))\} \\ &\coloneqq -\log P\{g(\mathbf{N})) \geq g(\mathbf{n})\}, \end{aligned} \tag{9}$$

where the second line is obtained by applying Eq. 6 and g is a composition $g = \tau \circ f$.

To compute the contribution of each noise term N_j to the IT score $S(x_n)$, we quantify the change in the log-likelihood of the tail event when we ask the counterfactual question, "Would the event x_n not have been an outlier had we assigned rather normal causal mechanisms at X_j by randomizing N_j?" In particular, we define the contribution of N_j, given that we have already randomized some noises $\mathbf{N}_{\mathcal{I}}$, as

$$\begin{aligned} C(j \mid \mathcal{I}) \coloneqq &-\log P^{rd(\mathbf{N}_{\mathcal{I}\cup\{j\}})}\{g(\mathbf{N}) \geq g(\mathbf{n})\} \\ &+\log P^{rd(\mathbf{N}_{\mathcal{I}})}\{g(\mathbf{N}) \geq g(\mathbf{n})\}). \end{aligned}$$

Rewriting the contribution, we get

$$C(j \mid \mathcal{I}) \coloneqq \log \frac{P^{rd(\mathbf{N}_{\mathcal{I}})}\{g(\mathbf{N}) \geq g(\mathbf{n})\})}{P^{rd(\mathbf{N}_{\mathcal{I}\cup\{j\}})}\{g(\mathbf{N}) \geq g(\mathbf{n})\}}, \tag{10}$$

which shows that the contribution measures the factor by which knowing the causal mechanism at X_j—by knowing the noise value n_j—increases the counterfactual tail probability of the target outlier.

But the contribution depends on the subset $\mathcal{I}$ according to which we randomize noises. For any ordering $\mathcal{U}_{\sigma(1)}, \ldots, \mathcal{U}_{\sigma(n)}$ of the index set $\mathcal{U}$, we could consider the contribution of $\mathcal{U}_{\sigma(j)}$ given $\mathcal{I}^{\text{ctx}} \coloneqq \{\mathcal{U}_{\sigma(1)}, \ldots, \mathcal{U}_{\sigma(j-1)}\}$ as the context. This dependence on ordering introduces arbitrariness in the attribution procedure.

To get rid of the arbitrariness, we leverage Shapley values (Shapley, 1953) from cooperative game theory. The key idea of Shapley value is to symmetrize over all orderings, i.e. consider all possible orderings, compute the contribution for each ordering, and then take the average.

Using the concept of Shapley values, the contribution of the noise term N_j to the target outlier score $S(x_n)$ is given by the average contribution over all possible orderings σ, i.e.

$$\phi(j) \coloneqq \frac{1}{n!} \sum_{\sigma} C(j \mid \mathcal{I}^{\text{ctx}}) \tag{11}$$

$$= \sum_{\mathcal{I} \subseteq \mathcal{U}\setminus\{j\}} \frac{1}{n\binom{n-1}{|\mathcal{I}|}} C(j \mid \mathcal{I}), \tag{12}$$

where the second summation follows from aggregating contributions for permutations of $\mathcal{U}$ with the same value of $\mathcal{I}^{\text{ctx}}$ before j in the ordering. Note that this contribution can be negative: one value being extreme can certainly decrease the likelihood of the outlier event, and a more common value at that node would have made the outlier even stronger.

The Shapley value is also desirable because it gives a unique solution to a set of axioms that capture the notion of fairness when dividing a pay-off among players in the coalition game (Sundararajan & Najmi, 2020). The theorem below follows directly from the "efficiency" property of Shapley values, by virtue of which Shapley values of all players sum up to the payoff (i.e., IT outlier score subject to $rd(\mathbf{N}_{\mathcal{I}})$ operation) of the grand coalition (i.e., $\mathcal{I} = \mathcal{U}$) with a nuance that we use the true joint distribution $P_{\mathbf{N}}$ when $\mathcal{I} = \mathcal{U}$.

Theorem 3.1 (Decomposition of target outlier score)**.** *The outlier score of an event x_n from any target variable X_n decomposes into the Shapley contribution of each of its ancestors plus itself, i.e., $S(x_n) = \sum_{j=1}^{n} \phi(j)$, where n is the number of ancestors of X_n including itself.*

We illustrate the interpretation behind our quantification of contribution through a simple example below.

Example 2 (Interpretation of contribution from co-occurrence of independent events)**.** *Suppose that our target variable X_n is a logical AND of independent binary random variables $X_1, \ldots, X_{n-1}$ (e.g., tosses from biased coins) with a binary noise N_n, i.e.*

$$\begin{aligned} X_n &\coloneqq X_1 \wedge X_2 \wedge \cdots \wedge X_{n-1} \wedge N_n \\ &\coloneqq N_1 \wedge N_2 \wedge \cdots \wedge N_{n-1} \wedge N_n, \end{aligned}$$

with $p_j \coloneqq P_{N_j}\{N_j = 1\}$. Suppose that we observe $X_n = 1$, which is a rare event as this can only happen for one combination $\mathbf{n} = (1, \ldots, 1)$ of noise values (out of 2^n). Let us quantify the contribution of each N_j to $S(x_n)$ using an identity feature map $\tau(x) = x$. As noises N_j are independent, we have $P_{X_n}\{X_n = 1\} = \prod_{j=1}^{n} p_j$. For a binary r.v. X_n, the tail probability coincides with pmf at one, i.e. $P_{X_n}\{X_n \geq 1\} = P_{X_n}\{X_n = 1\}$.

The contribution of any noise N_j given that we have randomized noise terms $\mathbf{N}_{\mathcal{I}}$ is given by (from Eq. 10):

$$C(j \mid \mathcal{I}) = \log \frac{P^{rd(\mathbf{N}_{\mathcal{I}})}\{X_n \geq 1\}}{P^{rd(\mathbf{N}_{\mathcal{I}\cup\{j\}})}\{X_n \geq 1\}} \tag{13}$$

$$= \log \frac{\prod_{i\in\mathcal{I}} p_i}{\prod_{i\in\mathcal{I}\cup\{j\}} p_i} = -\log p_j. \tag{14}$$

The contribution of N_j remains $-\log p_j$ regardless of the subset $\mathcal{I}$ as the noises are independent. Hence the Shapley value contribution of each noise N_j is also $\phi(j) = -\log p_j$.

Takeaway. *The lower the probability p_j, the higher the contribution $-\log p_j$ of noise term N_j. Thus, rare necessary conditions have high contribution, and are hence likely to be the root causes of the target outlier. A rare event can only*

be explained by other rare events. For example, a strong unexpected drop in Dow Jones index cannot be explained by an event that happens every week.

3.4. Implications of technical assumptions in practice

Next, we discuss the implications of our key assumptions.

On Markovian causal models. The causal model (DAG and joint distribution) does not have to be Markovian. Note that the independence of noise terms in the FCM implies that the causal model is Markovian (Shpitser & Pearl, 2006), i.e., the joint distribution factorises into a product of conditional distributions of each variable given its parents in the causal graph. Our method also works with semi-Markovian models, where there are unmeasured confounders (Shpitser & Pearl, 2006). In the FCM, this means some noise terms are confounded by unobserved common causes Z. A randomized intervention on noises $\mathbf{N}_{\mathcal{I}}$ hence blocks all back-door paths between $\mathbf{N}_{\mathcal{I}}$ and the target X_n via Z. Thus we still obtain *causal* counterfactuals in semi-Markovian models.

On causal graph and FCM. Our method does not solve the hard problem of causal discovery (Spirtes et al., 2000), i.e., how to obtain the causal graph. Instead, we provide a method to formally talk about attributing outliers to root causes when the causal graph is given and the structural equations are either given or inferred from data. To obtain the causal graph, it is typical to apply a combination of domain knowledge, interventional analysis and causal structure learning. Although getting the DAG is often difficult, it seems unavoidable for causal attribution problems.

For generic choices of functions and noises, FCMs do not uniquely follow from observed joint distributions even when the causal graph is given. But they can be inferred subject to appropriate assumptions, e.g., additive noise (Peters et al., 2017, Chap 4). We admit that inferring functions and noise from data is often an issue, but our example with river flows shows that domain knowledge can help. Using counterfactuals seems unavoidable for causal attribution at the unit level, and, more generally, Pearl considers "rung 3" causality in his ladder of causation (Pearl & Mackenzie, 2018) as crucial for understanding the world.

Both causal graph as well as FCMs can be misspecified. Although not exhaustive, we empirically investigate this concern for one type of misspecification through simulations (Section 5.1).

On Shapley values. To compute Shapley value contributions *numerically*, the contributions (Eq (10)) have to be averaged over all orderings $\mathcal{I}$. Up to tens of variables, we obtain the exact solution quite fast (within minutes). When the number of variables is larger than that, exact numerical solution is intractable. In such cases, we can trade-off the accuracy of Shapley value contributions for speed by applying sampling approximations to Eq. (11), see Strumbelj & Kononenko (2014) for example. The key idea is to sample orderings, instead of using all orderings.

On the role of noise. An outlier is not necessarily based on an outlier of the noise and emphasise that our framework does not assume this. The rationale is slightly more subtle: whenever an unlikely noise value would be required to explain the observed value x_j from pa_j, we conclude that either the structural equation did not hold for that particular statistical unit (due to a corrupted mechanism), or the noise behaved in an unexpected way.

4. Related Work

The vast body of literature on outliers focuses on *detecting* outliers (Breunig et al., 2000; Liu et al., 2008; Aggarwal, 2013; Hawkins, 1980; Chandola et al., 2009; Blázquez-García et al., 2021; Gupta et al., 2018). We refer the reader to Akoglu (2021) for a comprehensive overview.

The earliest work on explaining outliers dates back several decades (Knorr & Ng, 1999), which provides an explanation for "exceptionality" of outliers in terms of feature subspace. Most existing work in *explaining* the detected outliers are recent, and follow a similar pursuit (Micenková et al., 2013; Macha & Akoglu, 2018; Gupta et al., 2018; Liu et al., 2018). These methods describe outliers or their group by feature subspaces that separate them from "normal" observations.

There are at least two reasons why those methods do not explain *root causes* of target outliers. First they capture features that are statistically dependent on the target outliers. But those features do not necessarily cause the outliers as there can be a common cause of the features as well as the outliers. Second, and importantly, even if we consider causal ancestors of the target as the features, not *all* features from feature subspaces that stand out relative to normal observations *cause* the outlier.

A recent proposal by Idé et al. (2021) explains anomalous deviations of prediction from the *actual* output in terms of actions that might be taken to bring back the outlying sample to normalcy. Those actions are with respect to a prediction model, however, and hence not necessarily causal.

In this work, we identify root causes of an outlier event by quantifying the contribution of upstream nodes when the causal graph is known. We assume that the causal graph does not change when we observe outliers. In some scenarios, this can still happen. For the problem of inferring the causal graph from outlier statistics, we refer the reader to Gnecco et al. (2021); Gissibl et al. (2021).

5. Experiments

First, through simulations—where we can establish the ground truth—we evaluate the performance of our method for identifying root causes of outliers (Section 5.1). Then we assess whether results are sensible on real-world case study identifying root causes of extreme river flows (Section 5.2).

5.1. Simulations

Next, we generate synthetic data and establish the ground truth. We then evaluate the performance of the proposed method against the ground truth. In particular, we aim to answer the following two questions with simulations:

Q1. Can the proposed method identify top-k root causes when model assumptions hold?

Q2. Can the proposed method identify top-k root causes when model assumptions *do not* hold?

Experiment design. We randomly generate causal graphs and associated FCMs. From the FCMs, we generate training samples. To establish the ground truth ranking of root causes, we obtain target outliers in test samples by perturbing mechanisms upstream of the target with different strengths. We use the training samples to learn the FCMs. In the test samples, we get the rankings of root causes by applying various methods, which we then evaluate against the ground truth ranking. We compare our method against a baseline method based on existing outlier score:

- NaiveRCA. The ranking of root causes is based on the existing outlier score. In particular, we use **z**-score, as the variables we consider have unimodal marginal distributions. To compute the **z**-score for an event x_j of X_j, we use the marginal distribution of X_j. The higher the **z**-score, the higher the ranking.
- CausalRCA. The ranking of root causes is based on Shapley values computed from Eq. 11. The higher the contribution, the higher the ranking. In particular, we set **z**-score as the feature map, i.e., $\tau(x) \coloneqq \mathbf{z}(x)$.

We measure the quality of rankings by NDCG@k (Järvelin & Kekäläinen, 2000), which is a widely used metric for measuring rankings with graded relevance of outcomes like ours. The NDCG@k is higher if highly "relevant" root causes have higher ranks. A perfect ranking method will have an NDCG@k score of 1.0. To establish the ground truth relevance of all nodes, we assign zero relevance scores to non-root causes, and invert the ranking of injected root causes (i.e., the rank 1 root cause gets the highest relevance score of p if we injected p root causes in the test sample). When we ask for top-k results from a method, we assign the relevance scores to its ranking using the ground truth, which is then used to compute its NDCG@k score.

Data Generation & Ground Truth. To generate causal graphs, we follow the procedure described in Janzing et al. (2012). In particular, we generate causal graphs with at least 10 upstream nodes of the target node. To each node X_j, we assign a random linear structural equation of the form

$$X_j \coloneqq \sum_i \beta_{ij} \mathrm{PA}_{ij} + N_j, \tag{15}$$

where PA_{ij} is the i-th component of X_j's parents PA_j, $\beta_{ij} \sim \mathrm{Uniform}(0, 5)$ and $N_j \sim \mathrm{Gaussian}(0, 1)$. We draw training samples from these FCMs. We do not use non-linear FCMs in simulations because analytical solution of the Shapley value contribution (Eq. 11) is non-trivial to compute already for linear FCMs (to get the ground truth).

We then generate the test samples with the target outlier and its root causes by modifying the FCMs. In the test samples, we inject outliers in 1 to 5 noise terms randomly upstream of the target node, including its own, in the causal graph. Those noise terms are the root causes of the target outlier. In particular, we obtain an outlier $\tilde{x}_j$ at a node X_j by perturbing its noise value to $\lambda \sim \mathrm{Uniform}(3, 5)$, which is at least 3 standard deviation away from the mean of the marginal distributions of noise terms N_j (which is 0):

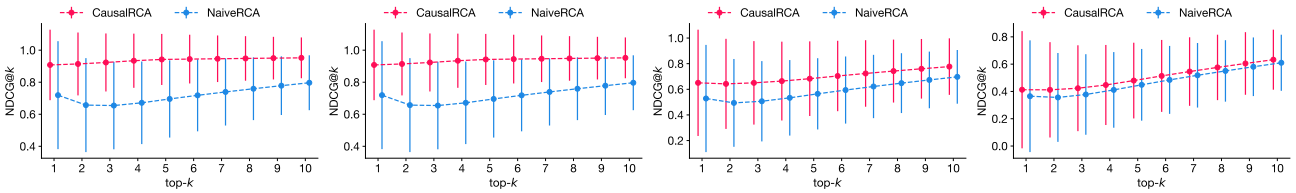
$$\tilde{x}_j \coloneqq \sum_i \beta_j^{(i)} \mathrm{pa}_j^{(i)} + \lambda. \tag{16}$$

We obtain the ground truth ranking of root causes by their contributions to the target outlier. As each node has a linear FCM, we can reduce the FCM of the target variable X_n as a linear combination of upstream noise terms, i.e. $X_n \coloneqq \sum_{j=1}^{n} \alpha_j N_j$, where n is the number of nodes. For the root causes, noise values are fixed to λ. Thus, root causes with larger values of α_j contribute more to the value attained by the target node, and hence its extremeness.[3]

Methodology. We draw 1K random causal graphs. Each causal graph has a randomly chosen linear FCM at each of its nodes as described in Eq. 15. From the linear FCMs of each graph, we draw 2000 training samples. From the FCMs, we then draw 10 random test samples, each containing a target outlier and between 1 to 5 root causes. Each method therefore yields 10K rankings.

For the case when model assumptions hold (i.e., Question **Q1**), we take the ground truth causal graph and then estimate its FCMs from its training samples assuming a linear additive noise model (Peters et al., 2017, Ch. 7). In particular, we use the linear regression to estimate the functions, i.e., $f_j \coloneqq \mathbb{E}(X_j \mid \mathrm{PA}_j)$ and use the empirical distribution of the residual for the noise term.

[3] This heuristic can yield some paradoxes, although it works in overwhelming cases, which we explain in the Appendix.



(a) No edge missing (b) 10% of edges missing (c) 20% of edges missing (d) 40% of edges missing

Figure 6: NDCG@k for various values of k, when there are (a) no missing edges, and (b–d) respectively 10%, 20% and 40% of edges in the causal graph are missing when estimating the FCMs. The higher values of NDCG@k indicate that CausalRCA is better at identifying root cases of outliers than NaiveRCA, even when model assumptions do not hold.

For the cases where model assumptions do not hold (i.e., Question **Q2**), we drop edges randomly from the ground truth causal graph and then estimate the FCMs from the imputed graph—assuming a linear additive noise model. By doing so, in our method CausalRCA, we neither use the exact causal graph, nor the exact inputs to the FCMs.

In both settings, we use empirical distributions for the root nodes. For our method, we estimate the tail probabilities empirically by drawing samples from the estimated FCMs, applying the feature map, and then counting the tail events.

Results In Figure 6 (a), we show the average NDCG@k and its standard error over 10K rankings when model assumptions hold. We observe that CausalRCA has higher NDCG@k scores on average for all values of k than the baseline NaiveRCA. The upward trend of NDCG@k scores and decreasing standard errors for both methods can be explained by the fact that relevant root causes are more likely to appear in the top-k results as k increases.

In Figure 6 (b–d), we show the results when we respectively drop 10%, 20% and 40% of the edges randomly from the ground truth causal graph and learn FCMs from the imputed graph. As expected, the NDCG@k scores drop as we drop more edges and strongly violate model assumptions. Even when model assumptions are strongly violated, CausalRCA does not perform worse than the baseline method NaiveRCA.

We have to be careful here not to generalize this observation, however. If functional forms of FCMs, for instance, are completely off, there is no reason to believe that our method will still perform sensibly. There is simply too many degrees of freedom to answer Question Q2 empirically. But we can reasonably say that the performance of our method drops as our model assumptions are violated.

5.2. Case Study: Root causes of extreme river flows

Next, we apply our method to a real-world scenario to see if the results we get are sensible.

Our goal here is to identify the root causes of extreme river flows at the New Jumbles Rock (NJR) station that is located right after the confluence of 3 rivers in England. As candidate causes, we consider river flows measured at 3 stations upstream of the confluence along each tributary river, namely Hodder Place (HP), Henthorn (HT) and Whalley Weir (WW) (see the map of stations in Figure 7 left).

As river flow downstream of the confluence is the result of river flows upstream, we can reasonably assume the causal graph in Figure 7 (right), with unobserved common causes like weather conditions (e.g. precipitation and temperature) represented by the dotted node. Note that unobserved common causes only affect how well we learn the FCM; the proposed framework works as long as we have the FCM (see Section 3.4). But estimating the FCM from data (e.g., estimating causal regression coefficients and noise by OLS) suffers from a confounding bias here, without further structural assumptions. Hence we take the following approach using domain knowledge: assuming that most of the water flow at each station will also reach downstream stations, we estimate the noise (i.e., the hidden influx) at each station simply by the difference between the flow recorded at the station and sum of flows upstream.

As training samples for estimating the FCM, we use daily river flows, measured in each station at 9:00 daily, from 1 January 2010 till 31 December 2018 (i.e., 3267 observations).[4] The FCM of the flow at the NJR station is simply

$$\mathrm{X}_{\mathrm{NJR}} := \mathrm{X}_{\mathrm{HP}} + \mathrm{X}_{\mathrm{HT}} + \mathrm{X}_{\mathrm{WW}} + N_{\mathrm{NJR}},$$

where X_j is the river flow measured at station j. We obtain the noise term N_{NJR} by a simple algebraic operation

$$N_{\mathrm{NJR}} = \mathrm{X}_{\mathrm{NJR}} - \mathrm{X}_{\mathrm{HP}} + \mathrm{X}_{\mathrm{HT}} + \mathrm{X}_{\mathrm{WW}}$$

We model the distribution of noise N_{NJR}, and the marginal distribution of the other stations, namely HP, HT and WW, by their empirical distributions.

We want to identify the root causes of outliers in the daily measurements between 2019 January until 2019 March. To

[4]Data Source: `https://tinyurl.com/ukriverdata`

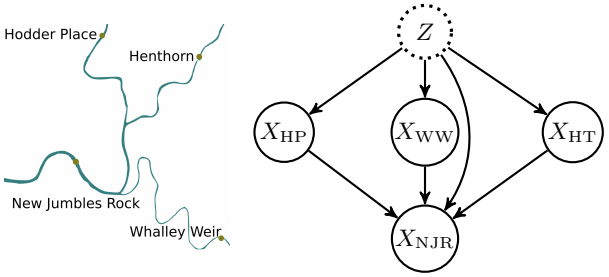


Figure 7: (left) Map of station locations. New Jumbles Rock (NJR) is the station downstream of other stations, namely namely Hodder Place (HP), Henthorn (HT) and Whalley Weir (WW). (right) Causal graph of river flows at stations. The dotted node Z represents unobserved common causes like weather conditions (e.g., precipitation, temperature).

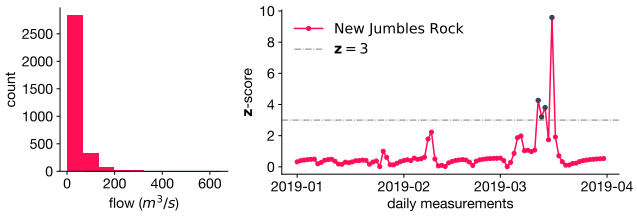


Figure 8: (left) Histogram of daily river flows between 2010 January and 2018 December at the NJR station, and (b) **z**-scores of daily river flows at the NJR station in 2019. We want to identify the root causes of the 4 outliers (black dots).

detect the outliers, we use **z**-scores, as the histogram of the measurements at the NJR station before 2019 shows a unimodal distribution (see Figure 8 left). With a threshold of $\mathbf{z} = 3$, we identify four outliers (see Figure 8 right).[5] We then use our method to identify their root causes.

In Figure 9 (top), we show the results of our method. In particular, we convert the absolute Shapley value contributions to percentages by dividing them by the total Shapley values (which is also equal to the IT score of the target outlier). We observe that flows upstream are the main contributors (root causes) to the flow at the NJR station, which is in agreement with the intuition and raw data in Figure 9 (bottom). For example, the peak flow on March 16 at the NJR station is explained almost completely by peak flows upstream.

Overall, our method is better at identifying top-k root causes, when modelling assumptions hold. Even when modelling assumptions do not hold, results of our method are comparable to the baseline method based on an existing outlier score. In real-world case study, the results are sensible.

[5]We can set a different threshold to obtain more or less outliers; our method will explain those anyway. But our focus here is to explain the detected outliers, not to detect outliers.

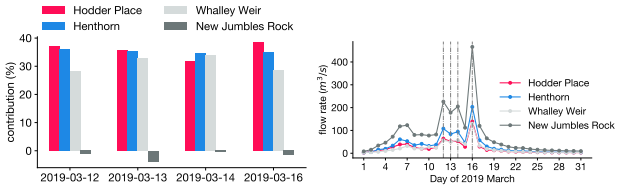


Figure 9: (left) Result of our method identifying root causes for extreme river flows at the NJR station. Extreme river flows at the NJR station are explained almost completely by peaks upstream. (right) Daily river flows on March 2019.

6. Conclusion

We have presented a formal method to identify "root causes" of outliers when we know the causal graph of variables, along with associated functional causal models. To generalise the method to existing outlier scores, we introduced information-theoretic (IT) outlier scores that probabilistically calibrate existing outlier scores (in Section 2). To identify root causes of an outlier event x_n, we attributed the outlier score $S(x_n)$ to unexpected behaviour of its ancestors using Shapley values from game theory (in Section 3). We illustrated our method on both synthetic and real datasets. As our method rests on causal assumptions, a systematic theoretical analysis of the robustness of our method against violation of assumptions requires further research.

Acknowledgements

The authors thank Claudia Shi, Jean-François Ton, Atalanti Mastakouri, Florian Saupe and the anonymous reviewers for their comments.

References

Aggarwal, C. C. *Outlier Analysis*. Springer, 2013.

Akoglu, L. Anomaly mining - past, present and future. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pp. 4932–4936. International Joint Conferences on Artificial Intelligence Organization, 2021.

Balke, A. and Pearl, J. Counterfactual probabilities: Computational methods, bounds and applications. In *Proceedings of the Tenth International Conference on Uncertainty in Artificial Intelligence*, UAI'94, pp. 46–54, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

Blázquez-García, A., Conde, A., Mori, U., and Lozano, J. A. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys*, 54(3), 2021.

Blöbaum, P., Götz, P., Budhathoki, K., Mastakouri, A. A., and Janzing, D. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *arXiv preprint arXiv:2206.06821*, 2022.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., and Sander, J. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, may 2000.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3): 15, 2009.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

Gissibl, N., Klüppelberg, C., and Lauritzen, S. Identifiability and estimation of recursive max-linear models. *Scandinavian Journal of Statistics*, 48(1):188–211, 2021.

Gnecco, N., Meinshausen, N., Peters, J., and Engelke, S. Causal discovery in heavy-tailed models. *The Annals of Statistics*, 49(3):1755 – 1778, 2021.

Gupta, N., Eswaran, D., Shah, N., Akoglu, L., and Faloutsos, C. Beyond outlier detection: Lookout for pictorial explanation. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2018, Dublin, Ireland, September 10-14, 2018, Proceedings, Part I*, volume 11051 of *Lecture Notes in Computer Science*, pp. 122–138. Springer, 2018.

Halpern, J. and Hitchcock, C. Graded causation and defaults. *The British Journal for the Philosophy of Science*, 66:413–457, 2013.

Halpern, J. and Pearl, J. Causes and explanations: A structural-model approach. part ii: Explanations. *The British Journal for the Philosophy of Science*, 56(4):889–911, 2005.

Hawkins, D. M. *Identification of Outliers*. Chapman and Hall, 1980.

Idé, T., Dhurandhar, A., Navrátil, J., Singh, M., and Abe, N. Anomaly attribution with likelihood compensation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):4131–4138, May 2021.

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*, 182–183:1–31, 2012.

Järvelin, K. and Kekäläinen, J. Ir evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pp. 41—-48, New York, NY, USA, 2000. Association for Computing Machinery.

Knorr, E. M. and Ng, R. T. Finding intensional knowledge of distance-based outliers. In *Proceedings of the 25th International Conference on Very Large Data Bases*, VLDB '99, pp. 211–222, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422. IEEE, 2008.

Liu, N., Shin, D., and Hu, X. Contextual outlier interpretation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 2461–2467. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

Macha, M. and Akoglu, L. Explaining anomalies in groups with characterizing subspace rules. *Data Mining and Knowledge Discovery*, 32(5):1444–1480, 2018.

Micenková, B., Ng, R. T., Dang, X.-H., and Assent, I. Explaining outliers by subspace separability. In *2013 IEEE 13th International Conference on Data Mining*, pp. 518–527, 2013.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

Pearl, J. and Mackenzie, D. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018.

Peters, J., Janzing, D., and Schlkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

Shapley, L. S. A value for n-person games. Technical report, Rand Corporation, 1953.

Shpitser, I. and Pearl, J. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pp. 1219–1226. AAAI Press, 2006.

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

Strumbelj, E. and Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*, 41(3):647—665, 2014.

Sundararajan, M. and Najmi, A. The many shapley values for model explanation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9269–9278. PMLR, 13–18 Jul 2020.

Zhang, K., Wang, Z., Zhang, J., and Schölkopf, B. On estimation of functional causal models: General results and application to the post-nonlinear causal model. *ACM Trans. Intell. Syst. Technol.*, 7(2), 2015.

A. Additional examples of IT outlier scores

Example 3 (Rarity)**.** *For unimodal distributions, it is also common to use negative logarithm of probability density function, which essentially measures the information content of an event w.r.t. a reference distribution. Suppose that probability density $p_X =: p$ of probability measure P_X exists. Then rarity of point x w.r.t. density p is given by*

$$r(x) \coloneqq -\log p(x).$$

By setting $\tau(x) \coloneqq r(x)$, we obtain the corresponding IT outlier score as

$$\begin{aligned} S_X^\tau(x) &= -\log P\{r(X) \geq r(x)\} \\ &= -\log P\{-\log p(X) \geq -\log p(x)\} \\ &= -\log P\{\log p(X) \leq \log p(x)\} \\ &= -\log P\{p(X) \leq p(x)\}. \end{aligned}$$

Note that $S(x) \coloneqq -\log p(x)$ would not define an IT outlier score because the total probability of points with small probability density need not be small.

Example 4 (Log Quantile)**.** *Another outlier score commonly in use is the right-sided log quantile. The right-sided log quantile measures the information content of events that are extreme than a given event, and given by*

$$q^{\geq}(x) \coloneqq -\log P\{X \geq x\}.$$

Unlike the previous examples, we need not plug in $q^{\geq}(x)$ to $\tau(x)$ to obtain the corresponding IT outlier score. Right-sided log quantile score is already an IT outlier score. To see this, we can simply set $\tau(x)$ to an identity map, i.e. $\tau(x) \coloneqq x$, in the definition of IT outlier score to obtain

$$S_X^\tau(x) \coloneqq -\log P\{X \geq x\} = q^{\geq}(x).$$

B. Paradox

Consider a bivariate causal graph $X \to Y$ with a linear FCM with following specifications, i.e.

$$X \coloneqq N_X \tag{17}$$

$$Y \coloneqq 10X + N_Y, \tag{18}$$

where $N_X, N_Y \sim \mathcal{N}(0,1)$. Observe that the coefficient of X is larger than that of N_Y in determining the value of Y. Therefore, intuitively, we expect the contribution of N_X to be higher than that of N_Y. Is this the case? Let us compute the Shapley value contribution of N_X and N_Y to the outlier $y = 11$. Suppose that we observe $n_X = 1$ corresponding to that value of Y. Then, we can recover corresponding noise value at Y as $n_Y = 11 - 10 \times 1 = 1$. Let $\mathbf{n} \coloneqq (n_X, n_Y)$.

Let us use an identity feature map, i.e. $\tau(x) \coloneqq x$. Then the contribution of N_X to $S(y)$ is given by

$$\begin{aligned} \phi(X) = \frac{1}{2}\log\frac{P^{do(\mathbf{n})}\{Y \geq 11\}}{P^{do(n_Y)}\{Y \geq 11\}} + \\ \frac{1}{2}\log\frac{P^{do(n_X)}\{Y \geq 11\}}{P^{do(\mathbf{n}_\emptyset)}\{Y \geq 11\}} \\ = \frac{1}{2}\log\frac{1}{0.46} + \frac{1}{2}\log\frac{0.15}{0.45} = -0.23, \end{aligned}$$

where we compute the tail probability $P^{do(\mathbf{n}_Y)}\{Y \geq 11\}$, for example, using the distribution of Y when we perform $do(N_Y \coloneqq 1)$ in the FCM above, i.e. $Y \sim \mathcal{N}(1, 100)$. Likewise the contribution of N_Y to $S(y)$ is given by

$$\begin{aligned} \phi(Y) = \frac{1}{2}\log\frac{P^{do(\mathbf{n})}\{Y \geq 11\}}{P^{do(n_X)}\{Y \geq 11\}} + \\ \frac{1}{2}\log\frac{P^{do(n_Y)}\{Y \geq 11\}}{P^{do(\mathbf{n}_\emptyset)}\{Y \geq 11\}} \\ = \frac{1}{2}\log\frac{1}{0.15} + \frac{1}{2}\log\frac{0.46}{0.45} = 1.38. \end{aligned}$$

Clearly this defies our intuition. Even when the value of n_X and n_Y are the same, but the coefficient of N_X (or X) is 10 times larger than that of N_Y, the contribution of N_Y to the outlier score $S(y)$ is significantly larger than that of N_X.

We will explain possible ways to solve this puzzle. Assume instead of defining the outlier event by $|Y| \geq |y|$ but $|Y| \geq (1-\epsilon)|y|$ instead. In words, we consider some value y' still 'the same' outlier event but slightly smaller than y. If ϵ is not too small, N_X has a significantly higher contribution than N_Y because changing n_Y to a more normal value changes y by such a small amount that we still get $|Y| \geq (1-\epsilon)|y|$. In other words, our judgement that N_X contributes more to the outlier event is implicitly based on a coarse-grained perspective, for which a slightly smaller outlier is still the same event.

This may suggest to redefine contribution analysis by replacing the event $\{\tau(X_n) \geq \tau(x_n)\}$ with $\{\tau(X_n) \geq (1-\epsilon)\tau(x_n)\}$ in contribution analysis as well as the definition of the outlier score. We will not follow up on this option here to avoid introducing another free parameter. Instead,

we emphasize that the counterintuitive behaviour degrades quickly with more nodes.

Consider the simple model with $Y = \sum_{j=1}^{n} \alpha_j N_j$ with independent Gaussians N_j. For $n \gg 2$ variables the contribution of N_j consists of many different conditional contributions $c(j|T)$, where most of the different subsets T do not contain all indices other than j. In other words, some other N_i are randomized, and changing n_j does not change the probability for the outlier event significantly since this change disappears in the random signal of the other variable if they have larger α_i.

C. Proofs

In the proofs, we will drop X and τ in S_X^τ whenever it is clear from the context to which variable and feature map we refer to.

Lemma 2.1 (Tail probability of IT outlier scores)**.** *Let X be a random variable (r.v.) with values in $\mathcal{X}$ and distribution P_X. Every information-theoretic outlier score satisfies the following properties:*

$$P\{S_X^\tau(X) \geq c\} \leq e^{-c} \quad \forall c \in \mathbb{R}_0^+ \tag{3}$$

$$P\{S_X^\tau(X) \geq S_X^\tau(x)\} = e^{-S_X^\tau(x)} \quad \text{for } P_X\text{-almost all } x \in \mathcal{X}. \tag{4}$$

where P_X denotes the distribution of X. Conversely, if $S_X^\tau : \mathcal{X} \to \mathbb{R}_0^+$ is measurable and satisfies (3) *and* (4)*, then there exists a measurable function $\tau : \mathcal{X} \to \mathbb{R}$ such that*

$$S_X^\tau(x) = -\log P\{\tau(X) \geq \tau(x)\}, \quad x \in \mathcal{X}.$$

If S_X^τ is surjective then equality holds in (3)*.*

Proof. First we show that S_X is measurable. Set $\tau(x) =: y$ and $\tau(X) =: Y$. Denote the cdf of Y by G. Then we have

$$\begin{aligned} S_X(x) &= -\log(1 - G(y)) \\ &= -\log(1 - G(\tau(x))) \end{aligned}$$

As S_X is a composition of measurable functions G and τ, it is measurable. Let G^{-1} be the generalized inverse of cdf of Y. Then it holds that

$$P\{S_X(X) \geq c\} = P\{G^{-1}(1 - e^{-c}) \leq G^{-1}(G(\tau(X)))\},$$

where the last equation holds since the generalized inverse is monotonically non-decreasing. Further, notice that for the generalized inverse, it holds $G^{-1}(G(x)) \leq x$ for all x, and hence:

$$\begin{aligned} P\{G^{-1}(1 - e^{-c}) &\leq G^{-1}(G(\tau(X)))\} \\ &\leq P\{G^{-1}(1 - e^{-c}) \leq \tau(X)\} \\ &= 1 - G(G^{-1}(1 - e^{-c})) \\ &\leq 1 - (1 - e^{-c}) = e^{-c}\,, \end{aligned}$$

again, the last inequality holds since $G^{-1}(G(x)) \leq x$ for all x. Therefore, we have shown that IT outlier score satisfies relation (3). Now to show that IT outlier score also satisfies relation (4), note that $G^{-1}(G(Y)) = Y$ holds almost surely for arbitrary r.v. Y with its cdf G. Hence, we have

$$\begin{aligned} P\{S_X(X) \geq S(x)\} &= P\{e^{-S_X(X)} \leq e^{-S_X(x)}\} \\ &= P\{\tau(X) \geq G^{-1}(G(\tau(x)))\}, \end{aligned}$$

where we applied in the last equality G^{-1} on both sides (note that G^{-1} is non-decreasing) and used the fact that $G^{-1}G(\tau(X)) = \tau(X)$ a.s.

By definition, this means that with $A := \{\omega \in \Omega : G^{-1}G(\tau(X(\omega))) = \tau(X(\omega))\}$ it holds $P(A) = 1$ (note that A is measurable since τ and X are measurable) and since

$$A = (\tau(X))^{-1}(\{y \in \mathbb{R} : G^{-1}(G(y)) = y\})$$

it holds

$$\begin{aligned} 1 &= P(A) \\ &= P((\tau(X))^{-1}(\{y \in \mathbb{R} : G^{-1}(G(y)) = y\})) \\ &= P(X^{-1}(\tau^{-1}((\{y \in \mathbb{R} : G^{-1}(G(y)) = y\})))) \\ &= P_X(\{x \in \mathcal{X} : G^{-1}(G(\tau(x))) = f(x)\} \end{aligned}$$

and therefore, $G^{-1}G(\tau(x)) = \tau(x)$ for P_X-almost all x and thus,

$$\begin{aligned} &P\{\tau(X) \geq G^{-1}(G(\tau(x)))\} \\ &= P\{\tau(X) \geq \tau(x)\} \qquad P_X - a.s \end{aligned}$$

Finally, by taking the exponents on the expression of $S_X(x)$, we get

$$P\{\tau(X) \geq \tau(x)\} = e^{-S_X(x)}\,.$$

Conversely, assume that S satisfies (3) and (4), then set $\tau := S$. Using property (4) we have $P\{S(X) \geq S(x)\} = e^{-S(x)}$ and hence $S(x) = -\log P\{\tau(X) \geq \tau(x)\}$. □

Lemma 2.2 (Relations between outlier scores)**.** *For any $\delta \in \mathbb{R}_0^*$, for almost all $c \in S_X^\tau(\mathcal{X})$, we have*

$$P\{S_Y^\tau(Y) \geq c + \delta \,|S_X^\tau(X) \geq c\} \leq e^{-\delta}.$$

Proof. We have: $P\{S(Y) \geq c + \delta | S(X) \geq c\} = P\{S(Y) \geq c + \delta, S(X) \geq c\}/P\{S(X) \geq c\} \leq P\{S(Y) \geq c + \delta\}/P\{S(X) \geq c\} \leq \frac{e^{-c-\delta}}{e^{-c}}$, where we used that $P\{S(X) \geq c\} = e^{-c}$ holds for all $c \in S(\mathcal{X})$. □

Theorem 3.1 (Decomposition of target outlier score)**.** *The outlier score of an event x_n from any target variable X_n decomposes into the Shapley contribution of each of its ancestors plus itself, i.e., $S(x_n) = \sum_{j=1}^{n} \phi(j)$, where n is the number of ancestors of X_n including itself.*

The proof to this theorem follows directly from the efficiency property of Shapley values (Shapley, 1953). But we will still provide some intuitions here.

The IT outlier score of x_n in terms of noises is given by

$$S(x_n) \coloneqq -\log P\{g(\mathbf{N})) \geq g(\mathbf{n})\}.$$

When none of the noise terms $\mathbf{N}_\emptyset$ are randomized, the tail probability is surely 1.0, and hence we obtain the outlier score of zero, i.e.,

$$-\log P^{rd(\mathbf{N}_\emptyset)}\{g(\mathbf{N}) \geq g(\mathbf{n})\}) = 0. \tag{19}$$

When all noise terms $\mathbf{N}_\mathcal{U}$ are randomized, with $\mathcal{U} = \{1, \ldots, n\}$, according to the true joint distribution of noise terms $P_\mathbf{N}$, we obtain the usual outlier score of x_n, i.e.,

$$-\log P^{rd(\mathbf{N}_\mathcal{U})}\{g(\mathbf{N}) \geq g(\mathbf{n})\}) = S(x_n). \tag{20}$$

We can thus change from Eq. 19 to Eq. 20 step by step from 0 to $S(x_n)$ by randomizing more and more of the noise terms.

It is important to note here that for the randomization operation $rd(\mathbf{N}_\mathcal{I})$ with $\mathcal{I} \subset \mathcal{U}$, we need not randomize noise terms $\mathbf{N}_\mathcal{I}$ according to their true joint distribution $P_{\mathbf{N}_\mathcal{I}}$; we can use any joint distribution $\tilde{P}_{\mathbf{N}_\mathcal{I}}$. When $\mathcal{I} = \mathcal{U}$ (i.e., when we randomize all noise terms), it is crucial to use the true joint distribution $P_\mathbf{N}$ to obtain the total score $S(x_n)$.