Root Cause Analysis of Outliers in Unknown Cyclic Graphs

Daniela Schkoda
Amazon Research Tübingen, Germany
Technical University of Munich, Germany

Dominik Janzing
Amazon Research Tübingen, Germany

Abstract

We study the propagation of outliers in cyclic causal graphs with linear structural equations, tracing them back to one or several "root cause" nodes. We show that it is possible to identify a short list of potential root causes provided that the perturbation is sufficiently strong and propagates according to the same structural equations as in the normal mode. This shortlist consists of the true root causes together with those of its parents lying on a cycle with the root cause. Notably, our method does not require prior knowledge of the causal graph and yields encouraging results on simulated data and real data from biology and cloud computing.

1 INTRODUCTION

Root cause diagnosis is a fundamental challenge across a wide range of domains. Whether tracing the origin of a failure in a cloud-based microservice architecture, identifying a perturbed gene in a gene-regulatory network, locating the source of a shock in a financial system, or detecting faults in industrial manufacturing processes, the need to accurately pinpoint the source of disruptions is critical. For instance, in microservice-based cloud applications, when a failure occurs in one service, its effects can propagate across others, creating a cascade of symptoms and thousands of alerts. Root cause analysis (RCA) is thus critical for maintaining reliability and availability.

RCA closely aligns with the task of intervention targets estimation (ITE). While the framing differs slightly, in both cases, interventions on a few nodes in a causal system explain shifts in the overall distribution, and the goal is to identify these nodes.

Existing methods for both tasks largely fall into two categories. The first one includes graph-based methods, which require (or first learn) a ground truth causal graph (Kim et al., 2013; Assaad et al., 2023; Tao et al., 2024; Xie et al., 2024; Tonon et al., 2025). However, in practice, obtaining a complete ground-truth model for large systems is typically infeasible, and causal discovery methods to learn them from data often rely on restrictive and unverifiable assumptions. To address this, a second line of work performs RCA without relying on a known ground truth: For instance, Ikram et al. (2022) only partially reconstructs the underlying graph to effectively identify root causes, while Shan et al. (2019) leverages time series data for root cause detection. Varici et al. (2021, 2022); Wang et al. (2018); Ghoshal et al. (2019); Malik et al. (2024) learn the intervention targets by inspecting the difference of the precision matrices of the observational and interventional data, and Yang et al. (2024) by applying contrastive learning to recover exogenous noises that shift between both data sets. Rothenhäusler et al. (2015) requires data from at least three distinct environments and recovers the targets by joint matrix diagonalization of differences between the covariance matrices.

A shared limitation across all these works is the reliance on multiple anomalous samples. In contrast, Okati et al. (2025) introduce two RCA approaches that operate with only a single anomalous sample: *Smooth Traversal*, which again requires ground-truth knowledge of the causal DAG (but not of the conditional distributions or even SCMs), and *Score Ordering*, which avoids such knowledge but assumes that the unknown ground truth follows a polytree structure. The work most closely related to ours is Li et al. (2025), which assumes a linear structural equation model (SEM) and demonstrates that a permutation search combined with Cholesky decomposition of the covariance matrix can uncover the root cause, even in the case of a single interventional sample. However, their approach still requires an *acyclic* ground-truth structure, which is again

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

often violated in practice: gene regulatory networks routinely contain cyclic feedback loops, and microservice calls can form cycles, e.g., through repeated calls after failures (Genfer and Zdun, 2021; Sedghpour et al., 2023).

In this work, we address exactly this gap. We consider the task of root cause analysis in settings described by SEMs that may include cycles without requiring ground truth. Our approach is heavily inspired by Li et al. (2025) as it also uses the covariance matrix to derive a simple linear transformation that reveals root causes when applied to the vector of outlier observations. However, it is conceptually more straightforward and avoids the expensive combinatorial search needed there, making it more computationally efficient; see the appendix for a more detailed comparison of the ideas.

2 NOTATION AND MODEL

Our method assumes access to data under the normal regime, along with one anomalous sample for which we seek to identify the root causes of the abnormality. We model the p-variate distribution under the normal regime, denoted by P^X, using a linear cyclic SEM (Richardson, 1996; Spirtes et al., 2001; Bongers et al., 2021). Specifically, each feature X_i of $X = (X_1, \dots, X_p)$ satisfies

$$X_i = \sum_{j=1}^{p} a_{ij} X_j + N_i, \tag{1}$$

or, in matrix-vector-form,

$$X = AX + N, \tag{2}$$

where A is a coefficient matrix with zero diagonal entries, and N is a noise vector with mean zero and uncorrelated components[1]. We further assume that $I - A$ is invertible. The matrix A defines the system's causal structure, with a non-zero entry a_{ij} indicating a direct causal effect from variable X_j to variable X_i. This structure is naturally represented by a directed graph $\mathcal{G}$, where an edge $X_j \to X_i$ exists if and only if $a_{ij} \neq 0$. We assume that the same linear relations still hold in the anomalous mode up to a sparse additive perturbation term Δ:

$$\tilde{X} = A\tilde{X} + N' + \Delta, \tag{3}$$

where N' is an i.i.d. copy of N and Δ has non-zero entries for the root cause nodes, which we denote by $\mathcal{R}$.

[1]Unlike many works in causal discovery, we do not require the noise terms to be mutually independent; uncorrelatedness suffices.

We may also absorb the perturbation into the noise by defining $\tilde{N} = N' + \Delta$.

Note that modelling the perturbation as a change of the noise does not mean that we assume a perturbation affecting the exogenous noise in any "material" sense. For example, assume that the root cause manifests as a change in the causal mechanisms, $\tilde{X}_r = f(\mathrm{pa}(r), \tilde{N}_r)$, whereas in the normal mode, $X_r = \sum_{i \in \mathrm{pa}(r)} a_{ri} X_i + N_r$. This modification can be "absorbed" into a perturbation term δ such that it matches our framework: $\tilde{X}_r = \sum_{i \in \mathrm{pa}(r)} a_{ri} X_i + N'_r + \delta$, where $\delta = f(\mathrm{pa}(r), \tilde{N}_r) - \sum_{i \in \mathrm{pa}(r)} a_{ri} X_i - N_r$; compare also Von Kügelgen et al. (2023).

Within this setup, the task of RCA lies in identifying which entries of the noise vector in $\tilde{N}$ are extreme when compared to the original noise vector N. To make the connection between X and N more explicit, we can rearrange (2) as

$$X = (I - A)^{-1} N.$$

The entries of $(I - A)^{-1}$ capture *total causal effects* and are related to paths in the graph $\mathcal{G}$. A *path* from i_1 to i_k, denoted by $i_1 \rightsquigarrow i_k$, is a sequence of edges $(i_1 \to i_2, i_2 \to i_3, \dots, i_{k-1} \to i_k)$ such that neither nodes nor edges are repeated, except possibly for a return to the starting node, $i_1 = i_k$, forming a cycle. We allow paths of length 0. The following standard result will become important later:

Lemma 2.1 (Path matrix, Lauritzen (1996))**.** *The total causal effect from j on i multiplies direct causal effects along paths from j to i:*

$$((I-A)^{-1})_{ij} = \frac{1}{\det(I-A)} \sum_{\pi : j \rightsquigarrow i} \prod_{k \to h \in \pi} a_{hk} \cdot \left(1 + \sum_{\substack{(C_1, \dots, C_q) \in \mathcal{C}_\pi, \\ q \geq 1.}} (-1)^q \prod_{m=1}^{q} \prod_{k \to h \in C_m} a_{hk} \right),$$

where $\mathcal{C}_\pi = \{(C_1, \dots, C_q) :$ collection of disjoint cycles in G not using any node in $\pi; q \in \mathbb{N}_0.\}$.

While the proof is provided in the supplementary material, especially for acyclic graphs, the intuition is quite straightforward: the total causal effect accounts for all paths - direct and indirect - through which one variable can influence another. Each path represents a potential causal route, and its contribution is the product of the direct effects along it. This motivates to call $(I - A)^{-1}$ the *path matrix*. As our primary interest lies in the sparsity pattern of the matrix, we omit the explicit determinant formula here and refer the reader to the proof of Lemma 4.5 in Amendola et al. (2020).

Many of the results proven in this paper assume *genericity* of the parameters A and Σ_{NN}, meaning that the results might fail on a Lebesgue measure zero set of matrices A (consistent with a fixed graph $\mathcal{G}$) and choices of Σ_{NN}. This assumption rules out degenerate cases, such as interventions on two nodes that cancel each other. For example, in the chain

$$X_1 \to X_2 \to X_3$$

with all edge weights being one and two root causes X_1, X_2 with $\Delta_1 = 1, \Delta_2 = -1$, we obtain $\tilde{X}_2 = N_1 + N_2 + \Delta_1 + \Delta_2 = N_1 + N_2$, which is not anomalous although X_2 is a root cause. Our method cannot identify X_2 as the root cause in this pathological case. A precise characterization of these exceptional situations, as well as most proofs, are deferred to the appendix. Finally, we denote the children, parents, descendants, and ancestors of a node X_i by $\mathrm{ch}(X_i), \mathrm{pa}(X_i), \mathrm{de}(X_i)$, and $\mathrm{an}(X_i)$, respectively.

Having established the model, we discuss how Root Cause Analysis can be solved in this framework.

3 ROOT CAUSE ANALYSIS VIA INVARIANT STRUCTURAL EQUATIONS

3.1 Causally Sufficient Cyclic Model

We first consider the case of a single root cause, that is $\Delta = (0, \ldots, 0, \delta, 0, \ldots, 0)$ with $\delta \in \mathbb{R}$ representing a significant perturbation at the root cause r. A key ingredient of our method is the precision matrix, which can be conveniently expressed in terms of the path matrix. Specifically,

$$\Sigma_{XX} = (I - A)^{-1} \Sigma_{NN} (I - A)^{-T},$$

where Σ_{NN} denotes the covariance matrix of the noise, which is diagonal by assumption: $\Sigma_{NN} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. Inverting this expression yields the precision matrix:

$$\Theta_{XX} := \Sigma_{XX}^{-1} = (I - A)^T \Theta_{NN} (I - A),$$

where Θ_{NN} denotes the precision matrix of N. To see how this relationship supports root cause analysis, consider

$$\xi := \Theta_{XX} \tilde{X} = \Theta_{XX} (I - A)^{-1} (\Delta + N').$$

Inspecting the first summand, we observe

$$\begin{aligned} \Theta_{XX}(I - A)^{-1}\Delta &= (I - A)^T \Theta_{NN} (I - A)(I - A)^{-1}\Delta \\ &= (I - A)^T \Theta_{NN} \Delta \\ &= \frac{1}{\sigma_r^2} \delta (I - A)^T e_r =: \xi', \end{aligned} \tag{4}$$

where e_r denotes the rth canonical basis vector. Since the rth column of $(I - A)^T$ is non-zero precisely at the positions of r and its parents, also ξ' is non-zero only at these positions. Turning to the second summand, $\Theta_{XX}(I - A)^{-1} N'$, note that it has the same distribution as $\Theta_{XX} X$ since $N' \overset{d}{=} N$. Therefore, by evaluating which ξ_i has the same distribution as $\Xi_i := (\Theta_{XX} X)_i$, we can separate the root cause and its parent from the remaining nodes. Especially, if δ has high absolute value, we can expect the entries of the root cause and its parents to differ significantly from all other entries. Overall, we derived:

> Applying the precision matrix to the outlier vector leaves only the root causes and their parents as extreme, thereby revealing them.

Theorem 3.1 (Shortlist of root causes)**.** *For generic A and Σ_{NN}, all entries in ξ_i follow their usual distribution, except the root causes and potentially their parents:*

$$\mathcal{R} \subseteq \{\xi_i \overset{d}{\neq} \Xi_i\} \subseteq \mathcal{R} \cup \mathrm{pa}(\mathcal{R}).$$

To exemplify this result, we consider a simple two-dimensional toy example.

Example 1. *Assume the data-generating process is described by the causal DAG $X_1 \to X_2$, and both X_1 and X_2 are standard normally distributed. The structural equations are*

$$\begin{aligned} X_1 &= N_1, \\ X_2 &= \rho X_1 + N_2, \end{aligned}$$

where ρ denotes the Pearson correlation between X_1 and X_2. Further, the covariance matrix and its inverse read

$$\Sigma_{XX} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad \Theta_{XX} = \frac{1}{1 - \rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

If X_1 is the root cause, i.e., $\Delta = (\delta, 0)^T$, then $\tilde{X} \overset{d}{=} X + (\delta, \delta \cdot \rho)^T$. Multiplying $\tilde{X}$ by Θ_{XX} yields

$$\Theta_{XX} \tilde{X} \overset{d}{=} \Theta_{XX} X + \Theta_{XX} \delta \begin{pmatrix} 1 \\ \rho \end{pmatrix} = \Theta_{XX} X + \begin{pmatrix} \delta \\ 0 \end{pmatrix},$$

where only the first entry is extreme for $\delta \to \infty$. On the other hand, if X_2 is the root cause, we have $\Delta = (0, \delta)^T$ and $\tilde{X} \overset{d}{=} X + (0, \delta)^T$ and obtain

$$\begin{aligned} \Theta_{XX} \tilde{X} &\overset{d}{=} \Theta_{XX} X + \Theta_{XX} \begin{pmatrix} 0 \\ \delta \end{pmatrix} \\ &= \Theta_{XX} X + \delta \cdot \frac{1}{1 - \rho^2} \begin{pmatrix} -\rho \\ 1 \end{pmatrix}, \end{aligned}$$

where both entries of the vector get extreme for high δ, hence we can tell the difference between the two scenarios.

This result gives a first short list of candidate root causes. To restrict the list further, we can make use of the fact that only the descendants of the root causes are anomalous.

Theorem 3.2 (Even shorter list of root causes)**.** *Assuming generic parameters A, Σ_{NN}, an abnormal distribution of both, ξ and $\tilde{X}$, occurs only for the root cause and for its parents that are simultaneously descendants:*

$$\mathcal{R} \subseteq \{\xi_i \overset{d}{\neq} \Xi_i \text{ and } \tilde{X}_i \overset{d}{\neq} X_i\} \subseteq \mathcal{R} \cup (\mathrm{pa}(\mathcal{R}) \cap \mathrm{de}(\mathcal{R})).$$

Note that Theorem 3.2 allows for unique identification of the root cause (if there is only one) in acyclic graphs as the parents of the root cause cannot be descendants at the same time.

Another challenge is dealing with latent variables – factors that influence the system but are not measured.

3.2 Latent Cyclic Model

Latent variables can be incorporated into the model by extending the variable set. We consider a random vector $Z = (X, L)$, where $X = (X_1, \dots, X_p)$ collects the observed variables and $L = (L_1, \dots, L_q)$ the latent variables. Overall, Z captures all variables relevant to the underlying causal system and satisfies a linear SEM

$$Z = AZ + N \quad \text{and} \quad \tilde{Z} = A\tilde{Z} + N' + \Delta, \tag{5}$$

which can be partitioned as

$$\begin{pmatrix} X \\ L \end{pmatrix} = \begin{pmatrix} A_{XX} & A_{XL} \\ A_{LX} & A_{LL} \end{pmatrix} \begin{pmatrix} X \\ L \end{pmatrix} + \begin{pmatrix} N_X \\ N_L \end{pmatrix},$$

similarly for the interventional equation. We aim to express the system in terms of X only, by incorporating the latent components into the noise term, at the cost of losing noise uncorrelatedness. To this end, denote

$$S = A_{XX} + A_{XL}(I - A_{LL})^{-1} A_{LX},$$

and let D the diagonal matrix of the same shape with entries $d_{ii} = \frac{1}{1-s_{ii}}$. Further define

$$\begin{aligned} \bar{A}_{ij} &= (DS)_{ij} \text{ for } i \neq j, \text{ and 0 otherwise,} \\ \bar{N} &= D(N_X + A_{XL}(I - A_{LL})^{-1} N_L), \\ \bar{\Delta} &= D(\Delta_X + A_{XL}(I - A_{LL})^{-1} \Delta_L). \end{aligned}$$

Lemma 3.3 (Remove nodes)**.** *If P^Z satisifies the SEM (5), then P^X and $P^{\tilde{X}}$ fulfill*

$$X = \bar{A}X + \bar{N} \quad \text{and} \quad \tilde{X} = \bar{A}\tilde{X} + \bar{N}' + \bar{\Delta}.$$

Proof. We show the observational SEM; the interventional one works analogously. As a first step, we prove that

$$X = SX + M, \tag{6}$$

where $M = N_X + A_{XL}(I - A_{LL})^{-1} N_L$. Note that

$$I - S = \big((I - A_{XX}) - A_{XL}(I - A_{LL})^{-1} A_{LX}\big)$$

is precisely the Schur complement (Schur, 1917) of the block $I - A_{LL}$ in $I - A$, yielding that

$$((I - A)^{-1})_{XX} = (I - S)^{-1},$$

and,

$$((I - A)^{-1})_{XL} = (I - S)^{-1} A_{XL} (I - A_{LL})^{-1}.$$

Combining both yields

$$\begin{aligned} (I - S)^{-1} M =& ((I - A)^{-1})_{XX} N_X \\ &+ (I - S)^{-1} A_{XL} (I - A_{LL})^{-1} N_L \\ =& ((I - A)^{-1})_{XX} N_X + ((I - A)^{-1})_{XL} N_L \\ =& X, \end{aligned}$$

which confirms the intermediate equation (6). While this representation accurately describes X, it does not yet constitute a valid SEM since S may feature nonzero diagonal entries. To enforce zero diagonals, we apply a rescaling. Writing out the i-th component,

$$X_i = a'_{ii} X_i + \sum_{j \neq i} a'_{ij} X_j + M_i,$$

then subtracting $a'_{ii} X_i$, and rescaling both sides gives the SEM stated in the lemma:

$$X_i = \frac{1}{1 - a'_{ii}} \left(\sum_{j \neq i} a'_{ij} X_j + M_i \right).$$

□

Example 2. *For the SEM depicted in Figure 1a, that is,*

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ L_1 \\ L_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & a_{14} & 0 \\ 0 & 0 & 0 & 0 & a_{25} \\ a_{31} & a_{32} & 0 & a_{34} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ a_{51} & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ L_1 \\ L_2 \end{pmatrix} + \begin{pmatrix} N_1 \\ N_2 \\ N_3 \\ N_4 \\ N_5 \end{pmatrix}$$

the projected SEM, visualized in Figure 1b, reads

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ a_{25}a_{51} & 0 & 0 \\ a_{31} & a_{32} & 0 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} + \begin{pmatrix} N_1 + a_{14}L_1 \\ N_2 + a_{25}L_2 \\ N_3 + a_{34}L_3 \end{pmatrix}.$$

Here, each latent variable is absorbed into the noise terms of its children, making the noises of nodes with a common latent parent correlated. Moreover, removing L_2 from the system introduces a direct link $X_1 \to X_2$, reflecting what was previously an indirect, mediated path.

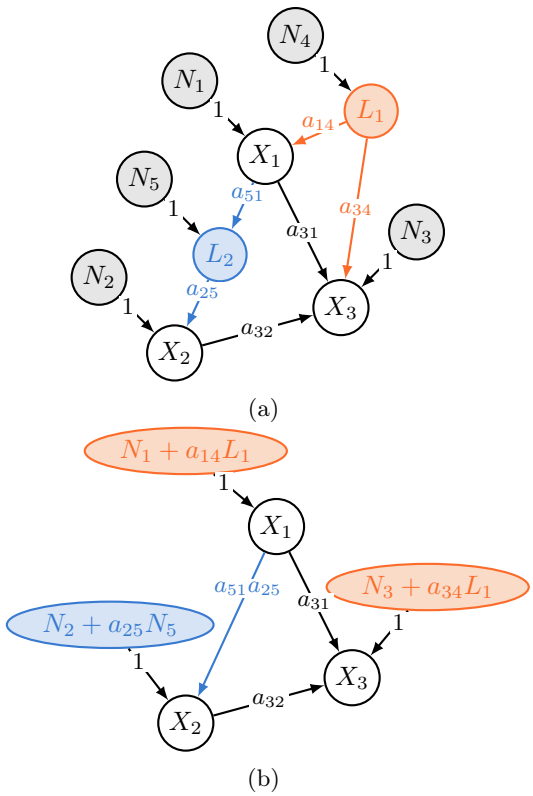

Figure 1: Projecting a linear SEM to an SEM for only the observed nodes.

In general graphs, $\bar{N}_i$ and $\bar{N}_j$ are correlated whenever they have a joint latent ancestor that connects to both through a path in which all inner nodes are latent, called *latent path* henceforth, and represented by $Z_k \rightsquigarrow Z_l$. Similarly, $\bar{\Delta}_i \neq 0$ for all observed root causes and all i which lie downstream of a latent root cause through a path consisting only of latent nodes. We refer to all these nodes i as *observable root causes*, aligning with the intuition that we cannot identify a root cause that was never observed. Instead, the observable root causes are the earliest observed nodes infected by the anomalies.

From the projected SEM over X, the precision matrix can be computed as:

$$(\Sigma_{XX})^{-1} = (I - \bar{A})^T (\Sigma_{\bar{N}\bar{N}})^{-1} (I - \bar{A}).$$

In contrast to the case without latent variables, the noise precision matrix $(\Sigma_{\bar{N}\bar{N}})^{-1}$ is not necessarily diagonal due to correlations introduced by marginalizing out latent nodes. Instead, its sparsity pattern can be described as follows.

Lemma 3.4 (Sparse noise precision matrix)**.** $(\Theta_{\bar{N}\bar{N}})_{ij} = 0$ *if there exists no sequence of latent paths*

$$X_i \leftsquigarrow L_{l_1} \rightsquigarrow X_{k_1} \leftsquigarrow L_{l_2} \rightsquigarrow X_{k_2} \leftsquigarrow \cdots \rightsquigarrow X_j,$$

with $l_\iota \in [q], k_\iota \in [p]$.

Following Boege et al. (2025), we call such sequences of paths *zig-zag-structures* and denote them by $X_i \leftsquigarrow \cdots \rightsquigarrow X_j$. For example, for $G = X_1 \leftarrow L_1 \rightarrow X_2 \leftarrow L_2 \rightarrow X_3$,

$$\bar{N} = \begin{pmatrix} N_1 + a_{14}L_1 \\ N_2 + a_{24}L_1 + a_{25}L_2 \\ N_3 + a_{35}L_2 \end{pmatrix}.$$

While $\bar{N}_1$ and $\bar{N}_3$ are marginally independent, conditioning on $\bar{N}_2$ introduces dependence.

As before, we define the score vector as

$$\begin{aligned} \xi &= \Theta_{XX}\tilde{X} \\ &= (I - \bar{A})^T \Theta_{\bar{N}\bar{N}} \bar{\Delta} + (\Sigma_{XX})^{-1}(I - \bar{A})^{-1}\bar{N}' \end{aligned}$$

Since the first term dominates for large perturbations, we are again interested in its support. From the Lemma, $\Theta_{\bar{N}\bar{N}}\bar{\Delta}$ is non-zero for all X_j having a zig-zag-connection to an observable root cause. Furthermore, $\bar{A}$ encodes the child-parent relationships in the projected graph, that is, $a_{ij} \neq 0$ if and only if $X_j \rightsquigarrow X_i$.

Theorem 3.5 (Shortlist of root causes)**.** *If A and Σ_{NN} are generic, then $\xi_i \overset{d}{\neq} (\Theta_{XX}X)_i$ if and only if X_i is connected to an observable root cause X_k through the structure*

$$X_i \rightsquigarrow X_j \leftsquigarrow \cdots \rightsquigarrow X_k.$$

As before, this shortlist can be restricted further by removing non-anomalous nodes. Besides the obvious application of RCA in systems including latent variables, this theory also enables tracing the routes through which anomalies propagate.

Example 3 (Find root cause propagation route)**.** *Assume that G is causally sufficient and that there is a unique root cause X_r. In a first iteration, we apply our procedure to identify it. Then, a natural next step is identifying which nodes were subsequently affected, i.e., infer the children of X_r. To this end, we remove X_r from the data, making it a latent node. On this reduced data set, we apply our method again, yielding a new score vector $\breve{\xi} \in \mathbb{R}^{p-1}$. From Theorem 3.5, it is non-zero precisely at the positions i where there is a route*

$$X_i \rightsquigarrow X_j \leftsquigarrow \cdots \rightsquigarrow X_k \leftsquigarrow X_r.$$

Since X_r is the sole latent node in this zigzag-structure, the latter collapses to either

$$X_i = X_j \leftarrow X_r \rightarrow \cdots \leftarrow X_r \rightarrow X_k \leftarrow X_r,$$

$$X_i \rightarrow X_j \leftarrow X_r \rightarrow \cdots \leftarrow X_r \rightarrow X_k \leftarrow X_r,$$

or to

$$X_i \rightarrow X_r \rightarrow X_j \leftarrow X_r \rightarrow \cdots \leftarrow X_r \rightarrow X_k \leftarrow X_r.$$

Thus, X_i is a child of X_r, or a parent of either X_r or of another child of X_r. So, again, we have a candidate set for the children. Reiterating this approach, we can find supersets of all routes through which the anomaly propagated, which, however, can grow exponentially.

4 IMPLEMENTATION WITH FALSE DISCOVERY RATE CONTROL

Turning to the practical implementation, we assume to have access to one anomalous sample $\tilde{X}$, alongside m i.i.d samples $X^{(1)}, \ldots, X^{(m)} \sim P^X$, collected into a data matrix $\boldsymbol{X} \in \mathbb{R}^{m \times p}$. We first focus on Theorems 3.1 and 3.5, which identify candidate root causes as the nodes that violate the hypothesis

$$H_0 : \xi_i \overset{d}{=} \Xi_i.$$

For simplicity, we henceforth refer to these violating nodes as *positive*, while the others are termed *negative*. A straightforward approach to assess H_0 is to use the empirical p-value, defined as the proportion of samples $|\Xi_i|$ exceeding $|\xi_i|$. While this guarantees valid Type I error control for incorrectly classifying a *single* negative node as positive, it is challenging to use p-values with arbitrary dependencies to control the total number of false positives within a False Discovery Rate Control (FDRC) framework. Therefore, we instead turn to *e-values*, where an e-value is a non-negative random variable e satisfying $\mathbb{E}(e) \leq 1$ under H_0. High values of e indicate stronger evidence against the null hypothesis. Unlike p-values, e-values are particularly well-suited for multiple testing and aggregation, making them effective for FDRC; see Vovk and Wang (2021); Grünwald et al. (2024) for further details. In our setting,

$$e_i = \frac{\xi_i^2}{(\Theta_{XX})_{ii}} \tag{7}$$

is a valid e-value since under H_0, ξ_i is centered and has covariance $(\Theta_{XX})_{ii}$, implying $\mathbb{E}(e) = 1$. Note that this e-value coincides with a commonly used outlier score, namely the squared Z-Score defined via

$$\text{Z-Score}(\xi_i; \boldsymbol{\Xi}_i)^2 = \left(\frac{\xi_i - \mu_{\boldsymbol{\Xi}_i}}{\sigma_{\boldsymbol{\Xi}_i}} \right)^2.$$

The equivalence follows from the fact that $\boldsymbol{\Xi}$ is centered and has covariance $\Sigma_{\boldsymbol{\Xi}\boldsymbol{\Xi}} = \Theta_{XX}$. We denote by $e_{[1]}, \ldots, e_{[p]}$ the e-values $e_1, \ldots, e_p$ arranged in decreasing order. Then, the e-value–based FDRC approach of Wang and Ramdas (2022) yields:

Lemma 4.1 (Benjamini-Hochberg for e-values)**.** *Let $\alpha > 0$ be a desired FDR control level and define k^* as the largest index k such that $ke_{[k]} \geq 1/\alpha$. Then, selecting the k^* nodes corresponding to $e_{[1]}, \ldots, e_{[k^*]}$ as candidate root causes yields*

$$FDR = \mathbb{E}\left(\frac{\textit{false positives}}{\textit{total number of rejections}} \right) \leq \alpha.$$

where the expectation is over the variables $X^{(1)}, \ldots, X^{(m)} \sim P^X, \tilde{X} \sim P^{\tilde{X}}$.

Finally, to incorporate the fact that the root cause is anomalous as stated in Theorem 3.2, we assign a root cause score of zero to all nodes not qualifying as outliers, as measured by their squared Z-Scores falling below a threshold τ. The resulting procedure is summarized in Algorithm 1. The computational complexity of our algorithm breaks down to its bottleneck: estimation of the precision matrix. Using Graphical Lasso (Friedman et al., 2008) incurs a cost of $O(Tp^3)$, where T is the iteration limit. A simpler option is to invert the sample covariance, with complexity $O(mp^2 + p^3)$. In both cases, space complexity is $O(mp + p^2)$. Further estimation methods are discussed in Appendix D.2.

Algorithm 1 Cyclic RCA

Require: Normal samples $\boldsymbol{X} \in \mathbb{R}^{m \times p}$, anomalous sample $\tilde{X} \in \mathbb{R}^p$, threshold τ to characterize outliers.

$\hat{\Theta} \leftarrow$ estimated precision matrix of $\boldsymbol{X}$.
$\hat{\xi} \leftarrow \hat{\Theta}\tilde{X}$.
$\hat{\boldsymbol{\Xi}} \leftarrow \hat{\Theta}\boldsymbol{X}$
for $i = 1$ to p **do**
  $s_i \leftarrow \text{Z-Score}(\xi_i; \boldsymbol{\Xi}_i)^2$
  if $\text{Z-Score}(\tilde{X}_i; X_i)^2 < \tau$ **then**
    $s_i \leftarrow 0$.
return RCA scores $s = (s_1, \ldots, s_p)$.

Beyond the linear algebra formulation presented here, our method admits alternative (but slightly more involved) derivations, provided in appendix B. One such view links our score $\text{Z-Score}(\xi_i; \boldsymbol{\Xi}_i)$ to linear regression. Furthermore, under certain assumptions, our null hypothesis $\xi_i \overset{d}{=} \Xi_i$ can be rephrased as a conditional independence test, after introducing a binary indicator variable distinguishing anomalous from normal samples. This perspective clarifies the relation to conditional independence–based RCA (Kocaoglu et al., 2019; Jaber et al., 2020; Mooij et al., 2020; Ikram et al., 2022), while also highlighting a key difference: owing to our parametric assumptions, we can perform the test with only a single anomalous sample.

Table 1: Means and standard deviations of the compute times in seconds for all methods. Cyclic (G) indicates Cyclic (Graphical Lasso) and (I) Inverse Covariance.

METHOD	COMPUTE TIME $p = 50$	$p = 100$
Cyclic (G)	2.82±1.96	781.93±463.66
Cyclic (I)	0.13±0.04	1.64±0.36
Cholesky	1168.04±1078.36	11853.67±10384.63
Z-Score	0.00±0.01	0.01±0.03
BARO	0.08±0.03	0.30±0.17

5 EXPERIMENTS

More details on the experiments, the entailed statistical challenges, and further simulation studies can be found in the appendix. All experiments were executed on a MacBook Pro (Apple M1, 36 GB RAM) or an internal cluster with 80 logical cores and 754 GB RAM, and the code is available at https://github.com/DanielaSchkoda/CyclicRCA.

5.1 Simulation Studies

We assess our method by comparing it against the Cholesky-decomposition-based approach introduced in Li et al. (2025) and the BARO method proposed in Pham et al. (2024a), which achieved the strongest performance in the benchmark study of Pham et al. (2024b). The latter evaluates 21 RCA methods with respect to their effectiveness in microservice environments. In addition, as in the study of Li et al. (2025), we consider a Z-Score baseline that directly employs Z-Score$(x_i; \boldsymbol{X}_i)^2$ as the root cause score.[2] For the Graphical Lasso variant of our algorithm, we fix the regularization parameter to 0.1; additional results for different values of the parameter are provided in Appendix Figure 5.

To generate data, we first simulate graphs with 50 or 100 observed nodes across four settings: cyclic or acyclic, each either with or without five additional latent nodes. Edges are sampled independently at random with a probability calibrated such that the expected node degree is 3. Given the graph, data with $m = 10 \cdot p$ is synthetically generated via the corresponding linear SEM, where the non-zero entries of the coefficient matrix B are uniformly sampled from $[-2, -0.5] \cup [0.5, 2]$. Noise follows one of four distributions (Gaussian, uniform, exponential, lognormal). The number of root causes varies between 1 and 4, and intervention strength δ ranges from 5 to 20.

[2]Both, BARO and the Z-Score baseline simply rank nodes by their anomalousness. This approach offers no theoretical guarantees for detecting the actual root cause except for the true graph being a collider-free polytrees. Nevertheless, some justification for inferring highly scored nodes as root cause candidates has been provided by Okati et al. (2025) when scores are obtained from p-values.

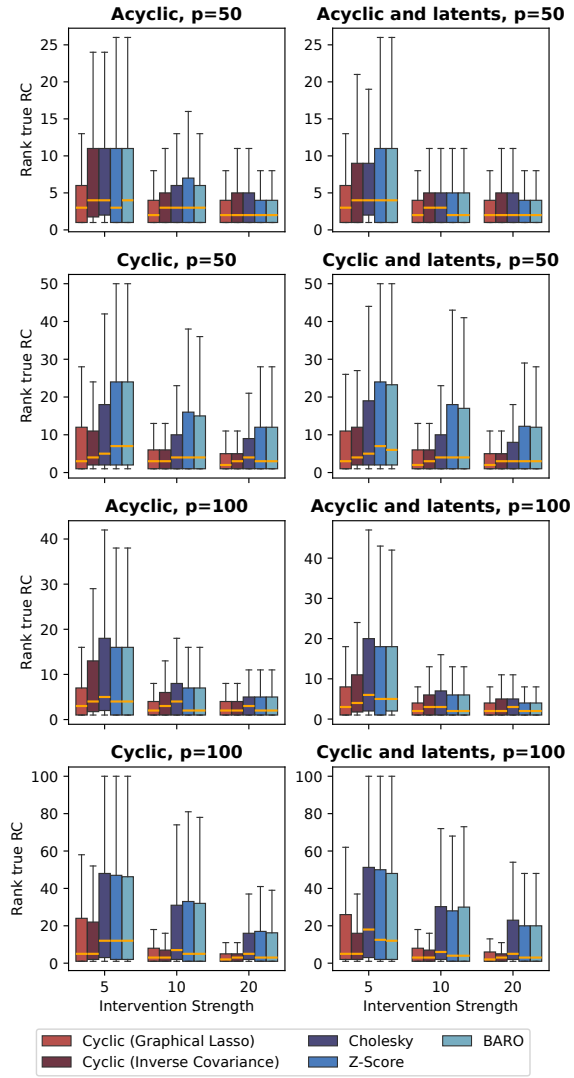

Figure 2: The boxplots illustrate the rank of the true root cause score among the scores of all nodes. Cyclic (Graphical Lasso) achieves the strongest performance, followed by Cyclic (Inverse Covariance). Cholesky performs similarly to Cyclic for the acyclic settings with $p = 50$, but worse for the remaining settings. Z-Score and BARO yield weaker results, especially in cyclic settings, probably because they simply treat the most anomalous node as the root cause.

Figure 2 reports the aggregated results across all noise distributions and numbers of root causes. Each box summarizes 400 replications with varying random seeds. The median is indicated by the yellow line, the box rep-

Table 2: Overview of the number of features (p), observational samples (m_{obs}), interventional samples (m_{int}), and number of interventional datasets (n_{int}) per benchmark. The datasets occur in the same order as in Figure 3. Differences in the number of features stem entirely from preprocessing, while variations in sample sizes are largely inherited from the original datasets.

	p	m_{obs}	m_{int}	n_{int}
C10	10	3000 – 3060	940 – 1000	100
C50	50	3000 – 3060	940 – 1000	100
R10	10	2000 – 2060	1940 – 2000	100
R50	50	2000 – 2060	1940 – 2000	100
S1	28 – 31	265 – 358	209 – 297	50
S2	225 – 277	360 – 420	187 – 361	125
OB	44 – 56	360 – 2160	301 – 2101	125
TT	533 – 828	360 – 540	301 – 481	125
GE	19737	364	1	57
PS	10 – 20	542 – 1575	4 – 5	24

resents the interquartile range (IQR), and the whiskers extend to 1.5 times the IQR. Table 1 shows the corresponding compute times. Both Cyclic methods match Cholesky in accuracy for $p = 50$ in the acyclic case, but are substantially faster and also more accurate for $p = 100$ and all cyclic cases. Detailed breakdowns by noise type and number of root causes, as well as experiments with varying expected node degree, are deferred to the appendix.

5.2 Real-World Experiemnts

For a case study on real-world data, we use three benchmark collections. Each benchmark consists of one or more datasets from the normal mode alongside data from several incidents. In contrast to the simulation setup described above, often, multiple interventional samples are available.

For instance, the PetShop benchmark (Hardt et al., 2024) contains latency, request, and availability metrics for 41 services in a microservice-based application that mimics a pet adoption platform. The metrics are collected at 5-minute intervals. In addition to the normal data, the dataset includes several injected performance issues affecting different services. Each issue lasts approximately half an hour, resulting in 5 interventional samples per incident. The dataset contains substantial missing data and exhibits strong collinearity among features. We therefore preprocess the data, which reduces the number of features to less than half of the original set; see the appendix for details. An overview of the sample and feature sizes after the preprocessing is provided in Table 2.

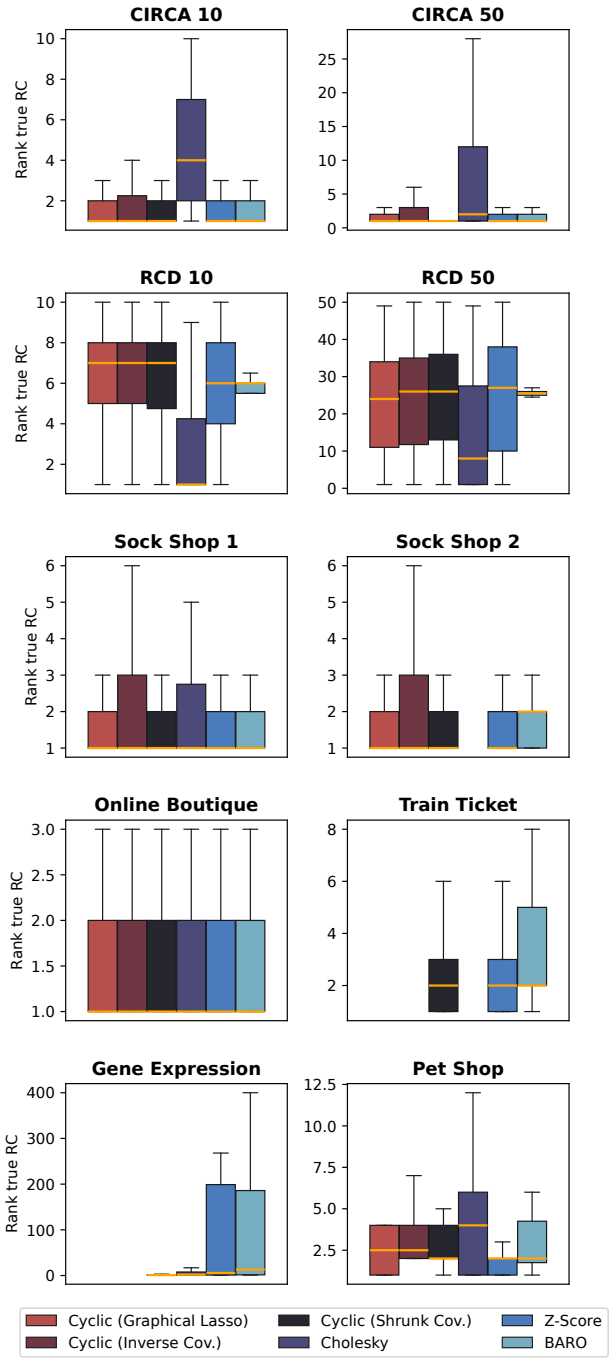

Figure 3: Results on the real-world datasets. All methods struggle with the discrete RCD dataset, although Cholesky still performs best. In contrast, on the continuous CIRCA data, Cholesky performs worse. On the gene expression data, the Cyclic and Cholesky methods clearly outperform the anomaly scorers. Otherwise, performance is largely similar across methods.

In addition, following Li et al. (2025), we use the gene expression dataset collected by Yépez et al. (2022). The dataset contains gene expression measurements from 57

patients with a Mendelian disease, together with samples from healthy individuals that serve as a baseline. The baseline data is high-dimensional, with substantially more features than observations. Consequently, neither Graphical Lasso nor empirical covariance estimation can be applied for precision matrix estimation. Instead, we employ the Shrunk Covariance estimator. Also, the standard Cholesky algorithm is no longer applicable, so we rely on their high-dimensional version (Li et al., 2025, Algorithm 3)).

Finally, we evaluate our approach on the RCAEval benchmark (Pham et al., 2024b), which comprises eight datasets. Four of these are synthetic benchmarks (RCD 10, RCD 50, CIRCA 10, CIRCA 50) generated from random directed acyclic graphs. The first two datasets contain continuous variables, whereas the latter two consist of discrete variables; see Pham et al. (2024b) for details. The remaining datasets consist of metrics collected from real-world microservice applications at one-second intervals. These include the e-commerce system Sock Shop (in two variants), the Online Boutique application with 12 services supporting browsing and purchasing, and Train Ticket, a large-scale train booking system comprising 64 services. Compared to the other systems, Train Ticket exhibits longer, more complex failure propagation paths. Since multiple metrics are recorded per service, the number of features p exceeds the number of services. For each application, different types of incidents, such as high CPU load, are injected at various services. Cholesky is omitted for Sock Shop 2 and Train Ticket due to excessive runtime. For Train Ticket and Gene Expression, precision matrices are estimated exclusively using the Shrunk Covariance estimator, as the number of features exceeds the number of samples.

Figure 3 shows the rank of the true root causes across all benchmark datasets. Following Pham et al. (2024a), when multiple interventional samples are present, we first compute a score for each sample individually and then use the maximum of these scores. Since this approach is computationally prohibitive for the Cholesky method, there, the interventional samples are aggregated using the mean beforehand.

In the appendix, for the RCAEval benchmark, we include a more detailed breakdown of the results, distinguishing the different fault types and reporting the Avg@5 metric rather than the rank of the true root cause. Doing so, we mimic the setup in Pham et al. (2024b) and show that our method outperforms 18 of the 21 methods considered there. We also present experiments on semi-synthetic data illustrating that our method remains effective under moderate violations of the linearity assumption. Finally, we compare our approach with the ICE method Rothenhäusler et al. (2015), since it can identify intervention targets even for cyclic graphs. However, it requires data from at least three distinct environments to be available.

6 CONCLUSION AND LIMITATIONS

We introduce a simple yet effective approach for RCA in linear SEMs, which, to our knowledge, is the only method that provably works for unknown cyclic graphs using a single anomalous sample. In addition, it remains applicable in the presence of multiple root causes or latent variables, given that the graph is sufficiently sparse. It relies crucially on the assumption that, except for the root causes, the SEM describing the normal regime also holds for the anomalous regime. Whether this assumption holds depends heavily on the respective use case; we have shown a real-world example where the method at least clearly outperforms chance level. In simulation studies, our method outperforms existing approaches while offering faster computation. There are two complementary ways to justify the assumption that only one or a few structural equations change in the anomalous regime while the others remain unchanged. On the one hand, one may argue that the task of root cause analysis would otherwise be ill-defined, since all nodes could be considered root causes (Budhathoki et al., 2022). On the other hand, the *Independence of Mechanisms Principle* (see Sections 2.1 and 2.2 in Peters et al. (2017)) posits that mechanisms vary independently, making it unlikely that *rare* changes occur simultaneously at the same time stamp or for the same statistical unit. Moreover, the *Sparse Mechanism Shift Hypothesis* (Schölkopf et al., 2021) explicitly assumes that distributional shifts can be attributed to changes in only a small number of mechanisms.

Acknowledgements

We thank the reviewers for the insightful discussions and helpful feedback, which substantially improved this paper. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 883818). DS acknowledges support by the DAAD program Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research.

References

C. Amendola, P. Dettling, M. Drton, F. Onori, and J. Wu. Structure learning for cyclic linear causal models. In J. Peters and D. Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 999–1008. PMLR, 03–06 Aug 2020.

C. K. Assaad, I. Ez-Zejjari, and L. Zan. Root cause identification for collective anomalies in time series given an acyclic summary causal graph with loops. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8395–8404. PMLR, 25–27 Apr 2023.

P. Blöbaum, P. Götz, K. Budhathoki, A. A. Mastakouri, and D. Janzing. DoWhy-GCM: An extension of DoWhy for causal inference in graphical causal models. *Journal of Machine Learning Research*, 25 (147):1–7, 2024. URL http://jmlr.org/papers/v25/22-1258.html.

T. Boege, M. Drton, B. Hollering, S. Lumpp, P. Misra, and D. Schkoda. Conditional independence in stationary distributions of diffusions. *Stochastic Processes and their Applications*, 184:104604, 2025.

S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885 – 2915, 2021.

K. Budhathoki, L. Minorics, P. Bloebaum, and D. Janzing. Causal structure-based root cause analysis of outliers. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2357–2369. PMLR, 17–23 Jul 2022.

P. Dawid. Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9(1):39–77, 2021.

P. Forré and J. M. Mooij. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2018.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

P. Genfer and U. Zdun. Identifying domain-based cyclic dependencies in microservice APIs using source code detectors. In S. Biffl, E. Navarro, W. Löwe, M. Sirjani, R. Mirandola, and D. Weyns, editors, *Software Architecture*, pages 207–222, Cham, 2021. Springer International Publishing.

A. Ghoshal, K. Bello, and J. Honorio. Direct learning with guarantees of the difference DAG between structural equation models. arXiv preprint arXiv:1906.12024, 2019.

P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 03 2024.

M. Hardt, W. R. Orchard, P. Blöbaum, E. Kirschbaum, and S. Kasiviswanathan. The petshop dataset — finding causes of performance issues across microservices. In *Proceedings of the Third Conference on Causal Learning and Reasoning (CLeaR)*, volume 236 of *Proceedings of Machine Learning Research*, pages 957–978. PMLR, 2024.

A. Ikram, S. Chakraborty, S. Mitra, S. K. Saini, S. Bagchi, and M. Kocaoglu. Root cause analysis of failures in microservices through causal discovery. In *Advances in Neural Information Processing Systems*, 2022.

A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9551–9561. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6cd9313ed34ef58bad3fdd504355e72c-Paper.pdf.

M. Kim, R. Sumbaly, and S. Shah. Root cause detection in a service-oriented architecture. In *Proceedings of the ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '13, page 93–104, New York, NY, USA, 2013. Association for Computing Machinery.

M. Kocaoglu, A. Jaber, K. Shanmugam, and E. Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

S. L. Lauritzen. *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, 1996. ISBN 978-0198522195.

J. Li, B. B. Chu, I. F. Scheller, J. Gagneur, and M. H. Maathuis. Root cause discovery via permutations and Cholesky decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf066, 10 2025. ISSN 1369-7412. doi: 10.1093/jrsssb/qkaf03.66. URL https://doi.org/10.1093/jrsssb/qkaf066.

V. Malik, K. Bello, A. Ghoshal, and J. Honorio. Identifying causal changes between linear structural equation models. In N. Kiyavash and J. M. Mooij, editors, *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 2383–2398. PMLR, 15–19 Jul 2024.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020. URL http://jmlr.org/papers/v21/17-123.html.

N. Okati, S. Garrido Mejia, W. R. Orchard, P. Blöbaum, and D. Janzing. Root cause analysis of outliers with missing structural knowledge. In *Advances in Neural Information Processing Systems*, volume 39. Curran Associates, Inc., 2025. URL https://neurips.cc/virtual/2025/poster/119709.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press, 2017.

L. Pham, H. Ha, and H. Zhang. Baro: Robust root cause analysis for microservices via multivariate bayesian online change point detection. In *Proceedings of the ACM on Software Engineering*, volume 1, pages 2214–2237, 2024a.

L. Pham, H. Ha, and H. Zhang. Root cause analysis for microservice system based on causal inference: How far are we? In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pages 706–715, 2024b.

T. Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, page 454–461, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc.

D. Rothenhäusler, C. Heinze, J. Peters, and N. Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/92262bf907af914b95a0fc33c3f33bf6-Paper.pdf.

J. Schur. Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind. *Journal für die reine und angewandte Mathematik*, 147:205–232, 1917. URL http://eudml.org/doc/149467.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

M. R. S. Sedghpour, D. Garlan, B. R. Schmerl, C. Klein, and J. Tordsson. Breaking the vicious circle: Self-adaptive microservice circuit breaking and retry. In *IC2E*, pages 32–42, 2023.

H. Shan, Y. Chen, H. Liu, Y. Zhang, X. Xiao, X. He, M. Li, and W. Ding. ϵ-diagnosis: Unsupervised and real-time diagnosis of small-window long-tail latency in large-scale microservice platforms. In *Proceedings of The World Wide Web Conference*, pages 3215–3222, New York, NY, USA, 2019. Association for Computing Machinery.

A. Sharma and E. Kiciman. DoWhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

P. Spirtes. Conditional independence in directed cyclic graphical models representing feedback or mixtures. Technical Report CMU-PHIL-59, Department of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania, 1994.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, 01 2001.

H. Tao, H. Yu, and J. Li. DeepITE: Designing variational graph autoencoders for intervention target estimation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 89978–90008. Curran Associates, Inc., 2024.

A. Tonon, M. Zhang, B. Caglayan, F. Shen, T. Gui, M. Wang, and R. Zhou. RADICE: Causal graph based root cause analysis for system performance diagnostic. In *2025 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 465–475, 2025. doi: 10.1109/SANER64311.2025.00050.

B. Varici, K. Shanmugam, P. Sattigeri, and A. Tajer. Scalable intervention target estimation in linear models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1494–1505. Curran Associates, Inc., 2021.

B. Varici, K. Shanmugam, P. Sattigeri, and A. Tajer. Intervention target estimation in the presence of latent variables. In J. Cussens and K. Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 2013–2023. PMLR, 01–05 Aug 2022.

J. Von Kügelgen, A. Mohamed, and S. Beckers. Backtracking counterfactuals. In M. van der Schaar, C. Zhang, and D. Janzing, editors, *Proceedings of the Second Conference on Causal Learning and Reason-*

ing, volume 213 of *Proceedings of Machine Learning Research*, pages 177–196. PMLR, 11–14 Apr 2023.

V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49 (3):1736 – 1754, 2021.

R. Wang and A. Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 01 2022.

Y. Wang, C. Squires, A. Belyaeva, and C. Uhler. Direct estimation of differences in causal graphs. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/e1314fc026da60d837353d20aefaf054-Paper.pdf.

Z. Xie, Y. Zheng, L. Ottens, K. Zhang, C. Kozyrakis, and J. Mace. Cloud atlas: Efficient fault localization for cloud systems using language models and causal insight, 2024. URL https://arxiv.org/abs/2407.08694.

Y. Yang, S. Salehkaleybar, and N. Kiyavash. Learning unknown intervention targets in structural causal models from heterogeneous data. In S. Dasgupta, S. Mandt, and Y. Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3187–3195. PMLR, 02–04 May 2024.

V. A. Yépez, M. Gusic, R. Kopajtich, C. Mertes, N. H. Smith, C. L. Alston, R. Ban, S. Beblo, R. Berutti, H. Blessing, E. Ciara, F. Distelmaier, P. Freisinger, J. Häberle, S. J. Hayflick, M. Hempel, Y. S. Itkis, Y. Kishita, T. Klopstock, T. D. Krylova, C. Lamperti, D. Lenz, C. Makowski, S. Mosegaard, M. F. Müller, G. Muñoz-Pujol, A. Nadel, A. Ohtake, Y. Okazaki, E. Procopio, T. Schwarzmayr, J. Smet, C. Staufner, S. L. Stenton, T. M. Strom, C. Terrile, F. Tort, R. Van Coster, A. Vanlander, M. Wagner, M. Xu, F. Fang, D. Ghezzi, J. A. Mayr, D. Piekutowska-Abramczuk, A. Ribes, A. Rötig, R. W. Taylor, S. B. Wortmann, K. Murayama, T. Meitinger, J. Gagneur, and H. Prokisch. Clinical implementation of rna sequencing for mendelian disease diagnostics. *Genome Medicine*, 14(1):38, 2022. doi: 10.1186/s13073-022-01019-9.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes, all assumptions are clearly stated alongside the theorems, the model is described 'Notation and Model' section and the algorithm in Algorithm 1 and the subsequent text.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes, in Section 4, we derive a Type I error control guarantee as well as computational and memory complexity. Moreover, our simulations suggest that the algorithm performs reliably with sample sizes on the order of ten times the number of features.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes, source code is publicly available.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes, the full set of assumptions is given alongside the theorems, and if applicable, the details on the genericity assumption required is given in the supplement.
   (b) Complete proofs of all theoretical results. Yes, all proofs can be found either in the main paper or in the supplement.
   (c) Clear explanations of any assumptions. Yes, all assumptions are clearly explained. Our most involved assumption is the genericity assumption, which is introduced in the 'Notation and Model' section, motivated by an example and the precise exception sets are given in the appendix.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes, the code to reproduce our plots and experimental results is attached in the supplemental material.
   (b) All the training details. Yes, while we do not train a model, all hyperparameters and all preprocessing details are clearly explained in 5 and the appendix.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes, a clear definition is given in Section 5.

(d) A description of the computing infrastructure used. Yes, the computing infrastructure used is described in the beginning of Section 5

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator if your work uses existing assets. Yes, we include the citation to all existing code and data used (PetShop, RCAEval and gene expression datasets, BARO method, Cholesky method, Graphical Lasso; dowhy package in the appendix).

   (b) The license information of the assets, if applicable. Yes, we include the license information of our own code as well as of all existing assets we use.

   (c) New assets either in the supplemental material or as a URL, if applicable. Yes, we include all new assets in the supplemental material.

   (d) Information about consent from data providers/curators. Yes, the licenses of all used assets are included in our code repository.

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable.

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. Not Applicable.

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable.

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable.

Supplementary Material

A MISSING PROOFS AND DETAILS ON GENERICITY ASSUMPTIONS

A.1 Proof of Lemma 2.1

We begin by expressing the inverse in terms of the adjugate:

$$\left((I-A)^{-1}\right)_{ij} = \frac{1}{\det(I-A)}(\mathrm{adj}(I-A))_{ij}.$$

To expand the adjugate entry $\mathrm{adj}(I-A)_{ij}$, fix i and j, and let $I-\tilde{A}$ the matrix obtained from $I-A$ by setting all entries in column i and row j to zero, except $(I-\tilde{A})_{ji}$, which is set to 1. Then, by Leibniz-Expansion,

$$(\mathrm{adj}(I-A))_{ij} = \det(I-\tilde{A}) = \sum_{\sigma\in S_p}\left(\mathrm{sgn}(\sigma)\prod_{m=1}^{p}(I-\tilde{A})_{m,\sigma(m)}\right) \tag{8}$$

Note that any permutation can be decomposed into disjoint cyclic components[3], say $\sigma = C_1\circ\cdots\circ C_s$. These components can be understood as graphical cycles: The matrix $I-\tilde{A}$ is the adjacency matrix of the graph $\tilde{G}$ constructed from G with these modifications:

1. Remove all incoming edges to node j and all outgoing edges from node i.
2. Multiply all edge weights by -1.
3. Add self-loops at each node except for i, j.
4. Add a single edge $i\to j$ with weight 1.

In (8), the summand for a permutation $\sigma\in S_n$ is non-zero only if for all k, the edge $k\to\sigma(k)$ exists in $\tilde{G}$. Therefore, we restrict to these permutations henceforth. For such a permutation σ, each cyclic component $\sigma(i_1) = i_2,\ \sigma(i_2)=i_3,\ \ldots,\ \sigma(i_{k-1})=i_k,\ \sigma(i_k)=i_1$ corresponds to the cycle $i_1\to i_k\to i_{k-1}\to\cdots\to i_2\to i_1\in\tilde{G}$. Also the sign of the permutation can be decomposed according to the cycles: $\mathrm{sgn}(\sigma)=(-1)^{p-s}=(-1)^{l_1+\cdots+l_s-s}$, l_* being the lengths of the cycles. Combined, this leads to

$$(\mathrm{adj}(I-A))_{ij} = \sum_{\substack{(C_1,\ldots,C_s)\text{: disjoint cycles}\\ \text{in }\tilde{G}\text{ covering all nodes }[p].}}(-1)^s\prod_{m=1}^{s}(-1)^{l_m}\prod_{k\to h\in C_m}(I-\tilde{A})_{hk}. \tag{9}$$

To relate this back to G, we go through the differences with $\tilde{G}$. Fix a permutation $\sigma = C_1\circ\cdots\circ C_s$ contributing a non-zero summand. First note that due to 1. and 4. the only outgoing edge of j leads to i, so $\sigma(j)=i$; otherwise the corresponding summand would be 0. In graphical terms, this means $i\to j$ is in one of the cycles, lets say in C_s. Because of 4., $i\to j$ is not necessarily in G. However, it has anyways weight 1 in $\tilde{G}$, so we may omit this factor in the product; in formulas

$$(-1)^{l_s}\prod_{h\to k\in C_s}(I-\tilde{A})_{hk} = (-1)^{l_\pi+1}\prod_{h\to k\in\pi}(I-\tilde{A})_{hk}$$

[3] A cyclic component of a permutation $\sigma\in S_n$ is a sequence of distinct elements $C=(i_1\ i_2\ \ldots\ i_k)$ such that $\sigma(i_1)=i_2,\ \sigma(i_2)=i_3,\ \ldots,\ \sigma(i_{l-1})=i_l,\ \sigma(i_l)=i_1$. It can be viewed as an element of S_n by defining $C(j)=j$ for all $j\in[p]\setminus\{i_1,\ldots,i_l\}$. Using this extension, one obtains that σ is the product of all its cyclic components.

where π is the path $j \rightsquigarrow i$ obtained from C_s by dropping $i \to j$, and l_π its length. Note that π does not feature any self-loops, since the cyclic component C_s consists of distinct elements. Therefore, $\pi \in G$, and the expression further simplifies to

$$\begin{aligned}(-1)^{l_\pi+1} \prod_{h\to k\in\pi} (I-\tilde{A})_{hk} &\overset{h\neq k}{=} (-1)^{l_\pi} \prod_{k\to h\in\pi} (\tilde{A})_{hk} \\ &\overset{2.}{=} (-1)^{l_\pi+1} \prod_{k\to h\in\pi} -a_{hk} \\ &= -\prod_{k\to h\in\pi} a_{hk}.\end{aligned} \tag{10}$$

Analogously, for every cycle neither using $i \to j$ nor self-loops,

$$(-1)^{l_m} \prod_{k\to h\in C_m} (I-\tilde{A})_{hk} = \prod_{k\to h\in C} a_{hk}.$$

Finally, some of the cycles might be self-loops, lets say $C_1, \dots, C_r$. Again, they have edge weight 1, and also the signs $(-1)^r \cdot \prod_{m=1}^r (-1)^{l_m} = 1$ cancel out, so we may omit them from the product in (9). This omission is reflected in the overall sum by no longer requiring that $(C_1, \dots, C_{s-1}, \pi)$ together cover all nodes. More precisely, define $\mathcal{C}_\pi = \{(C_1, \dots, C_q) : \text{collection of disjoint cycles in } \tilde{G} \text{ not using any node in } \pi;\ q \in \mathbb{N}_0.\}$. Then,

$$(\mathrm{adj}(I-A))_{ij} = -\sum_{\pi: j\rightsquigarrow i} \prod_{k\to h\in\pi} a_{hk} \cdot \left(\sum_{(C_1,\dots,C_q)\in\mathcal{C}_\pi} (-1)^{q+1} \prod_{m=1}^{q} \prod_{k\to h\in C_q} a_{hk} \right).$$

In this expression, the term $(-1)^s$ in (9) becomes $(-1)^{q+1}$ since π is extracted from the overall sum, therefore not counted by q. Multiplying with -1 and moving the summand for $q = 0$ outside of the sum, we obtain the formula stated in the lemma:

$$((I-A)^{-1})_{ij} = \frac{1}{\det(I-A)} \sum_{\pi: j\rightsquigarrow i} \prod_{k\to h\in\pi} a_{hk} \cdot \left(1 + \sum_{\substack{(C_1,\dots,C_q)\in\mathcal{C}_\pi, \\ q\geq 1.}} (-1)^q \prod_{m=1}^{q} \prod_{k\to h\in C_q} a_{hk} \right).$$

A.2 Proof of Theorem 3.1

As shown in Equation (4) in the main paper, $\xi \overset{d}{=} \Xi + ((I-A)^T \Theta_{NN} \Delta)$, implying that $\xi_i \overset{d}{\neq} \Xi_i$ if and only if $((I-A)^T \Theta_{NN} \Delta)_i \neq 0$. Using that Δ is non-zero only at the root cause positions, we obtain

$$((I-A)^T \Theta_{NN} \Delta)_i = ((I-A)_{:,i})^T \Theta_{NN} \Delta = \sum_{r\in\mathcal{R}} (I-A)_{r,i} \sigma_r^{-2} \delta_r = \sigma_i^{-2} \delta_i - \sum_{r\in\mathcal{R}\cap \mathrm{pa}(i)} a_{r,i} \sigma_r^{-2} \delta_r,$$

which is generically non-zero for all nodes except the root causes and their parents, and definitively zero for all other nodes. The exceptional parameter set on which $\xi_r \overset{d}{=} \Xi_r$ for some root cause r can be directly read off from the above equation: it consists of all parameter choices $a_{ij}, \sigma_{N_i}, \delta_i$ for which, for any root cause $r \in \mathcal{R}$,

$$((I-A)^T \Theta_{NN} \Delta)_r = 0.$$

That is, when the causal model is applied to Δ, the perturbations cancel exactly. This situation only arises if there are multiple root causes that are directly linked in the graph for a measure zero set of choices for all non-zero $a_{ij}, \sigma_{N_i}, \delta_i$.

A.3 Proof of Theorem 3.2

The claim follows directly from Theorem 3.1 together with the observation that root causes and their descendants are generically anomalous, i.e., $\tilde{X}_r \overset{d}{\neq} X_r$, while for all other nodes $\tilde{X}_i \overset{d}{=} X_i$. The exceptional set is the union of that from Theorem 3.1 and all parameter choices for which even a root cause is not anomalous, namely those satisfying

$$0 = ((I-A)^{-1}\Delta)_r.$$

Again, this can only happen if there are multiple directly connected root causes.

A.4 Proof of Lemma 3.4

Assume first that the noise vector N is jointly Gaussian, then $\bar{N} = N_X + (I - A)_{XL}(I - A_{LL})^{-1}N_L$ is also Gaussian, and the sparsity pattern of the precision matrix $K_{\bar{N}} := \Sigma_{\bar{N}}^{-1}$ corresponds to conditional independence relations among the components of $\bar{N}$. In particular, we have

$$(K_{\bar{N}})_{ij} = 0 \Leftrightarrow \bar{N}_i \perp\!\!\!\perp \bar{N}_i \mid N_{[p]\setminus\{i,j\}}.$$

Denoting $B = (I - A)_{XL}(I - A_{LL})^{-1}$, we observe that $\bar{N}$ combined with N_L fulfills the linear SEM

$$\begin{pmatrix} \bar{N} \\ N_L \end{pmatrix} = \begin{pmatrix} 0 & B \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \bar{N} \\ N_L \end{pmatrix} + \begin{pmatrix} N_X \\ N_L \end{pmatrix} \tag{11}$$

Conditional independencies among the components of $\bar{N}$ then follow from d-separation in the corresponding graph. Specifically, $\bar{N}_i \not\perp\!\!\!\perp \bar{N}_i \mid N_{[p]\setminus\{i,j\}}$. is equivalent so $\bar{N}_i$ and $\bar{N}_j$ being d-connected given $\bar{N}$. Since the SEM (11) only features edges from N_L to $\bar{N}$ this further breaks down to the existence of a path $\bar{N}_i \leftarrow N_{L_{l_1}} \rightarrow \bar{N}_{i_1} \leftarrow N_{L_{l_2}} \rightarrow \cdots \rightarrow \bar{N}_j$. Since $B = -A_{XL}(I - A_{LL})^{-1}$ and $(I - A_{LL})^{-1}$ encodes directed paths using only latents, $B_{il} \neq 0$ whenever in the graph belonging to $\boldsymbol{Z}$, there is a path from l to i using just latents. Combining both, the statement of the lemma follows.

Turning to general distributions, note that $K_{\bar{N}} = (\Sigma_{N_X} + B\Sigma_{N_L}B^T)^{-1}$ which depends only on the covariance of N and no other properties of the distribution. Thus, the lemma holds for arbitrary distributions of N.

A.5 Proof of Theorem 3.5

The deviation in the paper already proves that the theorem holds generically. The exception set can be obtained analogously to the one of Theorem 3.1. Specifically, $\xi_r \stackrel{d}{=} \Xi_r$ for an observable root cause r if

$$0 \neq ((I - \bar{A})^T \Theta_{\bar{N}\bar{N}} \bar{\Delta})_r.$$

B OTHER PERSPECTIVES ON OUR SCORE

Relation to linear regression As mentioned in the main paper, there are two other interpretations of our score $s = \text{Z-Score}(\xi; \Xi)$ worth mentioning. The first one is that it relates to linear regression. Specifically, consider the linear regression model

$$X_j = \sum_{i \neq j} \alpha_i X_i + E_j$$

with the data from the normal regime X as the training data set, i.e. the coefficients are learned by fitting on this distribution. Under infinite sample size, this yields

$$\alpha_i = -\frac{\Theta_{ji}}{\Theta_{jj}} \qquad (i \neq j),$$

and a residual with constant variance

$$\text{Var}(E_j) = \frac{1}{\Theta_{jj}}.$$

Next, testing the model on the anormal sample $\tilde{X}_j$, we obtain the residual

$$\tilde{E}_j = \tilde{X}_j - \sum_{i \neq j} \alpha_i \tilde{X}_i = \tilde{X}_j + \sum_{i \neq j} \frac{\Theta_{ji}}{\Theta_{jj}} \tilde{X}_i = \frac{1}{\Theta_{jj}}(\Theta \tilde{X}) = \frac{1}{\Theta_{jj}} \xi. \tag{12}$$

Combining this with $\text{Var}(\Xi) = \Theta$ so that $\text{Z-Score}(\xi_j; \Xi_j) = (\Theta \tilde{X})_j / \sqrt{\Theta_{jj}}$, yields the following result.

Lemma B.1. *Let E_j be the residual of the linear equation model predicting X_j in terms of X_{-j} and $\tilde{E}_j$ the residual when testing the model on $\tilde{X}_j$ as defined in* (12)*. Then,*

$$\textit{Z-Score}(\xi_j; \Xi_j) = \frac{\tilde{E}_j}{\textit{std}(E_j)}.$$

Assuming X to be Gaussian, this further simplifies: In this case, the linear predictor coincides with the conditional mean, that is

$$\mathbb{E}(X_j \mid X_{-j} = \tilde{X}_{-j}) = \sum_{i \neq j} \alpha_i \tilde{X}_i.$$

Moreover, the conditional variance is the residual's variance: $\mathrm{Var}(X_j \mid X_{-j} = \tilde{X}_{-j}) = \mathrm{Var}(E_j)$. Therefore, the Z-Score$(\xi_j; \Xi_j)$ can also be seen as a conditional Z-Score of $\tilde{X}$:

$$\text{Z-Score}(\xi_i; \Xi_i) = \text{Z-Score}(\tilde{X}_j; X_j \mid X_{-j} = \tilde{X}_{-j}) := \frac{|\tilde{X}_j - \mathbb{E}(X_j \mid X_{-j} = \tilde{X}_{-j})|}{\mathrm{Var}(X_j \mid X_{-j} = \tilde{X}_{-j})^{1/2}}.$$

RCA via conditional independence testing Many existing RCA or intervention target estimation approaches rely on conditional independence testing (Kocaoglu et al., 2019; Jaber et al., 2020; Mooij et al., 2020; Ikram et al., 2022). To see that, we first need to introduce a different perspective on modelling perturbations on root causes. For simplicity, we present it for the case of one root cause, but the approach can also be done for multiple ones. Specifically, one combines the normal and anomalous datasets $\boldsymbol{X} \in \mathbb{R}^{m \times p}$, $\tilde{X} \in \mathbb{R}^{1 \times p}$ into one joint dataset $\boldsymbol{Y} \in \mathbb{R}^{m+1 \times (p+1)}$ by introducing a 'regime indicator' (Dawid, 2021)[4] variable F, that is, a binary feature indicating whether a sample is anormal or normal:

$$\boldsymbol{Y} = \begin{pmatrix} X^{(1)} & 0 \\ \vdots & \vdots \\ X^{(m)} & 0 \\ \tilde{X} & 1 \end{pmatrix}$$

One can then show that the corresponding P^Y satisfies again linear SEM, specifically, the equations for all non-root cause features remain the same, while

$$Y_r = \sum_{Y_i \in \mathrm{pa}(i) \setminus \{F\}} a_{ri} Y_i + \delta F + \epsilon_r$$
$$F = \epsilon_0 \sim \text{Bernoulli}\Big(\frac{m_2}{m + m_2}\Big).$$

So, the corresponding graph is the same as G with one additional node F and an edge $F \to Y_r$. Note that this approach still works without assuming linear relations, letting each node be any function of its parents and noise term. In both cases, for DAGs, d-separation and the Markov property of the graph yields

$$Y_j \perp\!\!\!\perp F \mid Y_{-j} \iff Y_j \notin \mathrm{pa}(Y_r) \cup \{Y_r\}, \tag{13}$$

which again reveals the root cause and its parents. Since F is binary, testing this independence is the same as assessing whether the two distributions where $F = 0, 1$, respectively, are the same:

$$H_0 : X_j \mid X_{-j} \sim \tilde{X}_j \mid \tilde{X}_{-j};$$

compare Definition 2 in Jaber et al. (2020). Both our Z-Score(ξ_i, Ξ_i) as well as the conditional Z-Score can be seen as test statistics for this hypothesis, as they become 0 under H_0. For linear SEMs, Spirtes (1994) show that the d-separation based Markov property still holds in the presence of cycles, so the above conclusions remain valid. In contrast, for non-linear SEMs with cycles, d-separation must be replaced by σ-separation to recover a

[4] Also called 'F-node' (Jaber et al., 2020) or 'context variable' in the literature (Mooij et al., 2020).

valid Markov property Forré and Mooij (2018). Denoting by $\mathrm{SCC}(Y_r)$ the set of nodes lying on a directed cycle with Y_r, we obtain

$$Y_j \perp\!\!\!\perp F \mid Y_{-j} \iff Y_j \notin \mathrm{pa}(Y_r) \cup \mathrm{pa}(\mathrm{SCC}(Y_r)) \cup \mathrm{SCC}(Y_r),$$

thereby enlarging the root cause candidate set.

Moreover, without any parametric assumptions, estimating the conditional Z-Score is complex, and for both scores, without any assumptions, it is involved to derive the null distribution and assess the power. In contrast, as demonstrated in the main paper, under a linear SEM and for sufficiently large intervention strength δ, our test has statistical power even with a single anomalous sample. This reliance on just one anomalous sample constitutes the key distinction between our approach and other conditional independence–based RCA methods. For instance, while Jaber et al. (2020) also employs the test in (13) (among others), they do not provide a practical algorithm and establish guarantees only under the assumption of an oracle CI test. Nonetheless, since (13) also holds in the nonlinear case, this viewpoint may serve as a starting point for developing a generalization of our method for nonlinear data.

C CHOLESKY DECOMPOSITION-BASED RCA AND ITS RELATION TO OUR METHOD

The method of Li et al. (2025) assumes G to be a DAG and then identifies the root cause as follows: Like our approach, they start with the fact that under the linear SEM,

$$\Sigma_{XX} = (I - A)^{-1}\Sigma_{NN}(I - A)^{-T}.$$

Their core idea is that if X is arranged according to a topological order of the DAG, then $I - A$ is lower triangular. Therefore, also the product $L = (I - A)^{-1}\Sigma_{NN}^{1/2}$ is lower triangular and can be recovered from Σ_{XX}: Specifically, L is the so-called Cholesky factor of Σ_{XX}, defined as the lower triangular matrix satisfying $LL^T = \Sigma_{XX}$, which is unique and can be computed from Σ_{XX}. Knowing L, we can recover the noise up to standardization: Since $X = (I - A)^{-1}N$,

$$N_{\mathrm{std}} = \left((I - A)^{-1}\Sigma_{NN}^{1/2}\right)^{-1} X$$

where N_{std} is the standardized version of N. Similarly, the perturbed noise can be recovered (up to rescaling with the normal noise's variance) as $\Sigma_{NN}^{-1/2}\tilde{N}_{\mathrm{std}} = L^{-1}\tilde{X}$, whose only extreme entry directly identifies the root cause.

Because the correct ordering is not known, their method iterates over all permutations π, computing the Cholesky factor L_π of the permuted covariance and checking whether $L_\pi^{-1}\tilde{X}$ has exactly one extreme value. Based on non-trivial linear algebra insights, they show that as soon as they found one π yielding precisely one extreme component i, then i is indeed the root cause. Such a π does not necessarily represent the true causal ordering; instead, it must locally respect the structure around i insofar as the parents of i precede i, and its descendants appear after.

To compare their method to ours, note that we obtain a sparse vector via applying the transformation Θ_{XX} to the observed vector $\tilde{X}$, while they get a sparse vector via applying L_π^{-1} to $\tilde{X}$. Their method is more accurate in the sense that $L_\pi^{-1}\tilde{X}$ is only extreme for the root cause, while $\Theta_{XX}\tilde{X}$ only identifies a short list of root causes. However, in the acyclic case, by intersecting with the nodes that are actually anomalous, we also obtain the unique root cause; compare Theorem 3.2.

D FURTHER DETAILS ON EXPERIMENTS

D.1 Choice of hyperparameters

Our method involves two hyperparameters: the threshold τ that distinguishes normal from anomalous features, and the regularization parameter α in the Graphical Lasso. A suitable choice of τ can be expected to depend on the strength of the intervention. Indeed, as shown in Figure 4, for lower intervention strengths it is beneficial to select a smaller τ. This reflects the fact that an intervention strength of 5 only induces mild anomalies, since 5 is quite small compared to the noise variance in our simulation setup (for instance, for Gaussian noise, we sample

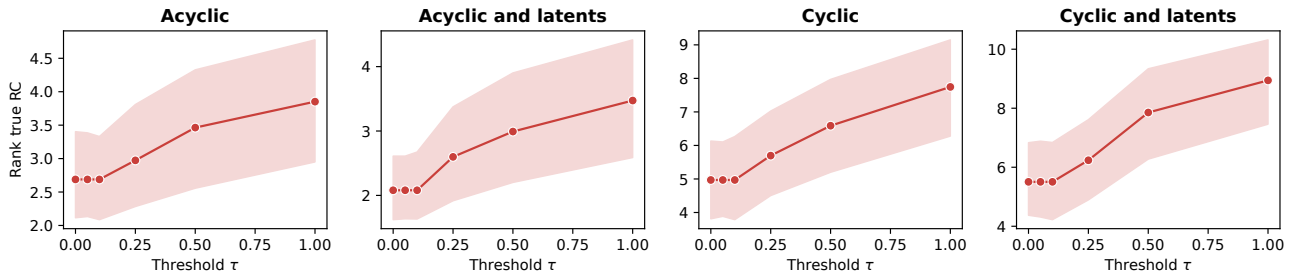

Figure 4: Effect of the threshold parameter τ on RCA performance when using Cyclic (Graphical Lasso) on 50 nodes with the same data simulation setup as in the main paper. The value $\tau = 0.1$ yields the best results.

the standard deviation of each noise term uniformly from $[0.5, 2]$). For our simulation studies, we set $\tau = 0.25$ to achieve reliable performance already at small intervention strengths. In practical applications, this threshold has the advantage of being intuitively interpretable: domain experts can often set it based on their expectation of how anomalous the data should appear in a given scenario. Otherwise, we recommend selecting τ after inspecting the anomaly scores of all features, potentially revealing two clusters of normal versus anormal features.

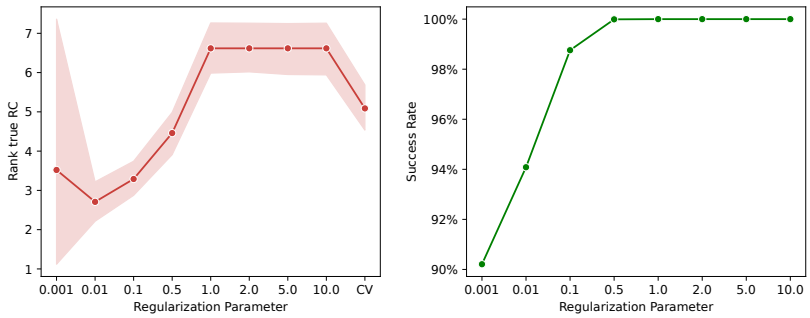

Figure 5: Left: Performance of Cyclic (Graphical) across different regularization parameters α in the Graphical Lasso. The data are generated as described in the main paper, but here all graph types and intervention strengths are combined into a single plot. Note that the Graphical Lasso may fail for small α, since the regularization is insufficient to stabilize inversion of ill-conditioned matrices. Only successful replications are included here. 'CV' denotes selection of α via cross-validation. Right: Corresponding success rate of Graphical Lasso.

Regarding the regularization parameter α, Figure 5 (left) shows that the best results are obtained for $\alpha = 0.01$. However, as seen in the right plot, Graphical Lasso frequently fails due to instability for such small α values. Therefore, we initialize with $\alpha = 0.1$ and add 0.5 in case of failure. If the method still fails after seven such increases, our implementation defaults to using the inverse covariance matrix instead. In practice, this fallback was never required in our experiments. Another possible strategy to ensure success is to select α via cross-validation. However, this approach is computationally more expensive, and also yields weaker performance, as shown in the rightmost column of Figure 5 (left).

D.2 Precision estimation

A challenge in both methods is the need to estimate the inverse of the covariance matrix (our method) or its Cholesky factor (Cholesky method), which might be ill-conditioned. While stronger regularization can alleviate this issue, as shown above, it may not resolve cases such as almost-perfect feature correlation, which shows up in our real-data experiment. In this case, we recommend applying PCA to merge such highly correlated features; compare Section D.4. While this merging may combine the root cause with other features, potentially obscuring its identity, we argue that if the root cause r is strongly correlated with another feature r', it becomes inherently difficult to distinguish between the two based on data alone. Thus, treating them as a single node is justifiable.

Even after mitigating high correlations, precision matrix estimation remains challenging and poses not only the computational but also the statistical bottleneck of our method. Therefore, in addition to the Graphical Lasso

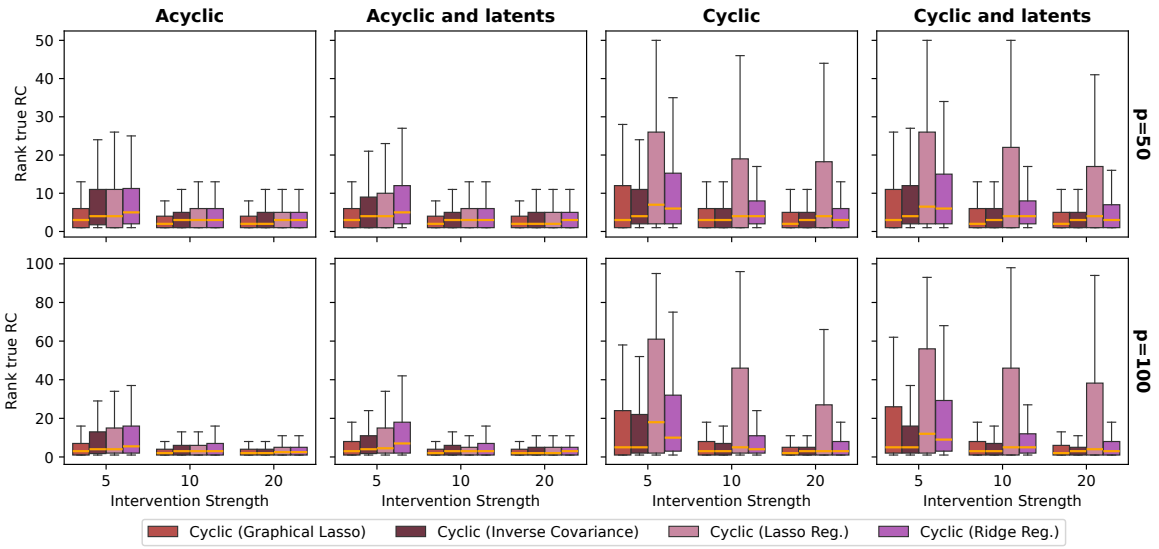

Figure 6: Ranks of the actual root causes under the same simulation settings as in Figure 2. The Lasso and Ridge regression variants both significantly underperform the methods presented in the main paper.

and inverting the sample covariance approaches described in the main paper, we explored alternative strategies. More specifically, we tried to bypass precision estimation entirely by first estimating the sample covariance matrix, and then solving $\hat{\Sigma}\hat{\xi} = \tilde{X}$ using Lasso or Ridge Regression with regularization parameter 10^{-4}. In both cases, the baseline Ξ is not estimated, and the scores s_i in line 5 of the algorithm are simply set to $|\hat{\xi}_i|$. However, as Figure 6 illustrates, both alternatives perform worse than Graphical Lasso. The variance of the Lasso regression approach is particularly high. A likely reason is that Lasso enforces sparsity in its solutions. This means that if the root cause's score ξ_r is incorrectly shrunk to zero, the corresponding root cause will be assigned the lowest possible score (i.e., 0), which is a severe misclassification. In contrast, if the other methods make statistical errors, these are less likely to produce the extreme outcome of assigning a score of 0 to the true root cause.

D.3 Further Simulation Studies

Results broken down by different noise distributions. Figure 7 shows the results of the same simulation study as in the main paper for $p = 50$ but with performance metrics reported separately for each noise distribution choice.

Results broken down by different numbers of root cause. Figure 8 again visualizes the results as in the main paper for $p = 50$ but broken down by how many root causes are chosen.

Varying the expected degree. We expect the performance of both our methods as well as the BARO and Z-Score baselines to depend heavily on the expected degree of the nodes: The higher the connectivity, the more nodes get infected by the root cause, are therefore anomalous, and have a high Z-Score and BARO score. In addition, the increased connectivity raises the probability that a node is a parent and descendants of the true root cause at the same time, thereby appearing in the shortlist described in Theorem 3.2. Indeed, the performance declines with higher degree as illustrated in Figure 9.

D.4 More details on Real-World Experiment

Preprocessing The PetShop dataset (Hardt et al., 2024) is organized with a three-level feature hierarchy in each CSV file: the top level corresponds to service names, the second to three core metrics (error rate, availability, and latency) and the third to statistical summaries used to aggregate these metrics over five-minute intervals (e.g., average, sum, and various percentiles). To reduce the data to the service level, we follow the PetShop package's

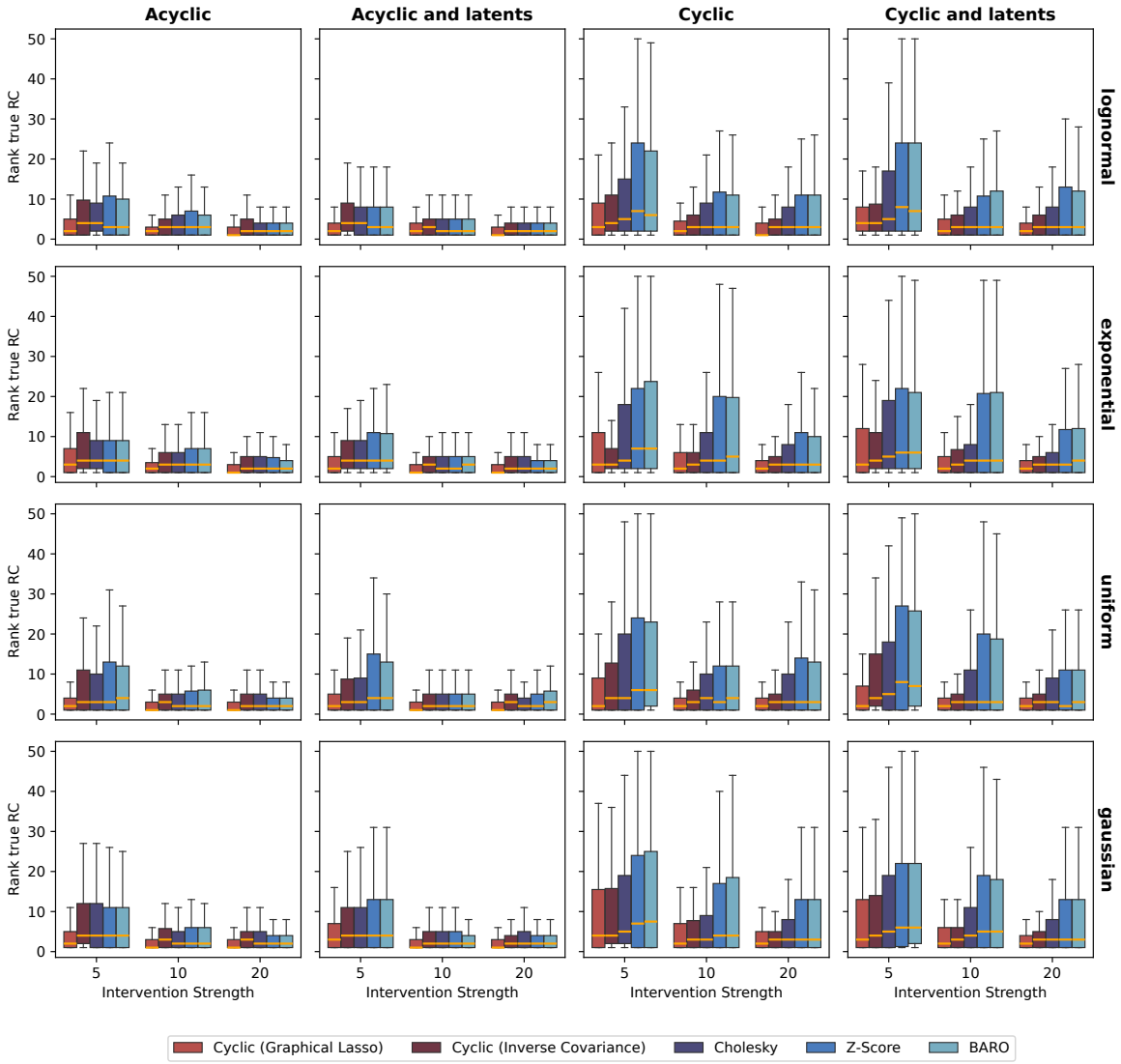

Figure 7: Performance for $p = 50$ across different noise distributions. All methods perform slightly weaker under uniform noise and $\delta = 5$, whereas for the other noise types and intervention strengths, no consistent patterns emerge.

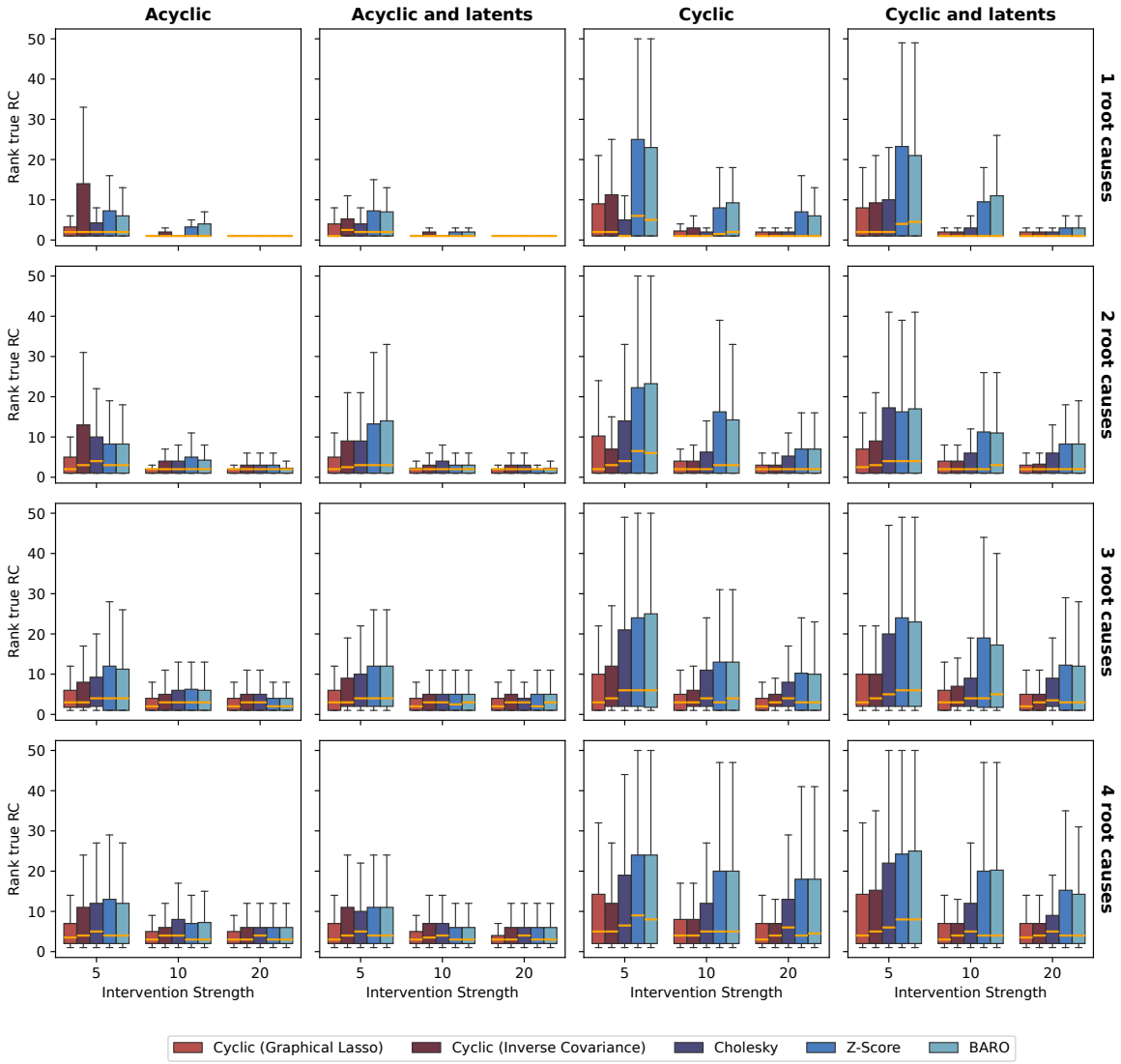

Figure 8: Performance for $p = 50$ when selecting 1, 2, 3 or 4 root causes. Unsurprisingly, performance decreases as the number of root causes increases, with the largest decline observed for Cholesky, which is designed to identify only a single root cause.

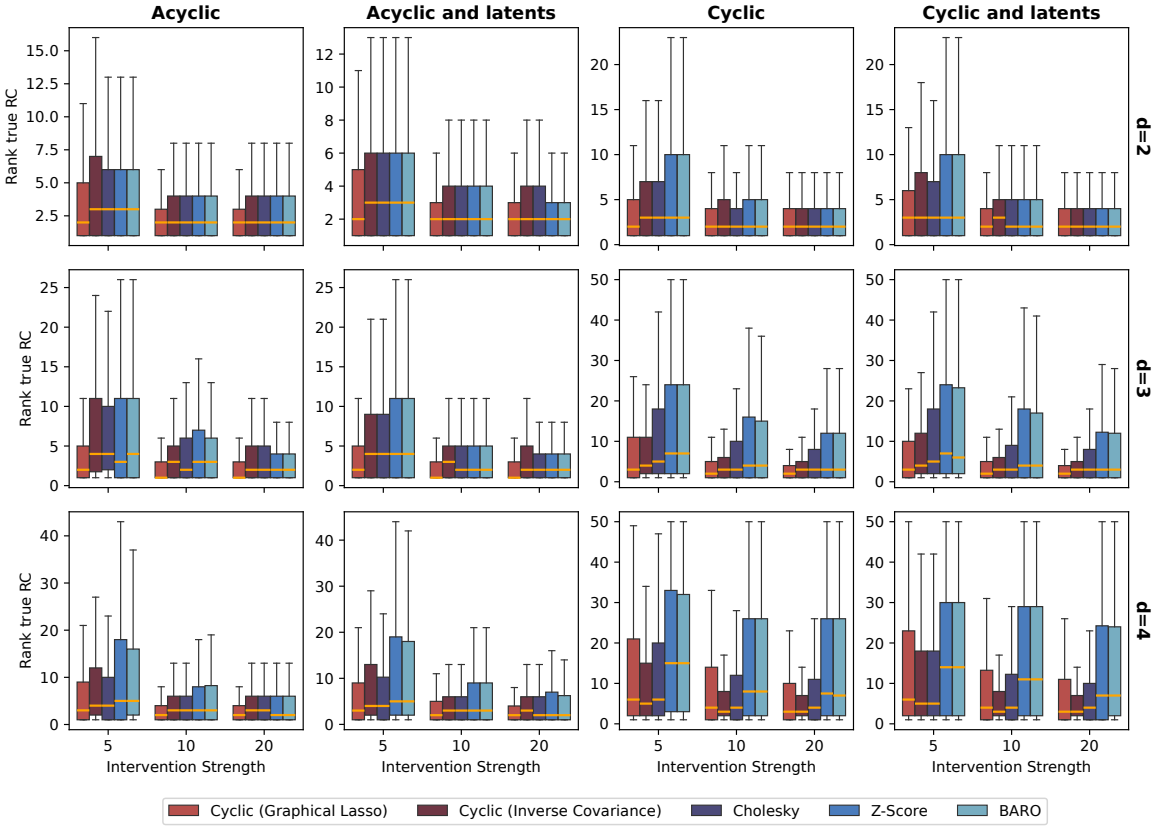

Figure 9: Performance at $p = 50$ across varying expected node degrees. All methods exhibit decreased performance as edge degree increases. This decline is minimal for Cyclic (Graphical Lasso), slightly more pronounced for Cyclic (Inverse Covariance) and Cholesky, and markedly strong for BARO and Z-Score.

recommendation to select only one relevant metric per incident, based on the nature of the incident. Furthermore, since the "average" statistic is the only one consistently available across all services, we restrict to this statistic.

After this filtering, we remove any features that are either constant or exact duplicates of others, as these provide no additional information. We then discard columns with a high proportion of missing values and afterwards remove any rows that still contain NaNs. To normalize the features, we apply a robust standardization, which scales both normal and anomalous data using the mean and standard deviation computed solely from the normal operation period.

Finally, we apply service-level dimensionality reduction using Principal Component Analysis (PCA). This step identifies and merges highly correlated services, retaining the principal components that explain at least 90% of the variance. Doing so, we mitigate the presence of highly correlated features, which would otherwise lead to numerical instability in estimating the precision.

For the remaining real-world datasets, there are no missing entries to fill, and also the PCA step is not necessary. Instead, it suffices to drop almost constant columns and standardize the data.

Further results on RCAEval As mentioned earlier, the benchmark study Pham et al. (2024b) evaluates 21 methods on the RCAEval benchmark dataset. In contrast to our experimental section, they do not report the rank of the true root cause. Instead, they report the Avg@5 score, which averages the five metrics counting how often the true root cause appears among the top 1, 2, 3, 4, or 5 ranked nodes. Furthermore, the results are broken down by fault type. In the anomalous datasets, different services are subjected to interventions, and each service is perturbed using one of several fault types: high CPU load, memory leak, disk I/O stress, network delay, or packet loss. Replicating this evaluation setup, we obtain the results reported in Table 3. In addition to evaluating methods on the original normal and anomalous datasets, the study also considers the scenario in which an incident is detected too late. In this case, a method may be applied under the assumption that the first portion of already anomalous data — for example the first minute — is still part of the normal dataset. The results for this scenario are shown in Table 4.

Moreover, since the benchmark contains interventions on different services, the data naturally provide three distinct environments. This allows us to apply the backShift method proposed by Rothenhäusler et al. (2015) to the datasets. More specifically, the data are organized into tuples consisting of one normal dataset and one anomalous dataset. Each tuple corresponds to a specific fault type and a particular service on which the intervention was performed. For every such combination, five replications are available, resulting in a total of 5^3 dataset tuples. Previously, we evaluated the algorithms separately on each dataset pair and averaged performance across all experiments. Alternatively, multiple interventions can be aggregated into a single dataset. In particular, by fixing the fault type and replication index and combining the data from the five intervention targets, we obtain a dataset suitable for applying backShift. The resulting performance is reported in Table 5.

D.5 Nonlinear data

Finally, we evaluate our approach using a semisynthetic dataset derived from a case study in the DoWhy package (Sharma and Kiciman, 2020; Blöbaum et al., 2024). The dataset is produced by a structural equation model featuring primarily linear but also some nonlinear relationships, designed to resemble patterns typical of microservice data. More specifically, the data is generated according to the graph shown in Figure 10 and each X_i is the sum of its parents plus a noise term, with the following two exceptions:

$$X_{\text{Caching Service}} = Z \cdot X_{\text{Product DB}} + \epsilon_{\text{Caching Service}}, \text{ with } Z \sim \text{Bernoulli}(p = 0.5);$$
$$X_{\text{Product Service}} = \max\left(X_{\text{Shipping Cost Service}},\ X_{\text{Caching Service}},\ X_{\text{Customer DB}}\right) + \epsilon_{\text{Product Service}}$$

For the noise distributions, truncated exponential noise (bounded above) is used for *Website*, *www*, *Order DB*, *Customer DB*, and *Product DB*, with varying truncation levels and a common scale parameter of 0.2. Half-normal noise (non-negative) is used for *API*, *Auth Service*, *Product Service*, *Order Service*, *Shipping Cost Service*, and *Caching Service*, with varying means and standard deviations.

In the original case study, the root cause is fixed as the *Caching Service*, perturbed by a shift of 2. By contrast, we extend this setup by conducting a series of experiments — one for every node as the root cause, which we also

Table 3: AVG@5 scores for each fault type are shown. CYC (G), (S), (I) refers to the Cyclic method with Graphical Lasso, Inverse Covariance, and Shrunk Covariance, respectively. Cholesky is omitted for Sock Shop 2 and Train Ticket due to excessive runtime. For Train Ticket, precision matrices are estimated only with the Shrunk Covariance estimator since the number of features exceeds the number of samples. The synthetic dataset does not include different fault types. Using the AVG@5 metric, the results are still similar to those shown in Figure 3, which reports the average rank. Among the fault types, the LOSS fault appears to be more challenging for all methods.

DATASET	FAULT	BARO	CHOLESKY	CYC (G)	CYC (I)	CYC (S)	Z-SCORE
CIRCA 10	SIM	0.82	0.44	0.84	0.79	0.85	0.81
CIRCA 50	SIM	0.81	0.57	0.88	0.74	0.92	0.82
RCD 10	SIM	0.05	0.69	0.19	0.19	0.19	0.23
RCD 50	SIM	0.04	0.40	0.08	0.08	0.09	0.08
Sock Shop 1	CPU	0.82	0.78	0.70	0.71	0.78	0.72
	MEM	0.97	0.86	0.91	0.88	0.90	0.96
Sock Shop 2	CPU	0.83	-	0.74	0.82	0.75	0.86
	DELAY	0.86	-	0.88	0.89	0.89	0.90
	DISK	0.72	-	0.76	0.62	0.75	0.83
	LOSS	0.78	-	0.77	0.68	0.81	0.82
	MEM	0.86	-	0.88	0.84	0.92	0.94
Online Boutique	CPU	0.77	0.96	0.90	0.86	0.89	0.90
	DELAY	0.90	0.90	0.91	0.91	0.90	0.91
	DISK	0.92	0.82	0.83	0.82	0.86	0.90
	LOSS	0.57	0.56	0.54	0.46	0.51	0.65
	MEM	0.99	1.00	1.00	1.00	1.00	0.99
Train Ticket	CPU	0.44	-	-	-	0.47	0.52
	DELAY	0.51	-	-	-	0.73	0.74
	DISK	0.90	-	-	-	1.00	1.00
	LOSS	0.35	-	-	-	0.50	0.50
	MEM	0.70	-	-	-	0.85	0.87

model by a shift intervention of 2. The RCA methods are then applied to the resulting data. Figure 11 shows that most methods rank the true root cause among the top two nodes, except for Cyclic (Inverse Covariance) and Cholesky performing worse when *Customer DB* is the root cause, and the Cyclic (Inverse Covariance) method drops in performance when intervening on *API*.

References

C. Amendola, P. Dettling, M. Drton, F. Onori, and J. Wu. Structure learning for cyclic linear causal models. In J. Peters and D. Sontag, editors, *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 999–1008. PMLR, 03–06 Aug 2020.

C. K. Assaad, I. Ez-Zejjari, and L. Zan. Root cause identification for collective anomalies in time series given an acyclic summary causal graph with loops. In F. Ruiz, J. Dy, and J.-W. van de Meent, editors, *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 8395–8404. PMLR, 25–27 Apr 2023.

P. Blöbaum, P. Götz, K. Budhathoki, A. A. Mastakouri, and D. Janzing. DoWhy-GCM: An extension of DoWhy for causal inference in graphical causal models. *Journal of Machine Learning Research*, 25(147):1–7, 2024. URL http://jmlr.org/papers/v25/22-1258.html.

T. Boege, M. Drton, B. Hollering, S. Lumpp, P. Misra, and D. Schkoda. Conditional independence in stationary distributions of diffusions. *Stochastic Processes and their Applications*, 184:104604, 2025.

Table 4: AVG@5 scores when the failure occurrence time is misspecified by 60 seconds, meaning the first 60 anomalous samples are incorrectly included in the normal data. On CIRCA, the performance of all methods drops sharply. In contrast, on the remaining datasets, BARO remains robust to this misspecification, while the performance of the other methods show a moderate decline.

DATASET	FAULT	BARO	CHOLESKY	CYC (G)	CYC (I)	CYC (S)	Z-SCORE
CIRCA 10	SIM	0.28	0.34	0.13	0.18	0.10	0.17
CIRCA 50	SIM	0.06	0.05	0.01	0.01	0.01	0.00
RCD 10	SIM	0.04	0.67	0.18	0.19	0.19	0.23
RCD 50	SIM	0.05	0.41	0.07	0.09	0.08	0.09
Sock Shop 1	CPU	0.82	0.71	0.78	0.73	0.75	0.24
	MEM	0.94	0.70	0.70	0.61	0.69	0.34
Sock Shop 2	CPU	0.76	-	0.35	0.38	0.30	0.41
	DELAY	0.87	-	0.55	0.48	0.48	0.53
	DISK	0.66	-	0.38	0.38	0.40	0.32
	LOSS	0.78	-	0.51	0.50	0.43	0.48
	MEM	0.85	-	0.34	0.52	0.36	0.30
Online Boutique	CPU	0.78	0.89	0.34	0.37	0.42	0.57
	DELAY	0.98	0.66	0.59	0.62	0.56	0.56
	DISK	0.89	0.68	0.50	0.50	0.45	0.49
	LOSS	0.54	0.53	0.41	0.39	0.48	0.46
	MEM	0.98	0.79	0.55	0.53	0.55	0.45
Train Ticket	CPU	0.41	-	-	-	0.02	0.01
	DELAY	0.50	-	-	-	0.10	0.10
	DISK	0.89	-	-	-	0.01	0.00
	LOSS	0.36	-	-	-	0.22	0.22
	MEM	0.70	-	-	-	0.05	0.04

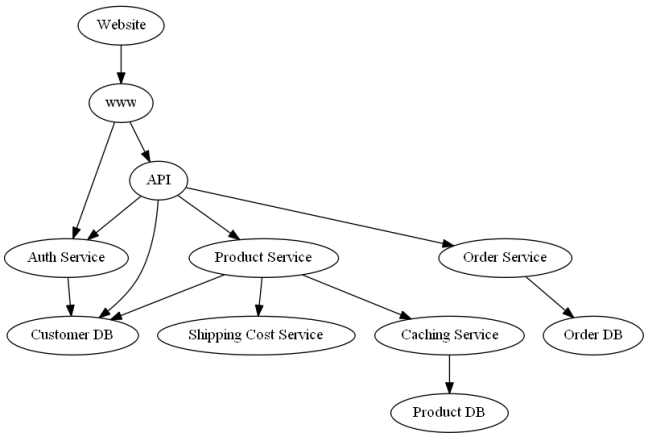

Figure 10: Microservice architecture underlying the semi-synthetic dataset. The graph represents a call graph, where an arrow $A \to B$ indicates that service A calls service B. Consequently, in the causal graph, the direction of the arrows are reversed. Figure taken from DoWhy documentation.

FAULT	SOCK SHOP 1	SOCK SHOP 2	TRAIN TICKET	ONLINE BOUTIQUE
CPU	0.44	0.40	0.29	0.35
MEM	0.44	0.45	0.26	0.33
DISK	-	0.42	0.27	0.27
DELAY	-	0.45	0.29	0.29
LOSS	-	0.48	0.29	0.26

Table 5: Avg@5 results for backShift. Despite being exposed to five times as many samples, the algorithm performs significantly worse than the top RCA approaches in Pham et al. (2024b) and our method.

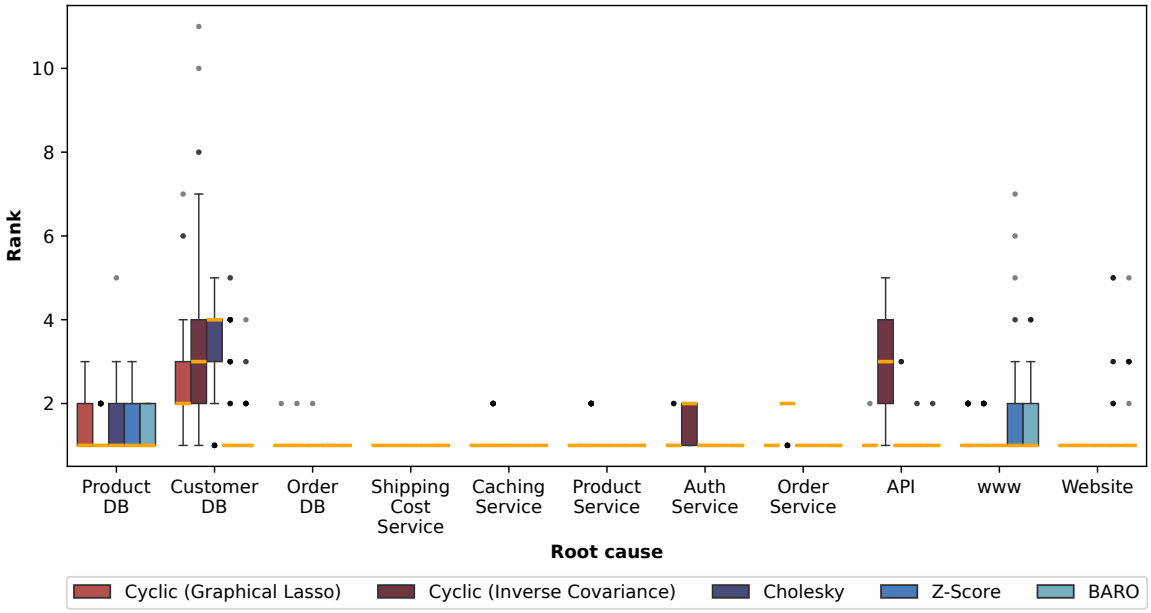

Figure 11: Root cause ranks across 100 replications, with the x-axis indicating the target node. Most methods place the true root cause in the top two, except for Cyclic (Inverse Covariance) and Cholesky on Customer DB and Cyclic (Inverse Covariance) on API.

S. Bongers, P. Forré, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885 – 2915, 2021.

K. Budhathoki, L. Minorics, P. Bloebaum, and D. Janzing. Causal structure-based root cause analysis of outliers. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2357–2369. PMLR, 17–23 Jul 2022.

P. Dawid. Decision-theoretic foundations for statistical causality. *Journal of Causal Inference*, 9(1):39–77, 2021.

P. Forré and J. M. Mooij. Constraint-based causal discovery for non-linear structural causal models with cycles and latent confounders. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, 2018.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

P. Genfer and U. Zdun. Identifying domain-based cyclic dependencies in microservice APIs using source code detectors. In S. Biffl, E. Navarro, W. Löwe, M. Sirjani, R. Mirandola, and D. Weyns, editors, *Software Architecture*, pages 207–222, Cham, 2021. Springer International Publishing.

A. Ghoshal, K. Bello, and J. Honorio. Direct learning with guarantees of the difference DAG between structural equation models. arXiv preprint arXiv:1906.12024, 2019.

P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 03 2024.

M. Hardt, W. R. Orchard, P. Blöbaum, E. Kirschbaum, and S. Kasiviswanathan. The petshop dataset — finding causes of performance issues across microservices. In *Proceedings of the Third Conference on Causal Learning and Reasoning (CLeaR)*, volume 236 of *Proceedings of Machine Learning Research*, pages 957–978. PMLR, 2024.

A. Ikram, S. Chakraborty, S. Mitra, S. K. Saini, S. Bagchi, and M. Kocaoglu. Root cause analysis of failures in microservices through causal discovery. In *Advances in Neural Information Processing Systems*, 2022.

A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9551–9561. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6cd9313ed34ef58bad3fdd504355e72c-Paper.pdf.

M. Kim, R. Sumbaly, and S. Shah. Root cause detection in a service-oriented architecture. In *Proceedings of the ACM SIGMETRICS/International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '13, page 93–104, New York, NY, USA, 2013. Association for Computing Machinery.

M. Kocaoglu, A. Jaber, K. Shanmugam, and E. Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

S. L. Lauritzen. *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. Oxford University Press, Oxford, 1996. ISBN 978-0198522195.

J. Li, B. B. Chu, I. F. Scheller, J. Gagneur, and M. H. Maathuis. Root cause discovery via permutations and Cholesky decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf066, 10 2025. ISSN 1369-7412. doi: 10.1093/jrsssb/qkaf03.66. URL https://doi.org/10.1093/jrsssb/qkaf066.

V. Malik, K. Bello, A. Ghoshal, and J. Honorio. Identifying causal changes between linear structural equation models. In N. Kiyavash and J. M. Mooij, editors, *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 2383–2398. PMLR, 15–19 Jul 2024.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020. URL http://jmlr.org/papers/v21/17-123.html.

N. Okati, S. Garrido Mejia, W. R. Orchard, P. Blöbaum, and D. Janzing. Root cause analysis of outliers with missing structural knowledge. In *Advances in Neural Information Processing Systems*, volume 39. Curran Associates, Inc., 2025. URL https://neurips.cc/virtual/2025/poster/119709.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press, 2017.

L. Pham, H. Ha, and H. Zhang. Baro: Robust root cause analysis for microservices via multivariate bayesian online change point detection. In *Proceedings of the ACM on Software Engineering*, volume 1, pages 2214–2237, 2024a.

L. Pham, H. Ha, and H. Zhang. Root cause analysis for microservice system based on causal inference: How far are we? In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering*, pages 706–715, 2024b.

T. Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI'96, page 454–461, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc.

D. Rothenhäusler, C. Heinze, J. Peters, and N. Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/92262bf907af914b95a0fc33c3f33bf6-Paper.pdf.

J. Schur. Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind. *Journal für die reine und angewandte Mathematik*, 147:205–232, 1917. URL http://eudml.org/doc/149467.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

M. R. S. Sedghpour, D. Garlan, B. R. Schmerl, C. Klein, and J. Tordsson. Breaking the vicious circle: Self-adaptive microservice circuit breaking and retry. In *IC2E*, pages 32–42, 2023.

H. Shan, Y. Chen, H. Liu, Y. Zhang, X. Xiao, X. He, M. Li, and W. Ding. ϵ-diagnosis: Unsupervised and real-time diagnosis of small-window long-tail latency in large-scale microservice platforms. In *Proceedings of The World Wide Web Conference*, pages 3215–3222, New York, NY, USA, 2019. Association for Computing Machinery.

A. Sharma and E. Kiciman. DoWhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

P. Spirtes. Conditional independence in directed cyclic graphical models representing feedback or mixtures. Technical Report CMU-PHIL-59, Department of Philosophy, Carnegie Mellon University, Pittsburgh, Pennsylvania, 1994.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, 01 2001.

H. Tao, H. Yu, and J. Li. DeepITE: Designing variational graph autoencoders for intervention target estimation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 89978–90008. Curran Associates, Inc., 2024.

A. Tonon, M. Zhang, B. Caglayan, F. Shen, T. Gui, M. Wang, and R. Zhou. RADICE: Causal graph based root cause analysis for system performance diagnostic. In *2025 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 465–475, 2025. doi: 10.1109/SANER64311.2025.00050.

B. Varici, K. Shanmugam, P. Sattigeri, and A. Tajer. Scalable intervention target estimation in linear models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 1494–1505. Curran Associates, Inc., 2021.

B. Varici, K. Shanmugam, P. Sattigeri, and A. Tajer. Intervention target estimation in the presence of latent variables. In J. Cussens and K. Zhang, editors, *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pages 2013–2023. PMLR, 01–05 Aug 2022.

J. Von Kügelgen, A. Mohamed, and S. Beckers. Backtracking counterfactuals. In M. van der Schaar, C. Zhang, and D. Janzing, editors, *Proceedings of the Second Conference on Causal Learning and Reasoning*, volume 213 of *Proceedings of Machine Learning Research*, pages 177–196. PMLR, 11–14 Apr 2023.

V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736 – 1754, 2021.

R. Wang and A. Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 01 2022.

Y. Wang, C. Squires, A. Belyaeva, and C. Uhler. Direct estimation of differences in causal graphs. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/e1314fc026da60d837353d20aefaf054-Paper.pdf.

Z. Xie, Y. Zheng, L. Ottens, K. Zhang, C. Kozyrakis, and J. Mace. Cloud atlas: Efficient fault localization for cloud systems using language models and causal insight, 2024. URL https://arxiv.org/abs/2407.08694.

Y. Yang, S. Salehkaleybar, and N. Kiyavash. Learning unknown intervention targets in structural causal models from heterogeneous data. In S. Dasgupta, S. Mandt, and Y. Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3187–3195. PMLR, 02–04 May 2024.

V. A. Yépez, M. Gusic, R. Kopajtich, C. Mertes, N. H. Smith, C. L. Alston, R. Ban, S. Beblo, R. Berutti, H. Blessing, E. Ciara, F. Distelmaier, P. Freisinger, J. Häberle, S. J. Hayflick, M. Hempel, Y. S. Itkis, Y. Kishita, T. Klopstock, T. D. Krylova, C. Lamperti, D. Lenz, C. Makowski, S. Mosegaard, M. F. Müller, G. Muñoz-Pujol, A. Nadel, A. Ohtake, Y. Okazaki, E. Procopio, T. Schwarzmayr, J. Smet, C. Staufner, S. L. Stenton, T. M. Strom, C. Terrile, F. Tort, R. Van Coster, A. Vanlander, M. Wagner, M. Xu, F. Fang, D. Ghezzi, J. A. Mayr, D. Piekutowska-Abramczuk, A. Ribes, A. Rötig, R. W. Taylor, S. B. Wortmann, K. Murayama, T. Meitinger, J. Gagneur, and H. Prokisch. Clinical implementation of rna sequencing for mendelian disease diagnostics. *Genome Medicine*, 14(1):38, 2022. doi: 10.1186/s13073-022-01019-9.