Journal of the Royal Statistical Society Series B: Statistical Methodology, 2025, **00**, 1–20
https://doi.org/10.1093/jrsssb/qkaf066

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology

Original Article

Root cause discovery via permutations and Cholesky decomposition

Jinzhou Li[1], Benjamin B. Chu[2], Ines F. Scheller[3], Julien Gagneur[3,4,5,6] and Marloes H. Maathuis[7]

[1]Department of Statistics and Data Science, National University of Singapore, Singapore, Singapore
[2]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
[3]School of Computation, Information and Technology, Technical University of Munich, Garching, Germany
[4]Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany
[5]Munich Data Science Institute, Technical University of Munich, Garching, Germany
[6]Institute of Human Genetics, School of Medicine, Technical University of Munich, Munich, Germany
[7]Seminar for Statistics, ETH Zürich, Switzerland

Address for correspondence: Jinzhou Li, Department of Statistics and Data Science, National University of Singapore, 117546 Singapore, Singapore. Email: jinzhouli@nus.edu.sg

Abstract

This work is motivated by the following problem: Can we identify the disease-causing gene in a patient affected by a monogenic disorder? This problem is an instance of root cause discovery. Specifically, we aim to identify the intervened variable in one interventional sample using a set of observational samples as reference. We consider a linear structural equation model where the causal ordering is unknown. We begin by examining a simple method that uses squared z-scores and characterize the conditions under which this method succeeds and fails, showing it generally cannot identify the root cause. We then prove, without additional assumptions, that the root cause is identifiable even if the causal ordering is not. Two key ingredients of this identifiability result are the use of permutations and the Cholesky decomposition, which allow us to exploit an invariant property across different permutations to discover the root cause. Furthermore, we characterize permutations that yield the correct root cause and, based on this, propose a valid method for root cause discovery. We also adapt this approach to high-dimensional settings. Finally, we evaluate our methods through simulations and apply the high-dimensional method to discover disease-causing genes in the gene expression dataset that motivates this work.

Keywords: Cholesky decomposition, identifiability, invariance, permutation, rare disease application, root cause discovery

1 Introduction

1.1 Motivation and problem statement

Rare diseases can stem from mutations in a single gene (also known as Mendelian or monogenic disorders). Typically, such mutations affect the expression of the gene itself, and of further downstream genes, ultimately leading to the disease (e.g. Yépez, Gusic, Kopajtich, Mertes, et al., 2022; Yépez, Mertes, et al., 2021). Discovering the disease-causing gene is important, as it enhances our understanding of the disease mechanism and is a critical step towards developing potential cures. This motivates us to consider root cause discovery, which, at a high level, is an instance of the reverse causal problem (see, e.g. Dawid et al., 2014; Gelman & Imbens, 2013; Pearl, 2015) and aims to identify the ultimate cause of an observed effect.

Specifically, we distinguish between observational samples (e.g. gene expression data from healthy individuals) and interventional samples (e.g. gene expression data from patients), with the former serving as a reference to detect the root cause variable in the latter. While we assume that observational samples are independent and identically distributed, we do not make this

Received: December 19, 2024. **Revised:** June 16, 2025. **Accepted:** September 11, 2025

assumption for interventional samples, since disease-causing genes often differ among rare disease patients, even among those with very similar phenotypes as in the case of mitochondrial disorders (Gusic & Prokisch, 2021). Consequently, we conduct root cause discovery for rare disease patients in a personalized manner, focusing on one interventional sample at a time. This presents a unique challenge, distinguishing the problem from most related work where identically distributed interventional samples are available.

To formalize the problem, we consider n i.i.d. observational samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of X (e.g. data of healthy individuals) and one interventional sample $\boldsymbol{x}^I$ of X^I (e.g. data of a patient) generated from the following two linear structural equation models (SEMs), respectively:

$$X \leftarrow b + BX + \varepsilon \tag{1}$$

and

$$X^I \leftarrow b + BX^I + \varepsilon + \delta, \tag{2}$$

where X, X^I, b, ε and $\delta \in \mathbb{R}^p$ and $B \in \mathbb{R}^{p\times p}$. Here, b is an intercept term. B encodes the underlying causal structure. It can be visualized in a causal directed acyclic graph (DAG) on vertices $\{1, \ldots, p\}$, where there is an edge from i to j if the edge weight $B_{ji} \neq 0$. Further, ε is an error term that follows an arbitrary distribution with mean 0 and diagonal covariance matrix. Finally, $\delta = (0, \ldots, 0, \delta_r, 0, \ldots, 0)^T$ represents a mean-shift intervention. It has only one nonzero entry δ_r in the rth position, indicating that the variable X_r^I is intervened upon by a mean shift δ_r, modelling a mutation that results in over- or under-expression of the gene. We refer to r or X_r^I as the *root cause* of X^I. We assume this model throughout the paper. The assumption of a unique root cause can be seen as an extreme case of the so-called 'sparse mechanism shift hypothesis' in Schölkopf et al. (2021).

Our goal is to discover the root cause of X^I by comparing $\boldsymbol{x}^I = (x_1^I, \ldots, x_p^I)^T$ with the observational samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Intuitively, when $|\delta_r|$ is large compared to the noise variance, x_r^I will stand out as prominently aberrant compared to the values of the rth variable in the observational samples. But the intervention effect on x_r^I can propagate to downstream variables, causing many other variables to appear aberrant as well. This propagation can make it difficult to identify the root cause.

For a quick illustration, we look at a simulated dataset shown in Figure 1b, where we use $b = (0, 0, 0, 0, 0)^T$, $B = \begin{pmatrix} 0 & 1 & -1 & -2 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -2 & 1 & 3 & 0 \end{pmatrix}$, $\mathrm{Cov}(\varepsilon) = \mathrm{Diag}(3, 2, 3, 2, 3)$, and $\delta = (0, 0, 10, 0, 0)^T$. The causal DAG in plot (a) represents the generating process. The vector δ indicates that X_3^I is the root cause for the interventional sample (visualized by the lightning symbol in the DAG). All variables except for X_4^I are descendants of X_3^I, meaning that the intervention effect propagates to X_1^I, X_2^I, and X_5^I. This is reflected in plot (b), where 100 observational samples (in grey) and one interventional sample (in colour) show that, except for X_4^I, all variables of the interventional sample appear aberrant compared to the observational samples.

1.2 Our contribution

In this paper, we start by studying a commonly used quantity for detecting aberrancy: the squared z-score (see (3) in Section 2). We characterize when the squared z-score can and cannot successfully identify the root cause, as the observational sample size and the intervention strength tend to infinity. In particular, we show that the squared z-score can consistently identify the root cause if $\mathrm{Var}(X_k) > \alpha_{r\to k}^2 \mathrm{Var}(X_r)$ for all k, where $\alpha_{r\to k}^2$ is the total causal effect of X_r on X_k. Three sufficient conditions on the generating mechanism that ensure this are: (i) for any descendants of the root cause, there is no common ancestor that has a directed path to the descendant without passing through the root cause; (ii) the causal DAG is a polytree (see, e.g. Jakobsen et al., 2022; Tramontano et al., 2022, 2023); and (iii) all entries of the matrix B are nonnegative. On the other

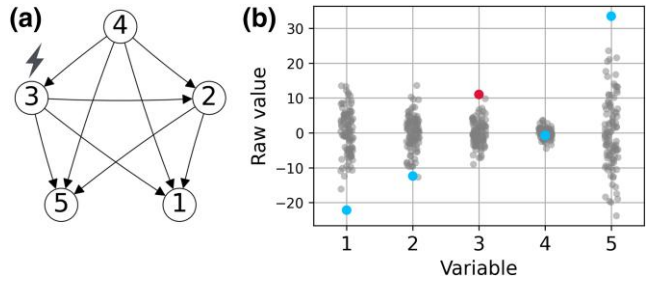

Figure 1. (a) The causal directed acyclic graph (DAG) that generates the samples shown in plot (b). (b) A simulated dataset based on the causal DAG in plot (a). Grey points represent 100 observational samples, while coloured points denote the data of one interventional sample.

hand, if there exists a k such that $\mathrm{Var}(X_k) < \alpha_{r\to k}^2 \mathrm{Var}(X_r)$, then the squared z-score cannot consistently identify the root cause.

In Section 3, we study the fundamental question of whether the root cause is identifiable based on the observational distribution and the first moment of the interventional distribution. We consider the observational distribution as it can theoretically be estimated with n i.i.d. observational samples, and we consider only the first moment of the interventional distribution because only one interventional sample is available. The answer to this question is not immediately clear, as even the causal ordering may be nonidentifiable in this setting. We prove, however, that the root cause is in fact identifiable. This identifiability result stems from an invariant property involving permutations of variables and the Cholesky decomposition.

The above identifiability result leads to a valid root cause discovery method, which requires trying all permutations of the variables. Considering $p!$ permutations, however, quickly becomes infeasible when the number of variables p increases. To address this problem, we characterize so-called 'sufficient' permutations that allow us to identify the root cause. In particular, sufficient permutations are those where the parents of the root cause are positioned before it and the true descendants are positioned after it. This leads to the second valid root cause discovery algorithm with a smaller computational burden (see Algorithm 2). At last, we develop a heuristic root cause discovery method designed for high-dimensional settings (online supplementary material, Algorithm 3), building on our previously proposed Algorithm 2.

We examine the performance of our methods through extensive simulations in Section 4. In Section 5, we revisit the problem of discovering disease-causing genes in the high-dimensional gene expression dataset that motivates our study. We apply the high-dimensional method (online supplementary material, Algorithm 3), yielding very promising results. We conclude the paper with discussions in Section 6.

1.3 Related ideas and work

Recall that we assume that the underlying causal structure is unknown. If the DAG or causal ordering were known, there are two natural methods for root cause discovery: (i) first identify all aberrant variables and then select the one that appears first in the causal ordering; (ii) leverage invariance: only the root cause will exhibit a changed conditional distribution given its parents (see, e.g. Haavelmo, 1944; Janzing et al., 2019; Li et al., 2022; Peters et al., 2016). See online supplementary material, Appendix E.2 for more details.

One may try to estimate the DAG or causal ordering from observational samples and then apply the ideas of the previous paragraph. However, it is well-known that these estimation problems are highly challenging and often requires very large sample sizes (see, e.g. Evans, 2020). More importantly, causal orderings and DAGs may be unidentifiable (see, e.g. Pearl, 2009b; Spirtes et al., 2001), which presents a fundamental issue that cannot be resolved regardless of sample size. Nevertheless, in Section 4, we implement methods based on these two ideas, using LiNGAM (Shimizu et al., 2006), and compare their performance to that of our method.

When the DAG is unknown, there is a line of work that combines both observational and interventional samples, with the primary goal of either estimating the intervened variables (i.e. root

causes) or learning the causal structure with the estimated root cause as a by-product. However, these methods rely on aspects of the interventional distributions like the second moments (Rothenhäusler et al., 2015; Varici et al., 2021, 2022), likelihoods (Eaton & Murphy, 2007; Taeb & Bühlmann, 2024), or the entire interventional distribution (Ikram et al., 2022; Jaber et al., 2020; Squires et al., 2020; Yang et al., 2024). These approaches cannot be applied to our problem, where only a single interventional sample is available.

Our method relies on the Cholesky decomposition, which has been used before in causal structure learning (Raskutti & Uhler, 2018; Ye et al., 2020) and for estimating the effects of joint interventions (Nandy et al., 2017). In particular, Raskutti and Uhler (2018) also combine permutations with the Cholesky decomposition to find the sparsest Cholesky factorization. Our approach differs fundamentally from these works: we combine permutations and the Cholesky decomposition to search for an invariant property related to the root cause.

Finally, there is related work that treats root cause analysis as a causal contribution problem (see, e.g. Budhathoki et al., 2021, 2022; Okati et al., 2024), as well as work that focuses on root cause analysis for microservice-based applications (see Hardt et al., 2023 and the references therein). These works either require knowledge of the causal DAG or estimate it from data. One exception is the method 'SCORE ORDERING' from Okati et al. (2024), which does not require the causal graph and is based on the heuristic that small outliers are unlikely to cause larger ones. This heuristic is justified in certain scenarios but does not hold in general (see the example in online supplementary material, Appendix A.1).

To the best of our knowledge, no formal method currently exists in the literature to address our problem, where only a single interventional sample is available, and the causal ordering may be unidentifiable.

2 Squared z-score is not generally valid

The squared z-score is commonly used to quantify aberrancy. Instead of examining raw values, it adjusts for each variable's marginal mean and variance. Intuitively, the squared z-score cannot identify the root cause in general, as it does not account for the causal relationships between variables. In this section, we formally verify this intuition by characterizing when the squared z-score succeeds and fails.

For $j \in [p] = \{1, \ldots, p\}$, the z-score of X_j^I is defined as

$$\widehat{Z}_{n,j} = \frac{X_j^I - \widehat{\mu}_{n,j}}{\widehat{\sigma}_{n,j}}, \tag{3}$$

where $\widehat{\mu}_{n,j} = \frac{1}{n}\sum_{i=1}^n x_{ij}$ and $\widehat{\sigma}_{n,j} = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (x_{ij} - \widehat{\mu}_{n,j})^2}$ are the sample mean and standard deviation of the observational samples.

Let $r \in [p]$ be the root cause and $k \in [p] \setminus \{r\}$. For the linear SEMs (1) and (2), we have

$$X_r^I = X_r + \delta_r \quad \text{and} \quad X_k^I = X_k + \alpha_{r\to k}\delta_r, \tag{4}$$

where $\alpha_{r\to k} = (I - B)^{-1}_{kr}$ is the total causal effect of X_r on X_k.

To gain some intuition about when the squared z-score works, we first consider the z-score Z_j with population mean μ_j and standard deviation σ_j of the observational distribution. Then, using (4), we have

$$Z_r = \frac{X_r^I - \mu_r}{\sigma_r} = \frac{X_r - \mu_r}{\sigma_r} + \frac{\delta_r}{\sigma_r} \quad \text{and} \quad Z_k = \frac{X_k^I - \mu_k}{\sigma_k} = \frac{X_k - \mu_k}{\sigma_k} + \frac{\alpha_{r\to k}\delta_r}{\sigma_k}.$$

Here, the first terms on the right-hand sides are standardized random variables. Hence, for δ_r large, the deterministic terms will dominant and $Z_r^2 > Z_k^2$ if $\sigma_k^2 > \alpha_{r\to k}^2\sigma_r^2$. The corresponding sample version result is given in Theorem 1.

Theorem 1 Let $r \in [p]$ be the root cause. For any $k \in [p]\setminus\{r\}$, the following two statements hold:

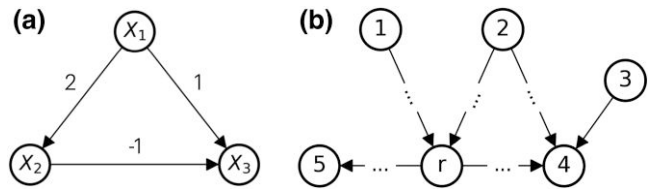

Figure 2. (a) The directed acyclic graph (DAG) corresponding to the structural equation model (SEM) that generates data for Figure 3. (b) Directed acyclic graph used to illustrate Proposition 2, where r denotes the root cause, and the edges with dots represent directed paths.

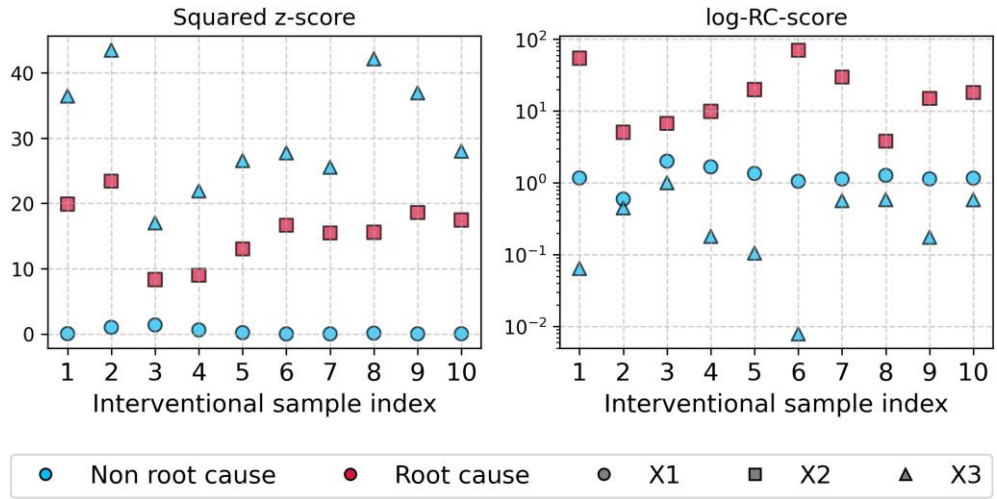

Figure 3. Simulation results based on a structural equation model (SEM) with the directed acyclic graph (DAG) in Figure 2a. The two plots show the squared z-scores and the RC-scores (in log scale) of 10 interventional samples with root cause X_2, respectively. For each interventional sample, red denotes the root cause (squares in this case) while blue represents nonroot cause variables (triangulars and circles in this case), and X_1, X_2, and X_3 are represented by circles, squares, and triangles, respectively.

$$\text{(i) If } \sigma_k^2 > \alpha_{r\to k}^2\sigma_r^2, \text{ then } \lim_{\substack{n\to\infty\\ \delta_r\to\infty}} \mathrm{P}(\widehat{Z}_{n,r}^2 > \widehat{Z}_{n,k}^2) = 1.$$

$$\text{(ii) If } \sigma_k^2 < \alpha_{r\to k}^2\sigma_r^2, \text{ then } \lim_{\substack{n\to\infty\\ \delta_r\to\infty}} \mathrm{P}(\widehat{Z}_{n,r}^2 < \widehat{Z}_{n,k}^2) = 1.$$

In the above theorem, $n \to \infty$ allows us to obtain consistent estimators for the population mean and standard deviation of the observational distribution. However, since there is only one interventional sample, which is used to estimate μ_j^I, the variance of this estimator does not vanish. To account for this, we let $\delta_r \to \infty$. Based on our proof in online supplementary material, Appendix C.1, one can also obtain nonasymptotic high-probability results with a more precise characterization of δ_r, but we refrain from doing so here for simplicity. As a direct corollary, if the condition $\sigma_k^2 > \alpha_{r\to k}^2\sigma_r^2$ holds for all $k \in [p] \setminus r$, the squared z-score can consistently identify the root cause.

To give some intuition for conditions (i) and (ii), we rewrite $X_k^I = \alpha_{r\to k}X_r^I + N$, where N is an appropriate linear combination of the error terms of the ancestors of X_k^I excluding X_r^I. Then, we have $\sigma_k^2 = \alpha_{r\to k}^2\sigma_r^2 + \mathrm{Var}(N) + 2\alpha_{r\to k}\mathrm{Cov}(X_r^I, N)$. Hence, condition (ii) holds if the term $\alpha_{r\to k}\mathrm{Cov}(X_r^I, N)$ is sufficiently negative, which can only happen if there exist common ancestors of X_k^I and X_r^I (i.e. confounding). This is also related to Proposition 2 and its accompanying discussion. To give a quick illustration, we consider a linear SEM with equal error variance 1 and the DAG shown in Figure 2a. For $r = 2$ and $k = 3$, it is easy to verify that condition (ii) in Theorem 1 holds, so X_3 will likely have a larger squared z-score than X_2. This is verified in the left plot of Figure 3, which shows the squared z-scores for 10 interventional samples where X_2

is the root cause. See also Ebtekar et al. (2025) for an explanation of this phenomenon from the perspective of the so-called conditional outlier scores. For details of this example and further discussions, see online supplementary material, Appendix A.1.

A natural follow-up question is: Can we give sufficient conditions on the generating model that ensure $\sigma_k^2 > \alpha_{r\to k}^2\sigma_r^2$ for some or all $k \in [p]\backslash\{r\}$? To answer this question, we first introduce some notation. For $j \in [p]$ and a DAG G, let the set of its ancestors and descendants be

$$\mathrm{An}(j) = \{l \in [p]\backslash\{j\} : \text{ there is a directed path from} l \text{ to } j \text{ in } G\},$$
$$\mathrm{De}(j) = \{l \in [p]\backslash\{j\} : \text{ there is a directed path from } j \text{ to } l \text{ in } G\}.$$

In addition, for $k \in [p]\backslash\{r\}$, define

$$O(r,k) = \{j \in \mathrm{An}(r) \cap \mathrm{An}(k) : \text{there is a directed path from } j \text{ to } k \text{ in } G \text{ that bypasses } r\}.$$

Then, we have the following Proposition 2:

Proposition 2 Let $r \in [p]$ be the root cause and let $k \in [p]\backslash\{r\}$. If (i) $k \notin \mathrm{De}(r)$, or (ii) $k \in \mathrm{De}(r)$ and $O(r,k) = \emptyset$, then $\sigma_k^2 > \alpha_{r\to k}^2\sigma_r^2$.

Based on Proposition 2, a necessary condition for $\sigma_k^2 < \alpha_{r\to k}^2\sigma_r^2$ to hold, or equivalently for a variable X_k to be problematic in the sense of having a larger asymptotic squared z-score than X_r, is that $k \in \mathrm{De}(r)$ and $O(r,k) \neq \emptyset$. For an illustration, consider Figure 2b, where r is the root cause and the edges with dots represent directed paths, variables 1, 2, 3, and 5 are considered safe because $1, 2, 3 \notin \mathrm{De}(r)$ and $5 \in \mathrm{De}(r)$ with $O(r,5) = \emptyset$. The only unsafe variable is 4, as $4 \in \mathrm{De}(r)$ with $O(r,4) = \{2\} \neq \emptyset$. Similarly, in Figure 2a, all variables are safe when $r = 1$ or $r = 3$. But if $r = 2$, X_3 is unsafe as $3 \in \mathrm{De}(2)$ and $O(2,3) = \{1\}$. See online supplementary material, Appendix A.2 for further discussion with a concrete example.

Proposition 2 only provides a sufficient condition. Another sufficient condition is shown in the following proposition.

Proposition 3 Let $r \in [p]$ be the root cause. If all edge weights are nonnegative, then $\sigma_k^2 > \alpha_{r\to k}^2\sigma_r^2$ for all $k \in [p]\backslash\{r\}$.

Corollary 2.1 directly follows from Propositions 2 and 3.

Corollary 2.1 Let $r \in [p]$ be the root cause. If the underlying DAG is a polytree or if all edge weights are nonnegative, then $\lim_{\substack{n\to\infty \\ \delta_r\to\infty}} \mathrm{P}(\widehat{Z}_r^2 > \widehat{Z}_k^2) = 1$ for all $k \in [p]\backslash\{r\}$.

In summary, using the squared z-score identifies the root cause in some cases, but not in general.

3 Root cause discovery via permutations and Cholesky decomposition

3.1 Is the root cause identifiable?

At this point, it is unclear whether the root cause is identifiable based solely on the observational distribution and the first moment of the interventional distribution. As mentioned in the introduction, we focus on these two quantities because the former can be estimated from n i.i.d. observational samples, while a single interventional sample only allows us to estimate the first moment of the interventional distribution. This question is fundamental and not immediately clear, given that the causal ordering and the causal DAG are generally nonidentifiable.

In the following, we give a positive answer to this question. In particular, we prove that the root cause is identifiable from the mean and covariance matrix of the observational distribution (denoted by μ_X and Σ_X) and the mean of the interventional distribution (denoted by μ_{X^I}). We focus on population quantities to address the identifiability question for now, and the sample version algorithm will be introduced in the next section.

Note that the population version of the z-score can be written as $\mathrm{Diag}(\sigma_1, \ldots, \sigma_p)^{-1}(\mu_{X^I} - \mu_X)$. Considering the generalization that takes the full covariance matrix into account, we introduce the following key term:

$$\xi := L_X^{-1}(\mu_{X^I} - \mu_X), \tag{5}$$

where L_X is the lower-triangular matrix with real and positive diagonals obtained by the Cholesky decomposition of Σ_X, i.e. $\Sigma_X = L_X L_X^T$.

To motivate the reason of using the Cholesky decomposition, we look at the special case where $(X_1, \ldots, X_p)$ are sorted by a causal ordering. In this case, the matrix B in model (1) is lower triangular. From model (1), we have $\Sigma_X = (I - B)^{-1} D_\varepsilon (I - B)^{-T}$, where $D_\varepsilon = \mathrm{Diag}(\sigma^2_{\varepsilon_1}, \ldots, \sigma^2_{\varepsilon_p})$ is the diagonal covariance matrix of ε. Hence, $L_X = (I - B)^{-1} D_\varepsilon^{1/2}$ by the uniqueness of the Cholesky decomposition. Additionally, since $\mu_{X^I} = (I - B)^{-1}(b + \delta)$ and $\mu_X = (I - B)^{-1} b$, we have

$$\xi = L_X^{-1}(\mu_{X^I} - \mu_X) = D_\varepsilon^{-1/2}\delta = (0, \ldots, 0, \delta_r/\sigma_{\varepsilon_r}, 0, \ldots, 0)^T. \tag{6}$$

Thus, when $(X_1, \ldots, X_p)$ are sorted by a causal ordering, one may view ξ_i^2 as a conditional outlier score given the true parents of X_i (cf. Janzing et al., 2019), and we can distinguish the root cause based on ξ because it is the only entry with a nonzero value.

However, since $(X_1, \ldots, X_p)$ is generally not in a causal ordering, the matrix B is not lower triangular and L_X does not equal $(I - B)^{-1} D_\varepsilon^{1/2}$. As a result, ξ may not exhibit the informative pattern seen in (6). Therefore, the root cause cannot be identified using the Cholesky decomposition alone.

Remarkably, there is actually a (conditional) invariant property related to the root cause across different permutations of variables: conditional on $\xi(\pi)$ (see (7)) having only one nonzero entry (we refer to this pattern as *1-sparse*), the variable corresponding to this entry is invariant and must be the root cause. This is formalized in Theorem 4, which can be proved using linear algebra (see online supplementary material, Lemma D.1 in Appendix D.1).

Theorem 4 ((Identifiability of the root cause)). Let $(X_1, \ldots, X_p)$ be sorted in any ordering (not necessarily a causal ordering). The corresponding ξ must have at least one nonzero element. Furthermore, if ξ has exactly one nonzero element, then this element must be the root cause.

Theorem 4 enables us to exploit an invariant property to identify the root cause: by trying all permutations of $(X_1, \ldots, X_p)$, we are guaranteed to find at least one permutation (since a causal ordering must exist) for which the corresponding ξ is 1-sparse, in which case the nonzero entry must correspond to the root cause. This verifies the identifiability of the root cause.

Notably, Theorem 4 implies that the root cause is identifiable even if the causal ordering is not. From this perspective, discovering the root cause is a fundamentally simpler task than estimating the causal ordering or learning the causal DAG.

3.2 Root cause discovery algorithm based on all permutations

Theorem 4 inspires a root cause discovery algorithm by trying all permutations of variables. Before presenting this algorithm, we first introduce some notation.

For a permutation $\pi = (\pi(1), \ldots, \pi(p))$, let $X^\pi = (X_{\pi(1)}, \ldots, X_{\pi(p)})^T$ and $X^{I\pi} = (X^I_{\pi(1)}, \ldots, X^I_{\pi(p)})^T$, with their corresponding means and covariance matrix denoted by μ_{X^π}, $\mu_{X^{I\pi}}$ and Σ_{X^π}, respectively. We introduce the following notation to make the dependence of ξ (see (5)) on the permutation π explicit:

$$\xi(\pi) := L_{X^\pi}^{-1}(\mu_{X^{I\pi}} - \mu_{X^\pi}), \tag{7}$$

where L_{X^π} is the lower triangular matrix from the Cholesky decomposition of Σ_{X^π}. Note that $\xi(\pi)$ is not simply a permuted version of ξ due to $L_{X^\pi}^{-1}$.

By Theorem 4, if $\xi(\pi)$ is 1-sparse, the nonzero entry must be in the $\pi^{-1}(r)$th position, corresponding to the root cause's position under the permutation. We call a permutation π *sufficient* if $\xi(\pi)$ is

1-sparse, indicating that this permutation is sufficient to identify the root cause. Otherwise, we call it *insufficient*, in which case $\xi(\pi)$ contains more than one nonzero element. The existence of a sufficient π is clear because a causal ordering is sufficient. As we will show in Section 3.3, noncausal orderings can also be sufficient.

Now, we introduce the sample version algorithm. Let $\boldsymbol{x}_1^\pi, \ldots, \boldsymbol{x}_n^\pi$ be n i.i.d. observational samples of X^π, and let $\boldsymbol{x}^{I\pi}$ be one interventional sample of $X^{I\pi}$. Denote the estimators of L_{X^π} and μ_{X^π} based on these observational samples by $\widehat{L}_{X^\pi}$ and $\widehat{\mu}_{X^\pi}$, respectively. Then, we estimate $\xi(\pi)$ by

$$\widehat{\xi}(\pi) = \widehat{L}_{X^\pi}^{-1}(\boldsymbol{x}^{I\pi} - \widehat{\mu}_{X^\pi}). \tag{8}$$

To quantify the evidence that $\widehat{\xi}(\pi)$ has the 1-sparse pattern and that the corresponding largest entry is the root cause, we use

$$\hat{c}(\pi) = \frac{|\widehat{\xi}(\pi)|_{(1)} - |\widehat{\xi}(\pi)|_{(2)}}{|\widehat{\xi}(\pi)|_{(2)}}, \tag{9}$$

where $|\widehat{\xi}(\pi)|_{(i)}$ denotes the ith largest entry in $|\widehat{\xi}(\pi)|$. Note that $\hat{c}(\pi)$ is infinite if $\widehat{\xi}(\pi)$ is 1-sparse, so a large value of $\hat{c}(\pi)$ indicates that $\widehat{\xi}(\pi)$ is close to having the desired 1-sparse pattern.

By Theorem 4, if π is sufficient, then the root cause is likely to have the largest value in $|\widehat{\xi}(\pi)|$. Therefore, we define the set of all potential root causes as

$$\widehat{U} = \cup_{\pi \in \Pi_{\text{all}}} \hat{u}(\pi), \tag{10}$$

where $\hat{u}(\pi) = \pi(\text{argmax}_{j\in[p]}|\widehat{\xi}(\pi)|_j)$ denotes the potential root cause with respect to a permutation π and Π_{all} is the set of all permutations of $(1, \ldots, p)$.

Based on $\widehat{U}$, we assign a score to each variable i as a measure of its likelihood of being the root cause. This score, which we call the *RC-score*, is defined as:

$$\widehat{C}_i = \begin{cases} \max\limits_{\pi:\hat{u}(\pi)=i} \hat{c}(\pi) & \text{if} \quad i \in \widehat{U}, \\ \hat{w}_i \hat{c}_{\min} & \text{if} \quad i \in [p] \setminus \widehat{U}, \end{cases} \tag{11}$$

where $\hat{c}_{\min} = \min_{i\in\widehat{U}} \widehat{C}_i/2$ is half of the smallest RC-score among potential root causes, and $\hat{w}_i = \widehat{Z}_{n,i}^2 / \sum_{j\in[p]\setminus\widehat{U}} \widehat{Z}_{n,j}^2$ is the weight based on the squared z-scores. This ensures that RC-scores for variables unlikely to be the root cause (i.e. not in $\widehat{U}$) remain smaller than the scores for all potential root causes. A large $\widehat{C}_i$ value indicates that variable i is more likely to be the root cause.

We note that while an alternative approach that assigns a score of zero to all variables not in $\widehat{U}$ is simpler and does not affect the theoretical consistency of the method, using $\hat{w}_i\hat{c}_{\min}$ has an advantage of assigning the root cause a better score in the unfavourable case where it falls outside $\widehat{U}$. For example, when the root cause has no children, even if it happens to be outside $\widehat{U}$, using (11) may still assign it a large score compared to other variables, as it might be the only aberrant variable with the largest squared z-score.

The following theorem shows that if the estimators $\widehat{L}_{X^\pi}$ and $\widehat{\mu}_{X^\pi}$ are both consistent, then using the RC-score (11) will consistently identify the root cause as the sample size and intervention strength go to infinity.

Theorem 5 Let r be the root cause and $\widehat{C}_i$ be obtained by (11). If $\widehat{L}_{X^\pi} \xrightarrow[n\to\infty]{p} L_{X^\pi}$ and $\widehat{\mu}_{X^\pi} \xrightarrow[n\to\infty]{p} \mu_{X^\pi}$, then

$$\lim_{\substack{n\to\infty \\ \delta_r\to\infty}} \text{P}\left(\widehat{C}_r > \max_{k\in[p]\setminus\{r\}} \widehat{C}_k\right) = 1.$$

We apply the RC-score (11) to the previous example in Section 2. Figure 3 shows that our RC-score successfully identifies all root causes, which aligns with our expectations based on the theoretical results.

3.3 Characterization of sufficient permutations and an efficient root cause discovery algorithm

Calculating the RC-score using (11) requires evaluating $p!$ permutations, which is computationally infeasible for large p. The purpose of considering all permutations is to ensure at least one sufficient permutation is found. Therefore, to reduce the search space, it is crucial to understand which permutations are sufficient. To this end, we give a complete characterization in Theorem 6.

Before presenting the theorem, we introduce two notations to be used. For $j \in [p]$, we define its parents and real descendants as

$$\mathrm{Pa}(j) = \{k \in [p] \setminus \{j\} : B_{jk} \neq 0\} \quad \text{and} \quad \mathrm{rDe}(j) = \{k \in [p] \setminus \{j\} : (I - B)^{-1}_{kj} \neq 0\},$$

respectively. Recall that $(I - B)^{-1}_{kj} = \alpha_{j \to k}$ is the total causal effect of X_j on X_k.

Theorem 6 Let r be the root cause. A permutation π is sufficient if and only if the following two conditions hold:

(i) $\pi^{-1}(k) < \pi^{-1}(r)$ for all $k \in \mathrm{Pa}(r)$,
(ii) $\pi^{-1}(k) > \pi^{-1}(r)$ for all $k \in \mathrm{rDe}(r)$.

Theorem 6 shows that sufficient permutations are those where the parents of the root cause are positioned before it, and the real descendants of the root cause are positioned after it. This characterization enables us to develop a method that significantly reduces the search space of permutations.

Specifically, let $D = \{r\} \cup \mathrm{rDe}(r)$ be the set comprising the root cause and its real descendants. If D is known (though the exact root cause r within D does not need to be known), then we can generate $|D|$ permutations that are guaranteed to include a sufficient permutation. Specifically, we generate $|D|$ permutations by

$$\pi = ([p]\backslash D, i, D\backslash\{i\}), \quad i \in D, \tag{12}$$

where the orderings within $[p]\backslash D$ and $D\backslash\{i\}$ are arbitrary. Then, it is clear that the permutation with $i = r$ is sufficient, as it satisfies the conditions in Theorem 6 (see online supplementary material, Appendix A.3 for a concrete example). Although we cannot target this sufficient permutation due to not knowing which variable is r, it suffices for our purpose that one of the $|D|$ permutations is sufficient.

The set D is unknown in practice. However, since the root cause and its real descendants are expected to be aberrant due to the intervention, we can estimate D using the set of aberrant variables. We use the squared z-scores $\widehat{Z}^2_{n,i}$ (see (3)) to form an estimate $\widehat{D} = \{j \in [p] : \widehat{Z}^2_{n,j} \geq \tau\}$ for some threshold τ. It is important to note that our primary goal is to reduce the search space of permutations, and trying more permutations is generally beneficial for root cause discovery. Therefore, we are not restricted to using just $|D|$ permutations generated from a single estimator of D. Instead, we can use multiple thresholds to obtain several sets $\widehat{D}$ and generate more permutations accordingly. For instance, we could use all the squared z-scores $\widehat{Z}^2_{n,1}, \ldots, \widehat{Z}^2_{n,p}$ as thresholds. Because of this, we are not facing the difficult issue of choosing an optimal threshold.

For the same reason, when generating the permutation $\pi = ([p]\backslash D, i, D\backslash\{i\})$ for a certain i (see (12)), we can record additional permutations by using different random orderings of $[p]\backslash D$ and $D\backslash\{i\}$. Although using more permutations does not bring any advantage from an asymptotic perspective, it tends to be beneficial for finite sample performance. The reason is that using only one permutation might result in the undesired case where the root cause is out of $\widehat{U}$ (see (10)), while trying more permutations increases the possibility that the root cause is included in $\widehat{U}$. This benefit comes at the cost of increased computational expense, which grows linearly with the number of

Algorithm 1 Obtain permutations based on squared z-scores

Input: Observational samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the interventional sample $\boldsymbol{x}^I$, and the number of random permutations ν.

Output: A set of permutations.

1: Calculate the squared z-scores $\widehat{Z}^2_{n,1}, \ldots, \widehat{Z}^2_{n,p}$ based on formula (3) and initialize an empty set $\widehat{\Pi}$ to record permutations.

2: **for** $i = 1, \ldots, p$ **do**

Obtain the set of aberrant variables $D = \{j \in [p] : \widehat{Z}^2_{n,j} \geq \widehat{Z}^2_{n,i}\} := \{d_1, \ldots, d_u\}$.

3: **for** $l = 1, \ldots, u$ **do**

4: **for** $k = 1, \ldots, \nu$ **do**

Randomly permute $[p]\backslash D$ and $D\backslash\{d_l\}$ to generate a permutation $\pi = ([p]\backslash D, d_l, D\backslash\{d_l\})$, and add π to the permutation set $\widehat{\Pi}$.

5: **end for**

6: **end for**

7: **end for**

8: **Return** $\widehat{\Pi}$.

permutations ν. We suggest using a moderate value of ν, depending on the available computational resources. See also online supplementary material, Appendix E.1 for empirical run-times and some more discussion.

We summarize the method for obtaining permutations in Algorithm 1. This algorithm generates a total of $\nu p(p+1)/2$ permutations. In practice, to reduce computational time, one can use some reasonable thresholds such as $\{1.5, 2, 2.5, \ldots, 5\}$ instead of using all squared z-scores in step 2 of this algorithm. In addition, instead of squared z-scores, alternative approaches could be useful for estimating aberrant variable sets in Algorithm 1 in different scenarios. For example, estimated tail probabilities (or the information-theoretic score in Budhathoki et al., 2022) could be beneficial when the marginal distributions of variables exhibit strongly different tail behaviours.

In Theorem 7, we show that as the sample size and intervention strength tend to infinity, one of the permutations outputted by Algorithm 1 is guaranteed to be sufficient with a probability tending to one.

Theorem 7 Let $\widehat{\Pi}$ be the set of permutations obtained by Algorithm 1, then

$$\lim_{\substack{n\to\infty \\ \delta_r\to\infty}} \mathrm{P}\Big(\widehat{\Pi} \text{ contains at least one sufficient permutation}\Big) = 1.$$

By applying Algorithm 1, we replace the set of all permutations in (10) with those generated by this algorithm. This significantly reduces computational expense and leads to our main root cause discovery method (Algorithm 2). As shown in the following Theorem 8, if both $\widehat{L}_{X^\pi}$ and $\widehat{\mu}_{X^\pi}$ are consistent, Algorithm 2 is guaranteed to assign the root cause the largest score with a probability approaching one, as the sample size and intervention strength increase to infinity.

Theorem 8 Let r be the root cause and $\widehat{C}_i$ be obtained by Algorithm 2. If $\widehat{L}_{X^\pi} \xrightarrow[n\to\infty]{p} L_{X^\pi}$ and $\widehat{\mu}_{X^\pi} \xrightarrow[n\to\infty]{p} \mu_{X^\pi}$, then

$$\lim_{\substack{n\to\infty \\ \delta_r\to\infty}} \mathrm{P}\left(\widehat{C}_r > \max_{k\in[p]\backslash\{r\}} \widehat{C}_k\right) = 1.$$

3.4 An adapted root cause discovery algorithm for high-dimensional settings

For Algorithm 2 to perform well, it is important to obtain an accurate estimate of the covariance matrix, as it is used in the Cholesky decomposition to get $\widehat{L}_{X^\pi}$. We found that obtaining a good covariance matrix is particularly challenging for the high-dimensional gene expression data analysed in Section 5, which involves around 20,000 variables and 400 samples. Estimating

Algorithm 2 Root cause discovery

Input: Observational samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and the interventional sample $\boldsymbol{x}^I$, and the number of random permutations ν.

Output: RC-score $\widehat{C}_i$ for all variables.

1: Obtain $\widehat{\Pi}$ by implementing Algorithm 1 with the same inputs as this algorithm.

2: Initialize an empty set $\widehat{U}$.

3: **for** $\pi \in \widehat{\Pi}$ **do:**

(i) Obtain permuted samples $\boldsymbol{x}_1^\pi, \ldots, \boldsymbol{x}_n^\pi$ and $\boldsymbol{x}^{I\pi}$.

(ii) Obtain estimators $\widehat{L}_{X^\pi}$ and $\widehat{\mu}_{X^\pi}$ based on $\boldsymbol{x}_1^\pi, \ldots, \boldsymbol{x}_n^\pi$.

(iii) Get $\widehat{\xi}(\pi) = \widehat{L}_{X^\pi}^{-1}(\boldsymbol{x}^{I\pi} - \widehat{\mu}_{X^\pi})$, $\hat{c}(\pi) = \frac{|\widehat{\xi}(\pi)|_{(1)} - |\widehat{\xi}(\pi)|_{(2)}}{|\widehat{\xi}(\pi)|_{(2)}}$, and $\hat{u}(\pi) = \pi(\text{argmax}_{j\in[p]}|\widehat{\xi}(\pi)|_j)$.

(iv) Add $\hat{u}(\pi)$ into $\widehat{U}$.

4: **end for**

5: **for** $i \in \widehat{U}$: Let $\widehat{C}_i = \max_{\pi:\pi\in\widehat{\Pi} \text{ and } \hat{u}(\pi)=i} \hat{c}(\pi)$.

6: **end for**

7: **for** $i \in [p] \setminus \widehat{U}$ **do:** Let $\widehat{C}_i = \hat{w}_i \hat{c}_{\min}$, where $\hat{c}_{\min} = \min_{j\in\widehat{U}} \widehat{C}_j/2$ and $\hat{w}_i = \widehat{Z}_{n,i}^2 / \sum_{j\in[p]\setminus\widehat{U}} \widehat{Z}_{n,j}^2$ with $\widehat{Z}_{n,j}^2$ defined by (3).

8: **end for**

9: Return $\widehat{C}_i$ for all $i \in [p]$.

covariance matrices in high-dimensional settings is known to be difficult and is an independent research topic from our problem (see, e.g. Cai et al., 2016 and Fan et al., 2016 for reviews on this topic). We therefore propose a special adaptation of our method that circumvents estimating a high-dimensional covariance matrix. We build on Algorithm 2 and incorporating Lasso (Tibshirani, 1996) for dimension reduction in a node-wise manner.

We first present the main idea. Consider one variable as the response and focus on the subsystem containing that variable and its Markov blanket (i.e. all variables that are conditionally dependent on it given the rest). Then, two key observations are: (i) If the chosen response variable is not the root cause, then it is not the root cause within the subsystem. Specifically, if this variable is influenced by the intervention, its parents, which are included in the Markov blanket, must also be influenced, indicating it cannot be the root cause; if it is not influenced by the intervention, then it clearly cannot be the root cause. (ii) If the chosen response variable is the root cause, it remains the root cause within the subsystem.

Building on these two observations, we treat each variable i as the response and apply cross-validated Lasso to estimate its Markov blanket. We then implement the first three steps of Algorithm 2 on the subdataset corresponding to the subsystem containing variable i and its estimated Markov blanket. If i is indeed the root cause, it remains the root cause within this subsystem. Consequently, there should exist a permutation π for which $\xi(\pi)$ [see (8)] is 1-sparse, with $|\xi(\pi)|_{\pi^{-1}(i)}$ being the largest value. Hence, the corresponding $\hat{c}(\pi)$ [see (9)] is a reasonable measure of the likelihood that it is the root cause. By considering all possible permutations, we set the root cause score for variable i as the largest $\hat{c}(\pi)$ among those π for which $|\xi(\pi)|_{\pi^{-1}(i)}$ has the largest value, as implemented in Algorithm 2. Lastly, we assign scores to variables outside $\widehat{U}$ in a similar manner as (11). We summarize the root cause discovery method for high-dimensional settings in online supplementary material, Algorithm 3 (see online supplementary material, Appendix B).

We point out a caveat regarding the high-dimensional version of the root cause discovery algorithm: after dimension reduction, latent variables may be introduced into the subsystem before applying Algorithm 2. To assess the robustness of Algorithm 2 to latent variables, we conduct simulations in latent variable settings in online supplementary material, Appendix E.6. Simulation results indicate that this algorithm is quite robust to latent variables. We also evaluate the performance of the high-dimensional root cause discovery algorithm in high-dimensional settings, see online supplementary material, Appendix E.7.

We emphasize that this high-dimensional root cause discovery algorithm is a heuristic algorithm, and no theoretical guarantees are provided here. We present it because it performs very well in discovering the disease-causing genes in the genetic application (see Section 5). While it would be interesting to investigate its theoretical properties in high-dimensional settings, this would require studying the theoretical behaviour of Algorithm 2 in the latent variable settings, which is beyond the scope of this paper. So we leave this for future research.

4 Simulations

We now evaluate the finite sample performance of our proposed RC-score (Algorithm 2). All simulations are carried out in Python, and the code is available at GitHub (https://github.com/Jinzhou-Li/RootCauseDiscovery).

4.1 Simulation setup and implemented methods

We generate observational and interventional samples according to models (1) and (2), respectively. Specifically, for the intercept term b, we randomly sample its entries from the uniform distribution $U(-10, 10)$. For each component of the error term ε, we consider either a Gaussian distribution with mean zero or a uniform distribution $U(-a, a)$. The variance σ^2 of each error component is independently sampled from $U(1, 2)$, ensuring errors have nonequal variances (Ng et al., 2024). This variance is used directly for the Gaussian distribution, and for the uniform distribution, we set $a = \sqrt{3\sigma^2}$ to achieve the desired variance. For the matrix B, we consider two types corresponding to the random DAG and the hub DAG. The random DAG, commonly used in the literature, connects nodes randomly with a probability s. The hub DAG is motivated by genetic interactions, where certain genes (known as hub genes) act as central connectors and significantly influence the overall network behaviour. There are $p = 100$ and $p = 104$ variables for the random DAG and the hub DAG, respectively. See online supplementary material, Appendix E.2.1 for more details.

To prevent the marginal variances of $X = (X_1, \ldots, X_p)$ from increasing along the causal ordering (Reisach et al., 2021). we first sample targeted variances for each variable X_i from $U(10, 50)$. We then rescale the nonzero entries of the matrix B and update the error variances of source nodes (i.e. those without parents) to ensure that the final variances of X_i's are close to the targeted values. Finally, we randomly permute the ordering of the variables so that they are not sorted according to a causal ordering (with b, ε and B permuted accordingly).

For each setting, we randomly generate 20 matrices B. For each B, we generate n observational samples. Moreover, we randomly choose 50 root causes, and generate one interventional sample with intervention effect δ_r for each root cause. The n observational samples are then used for root cause discovery on each of the 50 interventional samples. In total, there are $m = 1{,}000$ interventional samples with possibly different root causes.

To investigate the effects of sample size n, the intervention strength δ_r, and the sparsity level s, we consider the following scenarios: (i) Vary $n \in \{100, 200, 300\}$ while fixing $s = 0.4$ and $\delta_r = 8$; (ii) Vary $\delta_r \in \{4, 8, 12\}$ while fixing $s = 0.4$ and $n = 200$; (iii) Vary $s \in \{0.2, 0.4, 0.6\}$ while fixing $n = 200$ and $\delta_r = 8$.

Given n observational samples and one interventional sample, we calculate the squared z-scores and our proposed RC-scores (Algorithm 2) for each variable in the interventional sample. In addition, we implement two methods mentioned in Section 1.3 that require estimating the causal ordering or DAG, for which we use LiNGAM (Shimizu et al., 2011; see online supplementary material, Appendix E.2.2 for details on their implementation). LiNGAM can consistently estimate a causal ordering or DAG in linear non-Gaussian settings, which are part of our simulation setup. However, in linear Gaussian settings, which are also part of our simulation setup, the causal ordering and DAG are generally nonidentifiable. Below, we summarize the implemented methods as follows:

1. Squared z-score: Based on formula (3). This method is denoted as 'Z-score' in the plots.
2. The approach based on an estimated causal ordering and aberrant set: The causal ordering is estimated using the Python package *lingam* (Ikeuchi et al., 2023). We denote the three methods with the optimal threshold and thresholds of 2 and 5 for obtaining the aberrant variable

set as 'LiNGAM-opt', 'LiNGAM-2', and 'LiNGAM-5', respectively, in the plots. In particular, 'LiNGAM-opt' uses the squared z-score of the root cause as the threshold, which requires oracle information and is thus infeasible in practice. We present its results to illustrate the best possible performance achievable by such approaches.

3. The approach based on an estimated DAG and residuals: The DAG is estimated using the same Python package *lingam* as above. We denote this method as 'LiNGAM-Inva' in the plots.
4. RC-score: Implemented using Algorithm 2 using $v = 10$ random permutations and thresholds $(0.1, 0.3, \ldots, 5)$ in its first step. The estimator $\widehat{L}_{X^\pi}$ in step 2 is obtained by applying the Cholesky decomposition on the estimated covariance matrix, for which the sample covariance matrix is used when $n > p$, and a shrinkage estimator is used when $n < p$ (Schäfer and Strimmer (2005), we use the Python function *sklearn.covariance.ShrunkCovariance* for its implementation). This method is denoted as 'RC-score' in the plots.

4.2 Simulation results

Based on the obtained scores, we calculate and record the rank of the root cause for each interventional sample. A smaller rank indicates a larger score for the root cause, so rank 1 is the best. In total, we obtain 1,000 ranking values for each method, where smaller values indicate better performance. To compare their performances, we plot the cumulative distribution function (CDF) of the root cause rank for each method. The value of the CDF at $x = k$ represents the percentage of times the root cause is ranked in the top k among the 1,000 interventional samples. The results for the hub DAG with Gaussian errors are shown in Figure 4. The results for the hub DAG with uniform errors and the random DAG with uniform or Gaussian errors are shown in online supplementary material, Appendix E.3.

Our proposed RC-score outperforms the squared z-score, LiNGAM-Inva, LiNGAM-2, and LiNGAM-5 in all considered settings, and it also outperforms LiNGAM-opt in most settings. In particular, LiNGAM-2 and LiNGAM-5 perform worse than or similarly to the squared z-score, whereas LiNGAM-opt consistently outperforms the squared z-score, as expected. We plot the optimal thresholds (i.e. the squared z-score of the root cause) used in LiNGAM-opt across all settings in online supplementary material, Appendix E.4, and these plots show significant variation in the optimal thresholds across the 1,000 interventional samples, indicating that no single fixed optimal threshold exists for all scenarios. Even with the optimal threshold, LiNGAM-opt does not seem to perform very well. The performance of LiNGAM-Inva can be good if the causal DAG is well-estimated, such as in the uniform errors setting with a large sample size (see the plot with $n = 300$ in online supplementary material, Figure 11 in Appendix E.3), which is particularly favourable for using LiNGAM. Lastly, all methods tend to perform better with larger sample sizes or stronger interventions.

The run-times for the methods are reported in online supplementary material, Appendix E.5. To give a quick idea, in the setting of a random DAG with Gaussian errors (with sample size $n = 200$, sparsity level $s = 0.4$, and intervention strength $\delta_r = 8$), one run takes on average 0.00034 s for the squared z-score method, about 17 s for our proposed RC score method, and about 200 s for LiNGAM-based approaches.

5 Real application on a gene expression dataset

5.1 Data description and implemented method

In this section, we analyse the gene expression dataset mentioned in Section 1, which motivates this paper. The dataset comprises gene expression data in the form of RNA-sequencing read counts from skin fibroblasts of 423 individuals with a suspected Mendelian disorder, including 154 nonstrand-specific and 269 strand-specific RNA-sequencing samples. The nonstrand-specific dataset is available at https://zenodo.org/records/4646823 (Yépez, Gusic, et al., 2021), and the strand-specific dataset is available at https://zenodo.org/records/7510836 (Yépez, Gusic, Kopajtich, Meitinger, et al., 2022). These two datasets were sequenced using different protocols, yielding somewhat different distributions (17,133 differentially expressed genes out of 45,960 at an FDR of 10% using DESeq2 (Love et al., 2014)). Combining them increases the sample size but also introduces some confounding, leading to a bias-variance trade-off. Hence, it is generally hard

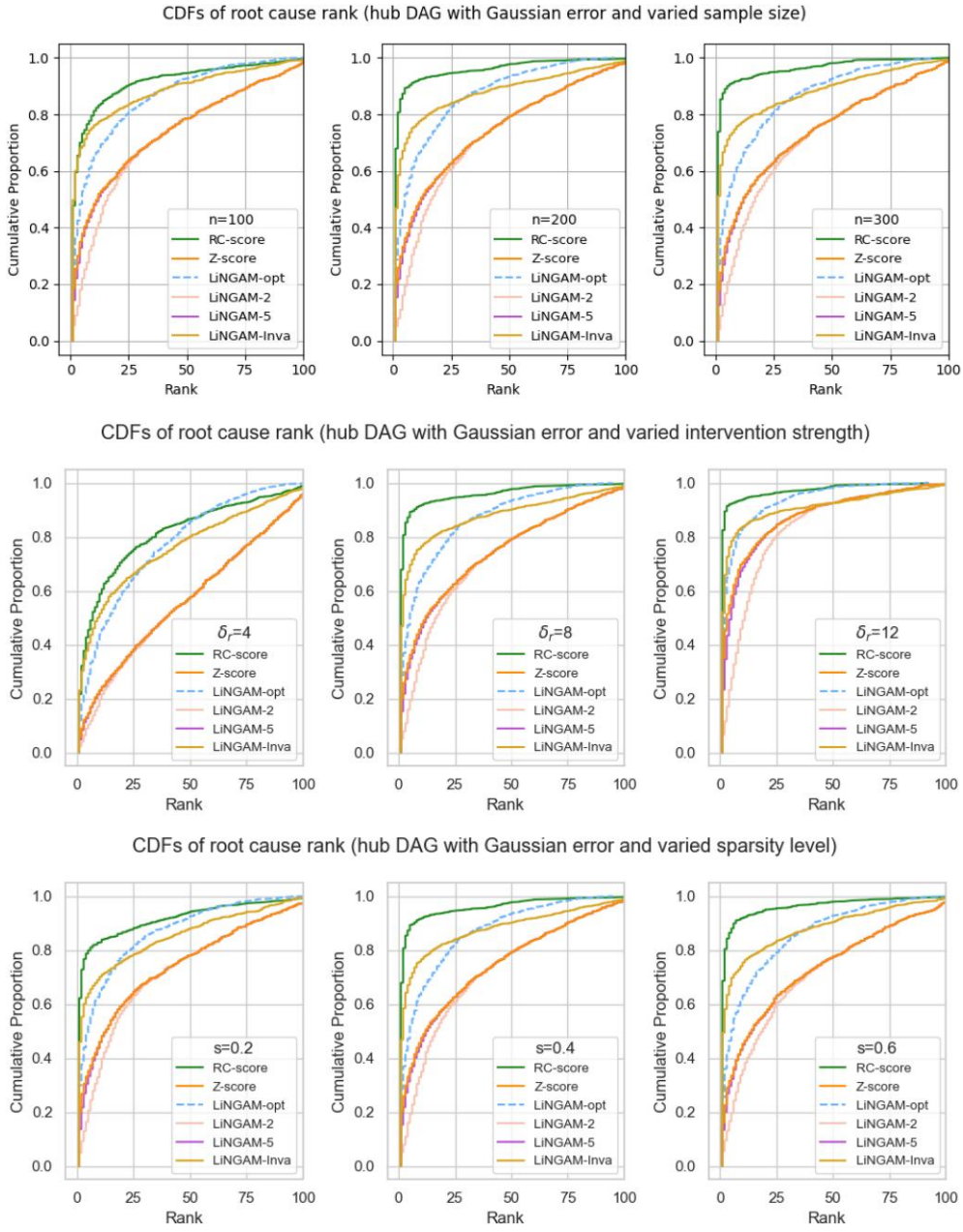

Figure 4. Cumulative distribution functions (CDFs) of the root cause rank using the squared z-score, RC-score, and LiNGAM-based approaches in the setting with a hub directed acyclic graph (DAG) and Gaussian errors. The top, middle, and bottom plots display results for varying sample sizes, intervention strengths, and sparsity levels, respectively.

to predict when combining is beneficial. In our application, however, we have a ground truth, and could therefore compare the results on the combined and separate datasets. We found that the combined dataset led to better results (see online supplementary material, Figure 19 in Appendix E.8). For future research, it would be interesting to explicitly model the sequencing protocol as a measured confounder and develop corresponding root cause discovery methods.

Among the 423 patients, 58 have (likely) known genetic mutations (see online supplementary material, Tables S2, S3 and S4 in Additional file 1 of Yépez, Gusic, Kopajtich, Mertes, et al., 2022).

We apply our method to identify the disease-causing gene in these patients and compare the results with the aforementioned genetic mutations, which serve as the ground truth.

When applying our method to one patient, we treat the other patients as observational samples. While this approach may not be ideal for detecting aberrancy and identifying the root cause, it is reasonable here because the aberrant genes in these rare disease patients are likely to be different. Better results are expected if gene expression data from healthy individuals were available as a reference.

We first apply some preprocessing and quality control steps to the raw gene expression data. Specifically, we filter out genes with counts <10 in more than 90% of the samples and remove genes that are highly correlated with others (marginal correlation >0.999). For the remaining genes, we follow the preprocessing procedure described by Brechtmann et al. (2018), applying a log-transformation to better satisfy the linearity assumption and dividing by a size factor (Anders & Huber, 2010) to account for sequencing depth. This results in $p = 19{,}736$ genes and $n = 423$ samples in the pre-processed gene expression data.

We apply online supplementary material, Algorithm 3 with $\nu = 20$ random permutations for each of the 58 patients for which we have a ground truth. To reduce computational time, in step 2 of online supplementary material, Algorithm 3, we treat only the variables with squared z-scores >1.5 as responses. For comparison, we implement the squared z-score method. We also implemented the LiNGAM-based methods (using *DirectLiNGAM* and its high-dimensional version *HighDimDirectLiNGAM*, both from the Python package *lingam*). However, the computation for a single patient did not complete even after 7 days on our university's computing clusters. As a result, we do not include these methods in our comparison.

5.2 Results

Figure 5 shows the raw transformed gene expression levels, squared z-scores, and RC-scores of genes for two representative patients, *R*16472 and *R*96820. In both cases, the raw gene expression values of the root cause are obscured by other genes due to the propagation of the intervention effect, which leads to many aberrant genes. For patient *R*16472, the squared z-score method successfully identifies the root cause. This is in line with our results in Section 2, which show that this method can be effective in some cases. The RC-score also assigns the root cause the highest score for this patient. However, for patient *R*96820, the squared z-score fails to distinguish the root cause while the RC-score was able to assign the root cause the second largest score. Plots for the other patients are provided in online supplementary material, Appendix E.9.

Figure 6 shows the root cause ranks based on the squared z-score and the RC-score for all 58 patients. In this table, for 17 out of 58 patients, both methods identify the root cause correctly. For one patient, they both assign rank 2. For 30 patients, the RC-score assigns a smaller rank to the root cause, indicating that it performs better than the squared z-score. In contrast, for 10 patients, the squared z-score outperforms the RC-score. It is worth noting that the RC-score often improves by a significant margin compared to the squared z-score. For example, for patients *R*59185 and *R*34834, the RC-score assigns the root cause a rank of 1, whereas the squared z-score gives ranks of 3,057 and 1,657, respectively. In cases where the squared z-score performs better, the rank differences tend to be smaller.

We also count how many patients have the true root cause ranked in the top k for each method, and show it in Figure 7. It is evident that the RC-score method outperforms the squared z-score. Specifically, using the RC-score, the root cause is ranked first in 28 out of 58 patients, in the top 5 in 38 patients, in the top 10 in 46 patients, and in the top 20 in 51 patients. Our method does not work well for the patients *R*18626, *R*46723, and *R*12128. One potential reason is that the genetic correlation structures around the root causes and their descendants in these patients are particularly complex, leading to inaccurate estimates of the associated Markov blankets and covariance matrices.

Overall, these results show that our proposed method can be useful for discovering disease-causing genes based on gene expression data. In particular, considering that we are not in the ideal scenario where gene expression data from healthy individuals are available and with a large sample size, we expect our method to be even more effective if such a better reference is available.

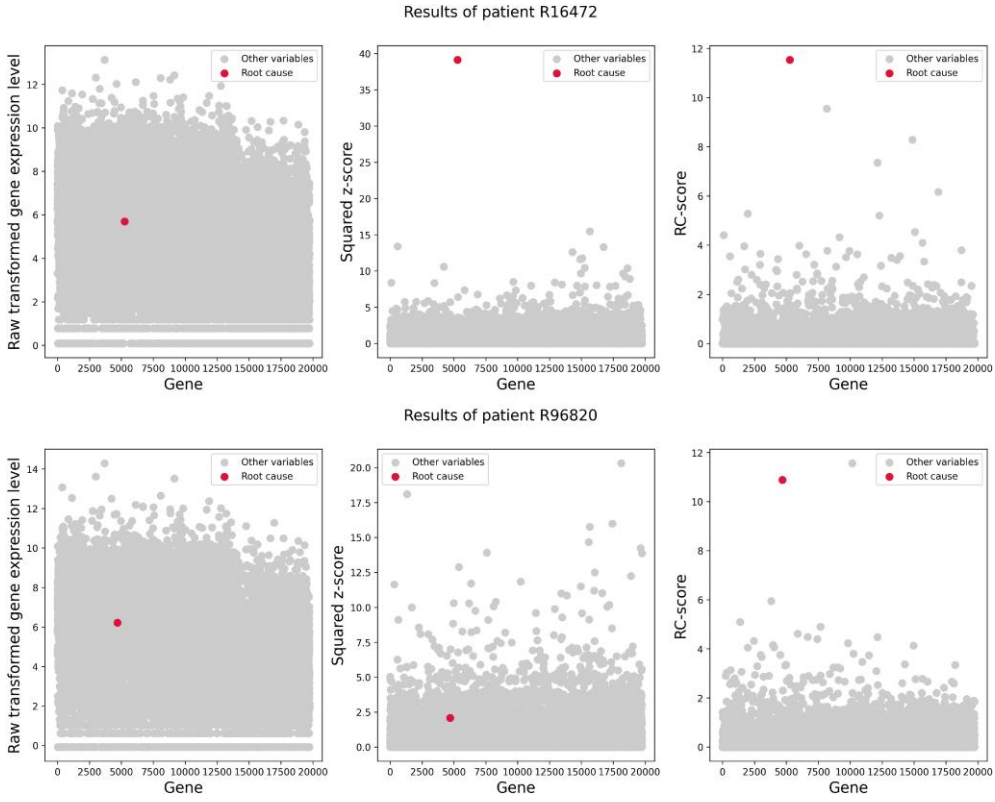

Figure 5. The raw transformed gene expression levels, squared z-scores, and RC-scores of genes for patients *R*16472 and *R*96820.

Patient ID	R19100	R61100	R77611	R16472	R28774	R64921	R80184	R59185	R91273	R60537
RC-score rank	1	1	1	1	1	1	1	1	1	1
Squared z-score rank	1	2	10	1	1	3	11	3057	2	1
Patient ID	R82353	R34834	R30367	R45867	R31640	R55237	R34820	R19907	R27473	R30525
RC-score rank	1	1	1	1	1	1	1	1	1	1
Squared z-score rank	1	1657	1	1	1	1	5	1	38	1
Patient ID	R64055	R15748	R66814	R77365	R42505	R64948	R21470	R47816	R62943	R96820
RC-score rank	1	1	1	1	1	1	1	1	2	2
Squared z-score rank	1	1	1	1	1	9	19	1	770	1633
Patient ID	R91016	R54158	R21147	R64046	R25473	R52016	R21993	R11258	R36605	R95723
RC-score rank	2	2	3	3	4	4	4	5	6	6
Squared z-score rank	2	1	1	15	91	1	2002	31	262	115
Patient ID	R15264	R24289	R76358	R89912	R44456	R20754	R72253	R70186	R98254	R75000
RC-score rank	7	7	7	8	8	9	13	14	17	18
Squared z-score rank	1	21	6	60	83	718	2	2	2289	585
Patient ID	R51757	R78764	R29620	R80346	R26710	R18626	R46723	R12128		
RC-score rank	19	25	52	71	95	610	1028	4475		
Squared z-score rank	29	4	268	1799	576	227	1326	1011		

Figure 6. Table showing the rank of the root cause for 58 patients based on the squared z-score and the RC-score. A smaller rank is better, with rank 1 being the best.

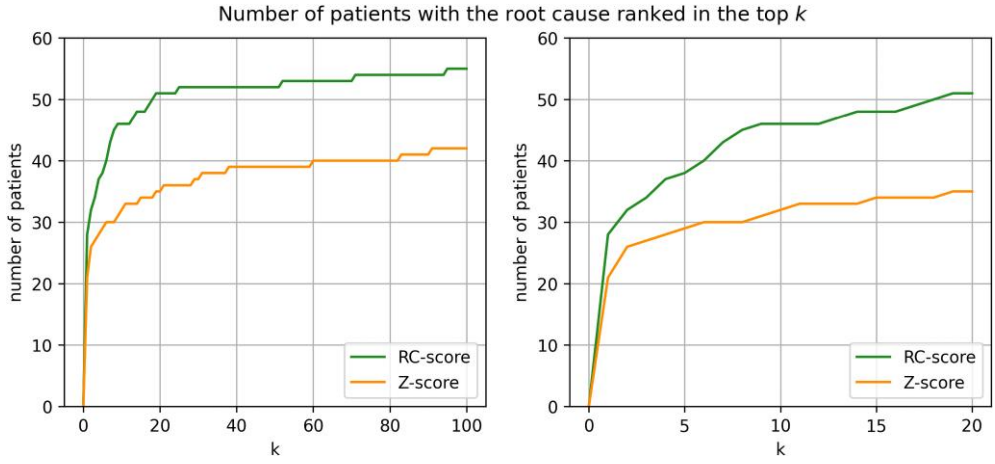

Figure 7. Number of patients with the root cause ranked in the top k based on the squared z-score and the RC-score. The right plot is the zoom-in version of the left plot for $k = 1, \ldots, 20$.

6 Discussion

There are many interesting directions for follow-up research. One direction involves cases with latent variables. Although our simulations (see online supplementary material, Appendix E.6) indicate some robustness of our method to latent variables, a formal study is important and desired. This is also closely related to a deeper understanding of the theoretical properties of online supplementary material, Algorithm 3, which is designed for high-dimensional settings.

With respect to the intervention types, we focus on mean-shift interventions in this paper because they are reasonable in genetic applications. There are also other types of interventions in the causal literature, such as do-intervention and variance-shift intervention (see, e.g. Eberhardt & Scheines, 2007; Pearl, 2009a). Investigating how to conduct root cause discovery in these settings is an interesting topic for future research.

Furthermore, in many applications, there may be multiple root causes, nonlinear relationships between variables, or feedback loops. Thus, generalizing the current methodology to address these complexities would be important.

Quantifying the uncertainty of our method would be very valuable. This seems challenging, because there are essentially two sources of uncertainty, coming from the unknown causal order and the unknown root cause. Ideally, we would like to develop a method that captures both, so that we can output a set of causes with a theoretical guarantee that the true root cause is inside this set with high probability.

Finally, the main idea of using the Cholesky decomposition and permutations to exploit an invariant property opens up new possibilities of utilizing interventional samples. It would be interesting to leverage this idea to develop new methods in structure learning (see, e.g. Huang et al., 2020), active causal learning, and experimental design where heterogeneous interventional samples are available.

Acknowledgments

We are grateful to the anonymous reviewers for their constructive and valuable comments, which greatly improved the manuscript. We thank Dominik Janzing and Daniela Schkoda for sharing with us a simplified proof of Lemma D.1, which has been included in online supplementary material D.1.

Conflicts of interest: None declared.

Funding

J.L. gratefully acknowledges support by the Swiss National Science Foundation (SNSF) Grant P500PT-210978. B.B.C. gratefully acknowledges support by the grants R01MH113078,

R56HG010812, R01MH123157, and the Stanford Biomedical Informatics National Library of Medicine (NLM) Training Grant T15 LM007033–40. J.G. and I.F.S. gratefully acknowledges support by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via the IT Infrastructure for Computational Molecular Medicine (project #461264291).

Data availability

The gene expression dataset analysed in Section 5 is publicly available and can be accessed through the references cited therein.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Anders S., & Huber W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10), R106. https://doi.org/10.1186/gb-2010-11-10-r106

Brechtmann F., Mertes C., Matusevičiūtė A., Yépez V. A., Avsec Ž., Herzog M., Bader D. M., Prokisch H., & Gagneur J. (2018). OUTRIDER: A statistical method for detecting aberrantly expressed genes in RNA sequencing data. *American Journal of Human Genetics*, *103*(6), 907–917. https://doi.org/10.1016/j.ajhg.2018.10.025

Budhathoki K., Janzing D., Bloebaum P., & Ng H. (2021). Why did the distribution change? In *International Conference on Artificial Intelligence and Statistics* (pp. 1666–1674). PMLR.

Budhathoki K., Minorics L., Blöbaum P., & Janzing D. (2022). Causal structure-based root cause analysis of outliers. In *International Conference on Machine Learning* (pp. 2357–2369). PMLR.

Cai T. T., Ren Z., & Zhou H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. *Electronic Journal of Statistics*, *10*, 1–59. https://doi.org/10.1214/15-EJS1081

Dawid A. P., Faigman D. L., & Fienberg S. E. (2014). Fitting science into legal contexts: Assessing effects of causes or causes of effects? *Sociological Methods & Research*, *43*(3), 359–390. https://doi.org/10.1177/0049124113515188

Eaton D., & Murphy K. (2007). Exact Bayesian structure learning from uncertain interventions. In *International Conference on Artificial Intelligence and Statistics* (pp. 107–114). PMLR.

Eberhardt F., & Scheines R. (2007). Interventions and causal inference. *Philosophy of Science*, *74*(5), 981–995. https://doi.org/10.1086/525638

Ebtekar A., Wang Y., & Janzing D. (2025). 'Toward universal laws of outlier propagation', arXiv, arXiv:2502.08593, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2502.08593

Evans R. J. (2020). Model selection and local geometry. *Annals of Statistics*, *48*(6), 3513–3544. https://doi.org/10.1214/19-AOS1940

Fan J., Liao Y., & Liu H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, *19*(1), C1–C32. https://doi.org/10.1111/ectj.12061

Gelman A., & Imbens G. (2013). *Why ask why? Forward causal inference and reverse causal questions* (Technical Report) National Bureau of Economic Research.

Gusic M., & Prokisch H. (2021). Genetic basis of mitochondrial diseases. *FEBS Letters*, *595*(8), 1132–1158. https://doi.org/10.1002/feb2.v595.8

Haavelmo T. (1944). The probability approach in econometrics. *Econometrica: Journal of the Econometric Society*, *12*, iii–115. https://doi.org/10.2307/1906935

Hardt M., Orchard W., Blöbaum P., Kasiviswanathan S., & Kirschbaum E. (2023). 'The PetShop dataset–finding causes of performance issues across microservices', arXiv, arXiv:2311.04806, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2311.04806

Huang B., Zhang K., Zhang J., Ramsey J., Sanchez-Romero R., Glymour C., & Schölkopf B. (2020). Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, *21*(89), 1–53. https://dl.acm.org/doi/abs/10.5555/3455716.3455805

Ikeuchi T., Ide M., Zeng Y., Maeda T. N., & Shimizu S. (2023). Python package for causal discovery based on LiNGAM. *Journal of Machine Learning Research*, *24*(14), 1–8. https://dl.acm.org/doi/abs/10.5555/3648699.3648713

Ikram A., Chakraborty S., Mitra S., Saini S., Bagchi S., & Kocaoglu M. (2022). Root cause analysis of failures in microservices through causal discovery. *Advances in Neural Information Processing Systems*, *35*, 31158–31170. https://dl.acm.org/doi/10.5555/3600270.3602529

Jaber A., Kocaoglu M., Shanmugam K., & Bareinboim E. (2020). Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in Neural Information Processing Systems*, *33*, 9551–9561. https://dl.acm.org/doi/10.5555/3495724.3496525

Jakobsen M. E., Shah R. D., Bühlmann P., & Peters J. (2022). Structure learning for directed trees. *Journal of Machine Learning Research*, *23*, 159. https://dl.acm.org/doi/10.5555/3586589.3586748

Janzing D., Budhathoki K., Minorics L., & Blöbaum P. (2019). 'Causal structure based root cause analysis of outliers', arXiv, arXiv:1912.02724, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.1912.02724

Li M., Li Z., Yin K., Nie X., Zhang W., Sui K., & Pei D. (2022). Causal inference-based root cause analysis for online service systems with intervention recognition. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 3230–3240).

Love M. I., Huber W., & Anders S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. https://doi.org/10.1186/s13059-014-0550-8

Nandy P., Maathuis M. H., & Richardson T. S. (2017). Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *Annals of Statistics*, *45*(2), 647–674. https://doi.org/10.1214/16-AOS1462

Ng I., Huang B., & Zhang K. (2024). Structure learning with continuous optimization: A sober look and beyond. In *Causal Learning and Reasoning* (pp. 71–105). PMLR.

Okati N., Mejia S. H. G., Orchard W. R., Blöbaum P., & Janzing D. (2024). 'Root cause analysis of outliers with missing structural knowledge', arXiv, arXiv:2406.05014, preprint: not peer reviewed. https://doi.org/10.48550/arXiv.2406.05014

Pearl J. (2009a). Causal inference in statistics: An overview. *Statistics Surveys*, *3*(none), 96–146. https://doi.org/10.1214/09-SS057

Pearl J. (2009b). *Causality*. Cambridge University Press.

Pearl J. (2015). Causes of effects and effects of causes. *Sociological Methods & Research*, *44*(1), 149–164. https://doi.org/10.1177/0049124114562614

Peters J., Bühlmann P., & Meinshausen N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *78*(5), 947–1012. https://doi.org/10.1111/rssb.12167

Raskutti G., & Uhler C. (2018). Learning directed acyclic graph models based on sparsest permutations. *Stat*, *7*(1), e183. https://doi.org/10.1002/sta4.v7.1

Reisach A., Seiler C., & Weichwald S. (2021). Beware of the simulated DAG! Causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, *34*, 27772–27784. https://dl.acm.org/doi/10.5555/3540261.3542388

Rothenhäusler D., Heinze C., Peters J., & Meinshausen N. (2015). BACKSHIFT: Learning causal cyclic graphs from unknown shift interventions. *Advances in Neural Information Processing Systems*, *28*. https://dl.acm.org/doi/10.5555/2969239.2969408

Schäfer J., & Strimmer K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1). https://doi.org/10.2202/1544-6115.1175

Schölkopf B., Locatello F., Bauer S., Ke N. R., Kalchbrenner N., Goyal A., & Bengio Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612–634. https://doi.org/10.1109/JPROC.2021.3058954

Shimizu S., Hoyer P. O., Hyvärinen A., Kerminen A., & Jordan M. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, *7*(10). https://dl.acm.org/doi/10.5555/1248547.1248619

Shimizu S., Inazumi T., Sogawa Y., Hyvarinen A., Kawahara Y., Washio T., Hoyer P. O., Bollen K., & Hoyer P. (2011). DirectLiNGAM: A direct method for learning a linear non-gaussian structural equation model. *Journal of Machine Learning Research: JMLR*, *12*, 1225–1248. https://dl.acm.org/doi/10.5555/1953048.2021040

Spirtes P., Glymour C., & Scheines R. (2001). *Causation, prediction, and search*. MIT Press.

Squires C., Wang Y., & Uhler C. (2020). Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence* (pp. 1039–1048). PMLR.

Taeb A., Gamella J., Heinze-Deml C., & Bühlmann P. (2024). Learning and scoring Gaussian latent variable causal models with unknown additive interventions. *The Journal of Machine Learning Research*, *25*(01). https://dl.acm.org/doi/abs/10.5555/3722577.3722870

Tibshirani R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tramontano D., Monod A., & Drton M. (2022). Learning linear non-Gaussian polytree models. In *Conference on Uncertainty in Artificial Intelligence* (pp. 1960–1969). PMLR.

Tramontano D., Waldmann L., Drton M., & Duarte E. (2023). Learning linear Gaussian polytree models with interventions. *IEEE Journal on Selected Areas in Information Theory*, *4*(72), 569–578. https://doi.org/10.1109/JSAIT.2023.3328429

Varici B., Shanmugam K., Sattigeri P., & Tajer A. (2021). Scalable intervention target estimation in linear models. *Advances in Neural Information Processing Systems*, *34*, 1494–1505. https://dl.acm.org/doi/10.5555/3540261.3540376

Varici B., Shanmugam K., Sattigeri P., & Tajer A. (2022). Intervention target estimation in the presence of latent variables. In *Conference on Uncertainty in Artificial Intelligence* (pp. 2013–2023). PMLR.

Yang Y., Salehkaleybar S., & Kiyavash N. (2024). Learning unknown intervention targets in structural causal models from heterogeneous data. In *International Conference on Artificial Intelligence and Statistics* (pp. 3187–3195). PMLR.

Ye Q., Amini A. A., & Zhou Q. (2020). Optimizing regularized Cholesky score for order-based learning of Bayesian networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(10), 3555–3572. https://doi.org/10.1109/TPAMI.2020.2990820

Yépez V. A., Gusic M., Kopajtich R., Meitinger T., Gagneur J., & Prokisch H. (2021). Gene expression and splicing counts from the Yépez, Gusic et al study - non strand-specific. https://doi.org/10.5281/zenodo.4646823

Yépez V. A., Gusic M., Kopajtich R., Meitinger T., Prokisch H., & Gagneur J. (2022). Gene expression and splicing counts from the Yépez, Gusic et al study - fibroblast, hg19, strand-specific, high seq depth. https://doi.org/10.5281/zenodo.7510836

Yépez V. A., Gusic M., Kopajtich R., Mertes C., Smith N. H., Alston C. L., Ban R., Beblo S., Berutti R., Blessing H., Ciara E., Distelmaier F., Freisinger P., Häberle J., Hayflick S. J., Hempel M., Itkis Y. S., Kishita Y., Klopstock T., Krylova T. D., & Prokisch H. (2022). Clinical implementation of RNA sequencing for Mendelian disease diagnostics. *Genome Medicine*, *14*(1), 38. https://doi.org/10.1186/s13073-022-01019-9

Yépez V. A., Mertes C., Müller M. F., Klaproth-Andrade D., Wachutka L., Frésard L., Gusic M., Scheller I. F., Goldberg P. F., Prokisch H., Prokisch H., & Gagneur J. (2021). Detection of aberrant gene expression events in RNA sequencing data. *Nature Protocols*, *16*(2), 1276–1296. https://doi.org/10.1038/s41596-020-00462-5