CAUSAL-STRUCTURE-PRESERVING TRANSFORMATION FOR ROOT-CAUSE DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Root-cause discovery (RCD) for a single interventional sample is important in many applications, including rare-disease diagnosis and cloud-service monitoring. Existing methods for one-sample RCD typically assume a linear structural equation model (SEM). In practice, however, observed common causes may affect variables through nonlinear or interaction effects, such as those involving age, gender, and experimental batch in RNA-sequencing studies, or incoming request load in cloud services. Ignoring these common causes can result in unaccounted-for confounding, whereas directly incorporating their nonlinear or interaction effects violates the linearity assumption underlying existing methods. These complications can obscure the sparse intervention signal and cause existing RCD methods to misidentify the root cause. To address this problem, we propose CSP-RCD, a two-stage procedure that first transforms the variables using flexible regression methods and then applies one-sample RCD algorithms to the transformed variables. We prove that this transformation preserves both the causal coefficient matrix and the sparse structural intervention of the original SEM, thereby providing a theoretical basis for the validity of downstream RCD methods. We provide a complete characterization of valid additive transformations and establish consistency of root-cause recovery for CSP-RCD. The proposed transformation can also be applied to other causal discovery tasks. For example, it can be combined with LiNGAM to estimate the causal structure when the variables are influenced by nonlinear effects of observed common causes. Extensive experiments on synthetic causal systems and real-world applications to the PetShop microservice benchmark and an RNA-sequencing dataset show that our proposed approach improves root-cause discovery.

1 INTRODUCTION

When a complex system becomes abnormal, it is important to identify the cause of the abnormality. A natural approach is to look for the variable with the largest observed deviation. However, this variable need not be the root cause, which is the variable at which the abnormality originated, because an abnormal effect can propagate through the causal system and create larger downstream anomalies. This problem arises in a wide range of applications, including rare-disease diagnosis, industrial systems, and cloud services.

Root-cause discovery (RCD; see, e.g., Yu & Rashid (2013); Kim et al. (2013); Shan et al. (2019); Budhathoki et al. (2022); Ikram et al. (2022); Nagalapatti et al. (2025); Xu et al. (2025); Han et al. (2025)) seeks to pinpoint the root cause among many abnormal variables by taking their causal relationships into account. A practically important setting arises when root-cause discovery must be performed for a single abnormal observation, such as a patient with a rare disease or a sudden system incident. This motivates the one-interventional-sample setting, where many normal reference samples are available and the goal is to identify the root cause of a single abnormal interventional sample. Existing permutation–Cholesky and precision-matrix-based methods assume a linear SEM and exploit this linear structure to identify the root cause without requiring prior knowledge or estimation of the underlying causal DAG (Li et al., 2026; Schkoda & Janzing, 2026).

In many applications, however, the variables may also be affected by observed common causes with nonlinear or interaction effects. Examples include age and experimental batch in gene-expression studies, ambient conditions in physical systems, and shared infrastructure or incoming request load

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

in cloud services. In such settings, it is more reasonable to consider models of the form

$$X \leftarrow b + BX + g(W) + \epsilon, \qquad X^I \leftarrow b + BX^I + g(W^I) + \epsilon^I + \delta, \tag{1}$$

where W denotes the observed common causes, none of which is a descendant of any variable in X, and g allows for nonlinear and interaction effects. See Section 3 for more details. Under this more general model, the linear SEM assumption no longer holds, and the validity of existing one-sample RCD methods is therefore no longer guaranteed. As a result, existing one-sample RCD methods may misidentify the root cause. In particular, omitting W leaves the contribution $g(W^I)$ unaccounted for, which can induce a dense shift that obscures the sparse intervention δ. Our goal is to identify the root causes while accounting for these nonlinear common-cause effects.

Our idea is to transform X to remove the contribution of the observed common causes. Importantly, we want to preserve both the structural coefficient matrix B and the sparse intervention δ, because the transformed variables then follow a standard linear SEM and existing one-sample RCD methods can be applied directly. In other words, we seek a causal-structure-preserving transformation.

A transformation that simply subtracts the common-cause contribution $g(W)$ from X does not preserve the causal structure. Subtracting $(I - B)^{-1}g(W)$ would preserve it, but requires estimating both the unknown structural coefficient matrix B, which may not be identifiable, and the possibly complex function g. Our key observation is that the transformation $X \mapsto X - m(W)$ is causal-structure-preserving, where $m(w) = \mathbb{E}[X \mid W = w] = (I - B)^{-1}\{b + g(w)\}$ is the conditional mean and can be estimated directly from observational data. Thus, although regressing variables on observed covariates and working with the resulting residuals is widely used in many areas (see Section 2), we show that in our setting this operation preserves the underlying causal structure and intervention. We further characterize all additive transformations with this property, showing that they are exactly the conditional-mean transformation up to an additive constant.

In practice, the conditional mean $m(w)$ must be estimated from data. We show that the conditional-mean estimation error upper bounds the resulting error in the transformed empirical covariance. For the permutation–Cholesky method, we further establish consistency of root-cause discovery under suitable estimation conditions. Experiments on synthetic data, the PetShop microservice benchmark, and an RNA-sequencing dataset show that the proposed approach improves root-cause discovery across a range of settings.

Beyond root-cause discovery, the proposed transformation can also be applied to other causal discovery and inference tasks. By preserving the structural coefficient matrix B, it broadens the applicability of methods developed for linear SEMs to settings with nonlinear effects of observed common causes (see equation 1), including methods for causal structure learning, ancestor identification (Schultheiss & Bühlmann, 2023), and causal effect estimation. For example, in the presence of such nonlinear common-cause effects, the transformation can be combined with LiNGAM for causal structure learning. Simulations in Appendix C show that this substantially improves estimation of both B and the underlying DAG.

2 Related work

Root-cause discovery and root-cause analysis. One line of work uses a known causal graph or first estimates the causal structure before identifying the root cause (Kim et al., 2013; Ikram et al., 2022; Assaad et al., 2023; Tao et al., 2024; Tonon et al., 2025). Other methods avoid requiring a known graph and instead identify intervention targets by comparing observational and interventional distributions, using changes in covariance or precision matrices, likelihoods, or more general distributional information (Rothenhäusler et al., 2015; Varici et al., 2021; 2022; Squires et al., 2020; Jaber et al., 2020; Yang et al., 2024). These approaches generally require multiple interventional or anomalous samples in order to estimate properties of the interventional distribution. Related work also studies causal explanations or attribution of outliers and distribution changes (Budhathoki et al., 2021; 2022). More recently, root-cause discovery from a single interventional observation has been considered by Li et al. (2026) and Schkoda & Janzing (2026). Our work focuses on this one-interventional-sample setting, but addresses the additional challenge that observed common causes may affect the variables through nonlinear or interaction effects.

Potential outcomes and covariate adjustment. In the potential-outcomes framework, observed pre-treatment covariates W are commonly used to adjust for confounding when estimating the causal

effect of a predefined treatment A on an outcome Y. Under standard identification conditions, this adjustment can be based on nuisance functions such as the outcome regression $\mu_a(w) = \mathbb{E}[Y \mid A = a, W = w]$ and the propensity score $e(w) = \Pr(A = 1 \mid W = w)$ (Rosenbaum & Rubin, 1983; Imbens, 2004; Chernozhukov et al., 2018). Modern methods allow these nuisance functions to be estimated flexibly, for example through double/debiased machine learning (Chernozhukov et al., 2018; Díaz, 2020; Chernozhukov et al., 2022). Our problem differs in both framework and goal. We consider a structural equation model for the multivariate X, with W representing observed common causes with nonlinear effects. In the potential-outcomes framework, outcome-regression methods use conditional means such as $\mu_a(w) = \mathbb{E}[Y \mid A = a, W = w]$ to adjust for covariates when estimating the effect of a predefined treatment A. Our problem has no predefined treatment variable, since the intervention target is itself unknown. We instead use $m(w) = \mathbb{E}[X \mid W = w]$ to transform X while preserving its causal structure.

Observed-covariate adjustment and residualization. Removing variation associated with observed covariates before downstream analysis is common in many areas. For example, covariate-adjusted graphical-model methods regress a multivariate response on external covariates and study the dependence structure of the residual variation (Li et al., 2012; Cai et al., 2013; Yin & Li, 2013). Related residualization procedures are widely used to remove batch effects in genomics (Ritchie et al., 2015) and measured confounds in neuroimaging (Snoek et al., 2019). Although our transformation has a similar form, its motivation and desired properties are fundamentally different: we seek a transformation that preserves the causal coefficient matrix B and the sparse intervention δ.

3 Problem formulation

We consider one-interventional-sample root-cause discovery with observed common causes. The observational and interventional samples are assumed to follow the SEMs:

$$X \leftarrow b + BX + g(W) + \epsilon, \tag{2}$$

$$X^I \leftarrow b + BX^I + g(W^I) + \epsilon^I + \delta. \tag{3}$$

Here, $\leftarrow$ denotes a structural assignment in the generative model rather than merely an equality. The vector $X \in \mathbb{R}^p$ denotes an observational random vector, $X^I \in \mathbb{R}^p$ denotes a single interventional observation, while $W \in \mathbb{R}^q$ and $W^I \in \mathbb{R}^q$ denote the observed common causes in the observational and interventional models, respectively. The vector $b \in \mathbb{R}^p$ is an intercept, $B \in \mathbb{R}^{p\times p}$ is the structural coefficient matrix encoding a directed acyclic graph (DAG), and $g : \mathbb{R}^q \to \mathbb{R}^p$ allows for nonlinear and interaction effects of the observed common causes. The noise ϵ has diagonal covariance D_ϵ, while ϵ^I is independent of the observational samples and satisfies $\epsilon^I \overset{d}{=} \epsilon$, with $\mathbb{E}[\epsilon \mid W] = \mathbb{E}[\epsilon^I \mid W^I] = 0$. We assume that the observed common causes W and W^I are not descendants of variables in X and X^I, respectively. Finally, the intervention δ is one-sparse, with $\delta_r \neq 0$ and $\delta_j = 0$ for $j \neq r$, where r indexes the root cause.

Given n i.i.d. observational samples $(X_1, W_1), \ldots, (X_n, W_n)$ and a single interventional observation (X^I, W^I), the goal is to recover the root cause r. The existing one-sample RCD methods of Li et al. (2026) and Schkoda & Janzing (2026) correspond to the special case $g = 0$. Even in this simpler setting, although the intervention is one-sparse in δ, its effect on the observed variables propagates through the causal system as $(I - B)^{-1}\delta$. Consequently, the variable with the largest marginal deviation in X^I need not be the root cause. These two existing methods exploit observational second-moment information to account for this propagation without requiring prior knowledge or estimation of B. When $g \neq 0$, however, the additional common-cause contribution can obscure the sparse intervention signal, and the validity of these methods is no longer guaranteed.

4 Causal-structure-preserving transformation

4.1 Causal-structure-preserving transformation and complete characterization

To discover the root cause r in the presence of nonlinear effects of observed common causes (see equations 2 and 3), our idea is to construct a transformation of X and X^I that removes the observed common-cause contribution while preserving both the structural coefficient matrix B and the sparse

intervention δ. The transformed variables then follow standard linear SEMs, allowing existing one-sample RCD methods to be applied directly. We now show how such a transformation can be obtained.

Rewriting equation 2 gives

$$X = (I - B)^{-1}b + (I - B)^{-1}g(W) + (I - B)^{-1}\epsilon. \tag{4}$$

Thus, if we transform X by subtracting the propagated common-cause contribution and intercept,

$$X^* = X - (I - B)^{-1}\{b + g(W)\},$$

then

$$X^* = BX^* + \epsilon,$$

which is again a standard linear SEM with the same structural coefficient matrix B.

At first glance, this transformation appears to require estimating the unknown matrix B and the possibly complex function g, as well as the intercept b. The key observation is that

$$m(w) := \mathbb{E}[X \mid W = w] = (I - B)^{-1}\{b + g(w)\}. \tag{5}$$

Therefore, there is no need to identify b, B, or g separately. It suffices to estimate the conditional mean m, which can be learned directly from the observational data.

Under an overlap condition ensuring that the interventional common-cause values are covered by the observational distribution (see Assumption 1 in Appendix A.1 for a rigorous statement), and assuming that W is not a descendant of any variable in X, the following theorem formalizes that the transformation preserves the causal structure. In particular, the transformed variables satisfy structural equations with the same B and δ, so the $\leftarrow$ relation is preserved rather than only the corresponding equality.

Theorem 1 (Causal-structure-preserving transformation)**.** *Let X, X^I follow the SEMs in equation 2 and equation 3. Assume that Assumption 1 holds and that the observed common causes W are not descendants of any variables in X. Define*

$$X^* = X - m(W), \qquad X^{I*} = X^I - m(W^I). \tag{6}$$

Then,

$$X^* \leftarrow BX^* + \epsilon, \quad X^{I*} \leftarrow BX^{I*} + \epsilon^I + \delta. \tag{7}$$

almost surely under the observational and interventional distributions, respectively.

The proof of Theorem 1 is given in Appendix A.2.

Then, a natural question is whether there are other causal-structure-preserving transformations. The following theorem gives a complete characterization and shows that, among additive transformations, the conditional-mean transformation is unique up to an additive constant.

Theorem 2 (Complete characterization of additive causal-structure-preserving transformations)**.** *Under the same assumptions as Theorem 1, consider measurable additive transformations*

$$T_h = X - h(W), \qquad T_h^I = X^I - h(W^I).$$

There exists a deterministic intercept $b_h \in \mathbb{R}^p$ such that

$$T_h \leftarrow b_h + BT_h + \epsilon, \quad T_h^I \leftarrow b_h + BT_h^I + \epsilon^I + \delta \tag{8}$$

almost surely if and only if

$$h(w) = m(w) + a$$

for some constant vector $a \in \mathbb{R}^p$. In this case, $b_h = -(I - B)a$.

The proof of Theorem 2 is given in Appendix A.3.

Different residualization procedures may use different intercept conventions, for example by subtracting the full conditional mean or by adding back a constant such as the marginal mean (Chernozhukov et al., 2018; Collinson & Ganong, 2018). Theorem 2 shows that all such choices are equivalent for our purpose: they preserve the same structural coefficient matrix B and sparse intervention δ. Thus, the treatment of the intercept does not affect the downstream root-cause discovery problem.

Theorem 2 also shows why omitting W or directly subtracting $g(W)$ cannot preserve the causal structure. Omitting W corresponds to $h = 0$, which belongs to the valid class $h = m + a$ only when $m(W)$ is constant. Directly subtracting $g(W)$ corresponds to $h = g$, which belongs to the valid class only in the special case that $Bg(W)$ is constant.

4.2 CSP-RCD: ERROR BOUNDS AND CONSISTENCY

Building on the causal-structure-preserving transformation above, we propose a two-step procedure for root-cause discovery with nonlinear common-cause effects (see equation 1), which we call causal-structure-preserving root-cause discovery (CSP-RCD):

1. Using the observational data to fit a feature-wise estimator $\widehat{m}$ of $m(w) = \mathbb{E}[X \mid W = w]$, and construct transformed samples $\widehat{X}_i^* = X_i - \widehat{m}(W_i)$ and $\widehat{X}^{I*} = X^I - \widehat{m}(W^I)$, $i = 1, \ldots, n$.
2. Apply a one-sample RCD method designed for the setting without observed common-cause effects to $(\widehat{X}_{1:n}^*, \widehat{X}^{I*})$.

When $\widehat{m}$ is estimated from the observational data, the transformed samples $(\widehat{X}_i^*, \widehat{X}^{I*})$ used by CSP-RCD differ from their population counterparts (X_i^*, X^{I*}). We therefore study the finite-sample behavior of CSP-RCD by first quantifying how the estimation error in $\widehat{m}$ propagates to the empirical mean and covariance of the transformed observational data. Since the covariance matrix is a key input to one-sample RCD methods such as Li et al. (2026) and Schkoda & Janzing (2026), this connects the accuracy of the estimated transformation to the downstream RCD procedure. We then focus on the method of Li et al. (2026) and establish an end-to-end root-cause recovery result for CSP-RCD.

For the finite-sample analysis, define the conditional-mean estimation error on the observational samples by

$$e_i = m(W_i) - \widehat{m}(W_i), \qquad r_n^2 = \frac{1}{n}\sum_{i=1}^{n} \|e_i\|_2^2,$$

so that $\widehat{X}_i^* = X_i^* + e_i$. Thus, r_n measures the estimation error introduced into the transformed observational samples.

Let $(\widetilde{\mu}, \widetilde{\Sigma})$ and $(\widehat{\mu}, \widehat{\Sigma})$ denote the empirical mean and covariance, with divisor n, of the oracle transformed samples $X_1^*, \ldots, X_n^*$ and the estimated transformed samples $\widehat{X}_1^*, \ldots, \widehat{X}_n^*$, respectively. Thus, $\widetilde{\Sigma}$ is the covariance available if $m(w)$ were known, while $\widehat{\Sigma}$ is the covariance used by CSP-RCD. Let $\Sigma = \mathrm{Cov}(X^*)$, and let $\|\cdot\|_2$ and $\|\cdot\|_{\mathrm{op}}$ denote the Euclidean and operator norms, respectively.

The following theorem quantifies how the estimation error in $\widehat{m}(w)$ affects the empirical moments of the transformed data.

Theorem 3 (Empirical moment error under the estimated transformation)**.** *For every realized sample,*

$$\|\widehat{\mu} - \widetilde{\mu}\|_2 \le r_n, \quad \|\widehat{\Sigma} - \widetilde{\Sigma}\|_{\mathrm{op}} \le 2\sqrt{\|\widetilde{\Sigma}\|_{\mathrm{op}}}\, r_n + r_n^2. \tag{9}$$

Consequently, if p is fixed and $r_n = o_p(1)$, then

$$\|\widehat{\mu}\|_2 = o_p(1), \qquad \|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} = o_p(1).$$

The proof of Theorem 3 is given in Appendix A.4. Theorem 3 shows that the errors in $\widehat{\mu}$ and $\widehat{\Sigma}$ induced by estimating $m(w)$ are controlled by the prediction error r_n. In particular, equation 9 bounds the additional covariance error due to the estimated transformation. For fixed p, $\widetilde{\Sigma}$ converges to Σ, so $r_n = o_p(1)$ implies that $\widehat{\Sigma}$ is also consistent for Σ.

To obtain end-to-end consistency of root-cause recovery, we also need to account for the transformation error of the interventional observation. Define

$$e_I = m(W^I) - \widehat{m}(W^I), \qquad r_I = \|e_I\|_2,$$

so that $\widehat{X}^{I*} = X^{I*} + e_I$. We now specialize to the permutation–Cholesky RCD method of Li et al. (2026). The following corollary establishes consistency of CSP-RCD with this downstream method.

Corollary 1 (Consistency of CSP-RCD with Cholesky RCD)**.** *Assume the setting of Theorem 1, with fixed p and $\Sigma \succ 0$. Suppose $r_n = o_p(1)$ and $r_I = o_p(1)$. Apply Algorithm 2 of Li et al. (2026) to $(\widehat{X}_{1:n}^*, \widehat{X}^{I*})$, and let $\widehat{C}_j$ denote the root-cause score for variable j. Let r be the root cause, then*

$$\lim_{\substack{n\to\infty \\ \delta_r\to\infty}} \mathbb{P}\left(\widehat{C}_r > \max_{k\in[p]\setminus\{r\}} \widehat{C}_k\right) = 1. \tag{10}$$

The conditions $r_n = o_p(1)$ and $r_I = o_p(1)$ require the fitted conditional mean to predict accurately on both the observational samples and the interventional common-cause value. These conditions can follow from standard prediction consistency together with suitable generalization or stability properties of the nuisance learner.

A similar consistency result may also be derived for the method of Schkoda & Janzing (2026). Since their analysis does not establish an asymptotic consistency result, we do not pursue an end-to-end extension here. Nevertheless, the population-level validity of applying their method to the transformed variables follows directly from Theorem 1.

4.3 Application to other causal discovery tasks

The causal-structure-preserving transformation is not specific to root-cause discovery. By Theorem 1, the transformed variable $X^* = X - m(W)$ follows a linear SEM with the same structural coefficient matrix B. Therefore, causal discovery methods developed for linear SEMs can also be applied to the transformed data in settings with nonlinear effects of observed common causes, provided that their other assumptions hold. In this sense, the proposed transformation extends the applicability of existing linear-SEM methods, including those for causal structure learning, ancestral relation recovery, and total causal effect estimation, to settings where the variables are also affected by nonlinear effects of observed common causes.

As a concrete example, consider causal structure learning with LiNGAM. The Python `lingam` package (Ikeuchi et al., 2023) already provides a utility for removing linear effects of selected variables. Our results show that nonlinear common-cause effects can likewise be addressed through a causal-structure-preserving transformation. After transformation, LiNGAM can be applied to X^* to estimate both B and the corresponding DAG. Simulation results in Appendix C show that this transformation substantially improves recovery of both B and the DAG compared with applying LiNGAM directly to the original data.

5 Experiments

Our experiments evaluate whether CSP-RCD improves root cause discovery when observed common causes violate the linear SEM assumed by the selected one-sample RCD methods. Specifically, each RCD task consists of n observational samples and one interventional sample. For each downstream RCD method, we compare the results obtained using raw inputs (X, X^I) with those obtained using the CSP-RCD transformed inputs $(\widehat{X}^*, \widehat{X}^{I*})$.

We use ordinary least squares (OLS) and generalized additive models (GAMs) (Hastie & Tibshirani, 1986) as the nuisance learners in the main synthetic experiments. For the real dataset, we use GAMs and random forests (Breiman, 2001) for PetShopt, and OLS for the RNA-sequencing dataset. For downstream RCD methods, We evaluate the Cholesky RCD method of Li et al. (2026), and the Cyclic RCA method of Schkoda & Janzing (2026). We also evaluate DirectLiNGAM (Shimizu et al., 2011) for estimating B from the transformed variables in low dimensional simulations. The results in Appendix C show that the proposed transformation can also improve causal structure learning.

5.1 Synthetic SEMs

We study linear and nonlinear common-cause effects with $q = 10$ and $p = 100$. The baseline setting is $(n, \delta_r, s_B) = (200, 8, 0.4)$. We vary one factor at a time over $n \in \{100, 200, 1000\}$, $\delta_r \in \{4, 8, 12\}$, or $s_B \in \{0.2, 0.4, 0.6\}$, while keeping the other two factors at their baseline values.

For each condition, we generate 20 independent structural replications. Within each replication, we independently generate a random DAG with exactly $\lfloor s_B p(p-1)/2 \rfloor$ edges, a diagonal noise covariance D_ϵ, an intercept, and a common-cause effect function g. To control variance growth through the DAG, we independently draw marginal-variance targets from $\mathrm{Unif}(10, 30)$. We then calibrate the rows of B for non-root nodes and the noise variances for root nodes so that $\mathrm{diag}\{(I-B)^{-1}D_\epsilon(I-B)^{-\top}\}$ matches these targets. After fixing B, D_ϵ, and the intercept, we multiply g by a single global constant so that its propagated contribution accounts for 50% of the total marginal variance $\mathrm{tr}(\mathrm{Cov}(X))$.

In the linear setting, $g(W) = \Gamma W$, and the nuisance functions are estimated by OLS. In the nonlinear setting, g is an additive combination of linear, centered quadratic, sine, and centered sigmoid terms. The nuisance functions are estimated using GAMs, with the smoothing parameter selected by generalized cross-validation.

Within each replication, we independently draw 50 intervention targets uniformly from $\{1, \ldots, p\}$ with replacement. For each condition, we pool the resulting $20 \times 50 = 1000$ one-sample RCD tasks to compute the rank distributions and root cause recovery metrics. Appendix B.1 provides the full simulation design and a linear $q = 30$ extension using cross-validated Lasso (Tibshirani, 1996).

Table 1: Top K recall across different observational sample sizes under linear and nonlinear common cause effects with $q = 10$, $\delta_r = 8$ and $s_B = 0.4$. Complete results for all inputs and for variations in δ_r and s_B, as well as different inputs, can be found in Tables 4, 5, and 6 in Appendix B.

Effect	n	Downstream method	Input	Top 1 recall	Top 3 recall	Top 5 recall
Linear	100	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	**0.105**	**0.179**	**0.234**
		Cyclic RCA	(X, X^I)	0.053	0.094	0.139
		Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	**0.522**	**0.686**	**0.748**
		Cholesky RCD	(X, X^I)	0.377	0.516	0.586
	200	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	**0.413**	**0.553**	**0.610**
		Cyclic RCA	(X, X^I)	0.412	0.545	0.596
		Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	**0.544**	**0.681**	**0.747**
		Cholesky RCD	(X, X^I)	0.492	0.646	0.695
	1000	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	**0.616**	**0.751**	**0.789**
		Cyclic RCA	(X, X^I)	0.576	0.716	0.771
		Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	**0.692**	**0.827**	**0.866**
		Cholesky RCD	(X, X^I)	0.593	0.727	0.787
Non-linear	100	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	**0.084**	**0.158**	**0.200**
		Cyclic RCA	(X, X^I)	0.044	0.090	0.134
		Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	**0.426**	**0.592**	**0.662**
		Cholesky RCD	(X, X^I)	0.343	0.498	0.581
	200	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	**0.360**	**0.509**	0.566
		Cyclic RCA	(X, X^I)	0.345	0.499	**0.569**
		Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	**0.478**	**0.637**	**0.706**
		Cholesky RCD	(X, X^I)	0.455	0.617	0.684
	1000	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	**0.605**	**0.727**	**0.778**
		Cyclic RCA	(X, X^I)	0.559	0.695	0.738
		Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	**0.676**	**0.803**	**0.848**
		Cholesky RCD	(X, X^I)	0.578	0.703	0.767

Table 1 and Figure 1 show that the transformed input generally gives better RCD results than the raw (X, X^I) input and the full observed $[(X, W), (X^I, W^I)]$ input under linear and nonlinear common-cause effects. Figure 5 in Appendix B further compares the fitted transformation with the oracle transformation across sample sizes, demonstrating that the fitted transformation approaches the oracle transformation as n grows. At $n = 1000$, its Top-1 recall is very close to the oracle in every simulation setting and for every downstream method. This suggests that more observational samples improve nuisance estimation and allow the fitted transformation to recover most of the oracle information.

5.2 MICROSERVICE INCIDENTS

The PetShop benchmark contains normal and abnormal measurements from a distributed microservice application, including 68 injected performance incidents with known service-level root causes (Hardt et al., 2024). We retain incidents whose target metric is latency and use the `Average` statistic of the latency metric for each service, following the benchmark suggestion. Pre-processing includes deleting features and samples containing missingness, and removing nearly constant features. We selected a set of four variables as candidate observed common causes. All four are root nodes in the service map provided with the Petshop dataset. In addition, from a distributed-systems perspective, they are either remote search endpoint serving multiple service paths, microservice handling authentication

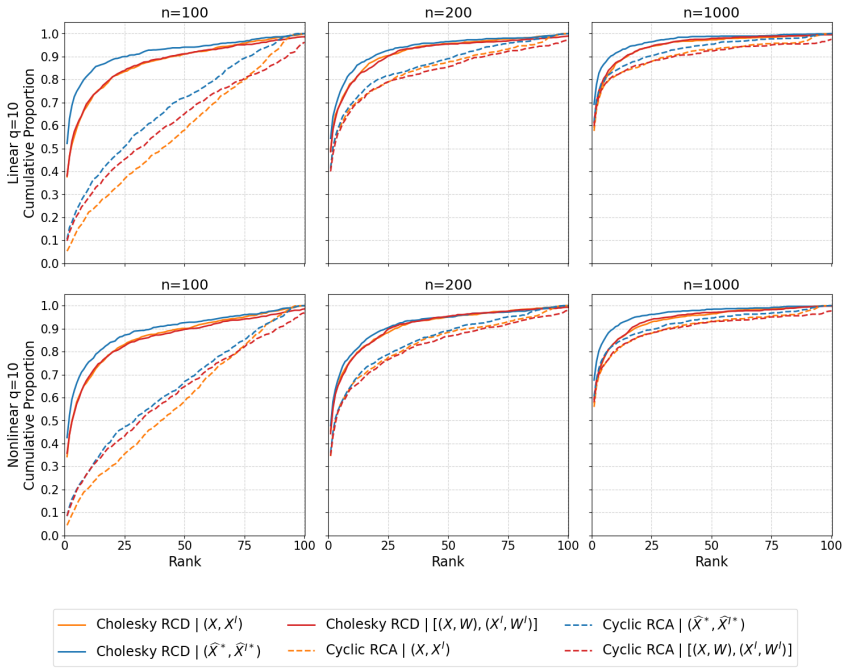


Figure 1: Root-cause recovery across observational sample sizes for linear and nonlinear common cause effect with $q = 10$. The complete tabular results of variation of δ_r and s_B, and different inputs can be found in Tables 10–12 and Tables 10–12 in Appendix B.

and providing configuration, or supporting feature evaluation, making them plausible common causes. For each timestamp, only candidate variables remaining after preprocessing are included in W.

Each eligible incident contains five interventional timestamps. We evaluate each timestamp separately as a one-sample RCD task. We also perform an incident-level evaluation by applying a softmax normalization to the service scores within each timestamp and then taking the maximum normalized score across the five timestamps for each service. The detailed list of candidate common-cause variables and the full preprocessing pipeline can be found in Appendix D.

Table 2: Incident-level root-cause ranks from Cyclic RCA under high and low traffic (lower is better). "Test0" and "Train0" are incident identifiers, not indicating train–test splits.

(a) High traffic

Nuisance model	Input	Test0	Test1	Test10	Test11	Test12	Test13	Test14	Test15	Test16	Test17
–	(X, X^I)	8	9	8	10	9	12	4	4	**3**	3
random forest	$(\widehat{X}^*, \widehat{X}^{I*})$	**5**	**4**	**1**	**8**	**6**	**7**	4	4	4	3
GAM	$(\widehat{X}^*, \widehat{X}^{I*})$	9	10	9	14	14	19	4	4	4	3
–	$[(X, W), (X^I, W^I)]$	10	11	10	13	14	11	4	4	4	3

(b) Low traffic

Nuisance model	Input	Test0	Test1	Test10	Test11	Test12	Test13	Test14	Test15	Test16	Test17	Train0	Train1	Train2	Train3
–	(X, X^I)	**9**	5	3	2	2	1	**10**	21	4	4	1	2	1	1
random forest	$(\widehat{X}^*, \widehat{X}^{I*})$	10	**2**	**1**	2	2	1	11	**6**	4	4	1	2	1	1
GAM	$(\widehat{X}^*, \widehat{X}^{I*})$	13	4	4	2	2	1	23	9	**3**	4	2	2	1	1
–	$[(X, W), (X^I, W^I)]$	10	5	4	2	2	1	13	23	4	4	1	2	1	1

Table 2 reports the incident-level root-cause ranks, while Table 3 summarizes timestamp-level performance across timestamps using Top-K recovery and rank statistics. Across both traffic settings, the random forest transformed input $(\widehat{X}^*, \widehat{X}^{I*})$ provides the largest improvements over the raw input (X, X^I) and the full observed input $[(X, W), (X^I, W^I)]$. The gains are stronger under low traffic. Random forest also generally outperforms GAM, consistent with its lower cross-validated prediction errors in Table 18 of Appendix D. It is also consistent with the theoretical link between nuisance estimation error and the empirical covariance convergence discussed in Section 4. Although the

Table 3: Timestamp-level root-cause recovery using Cyclic RCA on PetShop. #Top-K counts timestamps with the true root cause ranked among the top K (higher is better). The high- and low-traffic settings contain 50 and 70 timestamps, respectively.

Scenario	Nuisance model	Input	#Top 1	#Top 3	#Top 5	Mean rank ↓	Median rank ↓
High traffic	–	(X, X^I)	1	15	24	10.26	6.0
	random forest	$(\widehat{X}^*, \widehat{X}^{I*})$	**3**	15	**25**	**9.64**	**5.5**
	GAM	$(\widehat{X}^*, \widehat{X}^{I*})$	1	**16**	20	10.06	9.5
	–	$[(X, W), (X^I, W^I)]$	1	14	20	11.02	7.5
Low traffic	–	(X, X^I)	15	28	40	8.34	4.5
	random forest	$(\widehat{X}^*, \widehat{X}^{I*})$	**17**	**37**	**42**	**7.73**	**3.0**
	GAM	$(\widehat{X}^*, \widehat{X}^{I*})$	16	31	40	8.74	4.0
	–	$[(X, W), (X^I, W^I)]$	13	29	41	8.61	4.5

improvements are not uniform across all incidents, several improvements are substantial, whereas most non-improving cases are unchanged and the few deteriorations involve only one rank differences. Complete results for the temporal-traffic setting and Cholesky RCD are provided in Appendix D.

5.3 RNA-SEQUENCING DATASET

We also evaluate CSP-RCD on the RNA-sequencing dataset of Yépez et al. (2022), previously analyzed by Li et al. (2026) and Schkoda & Janzing (2026). After preprocessing, the dataset contains 59 target patients with Mendelian diseases and known genetic mutations and 364 reference samples used as observational data. For each target patient, the known mutated gene is treated as the ground-truth root cause. Among the 423 samples, 154 were generated using a non-strand-specific RNA-sequencing protocol and 269 using a strand-specific protocol. Because the sequencing protocol can induce systematic differences in gene expression, we treat it as a binary observed common cause. The final analysis therefore has $n = 364$, $p = 19736$, and $q = 1$. We use Cyclic RCA as the downstream method because it has better baseline performance than Cholesky RCD and is computationally faster, and use OLS as the nuisance learner. Further details are in Appendix E.

Table 19 in Appendix E.2 reports patient-level root-cause ranks from Cyclic RCA using the raw input (X, X^I) and transformed input $(\widehat{X}^*, \widehat{X}^{I*})$. Cyclic RCA already performs well, ranking the root cause first for many patients under both inputs. Nevertheless, the transformation improves several rankings while leaving the others unchanged. Specifically, the ranks of R80346 and R18626 improve from 46 to 36 and from 39 to 34, respectively, while the more difficult case R46723 improves from 6393 to 4997. These results suggest that, even when adjusting for a single binary common cause, the transformation can typically improve difficult patient-level rankings without degrading the others.

6 LIMITATIONS AND CONCLUSION

Limitations. Our method assumes that the relevant common causes are observed, are not descendants of the variables in X, and enter the SEM additively through $g(W)$. It therefore does not directly address unobserved common causes, post-intervention covariates, interactions between W and X, or nonlinear causal relations among the variables in X. Developing structure-preserving transformations for these settings is an important direction for future research.

Conclusion. We introduced a causal-structure-preserving transformation for linear SEMs with nonlinear effects of observed common causes. This yields CSP-RCD, which combines flexible regression methods with existing one-sample RCD algorithms without requiring estimation of the causal graph. We further characterize all valid additive transformations, quantify the effect of estimating m, and establish consistency of root-cause discovery. More broadly, our results show that nonlinear effects of observed common causes need not preclude the use of linear-SEM causal methods, as the proposed transformation preserves the underlying causal structure. Our results therefore extend the applicability of linear-SEM causal methods, including those for root-cause discovery and causal structure learning, to settings with nonlinear observed common-cause effects.

AI USE STATEMENT

In this work, we used generative AI tools for the following purposes: (1) polishing and improving the clarity of the writing; (2) identifying potentially relevant references for the related work section after the initial draft had been completed, in order to check whether important literature had been overlooked. Most citations in the paper were identified independently of AI tools; (3) checking the correctness and consistency of mathematical claims and derivations; and (4) improving the efficiency and readability of code. The main functions, algorithmic design, and code structure were developed by the authors rather than generated by AI. All AI-assisted content was reviewed and verified by the authors. AI-assisted code modifications were validated by confirming that they produced numerical results consistent with the original human-written implementations. References suggested by AI tools were read by the authors, and their relevance and support for the corresponding claims were independently verified. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Charles K. Assaad, Imad Ez-Zejjari, and Lei Zan. Root cause identification for collective anomalies in time series given an acyclic summary causal graph with loops. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 8395–8404. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/assaad23a.html.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

Kailash Budhathoki, Dominik Janzing, Patrick Bloebaum, and Hoiyi Ng. Why did the distribution change? In *International Conference on Artificial Intelligence and Statistics*, pp. 1666–1674. PMLR, 2021.

Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In *International conference on machine learning*, pp. 2357–2369. PMLR, 2022.

T Tony Cai, Hongzhe Li, Weidong Liu, and Jichun Xie. Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, 100(1):139–156, 2013.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 02 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022. doi: https://doi.org/10.3982/ECTA16294. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16294.

Robert Collinson and Peter Ganong. How do changes in housing voucher design affect rent and neighborhood quality? *American Economic Journal: Economic Policy*, 10(2):62–89, May 2018. doi: 10.1257/pol.20150176. URL https://www.aeaweb.org/articles?id=10.1257/pol.20150176.

Iván Díaz. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 21(2):353–358, 04 2020. ISSN 1465-4644. doi: 10.1093/biostatistics/kxz042. URL https://doi.org/10.1093/biostatistics/kxz042.

Xiao Han, Saima Absar, Lu Zhang, and Shuhan Yuan. Root cause analysis of anomalies in multivariate time series through granger causal discovery. In *International Conference on Learning Representations*, volume 2025, pp. 45137–45153, 2025.

Michaela Hardt, William R. Orchard, Patrick Blöbaum, Shiva Kasiviswanathan, and Elke Kirschbaum. The petshop dataset – finding causes of performance issues across microservices, 2024. URL https://arxiv.org/abs/2311.04806.

Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297 – 310, 1986. doi: 10.1214/ss/1177013604. URL https://doi.org/10.1214/ss/1177013604.

Takashi Ikeuchi, Mayumi Ide, Yan Zeng, Takashi Nicholas Maeda, and Shohei Shimizu. Python package for causal discovery based on lingam. *Journal of Machine Learning Research*, 24(14): 1–8, 2023.

Azam Ikram, Sarthak Chakraborty, Subrata Mitra, Shiv Saini, Saurabh Bagchi, and Murat Kocaoglu. Root cause analysis of failures in microservices through causal discovery. *Advances in Neural Information Processing Systems*, 35:31158–31170, 2022.

Guido W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86(1):4–29, 02 2004. ISSN 0034-6535. doi: 10.1162/003465304323023651. URL https://doi.org/10.1162/003465304323023651.

Amin Jaber, Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, 33:9551–9561, 2020.

Myunghwan Kim, Roshan Sumbaly, and Sam 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, pp. 93–104, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450319003. doi: 10.1145/2465529.2465753. URL https://doi.org/10.1145/2465529.2465753.

Bing Li, Hyonho Chun, and Hongyu Zhao. Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association*, 107(497):152–167, 2012.

Jinzhou Li, Benjamin B Chu, Ines F Scheller, Julien Gagneur, and Marloes H Maathuis. Root cause discovery via permutations and cholesky decomposition. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(3):779–798, 07 2026. ISSN 1369-7412. doi: 10.1093/jrsssb/qkaf066. URL https://doi.org/10.1093/jrsssb/qkaf066.

Lokesh Nagalapatti, Ashutosh Srivastava, Sunita Sarawagi, and Amit Sharma. Robust root cause diagnosis using in-distribution interventions. In *International Conference on Learning Representations*, volume 2025, pp. 25738–25765, 2025.

Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 04 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv007. URL https://doi.org/10.1093/nar/gkv007.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983. ISSN 0006-3444. doi: 10.1093/biomet/70.1.41. URL https://doi.org/10.1093/biomet/70.1.41.

Dominik Rothenhäusler, Christina Heinze, Jonas Peters, and Nicolai Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *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.

Daniela Schkoda and Dominik Janzing. Root cause analysis of outliers in unknown cyclic graphs, 2026. URL https://arxiv.org/abs/2510.06995.

Christoph Schultheiss and Peter Bühlmann. Ancestor regression in linear structural equation models. *Biometrika*, 110(4):1117–1124, 2023.

Huasong Shan, Yuan Chen, Haifeng Liu, Yunpeng Zhang, Xiao Xiao, Xiaofeng He, Min Li, and Wei Ding. ϵ - Diagnosis: Unsupervised and real-time diagnosis of small- window long-tail latency in large-scale microservice platforms. In *The World Wide Web Conference*, WWW '19, pp. 3215–3222, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313653. URL https://doi.org/10.1145/3308558.3313653.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. Directlingam: A direct method for learning a linear non-gaussian structural equation model. *J. Mach. Learn. Res.*, 12(null):1225–1248, July 2011. ISSN 1532-4435.

Lukas Snoek, Steven Miletić, and H. Steven Scholte. How to control for confounds in decoding analyses of neuroimaging data. *NeuroImage*, 184:741–760, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2018.09.074. URL https://www.sciencedirect.com/science/article/pii/S1053811918319463.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1039–1048. PMLR, 2020.

Hongyuan Tao, Hang Yu, and Jianguo 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 (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 89978–90008. Curran Associates, Inc., 2024. doi: 10.52202/079017-2857. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a3cf318fbeec1126da21e9185ae9908c-Paper-Conference.pdf.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 01 1996. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Andrea Tonon, Meng Zhang, Bora Caglayan, Fei Shen, Tong Gui, MingXue Wang, and Rong Zhou. Radice: Causal graph based root cause analysis for system performance diagnostic. In *2025 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pp. 465–475, 2025. doi: 10.1109/SANER64311.2025.00050.

Burak Varici, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Scalable intervention target estimation in linear models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1494–1505. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0b94ce08688c6389ce7b68c52ce3f8c7-Paper.pdf.

Burak Varici, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Intervention target estimation in the presence of latent variables. In *Uncertainty in Artificial Intelligence*, pp. 2013–2023. PMLR, 2022.

Junjielong Xu, Qinan Zhang, Zhiqing Zhong, Shilin He, Chaoyun Zhang, Qingwei Lin, Dan Pei, Pinjia He, Dongmei Zhang, and Qi Zhang. Openrca: Can large language models locate the root cause of software failures? In *International Conference on Learning Representations*, volume 2025, pp. 84562–84590, 2025.

Yuqin Yang, Saber Salehkaleybar, and Negar Kiyavash. Learning unknown intervention targets in structural causal models from heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 3187–3195. PMLR, 2024.

Vicente A. Yépez, Mirjana Gusic, Robert Kopajtich, Christian Mertes, Nicholas H. Smith, Charlotte L. Alston, Rui Ban, Skadi Beblo, Riccardo Berutti, Holger Blessing, Elżbieta Ciara, Felix Distelmaier, Peter Freisinger, Johannes Häberle, Susan J. Hayflick, Maja Hempel, Yulia S. Itkis, Yoshihito Kishita, Thomas Klopstock, Tatiana D. Krylova, Costanza Lamperti, Dominic Lenz, Christine Makowski, Signe Mosegaard, Michaela F. Müller, Gerard Muñoz-Pujol, Agnieszka Nadel, Akira Ohtake, Yasushi Okazaki, Elena Procopio, Thomas Schwarzmayr, Joél Smet, Christian Staufner,

Sarah L. Stenton, Tim M. Strom, Caterina Terrile, Frederic Tort, Rudy Van Coster, Arnaud Vanlander, Matias Wagner, Manting Xu, Fang Fang, Daniele Ghezzi, Johannes A. Mayr, Dorota Piekutowska-Abramczuk, Antonia Ribes, Agnès Rötig, Robert W. Taylor, Saskia B. Wortmann, Kei Murayama, Thomas Meitinger, Julien Gagneur, and Holger Prokisch. Clinical implementation of RNA sequencing for Mendelian disease diagnostics. *Genome Medicine*, 14(1):38, 2022. ISSN 1756-994X. doi: 10.1186/s13073-022-01019-9. URL https://doi.org/10.1186/s13073-022-01019-9.

Jianxin Yin and Hongzhe Li. Adjusting for high-dimensional covariates in sparse precision matrix estimation by ℓ_1-penalization. *Journal of multivariate analysis*, 116:365–381, 2013.

Jie Yu and Mudassir M. Rashid. A novel dynamic bayesian network-based networked process monitoring approach for fault detection, propagation identification, and root cause diagnosis. *AIChE Journal*, 59(7):2348–2365, 2013. doi: https://doi.org/10.1002/aic.14013. URL https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.14013.

A Proofs

A.1 Formal statement of overlap assumption

Assumption 1 (Covariate overlap)**.** *Let P_W and Q_W denote the distributions of W in the observational regime and of W^I in the interventional regime, respectively. Assume that Q_W is absolutely continuous with respect to P_W, denoted by $Q_W \ll P_W$. Since $m(w) = \mathbb{E}[X \mid W = w]$ is defined P_W-almost surely, we focus on m such that*

$$(I - B)m(w) = b + g(w)$$

for P_W-almost every w. By $Q_W \ll P_W$, the same identity then holds for Q_W-almost every w.

The notation

$$Q_W \ll P_W$$

means that Q_W is absolutely continuous with respect to P_W. Formally, for every measurable set A,

$$P_W(A) = 0 \quad \Longrightarrow \quad Q_W(A) = 0.$$

This assumption is needed because the conditional mean $m(w) = \mathbb{E}[X \mid W = w]$ is learned from the observational distribution and is only defined P_W-almost surely. The condition $Q_W \ll P_W$ ensures that the interventional common-cause values do not occur in regions that have zero probability under the observational distribution, so the same conditional-mean relation remains valid when evaluating $m(W^I)$. This is weaker than assuming W and W^I have the same distribution: if $Q_W = P_W$, then $Q_W \ll P_W$ holds automatically. Intuitively, the interventional regime may shift the distribution of W, but it should not introduce entirely new regions of W that are unsupported by the observational regime.

A.2 Proof of Theorem 1

Proof. We first consider the observational SEM. Rearranging equation 2 gives $X = (I - B)^{-1}\{b + g(W) + \epsilon\}$. Taking the conditional expectation given W and using $\mathbb{E}[\epsilon \mid W] = 0$ yields $m(W) = \mathbb{E}[X \mid W] = (I - B)^{-1}\{b + g(W)\}$. Equivalently, $(I - B)m(W) = b + g(W)$ almost surely in the observational regime. We then have

$$\begin{aligned}(I - B)X^* &= (I - B)X - (I - B)m(W) \\ &= b + g(W) + \epsilon - \{b + g(W)\} \\ &= \epsilon.\end{aligned}$$

Expanding the left-hand side and rearranging establishes $X^* = BX^* + \epsilon$, which is equation 7.

We now argue that this equality also defines the structural assignment

$$X^* \leftarrow BX^* + \epsilon.$$

Indeed, $X^* = X - m(W)$ is obtained by a deterministic shift depending only on W. Since W is not a descendant of any variable in X, this shift is not affected by interventions on X. Substituting $X = X^* + m(W)$ into the original structural assignments therefore induces a new SEM for X^*, with the same structural coefficient matrix B and the same disturbance ϵ.

For the interventional SEM, Assumption 1 implies that

$$(I - B)m(W^I) = b + g(W^I)$$

almost surely under the interventional distribution. Thus,

$$\begin{aligned}(I - B)X^{I*} &= (I - B)X^I - (I - B)m(W^I) \\ &= b + g(W^I) + \epsilon^I + \delta - \{b + g(W^I)\} \\ &= \epsilon^I + \delta,\end{aligned}$$

and hence

$$X^{I*} = BX^{I*} + \epsilon^I + \delta.$$

Because W^I is unaffected by the intervention, the same change of variables can be applied to the interventional structural assignments, giving

$$X^{I*} \leftarrow BX^{I*} + \epsilon^I + \delta.$$

□

A.3 Proof of Theorem 2

Proof. Since W is not a descendant of any variable in X, the transformation $T_h = X - h(W)$ can be applied directly to the structural assignment in equation 2. Substituting $X = T_h + h(W)$ gives the induced structural equation

$$T_h \leftarrow b + BT_h + \epsilon + g(W) - (I - B)h(W).$$

For the forward direction, suppose there exists a deterministic vector $b_h \in \mathbb{R}^p$ such that

$$T_h \leftarrow b_h + BT_h + \epsilon$$

almost surely. Comparing the two structural assignments yields

$$b + g(W) - (I - B)h(W) = b_h$$

almost surely. Since

$$(I - B)m(W) = b + g(W),$$

we obtain

$$m(W) - h(W) = (I - B)^{-1}b_h.$$

Defining $a = -(I - B)^{-1}b_h$ gives

$$h(w) = m(w) + a$$

for P_W-almost every w, with $b_h = -(I - B)a$.

For the reverse direction, suppose $h(w) = m(w) + a$ for P_W-almost every w. Then the induced structural equation becomes

$$\begin{aligned}T_h &\leftarrow b + BT_h + \epsilon + g(W) - (I - B)h(W) \\ &= BT_h + \epsilon + (I - B)\{m(W) - h(W)\} \\ &= BT_h + \epsilon - (I - B)a.\end{aligned}$$

Thus, with $b_h = -(I - B)a$,

$$T_h \leftarrow b_h + BT_h + \epsilon.$$

By Assumption 1, the identity $h(W^I) = m(W^I) + a$ also holds almost surely under the interventional distribution. Since W^I is not affected by the intervention, applying the same transformation to the interventional structural assignment gives

$$\begin{aligned}T_h^I &\leftarrow b + BT_h^I + \epsilon^I + \delta + g(W^I) - (I - B)h(W^I) \\ &= b_h + BT_h^I + \epsilon^I + \delta.\end{aligned}$$

Hence the transformation preserves the structural coefficient matrix B and the sparse intervention δ in both regimes. □

A.4 Proof of Theorem 3

Proof. Let

$$\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i.$$

Since $\widehat{X}_i^* = X_i^* + e_i$, the corresponding empirical means satisfy $\widehat{\mu} - \widetilde{\mu} = \overline{e}$. By the Jensen inequality,

$$\|\widehat{\mu} - \widetilde{\mu}\|_2^2 = \left\|\frac{1}{n}\sum_{i=1}^{n} e_i\right\|_2^2 \le \frac{1}{n}\sum_{i=1}^{n}\|e_i\|_2^2 = r_n^2.$$

Therefore, $\|\hat{\mu} - \tilde{\mu}\|_2 \le r_n$. To compare the empirical covariances, define the centered vectors

$$u_i = X_i^* - \widetilde{\mu}, \qquad v_i = e_i - \overline{e}.$$

Then $\widehat{X}_i^* - \widehat{\mu} = u_i + v_i$. Let

$$A = \frac{1}{n}\sum_{i=1}^{n} u_i v_i^\top, \qquad E = \frac{1}{n}\sum_{i=1}^{n} v_i v_i^\top.$$

Expanding the empirical covariance of $u_i + v_i$ gives $\widehat{\Sigma} - \widetilde{\Sigma} = A + A^\top + E$. The matrix E is positive semidefinite, and hence

$$\|E\|_{\mathrm{op}} \le \mathrm{tr}(E) = \frac{1}{n}\sum_{i=1}^{n}\|e_i - \overline{e}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\|e_i\|_2^2 - \|\overline{e}\|_2^2 \le r_n^2.$$

For arbitrary unit vectors $a, b \in \mathbb{R}^p$, the scalar Cauchy–Schwarz inequality gives

$$\begin{aligned}|a^\top A b| &= \left|\frac{1}{n}\sum_{i=1}^{n}(a^\top u_i)(b^\top v_i)\right| \le \left\{\frac{1}{n}\sum_{i=1}^{n}(a^\top u_i)^2\right\}^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n}(b^\top v_i)^2\right\}^{1/2} \\ &= \left(a^\top \widetilde{\Sigma} a\right)^{1/2}\left(b^\top E b\right)^{1/2} \le \sqrt{\|\widetilde{\Sigma}\|_{\mathrm{op}}\|E\|_{\mathrm{op}}}.\end{aligned}$$

Taking the supremum over unit vectors a and b yields

$$\|A\|_{\mathrm{op}} \le \sqrt{\|\widetilde{\Sigma}\|_{\mathrm{op}}\|E\|_{\mathrm{op}}} \le \sqrt{\|\widetilde{\Sigma}\|_{\mathrm{op}}}\, r_n.$$

Consequently,

$$\|\widehat{\Sigma} - \widetilde{\Sigma}\|_{\mathrm{op}} \le \|A\|_{\mathrm{op}} + \|A^\top\|_{\mathrm{op}} + \|E\|_{\mathrm{op}} \le 2\sqrt{\|\widetilde{\Sigma}\|_{\mathrm{op}}}\, r_n + r_n^2.$$

This proves the two deterministic bounds.

It remains to establish the consistency statement. By Theorem 1, $X^* = (I - B)^{-1}\epsilon$. Hence $X_1^*, \ldots, X_n^*$ are i.i.d., with $\mathbb{E}[X^*] = 0$ and $\Sigma = \mathrm{Cov}(X^*)$. Since p is fixed and X^* has finite second moments, the law of large numbers gives

$$\widetilde{\mu} = o_p(1)$$

and entrywise convergence of $\widetilde{\Sigma}$ to Σ. Because p is fixed, entrywise convergence implies

$$\|\widetilde{\Sigma} - \Sigma\|_{\mathrm{op}} = o_p(1), \quad \|\widetilde{\Sigma}\|_{\mathrm{op}} = O_p(1).$$

If $r_n = o_p(1)$, the deterministic mean bound gives

$$\|\widehat{\mu} - \widetilde{\mu}\|_2 = o_p(1), \quad \|\widehat{\mu}\|_2 \le \|\widehat{\mu} - \widetilde{\mu}\|_2 + \|\widetilde{\mu}\|_2 = o_p(1).$$

Similarly, the deterministic covariance bound gives

$$\|\widehat{\Sigma} - \widetilde{\Sigma}\|_{\mathrm{op}} \le 2\sqrt{\|\widetilde{\Sigma}\|_{\mathrm{op}}}\, r_n + r_n^2 = o_p(1).$$

A final application of the triangle inequality yields

$$\|\widehat{\Sigma} - \Sigma\|_{\mathrm{op}} \le \|\widehat{\Sigma} - \widetilde{\Sigma}\|_{\mathrm{op}} + \|\widetilde{\Sigma} - \Sigma\|_{\mathrm{op}} = o_p(1).$$

□

A.5 Proof of Corollary 1

Proof. Theorem 3 gives $\widehat{\Sigma} \xrightarrow{p} \Sigma$ and $\|\widehat{\mu} - \widetilde{\mu}\|_2 \le r_n = o_p(1)$. Since $\widetilde{\mu} \xrightarrow{p} \mathbb{E}[X^*] = 0$ by the law of large numbers, we also have $\widehat{\mu} \xrightarrow{p} 0$. Since $\Sigma \succ 0$, continuity of the Cholesky factor gives consistent mean and Cholesky-factor estimates for every permutation. Moreover, $r_I = o_p(1)$ implies that the fitted transformed interventional input converges to its structure-preserving counterpart. Hence the score and maximizer convergences in Equations (23)–(24) of Li et al. (2026) also hold for the fitted transformed inputs, and their Theorem 3.5 gives the result. □

B Sythetic simulation details and results

B.1 Detailed synthetic simulation settings

Causal effect p is set to 100. For each simulation condition, we use 20 independent structural replications. The following generation and calibration steps are performed independently within each structural replication. We first generate a strictly lower-triangular matrix B_{raw}. Its support contains exactly $\lfloor s_B p(p-1)/2 \rfloor$ entries selected uniformly from the possible lower-triangular edges. Each nonzero coefficient is sampled independently from $\mathrm{Unif}(-1, 1)$. We draw the raw noise variances $d_j^{\mathrm{raw}} \sim \mathrm{Unif}(1, 5)$ independently and define $D_\epsilon^{\mathrm{raw}} = \mathrm{diag}(d_1^{\mathrm{raw}}, \ldots, d_p^{\mathrm{raw}})$.

Next, we sample independently for marginal-variance targets $v_j \sim \mathrm{Unif}(10, 30)$. The intercept b has entries independently sampled from $\mathrm{Unif}(-5, 5)$. In topological order (i.e. parent before its children), we retain $d_j = d_j^{\mathrm{raw}}$ for each non-root node and scale its row of B_{raw} until its clean marginal variance equals v_j. For a root node (i.e. a root without any parent), the row of B_{raw} is zero and cannot control the node's variance, so we set $d_j = v_j$. We then define $D_\epsilon = \mathrm{diag}(d_1, \ldots, d_p)$. The result satisfies

$$\mathrm{diag}\{(I - B)^{-1} D_\epsilon (I - B)^{-\top}\} \approx (v_1, \ldots, v_p)^\top,$$

where the maximum absolute difference is at most 0.1. This operation preserves the graph support. After calibration, we apply one random node permutation to B, D_ϵ, the intercept, and the target vector. We then freeze these quantities within the structural replication before generating the common-cause effect.

Linear common-cause. For the linear settings, $g(W) = \Gamma W$. Each raw entry of Γ is active with probability 0.5. An active entry has a magnitude sampled from $\mathrm{Unif}(0.5, 2)$ and an independent sign sampled uniformly from $\{-1, 1\}$. Within each structural replication, we generate one $p \times 30$ raw matrix. The $q = 10$ setting uses its first 10 columns, and the $q = 30$ setting uses all 30 columns.

Nonlinear common-cause. For the nonlinear setting, $q = 10$ and

$$g_j(W) = \sum_{k=1}^{q} M_{jk} c_{jk} h_{jk}(W_k),$$

where $M_{jk} \sim \mathrm{Bernoulli}(0.9)$. For each active term, h_{jk} is sampled uniformly from

$$w, \qquad \tfrac{1}{2}(w^2 - 1), \qquad \sin(\omega_{jk} w), \qquad \{1 + \exp(-w)\}^{-1} - \tfrac{1}{2}.$$

The coefficient c_{jk} has magnitude sampled from $\mathrm{Unif}(0.5, 2)$ and a uniform random sign. For a sine term, $\omega_{jk} \sim \mathrm{Unif}(0.5, 1.2)$. The common cause effect is generated independently and then fixed within each structural replication. We estimate its raw covariance from an independent sample of 50000 covariate vectors.

Common-cause calibration. Let $C_{g,\mathrm{raw}}$ be the exact covariance $\Gamma_{\mathrm{raw}}\Gamma_{\mathrm{raw}}^\top$ in a linear setting or its pilot estimate in the nonlinear setting. With $A = (I - B)^{-1}$, we set

$$c_g = \left[\frac{\rho \,\mathrm{tr}(A D_\epsilon A^\top)}{\mathrm{tr}(A C_{g,\mathrm{raw}} A^\top)}\right]^{1/2}, \qquad \rho = 1,$$

and replace the g_{raw} by $g(W) = c_g g_{\mathrm{raw}}$. For the nonlinear setting, the ratio is defined using the pilot covariance. This calibration is performed separately within each structural replication.

Observational and interventional samples. For each condition and structural replication, we generate

$$X_i = A\{b + g(W_i) + \epsilon_i\}, \quad X_i^I = A\{b + g(W_i^I) + \epsilon_i^I + \delta e_{r_i}\},$$

where $W_i, W_i^I \sim \mathcal{N}(0, I_q)$, $\epsilon_i, \epsilon_i^I \sim \mathcal{N}(0, D_\epsilon)$, and e_{r_i} is the unit vector for the intervention target. Within each structural replication, we draw 50 targets independently and uniformly from $\{1, \ldots, p\}$, with replacement, and generate one interventional sample for each target. Each interventional sample has an independent W_i^I and ϵ_i^I. Thus, the 20 structural replications provide $20 \times 50 = 1000$ interventional samples for each condition.

Within each structural replication, the same 50 interventional baselines and targets are reused across sample sizes and intervention strengths. For the sample-size study, we generate the largest observational sample once per structural replication and use nested prefixes for smaller n. We generate 30-dimensional master samples of W in each structural replication and use the first q columns in each setting.

The baseline is $(n, \delta_r, s_B) = (200, 8, 0.4)$. The sample-size study uses $n \in \{100, 200, 1000\}$; the intervention-strength study uses $\delta_r \in \{4, 8, 12\}$; and the graph-density study uses $s_B \in \{0.2, 0.6\}$. Only one factor changes in each study.

Nuisance estimation. Within each structural replication and sample-size setting, we fit one nuisance regression for each feature of X. Linear $q = 10$ uses OLS. Linear $q = 30$ uses Lasso with 5-fold shuffled cross-validation over 50 values in $\mathrm{logspace}(-4, 1)$. Nonlinear $q = 10$ uses an additive GAM with 10 splines for each coordinate of W. For each response coordinate, we select the smoothing parameter by minimizing the cross-validation criterion over 11 values in $\mathrm{logspace}(-3, 3)$. We standardize W before fitting each nuisance learner. For Lasso, the standardization is fitted within each cross-validation training fold and is refitted on the complete observational sample with the selected penalty.

Inputs, scoring, and metrics. We evaluate four inputs: (X, X^I), $(\widehat{X}^*, \widehat{X}^{I*})$, (X^*, X^{I*}), and $[(X, W), (X^I, W^I)]$. We apply Cholesky RCD with five shuffles, and Cyclic RCA with empirical covariance. For each condition and method, we pool the root-cause ranks from all 20 structural replications and report the empirical CDF and top-1/top-3/top-5 recall over the resulting 1,000 interventional samples.

B.2 Synthetic simulation results

Complete results for varying n, δ_r, and s_B across different inputs are reported in Tables 4–6, Tables 7–9, and Tables 10–12 for the linear $q = 10$, linear $q = 30$, and nonlinear $q = 10$ settings, respectively.

Figure 2 compares the performance of the raw (X, X^I), transformed, and full observed inputs across different sample sizes for the linear $q = 30$ setting. Figures 3 and 4 show the corresponding comparisons as δ_r and s_B vary, respectively, for the linear $q = 10$, linear $q = 30$, and nonlinear $q = 10$ settings.

In addition, we compare the transformed input with the oracle input to assess how closely the estimated transformation approaches the oracle transformation. Figures 5, 6, and 7 show the comparisons as n, δ_r, and s_B vary, respectively.

Across the synthetic experiments, the transformed input $(\widehat{X}^*, \widehat{X}^{I*})$ achieves higher Top-1 recall than the (X, X^I). This improvement is consistent across all settings for Cholesky RCD and in most settings for Cyclic RCA, with a few exceptions under the nonlinear $q = 10$ setting.

Note that at $n = 100, q = 10$, the transformed input slightly outperforms oracle ones under Cyclic RCA. This finite-sample reversal does not necessarily imply superior nuisance recovery: the oracle input removes the true propagated common cause, but Cyclic RCA must still estimate a high-dimensional precision matrix. With $p = n = 100$, this estimate is sensitive to the covariance changes induced by the transformation.

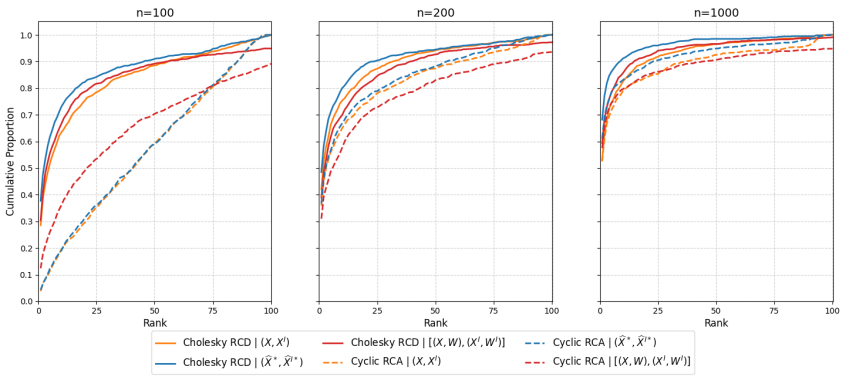


Figure 2: Root-cause recovery across observational sample sizes in the linear $q = 30$ setting, comparing the raw (X, X^I), transformed $(\widehat{X}^*, \widehat{X}^{I*})$, and full observed $[(X, W), (X^I, W^I)]$ inputs.

Table 4: Top-K root-cause recall as the observational sample size n varies in the linear $q = 10$ setting. Fix $\delta_r = 8$ and $s_B = 0.4$.

n	Downstream method	Input	Top 1 recall	Top 3 recall	Top 5 recall
100	Cholesky RCD	(X, X^I)	0.377	0.516	0.586
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.522	0.686	0.748
	Cholesky RCD	(X^*, X^{I*})	0.571	0.711	0.766
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.381	0.532	0.599
	Cyclic RCA	(X, X^I)	0.053	0.094	0.139
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.105	0.179	0.234
	Cyclic RCA	(X^*, X^{I*})	0.048	0.095	0.134
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.098	0.161	0.206
200	Cholesky RCD	(X, X^I)	0.492	0.646	0.695
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.544	0.681	0.747
	Cholesky RCD	(X^*, X^{I*})	0.572	0.715	0.781
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.486	0.629	0.693
	Cyclic RCA	(X, X^I)	0.412	0.545	0.596
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.413	0.553	0.610
	Cyclic RCA	(X^*, X^{I*})	0.451	0.585	0.639
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.400	0.525	0.581
1000	Cholesky RCD	(X, X^I)	0.593	0.727	0.787
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.692	0.827	0.866
	Cholesky RCD	(X^*, X^{I*})	0.699	0.825	0.870
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.613	0.751	0.804
	Cyclic RCA	(X, X^I)	0.576	0.716	0.771
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.616	0.751	0.789
	Cyclic RCA	(X^*, X^{I*})	0.623	0.758	0.794
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.597	0.727	0.773

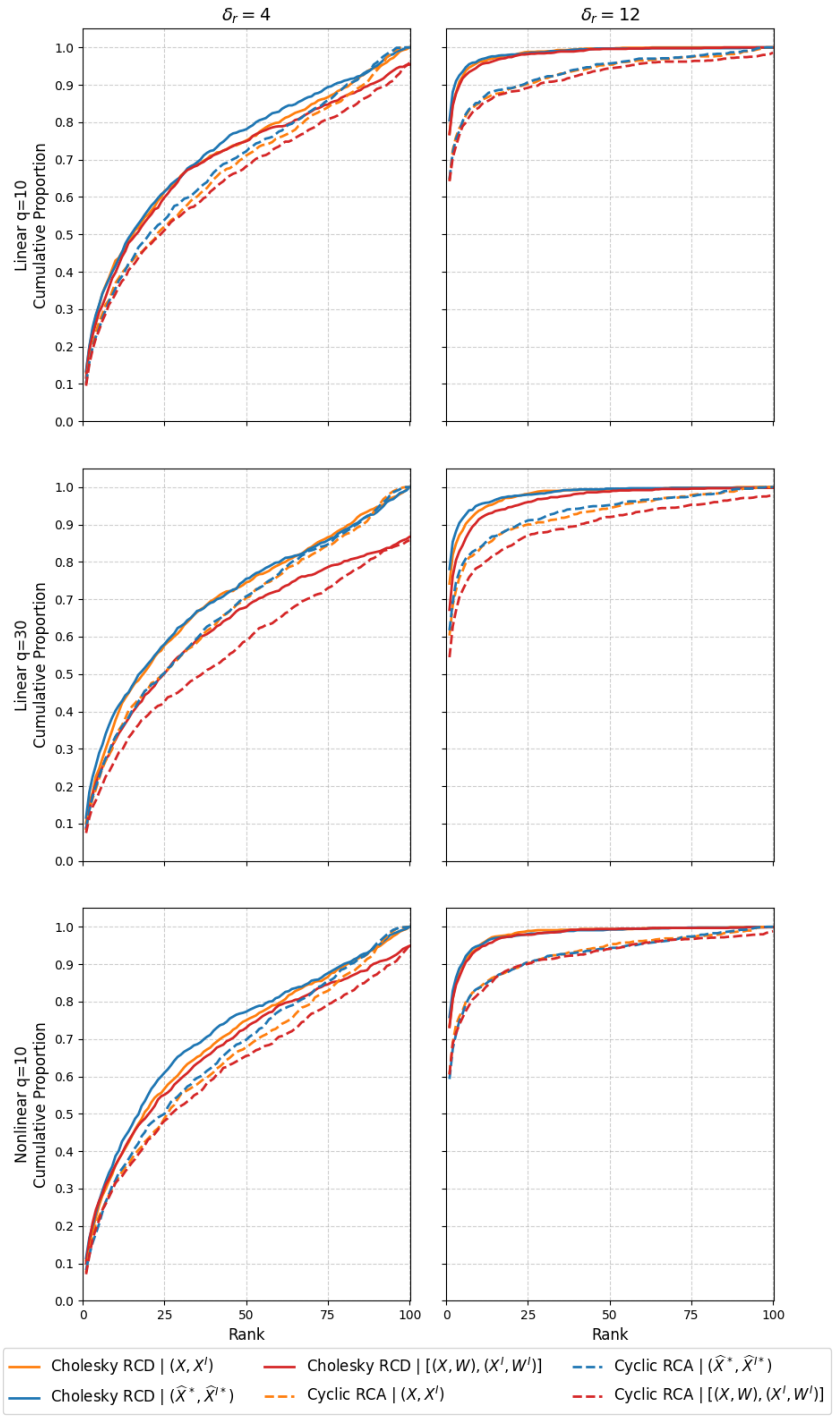


Figure 3: Root-cause recovery across the intervention strength δ_r in the linear $q = 10$, linear $q = 30$, and nonlinear $q = 10$ settings, comparing the raw (X, X^I), transformed $(\widehat{X}^*, \widehat{X}^{I*})$, and full observed $[(X, W), (X^I, W^I)]$ inputs.

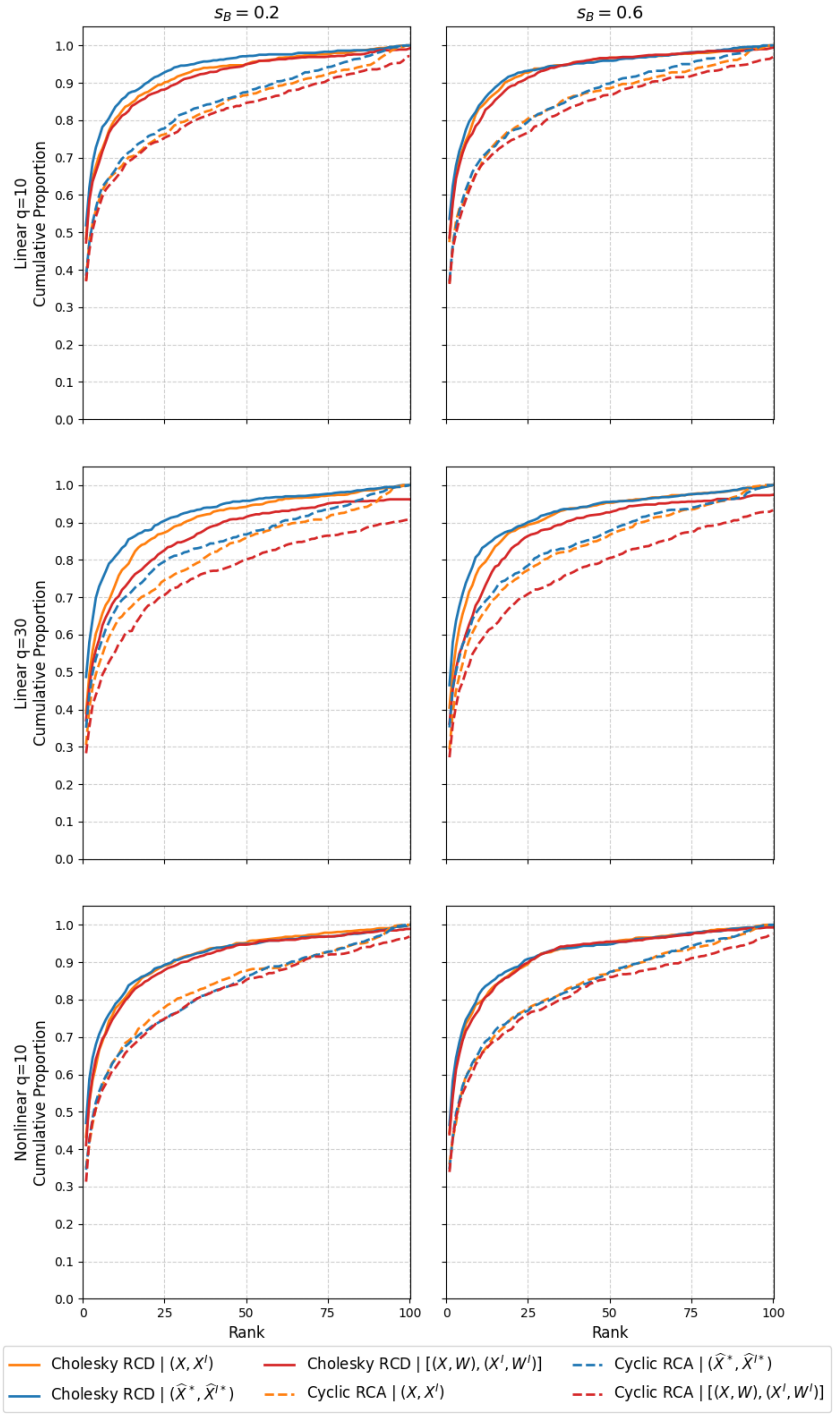


Figure 4: Root-cause recovery across s_B in the linear $q = 10$, linear $q = 30$, and nonlinear $q = 10$ settings, comparing the raw (X, X^I), transformed $(\widehat{X}^*, \widehat{X}^{I*})$, and full observed $[(X, W), (X^I, W^I)]$ inputs.

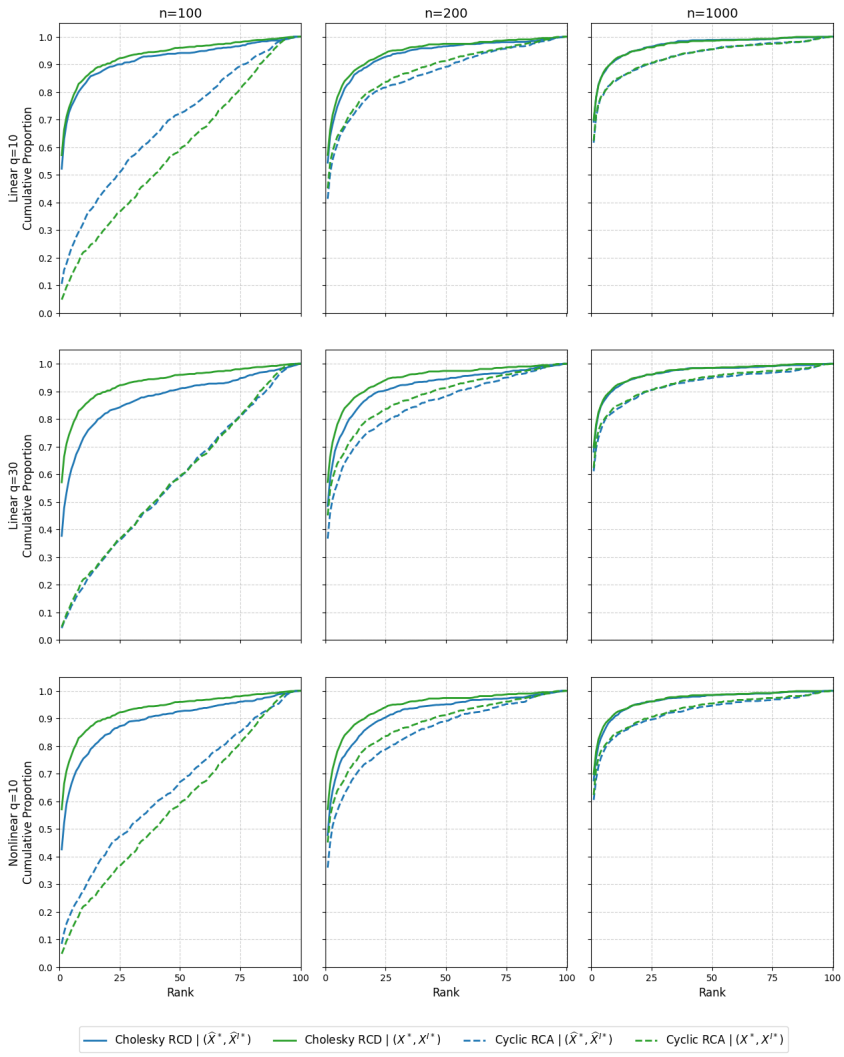


Figure 5: Root-cause recovery across observational sample sizes for the oracle (X^*, X^{I*}) and fitted transformed $(\widehat{X}^*, \widehat{X}^{I*})$ inputs in the linear $q = 10$, linear $q = 30$, and nonlinear $q = 10$ settings.

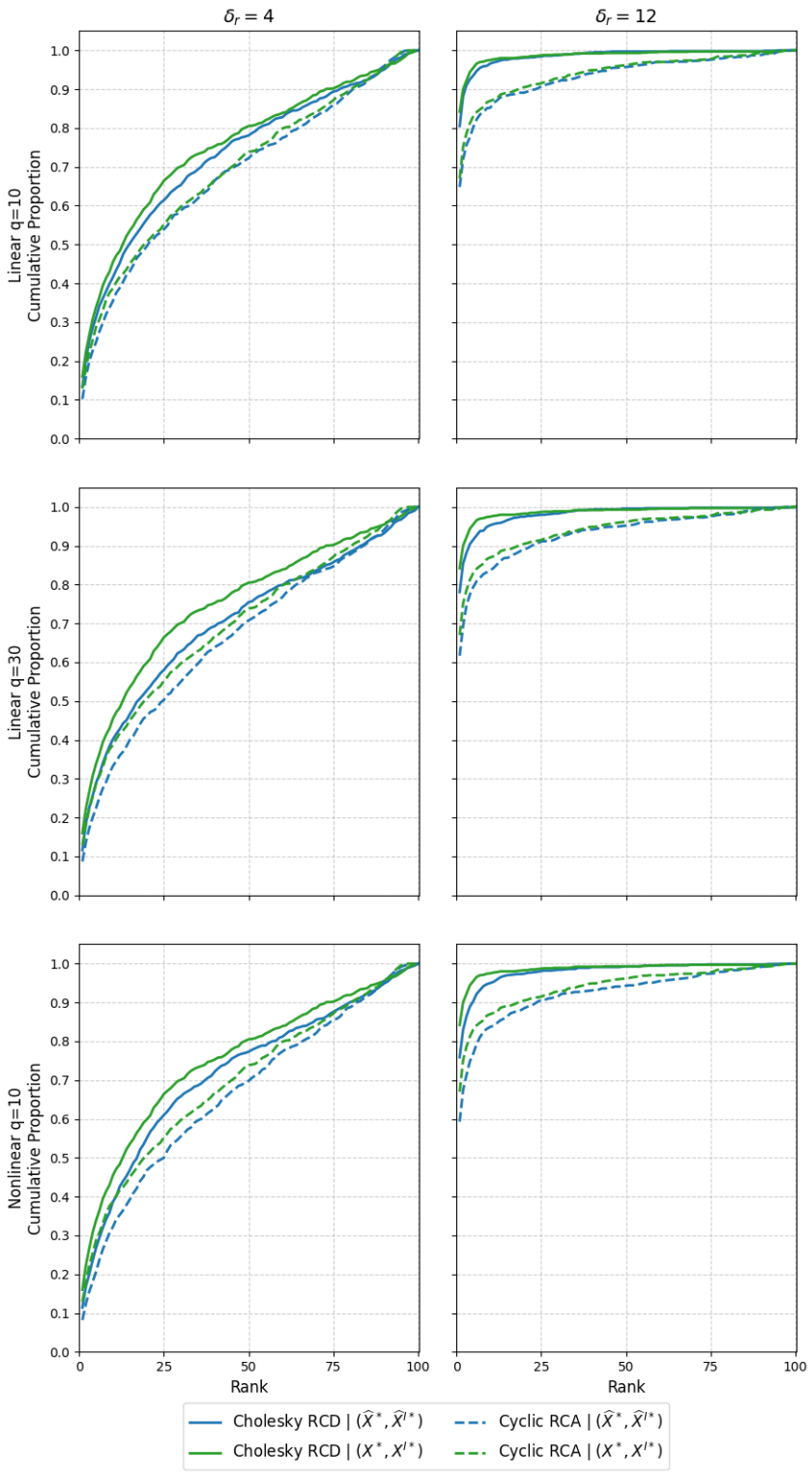


Figure 6: Root-cause recovery across the intervention strength δ_r for the oracle (X^*, X^{I*}) and fitted transformed $(\widehat{X}^*, \widehat{X}^{I*})$ inputs in the linear $q = 10$, linear $q = 30$, and nonlinear $q = 10$ settings.

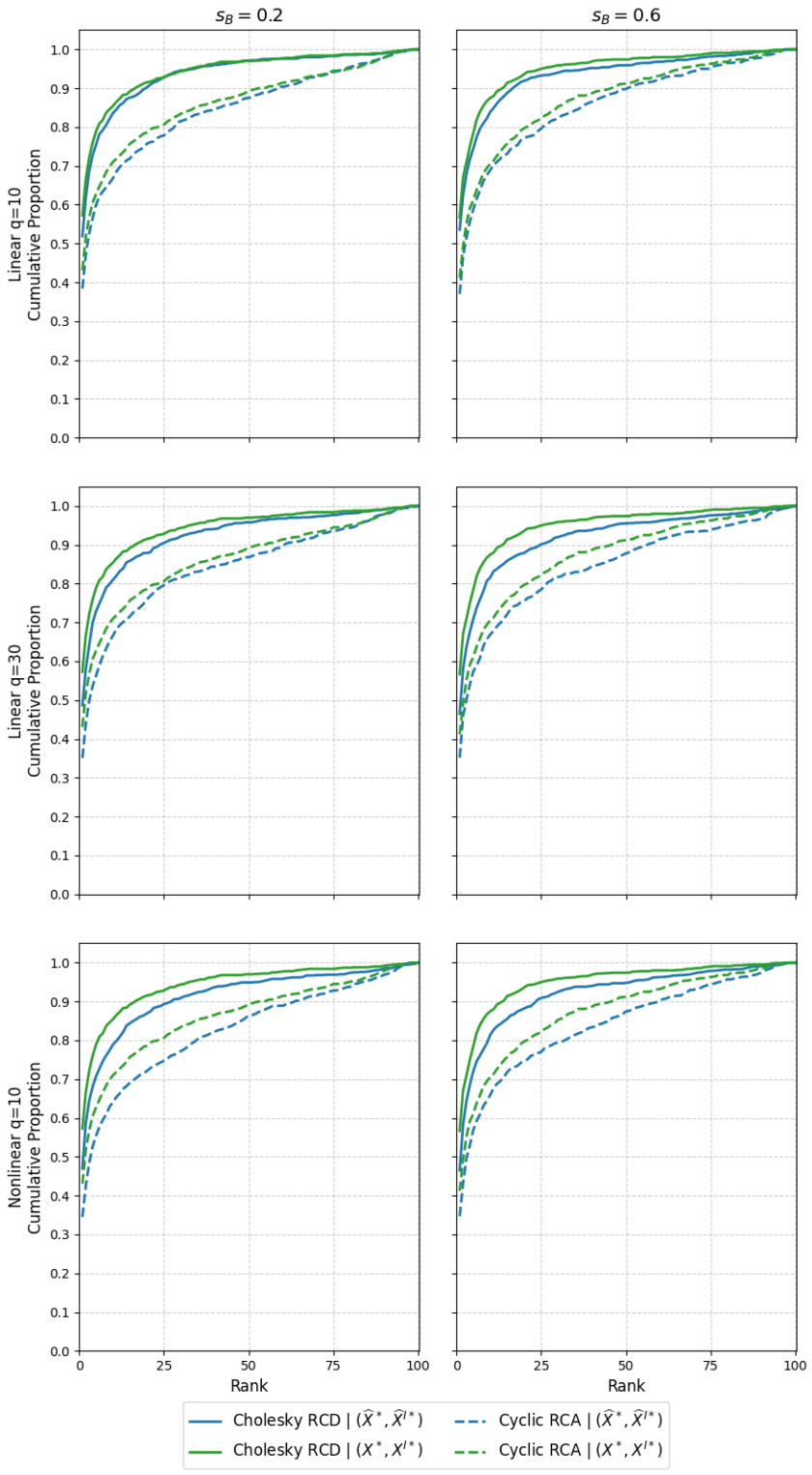


Figure 7: Root-cause recovery across s_B for the oracle (X^*, X^{I*}) and fitted transformed $(\widehat{X}^*, \widehat{X}^{I*})$ inputs in the linear $q = 10$, linear $q = 30$, and nonlinear $q = 10$ settings.

Table 5: Top-K root-cause recall as the intervention strength δ_r varies in the linear $q = 10$ setting. Fix $n = 200$ and $s_B = 0.4$.

δ_r	Downstream method	Input	Top 1 recall	Top 3 recall	Top 5 recall
4.0	Cholesky RCD	(X, X^I)	0.131	0.244	0.310
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.134	0.251	0.312
	Cholesky RCD	(X^*, X^{I*})	0.160	0.266	0.338
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.117	0.230	0.292
	Cyclic RCA	(X, X^I)	0.101	0.202	0.267
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.102	0.198	0.248
	Cyclic RCA	(X^*, X^{I*})	0.129	0.222	0.287
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.095	0.192	0.244
12.0	Cholesky RCD	(X, X^I)	0.769	0.879	0.922
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.805	0.910	0.935
	Cholesky RCD	(X^*, X^{I*})	0.842	0.922	0.955
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.768	0.877	0.917
	Cyclic RCA	(X, X^I)	0.644	0.761	0.798
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.648	0.756	0.802
	Cyclic RCA	(X^*, X^{I*})	0.670	0.786	0.830
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.642	0.737	0.789

Table 6: Top-K root-cause recall as the graph density s_B varies in the linear $q = 10$ setting. Fix $n = 200$ and $\delta_r = 8$.

s_B	Downstream method	Input	Top 1 recall	Top 3 recall	Top 5 recall
0.2	Cholesky RCD	(X, X^I)	0.484	0.645	0.706
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.519	0.686	0.754
	Cholesky RCD	(X^*, X^{I*})	0.573	0.721	0.789
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.473	0.636	0.688
	Cyclic RCA	(X, X^I)	0.372	0.525	0.587
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.384	0.526	0.599
	Cyclic RCA	(X^*, X^{I*})	0.431	0.567	0.627
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.369	0.506	0.569
0.6	Cholesky RCD	(X, X^I)	0.477	0.660	0.721
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.536	0.679	0.744
	Cholesky RCD	(X^*, X^{I*})	0.567	0.709	0.784
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.486	0.644	0.712
	Cyclic RCA	(X, X^I)	0.365	0.531	0.574
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.370	0.515	0.587
	Cyclic RCA	(X^*, X^{I*})	0.412	0.552	0.611
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.362	0.497	0.562

Table 7: Top-K root-cause recall as the observational sample size n varies in the linear $q = 30$ setting. Fix $\delta_r = 8$ and $s_B = 0.4$.

n	Downstream method	Input	Top 1 recall	Top 3 recall	Top 5 recall
100	Cholesky RCD	(X, X^I)	0.286	0.440	0.518
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.377	0.534	0.618
	Cholesky RCD	(X^*, X^{I*})	0.571	0.711	0.766
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.304	0.482	0.557
	Cyclic RCA	(X, X^I)	0.038	0.084	0.114
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.043	0.083	0.123
	Cyclic RCA	(X^*, X^{I*})	0.048	0.095	0.134
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.124	0.208	0.261
200	Cholesky RCD	(X, X^I)	0.420	0.579	0.662
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.485	0.637	0.710
	Cholesky RCD	(X^*, X^{I*})	0.572	0.715	0.781
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.390	0.531	0.616
	Cyclic RCA	(X, X^I)	0.359	0.501	0.560
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.367	0.508	0.566
	Cyclic RCA	(X^*, X^{I*})	0.451	0.585	0.639
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.310	0.426	0.487
1000	Cholesky RCD	(X, X^I)	0.531	0.679	0.744
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.682	0.818	0.862
	Cholesky RCD	(X^*, X^{I*})	0.699	0.825	0.870
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.594	0.729	0.785
	Cyclic RCA	(X, X^I)	0.525	0.667	0.718
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.611	0.733	0.784
	Cyclic RCA	(X^*, X^{I*})	0.623	0.758	0.794
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.575	0.699	0.745

Table 8: Top-K root-cause recall as the intervention strength δ_r varies in the linear $q = 30$ setting. Fix $n = 200$ and $s_B = 0.4$.

δ_r	Downstream method	Input	Top 1 recall	Top 3 recall	Top 5 recall
4.0	Cholesky RCD	(X, X^I)	0.102	0.183	0.247
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.116	0.225	0.289
	Cholesky RCD	(X^*, X^{I*})	0.160	0.266	0.338
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.087	0.176	0.232
	Cyclic RCA	(X, X^I)	0.082	0.167	0.221
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.087	0.172	0.224
	Cyclic RCA	(X^*, X^{I*})	0.129	0.222	0.287
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.075	0.145	0.184
12.0	Cholesky RCD	(X, X^I)	0.742	0.849	0.884
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.781	0.882	0.916
	Cholesky RCD	(X^*, X^{I*})	0.842	0.922	0.955
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.672	0.805	0.844
	Cyclic RCA	(X, X^I)	0.603	0.725	0.776
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.617	0.746	0.794
	Cyclic RCA	(X^*, X^{I*})	0.670	0.786	0.830
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.545	0.670	0.724

Table 9: Top-K root-cause recall as the graph density s_B varies in the linear $q = 30$ setting. Fix $n = 200$ and $\delta_r = 8$.

s_B	Downstream method	Input	Top 1 recall	Top 3 recall	Top 5 recall
0.2	Cholesky RCD	(X, X^I)	0.391	0.557	0.628
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.487	0.641	0.729
	Cholesky RCD	(X^*, X^{I*})	0.573	0.721	0.789
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.371	0.523	0.587
	Cyclic RCA	(X, X^I)	0.305	0.460	0.522
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.351	0.498	0.563
	Cyclic RCA	(X^*, X^{I*})	0.431	0.567	0.627
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.283	0.412	0.462
0.6	Cholesky RCD	(X, X^I)	0.405	0.568	0.655
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.465	0.637	0.709
	Cholesky RCD	(X^*, X^{I*})	0.567	0.709	0.784
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.362	0.509	0.572
	Cyclic RCA	(X, X^I)	0.293	0.451	0.522
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.352	0.501	0.574
	Cyclic RCA	(X^*, X^{I*})	0.412	0.552	0.611
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.272	0.409	0.473

Table 10: Top-K root-cause recall as the observational sample size n varies in the nonlinear $q = 10$ setting. Fix $\delta_r = 8$ and $s_B = 0.4$.

n	Downstream method	Input	Top 1 recall	Top 3 recall	Top 5 recall
100	Cholesky RCD	(X, X^I)	0.343	0.498	0.581
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.426	0.592	0.662
	Cholesky RCD	(X^*, X^{I*})	0.571	0.711	0.766
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.358	0.494	0.572
	Cyclic RCA	(X, X^I)	0.044	0.090	0.134
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.084	0.158	0.200
	Cyclic RCA	(X^*, X^{I*})	0.048	0.095	0.134
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.085	0.143	0.196
200	Cholesky RCD	(X, X^I)	0.455	0.617	0.684
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.478	0.637	0.706
	Cholesky RCD	(X^*, X^{I*})	0.572	0.715	0.781
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.443	0.614	0.671
	Cyclic RCA	(X, X^I)	0.345	0.499	0.569
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.360	0.509	0.566
	Cyclic RCA	(X^*, X^{I*})	0.451	0.585	0.639
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.347	0.487	0.556
1000	Cholesky RCD	(X, X^I)	0.578	0.703	0.767
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.676	0.803	0.848
	Cholesky RCD	(X^*, X^{I*})	0.699	0.825	0.870
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.592	0.716	0.772
	Cyclic RCA	(X, X^I)	0.559	0.695	0.738
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.605	0.727	0.778
	Cyclic RCA	(X^*, X^{I*})	0.623	0.758	0.794
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.580	0.703	0.741

Table 11: Top-K root-cause recall as the intervention strength δ_r varies in the nonlinear $q = 10$ setting. Fix $n = 200$ and $s_B = 0.4$.

δ_r	Downstream method	Input	Top 1 recall	Top 3 recall	Top 5 recall
4.0	Cholesky RCD	(X, X^I)	0.108	0.189	0.255
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.114	0.200	0.267
	Cholesky RCD	(X^*, X^{I*})	0.160	0.266	0.338
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.100	0.208	0.267
	Cyclic RCA	(X, X^I)	0.076	0.174	0.226
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.082	0.155	0.206
	Cyclic RCA	(X^*, X^{I*})	0.129	0.222	0.287
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.071	0.165	0.216
12.0	Cholesky RCD	(X, X^I)	0.734	0.854	0.899
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.759	0.864	0.903
	Cholesky RCD	(X^*, X^{I*})	0.842	0.922	0.955
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.732	0.845	0.887
	Cyclic RCA	(X, X^I)	0.606	0.738	0.776
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.593	0.713	0.768
	Cyclic RCA	(X^*, X^{I*})	0.670	0.786	0.830
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.605	0.716	0.757

Table 12: Top-K root-cause recall as the graph density s_B varies in the nonlinear $q = 10$ setting. Fix $n = 200$ and $\delta_r = 8$.

s_B	Downstream method	Input	Top 1 recall	Top 3 recall	Top 5 recall
0.2	Cholesky RCD	(X, X^I)	0.433	0.585	0.665
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.470	0.644	0.707
	Cholesky RCD	(X^*, X^{I*})	0.573	0.721	0.789
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.411	0.595	0.668
	Cyclic RCA	(X, X^I)	0.350	0.468	0.544
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.345	0.484	0.554
	Cyclic RCA	(X^*, X^{I*})	0.431	0.567	0.627
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.313	0.476	0.534
0.6	Cholesky RCD	(X, X^I)	0.444	0.633	0.697
	Cholesky RCD	$(\widehat{X}^*, \widehat{X}^{I*})$	0.465	0.644	0.720
	Cholesky RCD	(X^*, X^{I*})	0.567	0.709	0.784
	Cholesky RCD	$[(X, W), (X^I, W^I)]$	0.440	0.613	0.690
	Cyclic RCA	(X, X^I)	0.359	0.471	0.561
	Cyclic RCA	$(\widehat{X}^*, \widehat{X}^{I*})$	0.347	0.496	0.569
	Cyclic RCA	(X^*, X^{I*})	0.412	0.552	0.611
	Cyclic RCA	$[(X, W), (X^I, W^I)]$	0.339	0.483	0.550

C SYNTHETIC EXPERIMENTS ON LINGAM

We conduct experiments to examine whether our proposed causal-structure-preserving transformation also helps with another causal task: causal graph estimation. Our transformation removes the estimated effect of W while preserving the same B. Therefore, a causal graph estimated from $\widehat{X}^*$ should be closer to the true graph than one estimated from the raw X. We test this task using DirectLiNGAM and compare graph recovery from the raw and transformed data. This experiment shows whether removing observed common-cause effects improves causal graph estimation in addition to improving root-cause discovery.

In our experiment, we use $p \in \{10, 20\}$, $q = 10$, $n = 200$, intervention strength $\delta_r = 8$, and DAG density $s_B = 0.4$, while keeping the random-DAG construction, variance calibration, and linear and nonlinear common-cause effects from our main simulation.To satisfy the non-Gaussian error assumption of LiNGAM, we then generate each structural error as $\epsilon_j \sim \mathrm{Unif}(-\sqrt{3d_j}, \sqrt{3d_j})$, which has variance d_j, where d_j is the calibrated variance defined in the main simulation in AppendixB. OLS estimates the transformation under linear common-cause effects, and GAM estimates it under nonlinear effects. The average MSE and normalized structural hamming distance are reported in Table 13. We can see that the transformed variables improve causal-graph estimation in every setting. The gains are larger in the linear setting which might be because OLS correctly represents the linear conditional mean, whereas estimating the nonlinear conditional mean with a GAM introduces more estimation error. These results show that removing the observed common-cause effects helps DirectLiNGAM recover both the edge coefficients and the graph structure, supporting the use of our transformation for causal graph estimation as well as root-cause discovery.

Table 13: Causal-structure estimation errors for DirectLiNGAM under linear and nonlinear common-cause effects, comparing the raw and transformed inputs. Entries are means, with standard deviations shown in parentheses on the following row. SHD denotes structural Hamming distance.

Effect	p	Input	B MSE	Normalized SHD
Linear	10	X	0.1186	0.5756
			(0.0431)	(0.1062)
	10	$\widehat{X}^*$	**0.0080**	**0.0933**
			(0.0143)	(0.0709)
	20	X	0.0768	0.6050
			(0.0151)	(0.0506)
	20	$\widehat{X}^*$	**0.0089**	**0.1584**
			(0.0100)	(0.0634)
Nonlinear	10	X	0.1032	0.5033
			(0.0454)	(0.0932)
	10	$\widehat{X}^*$	**0.0382**	**0.2100**
			(0.0358)	(0.1111)
	20	X	0.0644	0.5737
			(0.0126)	(0.0493)
	20	$\widehat{X}^*$	**0.0300**	**0.3013**
			(0.0133)	(0.0874)

D PETSHOP

D.1 PRE-PROCESSING

We predefined the candidate observed common-cause variables which are root nodes (i.e. no parent nodes) from the service map provided in the dataset, and are plausible as common cause from the distributed system structure. The candidate set comprised

- `servi-searc.us-west-2.elb.amazonaws.com_remote`,

- `STS_AWS::STS`,
- `SSM_AWS::SSM`,
- `Evidently_AWS::Evidently`

There are three scenarios dividing based on the traffic: high traffic, low traffic, and temporal traffic (which is collected over the day so the traffic can be both high and low). Under each scenario, there is a training set and a test set of incidents, as well as a normal reference as the observational data. Our method, however, does not require a train-test split so we pooled the incidents from the original training and test partitions and retained only incidents whose target metric was latency, as no candidate common cause remained for the availability metric after pre-processing. Each incident contains five timestamps which are treated as five interventional samples, corresponding to the same root cause. From both the normal-reference and incident, we selected the `Average` statistic of the latency metric as suggested by the original paper and retained columns present in both matrices. In the normal-reference data, rows containing missing values were removed before columns were removed, thereby retaining as many service features as possible; columns were removed only when necessary to eliminate the remaining missing values. We subsequently removed columns with normal-period variance below 10^{-7} and duplicate columns, and standardized every retained feature using its normal-reference sample mean and standard deviation. Finally, we defined W as the exact intersection between the retained service names and the candidate set above; all other retained service features formed X. We excluded a timestamp if pre-processing failed, fewer than 50 normal-reference observations remained, the annotated root-cause service was removed, or no candidate variable for W remained.

D.2 Detailed results and analysis

We first present the results from Cholesky RCA. Table 14 shows the incident level results, while Table 15 demonstrates the timestamp level results.

Table 14: Incident-level root-cause ranks produced by Cholesky RCD under the high-, low-, and temporal-traffic settings. Lower ranks indicate better recovery. Labels such as "Test0" and "Train0" are incident identifiers and do not indicate a train–test split. Likewise, "Temporal traffic 1" and "Temporal traffic 2" denote subfolders within the same temporal-traffic scenario, rather than separate scenarios.

(a) High traffic

Nuisance model	Input	Test0	Test1	Test10	Test11	Test12	Test13	Test14	Test15	Test16	Test17
–	(X, X^I)	18	20	26	12	20	**15**	2	1	2	2
random forest	$(\widehat{X}^*, \widehat{X}^{I*})$	**17**	**12**	**8**	12	20	18	**1**	1	**1**	**1**
GAM	$(\widehat{X}^*, \widehat{X}^{I*})$	17	18	19	22	22	18	2	1	2	2
–	$[(X, W), (X^I, W^I)]$	21	23	22	22	21	16	2	2	2	2

(b) Low traffic

Nuisance model	Input	Test0	Test1	Test10	Test11	Test12	Test13	Test14	Test15	Test16	Test17	Train0	Train1	Train2	Train3
–	(X, X^I)	12	7	1	1	1	9	11	**24**	3	1	1	2	1	1
random forest	$(\widehat{X}^*, \widehat{X}^{I*})$	14	7	1	1	1	9	14	26	3	1	1	2	1	1
GAM	$(\widehat{X}^*, \widehat{X}^{I*})$	**10**	7	1	1	1	9	12	26	3	1	1	2	1	1
–	$[(X, W), (X^I, W^I)]$	9	7	1	1	1	9	**9**	26	3	1	1	2	1	1

(c) Temporal traffic

		Temporal traffic 1				Temporal traffic 2			
Nuisance model	Input	Test0	Test1	Test2	Train0	Test0	Test1	Test2	Train0
–	(X, X^I)	1	28	**14**	1	**4**	6	1	1
random forest	$(\widehat{X}^*, \widehat{X}^{I*})$	1	28	16	1	5	**5**	1	1
GAM	$(\widehat{X}^*, \widehat{X}^{I*})$	1	28	15	1	8	6	1	23
–	$[(X, W), (X^I, W^I)]$	1	30	15	1	5	7	1	1

We next show the results from choosing Cyclic RCA as the downstream RCD method on the temporal traffic scenario. Table 16 shows the incident level rank, while Table 17 shows the timestamp level statistics. We can see that there are small improvements using the transformed $\widehat{X}^*$ as input.

Lastly, we present the cross-validated prediction error for each scenario with each nuisance estimator in Table 18.

Table 15: Timestamp-level root-cause recovery on the PetShop dataset under the high-, low-, and temporal-traffic settings using Cholesky RCD. The Top-K count is the number of timestamps for which the true root cause is ranked among the top K; higher counts indicate better recovery, whereas lower mean and median ranks are better. The high-, low-, and temporal-traffic settings contain 50, 70, and 40 evaluated timestamps, respectively.

Scenario	Nuisance model	Input	Top 1	Top 3	Top 5	Mean rank ↓	Median rank ↓
High traffic	–	(X, X^I)	2	11	27	**7.78**	**5.0**
	random forest	$(\widehat{X}^*, \widehat{X}^{I*})$	**7**	10	27	8.36	**5.0**
	GAM	$(\widehat{X}^*, \widehat{X}^{I*})$	3	11	21	8.92	8.0
	–	$[(X, W), (X^I, W^I)]$	2	11	24	8.86	6.0
Low traffic	–	(X, X^I)	23	**27**	**46**	**7.21**	5.0
	random forest	$(\widehat{X}^*, \widehat{X}^{I*})$	**24**	26	45	7.34	5.0
	GAM	$(\widehat{X}^*, \widehat{X}^{I*})$	23	25	41	7.60	5.0
	–	$[(X, W), (X^I, W^I)]$	21	25	44	7.77	5.0
Temporal traffic	–	(X, X^I)	**12**	18	**27**	6.93	4.0
	random forest	$(\widehat{X}^*, \widehat{X}^{I*})$	10	**21**	**27**	**6.73**	**3.0**
	GAM	$(\widehat{X}^*, \widehat{X}^{I*})$	7	11	20	8.10	5.5
	–	$[(X, W), (X^I, W^I)]$	11	15	26	7.70	4.0

Table 16: Incident-level root-cause ranks produced by Cyclic RCA in the temporal-traffic scenario. Lower ranks indicate better recovery. "Temporal traffic 1" and "Temporal traffic 2" denote subfolders within the same scenario, rather than separate scenarios. Labels such as "Test0" and "Train0" are incident identifiers and do not indicate a train–test split.

		Temporal traffic 1				Temporal traffic 2			
Nuisance model	Input	Test0	Test1	Test2	Train0	Test0	Test1	Test2	Train0
–	(X, X^I)	2	28	1	1	26	1	1	1
random forest	$(\widehat{X}^*, \widehat{X}^{I*})$	2	**27**	1	1	**25**	1	1	1
GAM	$(\widehat{X}^*, \widehat{X}^{I*})$	2	**27**	1	1	27	1	1	5
–	$[(X, W), (X^I, W^I)]$	2	30	1	1	27	1	1	1

Table 17: Timestamp-level root-cause recovery in the temporal-traffic scenario using Cyclic RCA. The Top-K count is the number of timestamps for which the true root cause is ranked among the top K; higher counts indicate better recovery, whereas lower mean and median ranks are better. The table summarizes 40 evaluated timestamps.

Nuisance model	Input	#Top 1	#Top 3	#Top 5	Mean rank ↓	Median rank ↓
–	(X, X^I)	13	20	21	9.95	3.5
random forest	$(\widehat{X}^*, \widehat{X}^{I*})$	13	20	21	**9.58**	3.5
GAM	$(\widehat{X}^*, \widehat{X}^{I*})$	9	18	20	10.98	5.0
–	$[(X, W), (X^I, W^I)]$	13	20	21	10.23	3.5

Table 18: Mean and median cross-validated MSE and RMSE for the GAM and random-forest nuisance estimators under the high-, low-, and temporal-traffic settings.

Scenario	Nuisance model	Mean CV MSE	Median CV MSE	Mean CV RMSE	Median CV RMSE
High traffic	GAM	0.935334	0.942562	0.967034	0.970857
	random forest	0.866915	0.866422	0.931079	0.930818
Low traffic	GAM	0.893063	0.892388	0.944985	0.944663
	random forest	0.884507	0.885060	0.940482	0.940776
Temporal traffic	GAM	0.891976	0.890597	0.944431	0.943714
	random forest	0.910928	0.911498	0.954425	0.954724

Tables 14 to 17 show that the proposed transformation (input with $(\widehat{X}^*, \widehat{X}^{I*})$ generally improves root-cause recovery. The improvement is especially clear when random forest is used as the nuisance model. Under high traffic, Cholesky RCD with the transformed input improves several incident-level ranks, for example from 20 to 12 for Test1 and from 26 to 8 for Test10. At the timestamp level, the number of Top-1 recoveries increases from 2 with the (X, X^I) input to 7 with the transformed input. Under low traffic, the improvement is smaller because the original inputs already perform well, but the transformed input still achieves the highest Top-1 count.

The improvement is less consistent under temporal traffic. In this setting, the observational data contain different traffic levels over time, so the relation between W and X is more variable than in the fixed high- or low-traffic settings. This makes $m(W)$ harder to estimate accurately, especially at the incident-specific W^I, and therefore can reduce the benefit of the transformation. Even so, with random forest, the transformed input improves the Top-3 count for Cholesky RCD from 18 to 21 and reduces the mean rank and the median rank.

The root-cause results are also consistent with the nuisance-model prediction errors in Table 18. Under high and low traffic, random forest has lower cross-validated MSE and RMSE than GAM, and it also gives better downstream results after transformation. Under temporal traffic, GAM has slightly lower average cross-validated error, but random forest still gives better root-cause recovery. This shows that average prediction error is related to, but does not fully determine, downstream performance, since root-cause recovery also depends on how accurately $m(W^I)$ is estimated at the incident points.

The overall improvement brought by the transformation is more obvious for Cyclic RCA. One possible explanation might be that Cyclic RCA directly uses the estimated precision matrix to construct its root-cause scores, so removing common-cause variation directly changes the covariance structure on which its scoring procedure relies. In contrast, Cholesky RCD uses the covariance structure indirectly through permutation-specific Cholesky factors, which may make the effect of the transformation less obvious.

E RNA-SEQUENCING DATASET DETAILS

E.1 PREPROCESSING

We followed the preprocessing procedure and released code of Li et al. (2026) for the RNA-sequencing dataset. Briefly, genes were retained if their raw count exceeded 10 in more than 10% of the samples; genes with absolute pairwise correlation greater than 0.999 were then removed. This produced $p = 19736$ genes, 364 observational samples, and 59 target samples with known root-cause genes which were treated as interventional samples. We compared an unadjusted input (X, X^I) with a protocol-adjusted input $(\widehat{X}^*, \widehat{X}^{I*})$. Specifically, let $P_i = 0$ denote the non-strand-specific protocol and $P_i = 1$ the strand-specific protocol. For each gene j, the nuisance function was estimated by the ordinary least-squares model $\widehat{m}_j(P_i) = \widehat{\alpha}_j + \widehat{\beta}_j P_i$. The fitted coefficients were applied to both the observational and target samples to obtain $\widehat{X}^*_{ij} = X_{ij} - \widehat{m}_j(P_i)$. Before applying Cyclic RCA, the X and $\widehat{X}^*$ inputs were standardized as well as their target samples, both using observational statistics. We applied the Shrunk-Covariance version of Cyclic RCA from Schkoda & Janzing (2026) using the shrinkage estimator with shrinkage set to 0.1, suggested in their paper.

E.2 DETAILED RESULTS

Table 19 shows the full ranking results for each patient using Cyclic RCA with the raw (X, X^I) and transformed $(\widehat{X}^*, \widehat{X}^{I*})$ as inputs.

Table 19: Patient-level root-cause ranks from Cyclic RCA using the raw and transformed inputs. Lower ranks indicate better recovery.

Input	**R62943**	**R98254**	**R89912**	**R19100**	**R15264**	**R36605**	**R61100**	**R77611**	**R16472**	**R51757**
(X, X^I)	2	4	6	3	1	3	1	1	1	2
$(\widehat{X}^*, \widehat{X}^{I*})$	2	4	6	3	1	3	1	1	1	2
Arm	**R80346**	**R20754**	**R25473**	**R28774**	**R96820**	**R21147**	**R64921**	**R52016**	**R46723**	**R80184**
(X, X^I)	46	1	1	1	1	2	1	1	6393	1
$(\widehat{X}^*, \widehat{X}^{I*})$	**36**	1	1	1	1	2	1	1	**4997**	1
Arm	**R59185**	**R44456**	**R24289**	**R91273**	**R60537**	**R72253**	**R75000**	**R91016**	**R82353**	**R34834**
(X, X^I)	2	2	5	1	1	3	8	1	1	1
$(\widehat{X}^*, \widehat{X}^{I*})$	2	2	5	1	1	3	8	1	1	1
Arm	**R78764**	**R30367**	**R76358**	**R12128**	**R45867**	**R31640**	**R95723**	**R55237**	**R18626**	**R34820**
(X, X^I)	1	1	6	2	1	1	1	1	39	1
$(\widehat{X}^*, \widehat{X}^{I*})$	1	1	6	2	1	1	1	1	**34**	1
Arm	**R25912**	**R11258**	**R64046**	**R54158**	**R19907**	**R27473**	**R70186**	**R26710**	**R30525**	**R29620**
(X, X^I)	3	1	1	1	1	1	4	33	1	1
$(\widehat{X}^*, \widehat{X}^{I*})$	3	1	1	1	1	1	4	33	1	1
Arm	**R64055**	**R15748**	**R66814**	**R77365**	**R21993**	**R42505**	**R64948**	**R21470**	**R47816**	
(X, X^I)	1	1	1	1	2	1	1	1	1	
$(\widehat{X}^*, \widehat{X}^{I*})$	1	1	1	1	2	1	1	1	1	