Sample-Specific Root Causal Inference with Latent Variables

Eric V. Strobl ERIC.STROBL@VUMC.ORG and **Thomas A. Lasko** TOM.LASKO@VUMC.ORG
Vanderbilt University Medical Center

Editors: Mihaela van der Schaar, Dominik Janzing and Cheng Zhang

Abstract

Root causal analysis seeks to identify the set of initial perturbations that induce an unwanted outcome. In prior work, we defined sample-specific root causes of disease using exogenous error terms that predict a diagnosis in a structural equation model. We rigorously quantified predictivity using Shapley values. However, the associated algorithms for inferring root causes assume no latent confounding. We relax this assumption by permitting confounding among the predictors. We then introduce a corresponding procedure called Extract Errors with Latents (EEL) for recovering the error terms – possibly up to contamination by other error terms lying on certain paths – under the linear non-Gaussian acyclic model. EEL also identifies the smallest sets of dependent errors for fast computation of the Shapley values. The algorithm bypasses the hard problem of estimating the underlying causal graph in both cases. Experiments highlight the superior accuracy and robustness of EEL relative to its predecessors.

Keywords: causal inference, root cause, confounding, LiNGAM

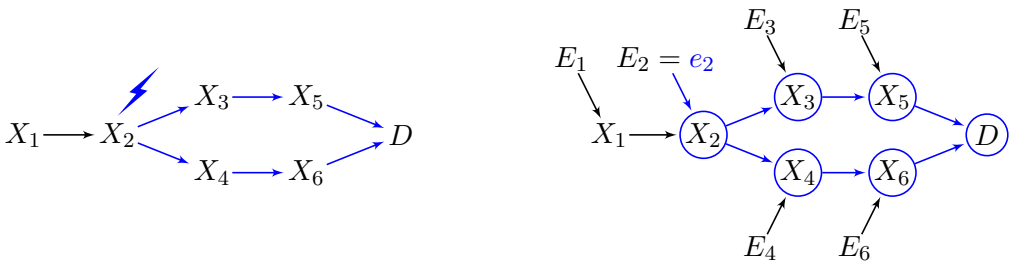
1. Introduction

Causal inference refers to the process of inferring causal relations from data. Most scientists identify causal relations by conducting randomized controlled trials (RCTs). RCTs nevertheless do not distinguish between a cause and a *root cause* of disease, or the initial perturbation to an otherwise healthy system that ultimately induces a diagnostic label. Identifying root causes is critical for (a) understanding disease mechanisms and (b) discovering drug targets that eliminate disease *at its onset* in a biological pathway.

Consider for example the directed graph in Figure 1 (a), where vertices in $\boldsymbol{X}$ represent random variables and directed edges their direct causal relations; we have $X_i \to X_j$ when X_i directly causes X_j. The lightning bolt in the figure denotes an exogenous perturbation of the root cause X_2, such as a virus, mutation or physical injury. This perturbation in turn affects many downstream variables, such as $\{X_3, X_4\}$, ultimately causing symptoms $\{X_5, X_6\}$ and physicians to label a patient with a diagnosis $D = 1$ indicating disease. The causes of D include $X_1, \dots, X_6$, but we only seek to identify the root cause X_2 that may lie arbitrarily far upstream from D in the general case.

Identifying root causes is further complicated by the existence of *complex disease*, where each patient may have multiple root causes, and root causes may differ between patients even within the same diagnostic category. The disease may also only affect certain tissues or cells in the body. We therefore more specifically seek to identify *sample-specific* root causes, where a sample may denote an arbitrary unit of granularity such as a patient, tissue or cell. Identifying sample-specific root causes has the potential to help experimentalists rapidly identify interventions that target the very beginnings of disease unique to each patient.

The above intuitive idea of a sample-specific root cause nevertheless lacks a rigorous mathematical definition. This in turn hinders the development of principled algorithms designed for their



(a) (b)

Figure 1: The lightning bolt in (a) denotes an exogenous perturbation of X_2 that eventually affects many downstream variables and causes a diagnosis D. In (b), we model the lightning bolt as an intervention of E_2 to the value e_2 that impacts the values of all of its descendants and ultimately D.

automated detection. As a result, we explicitly defined sample-specific root causes of disease as the error terms in a structural equation model that predict a diagnostic label in prior work (Strobl and Lasko, 2022a). We quantified predictivity using Shapley values. We also proposed methods to directly extract these error terms both in the linear and non-linear settings via regression residuals (Strobl and Lasko, 2022a,b). The methods do not require knowledge about the underlying graph and achieve sample efficiency by bypassing the hard problem of causal graph recovery (Chickering et al., 2004). These algorithms however rely on the unreasonable assumption that the dataset contain no unobserved confounders, which we relax in this paper by permitting confounding between the variables in $\boldsymbol{X}$.

We specifically make the following contributions in this paper:

- We introduce a strategy for identifying sample-specific root causes with confounding by extracting the error terms – possibly up to contamination by other error terms lying on specific paths.
- We propose an algorithm called Extract Errors with Latents (EEL) that recovers the above error terms and computes an undirected graph summarizing their statistical dependencies.
- We use the graph to efficiently compute Shapley values of the error terms by averaging over small neighborhoods of dependence.

Experiments in Section 7 highlight the superiority of EEL relative to existing methods in the presence of confounding.

2. Structural Equation Models

We can formalize causal inference under the framework of structural equation models (SEMs). An SEM over a set of p random variables $\boldsymbol{X}$ refers to a set of deterministic equations in the following form:

$$X_i = f_i(\text{Pa}(X_i), E_i), \quad \forall X_i \in \boldsymbol{X}.$$

where $\boldsymbol{E}$ denotes a random vector of p mutually independent error terms, and $\text{Pa}(X_i) \subseteq \boldsymbol{X}$ the *parents*, or direct causes, of X_i. We can equivalently set the equality sign in the above equation to algorithmic assignment $\leftarrow$ in order to emphasize that interventions on $\text{Pa}(X_i)$ induce changes in the marginal probability distribution of X_i.

We can associate an SEM with a *directed graph* $\mathbb{G}$ containing at most one directed edge between any two variables in $\boldsymbol{X}$. We have $X_i \to X_j$, when there exists a direct causal relation from $X_i \in \text{Pa}(X_j)$ to X_j. We always have $E_i \to X_i$ in $\mathbb{G}$ but only draw the vertices in $\boldsymbol{E}$ and their outgoing edges when informative. We use the notation $\text{Pa}_{\mathbb{G}}(X_j)$ to emphasize the underlying graph $\mathbb{G}$. A *directed path* is sequence of adjacent directed edges. X_i is an *ancestor* of X_j, and X_j a *descendant* of X_i, if a directed path exists from X_i to X_j or $X_i = X_j$. A *directed acyclic graph* (DAG) corresponds to a directed graph without *cycles*, where X_i is an ancestor of X_j and $X_j \to X_i$ with $X_i \neq X_j$. A vertex X_j is a *collider* on a path if we have $X_i \to X_j \leftarrow X_k$ on the path. Two vertices X_i and X_j are *d-connected* given $\boldsymbol{W} \setminus \{X_i, X_j\}$ when there exists a path between X_i and X_j such that every collider has a descendant in $\boldsymbol{W}$ and no non-collider is in $\boldsymbol{W}$. The two vertices are likewise *d-separated* when they are not d-connected.

An SEM with an associated DAG $\mathbb{G}$ can admit a density that factorizes according to the graph:

$$p(\boldsymbol{X}) = \prod_{i=1}^{p} p(X_i | \text{Pa}_{\mathbb{G}}(X_i)).$$

The above factorization implies that, if X_i and X_j are d-separated given $\boldsymbol{W}$ in $\mathbb{G}$, then the two vertices are also conditionally independent given $\boldsymbol{W}$, which we denote by $X_i \perp\!\!\!\perp X_j | \boldsymbol{W}$ for shorthand (Lauritzen et al., 1990). *D-separation faithfulness* refers to the converse: if $X_i \perp\!\!\!\perp X_j | \boldsymbol{W}$, then X_i and X_j are d-separated given $\boldsymbol{W}$.

In this paper, we focus on linear SEMs with an associated DAG:

$$X_i = \sum_{j=1}^{p} X_j \beta_{ji} + E_i, \quad \forall X_i \in \boldsymbol{X}, \tag{1}$$

comprised of a set of linear equations with coefficient matrix β where $\beta_{ji} \neq 0$ if and only if $X_j \in \text{Pa}_{\mathbb{G}}(X_i)$. We assume $\mathbb{E}(\boldsymbol{X}) = 0$ without loss of generality. The equations more specifically follow a *Linear Non-Gaussian Acyclic Model* (LiNGAM) when each error term is continuous non-Gaussian (Shimizu et al., 2006).

Most existing methods also assume that we observe all of the variables in $\boldsymbol{X}$. We relax this assumption by dividing $\boldsymbol{X}$ into a set of q observed variables $\boldsymbol{O}$ and a set of m latent – or unobserved – common causes $\boldsymbol{L}$. We can *always* write the following:

$$O_i = \sum_{j=1}^{q} O_j \beta_{ji} + \sum_{k=1}^{m} L_k \gamma_{ki} + E_i, \quad \forall O_i \in \boldsymbol{O}. \tag{2}$$

Each L_k must have at least two children, or else we can accommodate L_k into E_i. Without loss of generality, we may also assume that $\boldsymbol{T} = \boldsymbol{L} \cup \boldsymbol{E}$ denotes a set of mutually independent random variables with no parents (Hoyer et al., 2008). We refer to Equation (2) as the *canonical form*.

We can write Equation (2) in matrix notation:

$$\boldsymbol{O} = \boldsymbol{O}\beta + \boldsymbol{L}\gamma + \boldsymbol{E}.$$

Re-arranging terms yields:

$$\boldsymbol{O} = (\boldsymbol{L}\gamma + \boldsymbol{E})(I - \beta)^{-1} = \boldsymbol{E}\lambda + \boldsymbol{L}\gamma\lambda = \boldsymbol{T}\theta$$

where $\lambda = (I - \beta)^{-1}$ is assumed to exist and $\theta = [\lambda; \gamma\lambda]$. Notice that $\boldsymbol{T}$ is now ordered such that $T_j = E_j$ if $j \leq q$. The entry θ_{ji} quantifies the *total effect* of the latent variable or error term T_j on O_i.

3. Sample-Specific Root Causes

We consider LiNGAM over $\boldsymbol{X}$ and introduce an additional label D representing a diagnosis; we have $D = 1$ for samples deemed to have a disease, and $D = 0$ for healthy controls. We then assume a DAG over $\boldsymbol{X} \cup D$ such that D is a *terminal vertex*, or a vertex without descendants, and linked to $\boldsymbol{X}$ via a logistic function:

Assumption 1. *D is a terminal vertex such that* $\mathbb{P}(D|\boldsymbol{X}) = \text{logistic}(\boldsymbol{X}\beta_{\cdot D} + \alpha)$.

This is a reasonable assumption because a scientist who seeks to identify the causes of D will likely use datasets containing measurements of the non-descendants of the diagnosis, such as gene expression levels, clinical laboratory values or imaging. The logistic link also provides a natural extension of LiNGAM to handle a *noisy* binary variable.

We model a sample-specific perturbation first affecting the root cause $X_i \in \boldsymbol{X}$ as a change in the value of its error term E_i. We may write the following for any healthy control:

$$X_i = \sum_{j=1}^{p} X_j \beta_{ji} + \widetilde{e}_i, \tag{3}$$

where we have set the value of E_i in Equation (1) to $\widetilde{e}_i$. Suppose however that an exogenous perturbation – such as a virus, mutation or physical injury – changes the value of E_i from $\widetilde{e}_i$ to e_i. This intervention in turn effects the value of X_i and all of its downstream effects, ultimately changing the probability of developing disease $D = 1$ (Figure 1 (b)).

We can quantify the change in the probability of developing disease using logistic regression. We in particular consider:

$$f(\boldsymbol{E}) = \ln\left[\frac{\mathbb{P}(D = 1|\boldsymbol{E})}{\mathbb{P}(D = 0|\boldsymbol{E})}\right] = \boldsymbol{E}\theta_{\cdot D} + \alpha,$$

where the last equality follows by Assumption 1. Let $v(\boldsymbol{W})$ denote the conditional expectation of the logistic regression model $\mathbb{E}(f(\boldsymbol{E})|\boldsymbol{W})$, initially where $\boldsymbol{W} = \emptyset$. We then measure the change in probability when intervening on E_i via the following difference:

$$\gamma_{E_i \boldsymbol{W}} = \underbrace{v(E_i, \boldsymbol{W})}_{(a)} - \underbrace{v(\boldsymbol{W})}_{(b)} \tag{4}$$

We have $\gamma_{e_i \boldsymbol{W}} > 0$ when $E_i = e_i$ increases the probability that $D = 1$ because (a) is larger than (b).

Expression (4) unfortunately only quantifies the effect of E_i on D *in isolation*. We however also want to quantify the joint effect of E_i in conjunction with the other error terms in $\boldsymbol{E} \setminus E_i$ when $\boldsymbol{W} \neq \emptyset$. We therefore average over all possible combinations of the errors as follows:

$$S_i = \frac{1}{p} \underbrace{\sum_{\boldsymbol{W} \subseteq (\boldsymbol{E} \setminus E_i)} \frac{1}{\binom{p-1}{|\boldsymbol{W}|}}}_{\text{Average over all possible combinations of } \boldsymbol{E} \setminus E_i} \gamma_{E_i \boldsymbol{W}}. \tag{5}$$

The quantity corresponds precisely to the well-known Shapley value which, as the reader may recall, is the *only* value satisfying the linearity, efficiency, symmetry and null player properties (see e.g., (Lundberg and Lee, 2017; Štrumbelj and Kononenko, 2014)). Note that the Shapley "value" is actually a random variable, but we call it a value based on tradition.

The following result holds:

Proposition 1. *Under LINGAM over $\boldsymbol{X}$ and Assumption 1, the Shapley value S_i corresponds to the sample-specific total effect of E_i on D: $S_i = E_i \theta_{iD}$.*[1]

The proof follows directly from Corollary 1 of (Lundberg and Lee, 2017). This justifies the following definition of a sample-specific root cause:

Definition 1. *$X_i \in \text{Anc}_{\mathbb{G}}(D)$ is a sample-specific root cause of disease if $S_i = s_i > 0$.*

In other words, a sample-specific root cause of disease is a variable associated with an error term that increases the probability that $D = 1$ as quantified by the Shapley value $S_i > 0$. We do not consider $S_i \leq 0$ because E_i decreases the probability that $D = 1$ (or likewise increases the probability that $D = 0$) when $S_i < 0$. Similarly, E_i has no effect on increasing or decreasing the probability that $D = 1$ when $S_i = 0$. We have thus arrived at a concise definition of a sample-specific root cause as a variable associated with a positive Shapley value of its error.

4. Inducing Paths & Terms

The definition of a sample-specific root cause implies that we must develop methods that can accurately extract the error terms in order to compute the Shapley value. We however cannot identify the error terms $\boldsymbol{E}$ exactly when confounding exists. Consider for example the graph shown in Figure 2, where we cannot partial out L_1 from O_1 and O_2 because L_1 is unobserved.

We can however identify the error terms up to connection by directed inducing paths:

Definition 2. *A directed inducing path to O_i is a path between O_i and $T_j \in \boldsymbol{T}$ (possibly $i = j$) where every collider is an ancestor of O_i and every non-collider is in $\boldsymbol{L}$.*

All colliders are directed to O_i. We only consider directed inducing paths starting from the *error terms* or *latent variables* to O_i. We provide an example in Figure 2. Any error term incident on or latent variable lying on a directed inducing path to O_i also has a directed inducing path to O_i. Only E_i lies on a directed inducing path to O_i in the unconfounded setting, but more error terms may lie on the path when confounding exists. Finally, the above definition corresponds to the directed

1. If $D = \boldsymbol{X}\beta_{\cdot D} + E_D$ is terminal and continuous, then we arrive at the same result when $\gamma_{E_i \boldsymbol{W}} = \mathbb{E}(D|E_i, \boldsymbol{W}) - \mathbb{E}(D|\boldsymbol{W})$. We focus on a binary target because this is the most common situation encountered by far.

analogue of an (undirected) *inducing path* utilized in constraint-based search with latent variables, where every collider is an ancestor of either endpoint O_i or T_j (or both) (Spirtes et al., 2000).

The following result elucidates the limits of error term recovery when assessing statistical independence with regression residuals. Consider the ideal scenario where we have access to $F_j = E_j + \sum_{L_k \in \text{Pa}_{\mathbb{G}}(O_j) \cap \boldsymbol{L}} L_k \gamma_{kj}$ for each $O_j \in \boldsymbol{O}$, which we collect into the set $\boldsymbol{F}$. The notation $R_{O_i \boldsymbol{W}}$ denotes the residuals of O_i when linearly regressed on $\boldsymbol{W} \subseteq \boldsymbol{F} \setminus F_i$.

Lemma 1. *Under LiNGAM and d-separation faithfulness, if some entry in $\boldsymbol{W} \subseteq \boldsymbol{F} \setminus F_i$ corresponds to an observed vertex lying on a directed inducing path to O_i, then $R_{O_i \boldsymbol{W}} \not\perp\!\!\!\perp F_j$ for some $F_j \in \boldsymbol{W}$.*

We delegate proofs to Appendix 9.4. The latent common causes lying on a directed inducing path to O_i thus ensure that we cannot partial out the error terms incident on the path from O_i in general, even if we identified all entries in $\boldsymbol{F} \setminus F_i$.

We instead focus on identifying the error terms *up to* connection by a directed inducing path. Specifically, let $\boldsymbol{C}_i \subseteq \boldsymbol{T}$ denote the set of error terms and latent variables lying on any directed inducing path to O_i. We consider:

$$E_i^* = \boldsymbol{C}_i \theta_{\boldsymbol{C}_i i}, \tag{6}$$

for each $O_i \in \boldsymbol{O}$. For example, $E_1^* = L_1 \gamma_{11} + E_1$ and $E_2^* = E_1 \beta_{12} + L_1(\gamma_{12} + \gamma_{11}\beta_{12}) + E_2$ in Figure 2. This generalizes the unconfounded setting where $E_i^* = E_i \theta_{\boldsymbol{C}_i i} = E_i$ because we have $\boldsymbol{C}_i = E_i$ and $\theta_{E_i i} = 1$ in this case. We call the set $\boldsymbol{E}^*$ the *inducing terms*. The variable E_i^* represents a corrupted estimate of the original error term E_i in the sense that E_i^* is a linear combination of E_i and a small set of error terms and latent variables ancestral to O_i.

5. Extracting Inducing Terms

We now design an algorithm that identifies the inducing terms from the joint distribution of $\boldsymbol{O}$, without access to the ground truth DAG. We specifically build upon the DirectLiNGAM and EE algorithms explicated in Appendices 9.1 and 9.2 to handle cases where $\boldsymbol{L} \neq \emptyset$.

We identify the inducing term of O_j by regressing out as many of its ancestors in $\boldsymbol{T} \setminus E_j$ as possible. Let $\boldsymbol{W}$ denote a set of arbitrary linear combinations of error terms and latent variables in a minimal set $\boldsymbol{S} \subseteq \boldsymbol{T}$. We have:

Proposition 2. *Under LiNGAM, W_i is independent of the residuals $R_{O_j \boldsymbol{W}}$ for all $W_i \in \boldsymbol{W}$ if and only if O_j can be written as a linear function of $\boldsymbol{W}$ plus a linear function of $\boldsymbol{T} \setminus \boldsymbol{S}$. Thus, the residuals are a linear function of $\boldsymbol{T} \setminus \boldsymbol{S}$.*

The above proposition suggests that we should design an algorithm that iteratively replaces O_j with $R_{O_j \boldsymbol{W}}$ because $R_{O_j \boldsymbol{W}}$ depends on a fewer number of members in $\boldsymbol{T}$. We can also partial out ancestors by performing a series of *univariate and multivariate* regressions, progressively increasing the conditioning set size of $\boldsymbol{W}$. We partial out $\boldsymbol{W}$ from O_j once we find a large enough $\boldsymbol{W}$ such that $R_{O_j \boldsymbol{W}} \perp\!\!\!\perp O_i$ for all $O_i \in \boldsymbol{W}$.

Extract Errors with Latents (EEL) summarized in Algorithm 1 repeats the above procedure for each $O_j \in \boldsymbol{O}$. EEL proceeds just like EE but with additional steps highlighted in gray for increasing the conditioning set size. The algorithm first instantiates a complete undirected graph $\mathcal{G}$ over $\boldsymbol{O}$ in Line 1. The graph represents the statistical dependencies between the vertices in each iteration of the algorithm; $\mathcal{G}$ is not the DAG $\mathbb{G}$. The notation $\text{Adj}_{\mathcal{G}}(O_j)$ refers the observed variables adjacent

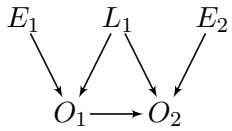


Figure 2: Example where we cannot recover E_1 and E_2 exactly. Also, E_1, L_1, E_2 and O_1 each lie on a directed inducing path to O_2.

to O_j in $\mathcal{G}$. The variable l denotes the size of the set $\boldsymbol{W}$. EEL gradually increases l until it finds a set $\boldsymbol{W} \subseteq \text{Adj}_{\mathcal{G}}(O_j)$ where $R_{O_j \boldsymbol{W}} \perp\!\!\!\perp O_i$ for all $O_i \in \boldsymbol{W}$ in Line 9. EEL then partials out $\boldsymbol{W}$ from O_j and removes the corresponding adjacencies from $\mathcal{G}$ in Lines 14-15. The algorithm finally resets the size of $\boldsymbol{W}$ in Line 16 by assigning $l \leftarrow 0$. This ensures that EEL proceeds with a fresh search after partialing out $\boldsymbol{W}$ from O_j.

EEL recovers the inducing terms in the oracle setting. We first require a new definition:

Definition 3. *A confounding path of O_i is a path between $T_j \in \boldsymbol{T}$ and O_k (possibly $i = j = k$) where every collider is an ancestor of O_i and every non-collider is in $\boldsymbol{L}$.*

A directed inducing path to O_i must end at O_i, whereas a confounding path of O_i may not end at O_i. We now formally have:

Theorem 1. *Under LiNGAM and d-separation faithfulness, if at most d observed variables lie on a confounding path of any member of $\boldsymbol{O}$, then EEL with $l \leq d$ recovers the inducing terms $\boldsymbol{E}^*$.*

EEL thus partials out variables at each iteration and then discovers all inducing terms by only searching over subsets of variables adjacent in $\mathcal{G}$. In contrast, algorithms that discover causal structure in the confounded setting, such as the FCI, do not partial out variables but must search over an often much larger set of variables that lie on sequences of (undirected) inducing paths (Spirtes et al., 2000; Zhang, 2008).

We must of course perform the necessary regressions and independence tests with n samples in practice. We assume that the independence test requires $O(n\log(n))$ time (Even-Zohar, 2020; Even-Zohar and Leng, 2021) and consider the standard $O(n^2 d + d^3)$ complexity of linear regression. The outer and innermost loops of EEL iterate over at most $\sum_{k=1}^{d} \binom{q-1}{k}$ combinations with an independence oracle, and the second loop over at most q variables. EEL therefore depends polynomially on the number of variables q because $O(q \sum_{k=1}^{d} \binom{q-1}{k}) = O(q^{d+1})$. We conclude that EEL theoretically takes $O(q^{d+1} n^2 d)$ time in the oracle setting. However, the independence tests take longer than linear regression in practice due to the existence of highly optimized linear algebra libraries, so EEL runs in $O(q^{d+1} n \log(n))$ time for realistic sample sizes.

6. Causal & Predictive Contributions

We want to quantify the sample-specific total effect of E_i on D, but EEL can only recover the inducing terms $\boldsymbol{E}^*$ when confounding exists. The variable E_i^* is a linear combination of E_i and some of the error terms and latent variables that are ancestors of O_i per Equation (6). The sample-specific total effect of E_i^* can therefore lie far from that of E_i. Even worse, the abstract quantity E_i^* may not correspond to any real-world entity that we can manipulate in practice.

Algorithm 1 Extract Errors with Latents (EEL)

Input: $\boldsymbol{O}$
Output: $\boldsymbol{E}^*, \mathcal{G}$

1: $\mathcal{G} \leftarrow$ complete undirected graph over $\boldsymbol{O}$
2: $l \leftarrow 0$
3: **repeat**
4: $\quad l = l + 1$
5: $\quad \boldsymbol{Y} \leftarrow \emptyset$
6: $\quad$ **for all** O_j s.t. $|\text{Adj}_{\mathcal{G}}(O_j)| \geq l$ **do**
7: $\quad\quad$ **repeat**
8: $\quad\quad\quad$ Choose a new $\boldsymbol{W} \subseteq \text{Adj}_{\mathcal{G}}(O_j)$ s.t. $|\boldsymbol{W}| = l$
9: $\quad\quad\quad$ **if** $R_{O_j \boldsymbol{W}} \perp\!\!\!\perp O_i, \forall O_i \in \boldsymbol{W}$ **then**
10: $\quad\quad\quad\quad \boldsymbol{Y} \leftarrow \boldsymbol{W}$; **break**
11: $\quad\quad\quad$ **end if**
12: $\quad\quad$ **until** all $\boldsymbol{W} \subseteq \text{Adj}_{\mathcal{G}}(O_j)$ with $|\boldsymbol{W}| = l$ have been considered
13: $\quad\quad$ **if** $\boldsymbol{Y} \neq \emptyset$ **then**
14: $\quad\quad\quad$ Partial out $\boldsymbol{Y}$ from O_j
15: $\quad\quad\quad$ Remove $\boldsymbol{Y}$ from $\text{Adj}_{\mathcal{G}}(O_j)$
16: $\quad\quad\quad$ $l \leftarrow 0$; **break**
17: $\quad\quad$ **end if**
18: $\quad$ **end for**
19: **until** all vertices satisfy $|\text{Adj}_{\mathcal{G}}(O_j)| < l$
20: $\boldsymbol{E}^* \leftarrow \boldsymbol{O}$

We instead seek a unified variable importance measure that (1) identifies the sample-specific total effects of the error terms *when possible* and (2) otherwise corresponds to a measure of predictivity rather than causality. The output of EEL must also clearly indicate when (1) or (2) holds.

We in particular consider the following Shapley value as a natural generalization of Equation (5), where we have replaced $\boldsymbol{E}$ with $\boldsymbol{E}^*$:

$$S_i^* = \frac{1}{q} \sum_{\boldsymbol{W} \subseteq (\boldsymbol{E}^* \setminus E_i^*)} \frac{1}{\binom{q-1}{|\boldsymbol{W}|}} \gamma_{E_i^* \boldsymbol{W}}. \tag{7}$$

We can gain a deeper understanding of S_i^* using the undirected graph $\mathcal{G}$ provided by EEL.

EEL instantiates the graph $\mathcal{G}$ over $\boldsymbol{O}$ in Line 1, but the graph summarizes the dependence relations between the inducing terms $\boldsymbol{E}^*$ when EEL terminates. We can construct the final form of $\mathcal{G}$ using the sets $\boldsymbol{C}_i$ with the following procedure:

1. Instantiate an empty graph $\mathcal{G}$ over $\boldsymbol{E}^*$;

2. Draw an undirected edge between E_i^* and E_j^* if and only if $\boldsymbol{C}_i \cap \boldsymbol{C}_j \neq \emptyset$ for all pairs $\{O_i, O_j\}$.

By construction:

Proposition 3. *Two inducing terms are adjacent in $\mathcal{G}$ if and only if they involve a common error term or latent variable.*

The graph $\mathcal{G}$ therefore implies only small groups of dependent inducing terms. Let $\boldsymbol{B}_i^*$ denote the inducing terms with corresponding vertices adjacent to E_i^* in $\mathcal{G}$. Consider:

$$\psi_k = \frac{q}{\binom{|\boldsymbol{B}_i^*|-1}{k}|\boldsymbol{B}_i^*|}.$$

Let δ denote the vector of coefficients obtained by logistically regressing D on $\boldsymbol{E}^*$ so that $E_i^*\delta_i = (\boldsymbol{C}_i\theta_{\boldsymbol{C}_i i})\delta_i$. Then:

Theorem 2. *The following relation holds under a linear model:* $\gamma_{E_i^*\boldsymbol{W}} = E_i^*\delta_i - \mathbb{E}(E_i^*|\boldsymbol{V})\delta_i$, *where* $\boldsymbol{W} \subseteq (\boldsymbol{E}^* \setminus E_i^*)$ *and* $\boldsymbol{V} = (\boldsymbol{B}_i^* \setminus E_i^*) \cap \boldsymbol{W}$, *so that:*

$$S_i^* = E_i^*\delta_i - \frac{\delta_i}{q}\sum_{\boldsymbol{V}\subseteq(\boldsymbol{B}_i^*\setminus E_i^*)} \psi_{|\boldsymbol{V}|}\mathbb{E}(E_i^*|\boldsymbol{V}). \tag{8}$$

In other words, $S_i^* = E_i^*\delta_i = E_i\theta_{iD}$ when $\mathcal{G}$ has no adjacencies because $E_i^* = E_i$ and $\delta_i = \theta_{iD}$. S_i^* thus corresponds to S_i when O_i has no adjacencies in $\mathcal{G}$ and to a measure of predictivity when O_i has adjacencies in $\mathcal{G}$. Furthermore, $\boldsymbol{S}^*$ is a unified measure still uniquely satisfying the linearity, efficiency, symmetry and null player properties.

The above result also implies that we can compute the Shapley value using subsets of $\boldsymbol{B}_i^* \setminus E_i^*$ rather than subsets of the much larger set $\boldsymbol{E}^* \setminus E_i^*$ in Equation (7). We estimate the expectations in Equation (8) quickly even when q is large, so long as $|\boldsymbol{B}_i^*|$ is small (e.g., $|\boldsymbol{B}_i^*| \leq 10$).

If $|\boldsymbol{B}_i^*|$ is also large, then we estimate S_i^* by Monte Carlo, where we sample the error terms with probabilities obeying the Shapley weights. We first sample K with probability $\mathbb{P}(K) = 1/|\boldsymbol{B}_i^*|$. We then sample a set $\boldsymbol{V}$ by choosing a random subset of $\boldsymbol{B}_i^* \setminus E_i^*$ with size $K = k$; in other words, we sample $\boldsymbol{V}$ with probability $1/\binom{|\boldsymbol{B}_i^*|-1}{k}$ uniformly. We do not need to resort to Monte Carlo for the vast majority of cases because $\mathcal{G}$ is sparse in practice.

7. Experiments

We compared EEL against the following algorithms representing the state of the art:

2. Root Causal Inference (RCI): extracts error terms from the top-down by regressing on root vertices using a localized version of DirectLiNGAM and then computes Shapley values (Strobl and Lasko, 2022a).

3. Generalized Root Causal Inference (GRCI): extracts error terms from the bottom-up by regressing on parents of sink vertices and then computes Shapley values (Strobl and Lasko, 2022b).

4. Independent Component Analysis (ICA): performs ICA to extract the independent error terms and ranks variables according to a random forest permutation measure (Lasko and Mesa, 2019).

5. Root Causal Analysis of Outliers (RCAO): defines root causes according to an outlier score and computes Shapley values using the outlier scores and an estimated DAG (Budhathoki et al., 2022b).

6. Model Substitution (MS): re-samples the underlying DAG after substituting causal conditionals in an estimated DAG (Budhathoki et al., 2021).

No existing method accounts for latent variables besides EEL. See Appendix 9.3 for a comprehensive review of related work. We fixed the Type I error rate of EEL to 0.05 and estimated the Shapley values using MARS regression (Friedman, 1991). We further standardized the data to prevent gaming of the marginal variances (Reisach et al., 2021).

We do not have access to the ground truth Shapley values due to the unknown conditional expectations in Equation (8). We therefore approximated the conditional expectations to high accuracy by training a committee of ten Linear Model Trees (Quinlan et al., 1992) – a different model class than MARS – on a sample of one hundred thousand ground truth inducing terms. We otherwise used the ground truth inducing terms, total causal effects and dependence graph $\mathcal{G}$ to compute Equation (8) for each $O_i \in \boldsymbol{O}$.

Reproducibility. All code and data for reproducing experimental results are available at https://github.com/ericstrobl/EEL.

7.1. Synthetic Data

7.1.1. Data Generation

We generated linear structural equation models using the following procedure. We first created a DAG with $p = 15$ variables and an expected neighborhood size of two by creating a random adjacency matrix with independent realizations from a Bernoulli$(2/(p-1))$ distribution in the upper triangle portion of the matrix. We then permuted the variable ordering. We chose D uniformly from the set of vertices without children and at least one parent. We selected 0, 10 or 20% of the vertices as unobserved confounders provided each had at least two observed children not including D and no parents. We then replaced the ones in the matrix by independent realizations of a uniform distribution on $[-1, -0.25] \cup [0.25, 1]$. We chose the distribution of each error term by uniformly sampling from the following set of possibilities: the t-distribution with five degrees of freedom, the chi-square distribution with three degrees of freedom, and the uniform distribution on -1 to 1. We finally drew instantiations of D according to a Bernoulli random variable with probabilities obeying a logistic function per Assumption 1. We repeated the above procedure 120 times for sample sizes of one, ten and one hundred thousand and latent variables of 0, 10 and 20%. We therefore generated a total of $120 \times 3 \times 3 = 1080$ independent datasets.

7.1.2. Evaluation Criteria

The output of the five algorithms differ, but we can convert the output of each algorithm to a ranked list of variables. The top ranked variables should correspond to the true root causes with the largest Shapley values. We therefore first compared the algorithms using rank biased overlap (RBO), a well-established measure designed to compare two ranked lists of potentially differing lengths (Webber et al., 2010):

$$\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{r_k}\widetilde{s}_i^k|\widehat{\mathcal{R}}_{1:i}^k \cap \mathcal{R}_{1:i}^k|/i,$$

where s_i^k denotes the true Shapley value of the variable O_i for sample k, $\widetilde{s}_i^k = \frac{s_i^k}{\sum_{i=1}^{r_k} s_i^k}$ the Shapley values normalized to sum to one, and r_k the total number of root causes for sample k. The notation $\mathcal{R}_{1:i}^k$ refers to the first i variables in the ranking $\mathcal{R}^k$ for sample k. RBO increases monotonically with depth and weighs top ranks more heavily. The metric equals one when the top ranks coincide

l	n	EEL	RCI	GRCI	ICA	RCAO	MS
0%	1,000	0.850	**0.918**	0.883	0.713	0.652	0.662
	10,000	0.962	**0.975**	**0.971**	0.779	0.669	0.673
	100,000	**0.980**	**0.993**	**0.992**	0.796	0.669	0.670
10%	1,000	0.806	**0.859**	0.826	0.678	0.574	0.579
	10,000	**0.931**	0.904	0.890	0.738	0.595	0.596
	100,000	**0.961**	0.910	0.899	0.756	0.595	0.592
20%	1,000	**0.781**	**0.784**	0.763	0.624	0.479	0.500
	10,000	**0.892**	0.806	0.796	0.677	0.507	0.519
	100,000	**0.938**	0.811	0.800	0.695	0.508	0.517

Table 1: RBO results with the synthetic data. EEL achieved the highest mean RBO values with larger sample sizes as highlighted in gray.

exactly with the true sample-specific root causes sorted in decreasing order by Shapley values, and zero when no overlap exists. Higher is therefore better.

We also compared the algorithms using the traditional mean squared error (MSE) to the true Shapley values:

$$\frac{1}{nq}\sum_{k=1}^{n}\sum_{i=1}^{q}(\widehat{s}_i^k - s_i^k)^2.$$

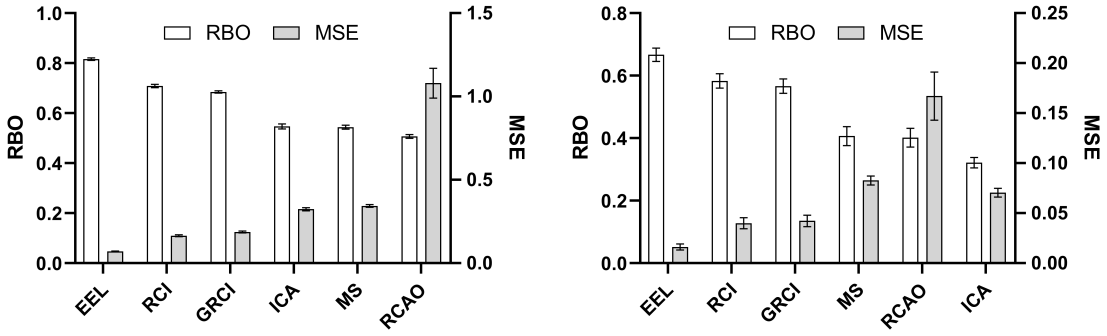
where lower is better. If an algorithm only outputs Shapley values for a subset of variables, then we set the estimated Shapley values to zero for the excluded subset.

7.1.3. RESULTS

We summarize the RBO results for the synthetic data in Table 2. Bolded values denote the best performance in each row according to one-sided paired t-tests at a Bonferroni corrected threshold of 0.05/6, since we compared a total of six algorithms. We present tables summarizing the MSE and timing results in Appendix 9.5. RBO and MSE results were similar.

EEL achieved the best performance in terms of both RBO and MSE with confounding once sample sizes reached ten thousand. The margin continued to widen with increasing sample size and confounding degree. EEL outperformed the second best algorithm by a 15.6% margin with $n = 100,000$ and $l = 20\%$. EEL therefore requires a sizable number of samples in order to achieve state of the art performance.

EEL underperformed both RCI and GRCI without confounding. The margin however was small, and we cannot expect EEL to outperform algorithms explicitly designed for the unconfounded case. For example, RCI exploits certain local properties in unconfounded LiNGAM to significantly reduce the search space. We conclude that EEL remains competitive when no latent common causes exist.



(a) (b)

Figure 3: RBO and MSE results for the (a) diabetes and (b) flow cytometry datasets. Error bars denote 95% confidence intervals. EEL again achieved the highest RBO and lowest MSE in both datasets.

7.2. Real Data

7.2.1. DIABETES

We ran the algorithms on a real clinical dataset to identify *patient*-specific root causes of diabetes. The dataset contains measurements of 8 variables related to the metabolic system among 768 patients of Pima Indian ancestry (Smith et al., 1988).[2] Diabetes is a well-studied disease, so we asked a physician to generate the ground truth causal graph shown in the Appendix. We derived the values of the error terms via linear regression on the parents. We chose one to two latent variables uniformly at random from age and the diabetes pedigree function. The target is a binary diagnostic label of diabetes.

We summarize the results as averaged over 200 bootstrapped datasets in Figure 3 (a) with algorithms sorted in decreasing order according to mean RBO value. EEL outperformed its nearest competitor by a 10.7 point RBO margin. EEL also achieved a 57.8% reduction in the MSE. Both results were significant at a Bonferroni corrected threshold of 0.05/6 by paired t-tests. The algorithm completed in 9.2 seconds on average. We conclude that EEL achieves the best performance in this dataset.

7.2.2. FLOW CYTOMETRY

We next ran the algorithms on a real flow cytometry dataset to identify *cell*-specific root causes. The dataset from (Sachs et al., 2005) contains measurements of 11 phosphoproteins and phospholipids from 7466 primary human immune system cells across 9 experimental conditions.[3] We log-transformed the data and standardized the samples in each experimental condition as recommended in (Ramsey and Andrews, 2018). We again derived the values of the error terms via linear regression on parents using the ground truth causal graph. We chose one to three latent variables uniformly at random from the options PKA, PKC and PIP3. We passed the mean of one to three observed

2. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
3. https://arxiv.org/src/1805.03108v1/anc/data.txt

variables, also chosen uniformly at random, through a logistic function for the binary target. We finally repeated the above process 200 times on bootstrapped samples.

We summarize the results in Figure 3 (b). EEL outperformed all other algorithms by at least an 8.3 point margin according to RBO. EEL similarly achieved a 59.3% reduction of the MSE from its nearest competitor. The algorithm took 64.6 seconds on average. We conclude that both real dataset results mimic those seen with the synthetic data.

8. Conclusion

We presented a novel algorithm called EEL that recovers the error terms of a structural equation model up to directed inducing paths. EEL also returns a sparse graph $\mathcal{G}$ summarizing the statistical dependencies between the recovered terms. We used the graph to quickly compute Shapley values, a unified measure corresponding to the sample-specific total effect when the inducing term E_i^* corresponds to its associated error term E_i. Experiments demonstrated considerable improvements in accuracy relative to existing methods. We conclude that the combination of EEL and Shapley values offers a principled framework for performing sample-specific root causal inference with latent variables.

Acknowledgments

We thank the reviewers for their helpful comments that improved this paper.

References

Bjørn Andersen and Tom Fagerhaug. *Root cause analysis: simplified tools and techniques*. Quality Press, 2006.

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

Kailash Budhathoki, George Michailidis, and Dominik Janzing. Explaining the root causes of unit-level changes. *arXiv preprint arXiv:2206.12986*, 2022a.

Kailash Budhathoki, Lenon Minorics, Patrick Blöbaum, and Dominik Janzing. Causal structure-based root cause analysis of outliers. In *International Conference on Machine Learning*, pages 2357–2369. PMLR, 2022b.

Max Chickering, David Heckerman, and Chris Meek. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5:1287–1330, 2004.

Chaim Even-Zohar. independence: Fast rank tests. *arXiv preprint arXiv:2010.09712*, 2020.

Chaim Even-Zohar and Calvin Leng. Counting small permutation patterns. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2288–2302. SIAM, 2021.

Jerome H Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

Patrik O Hoyer, Shohei Shimizu, Antti J Kerminen, and Markus Palviainen. Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378, 2008.

Aapo Hyvärinen and Stephen M Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *Journal of Machine Learning Research*, 14(Jan):111–152, 2013.

Thomas A Lasko and Diego A Mesa. Computational phenotype discovery via probabilistic independence. *KDD Workshop on Applied Data Science for Healthcare*, 2019.

Steffen L Lauritzen, A Philip Dawid, Birgitte N Larsen, and H-G Leimer. Independence properties of directed markov fields. *Networks*, 20(5):491–505, 1990.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, 2017.

John R Quinlan et al. Learning with continuous classes. In *Fifth Australian Joint Conference on Artificial Intelligence*, volume 92, pages 343–348. World Scientific, 1992.

Joseph Ramsey and Bryan Andrews. Fask with interventional knowledge recovers edges from the sachs model. *arXiv preprint arXiv:1805.03108*, 2018.

Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated dag! causal discovery benchmarks may be easy to game. *Advances in Neural Information Processing Systems*, 34, 2021.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

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. *The Journal of Machine Learning Research*, 12:1225–1248, 2011.

Jack W Smith, James E Everhart, WC Dickson, William C Knowler, and Robert Scott Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association, 1988.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Eric V Strobl. A constraint-based algorithm for causal discovery with cycles, latent variables and selection bias. *International Journal of Data Science and Analytics*, 8(1):33–56, 2019.

Eric V. Strobl and Thomas A. Lasko. Identifying patient-specific root causes of disease. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '22, New York, NY, USA, 2022a. Association for Computing Machinery. ISBN 9781450393867.

Eric V. Strobl and Thomas A. Lasko. Identifying patient-specific root causes with the heteroscedastic noise model. *arXiv preprint arXiv:2205.13085*, 2022b.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014.

William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.

Albert W Wu, Angela KM Lipshutz, and Peter J Pronovost. Effectiveness and efficiency of root cause analysis in medicine. *Jama*, 299(6):685–687, 2008.

Doron Zeilberger. The method of creative telescoping. *J. Symb. Comput.*, 11(3):195–204, 1991.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

9. Appendix

9.1. DirectLiNGAM

The DirectLiNGAM (DL) algorithm is a well-known method for estimating the error terms assuming LiNGAM and no confounding where $\boldsymbol{X} = \boldsymbol{O}$ (Shimizu et al., 2011). We will build upon DL in the next section, so we require a deep understanding of the algorithm's inner workings. More accurate and much faster variants of DL exist (Strobl and Lasko, 2022a), but we present the simplest version here to emphasize general concepts rather than algorithmic details.

DL capitalizes on the following result:

Proposition 4. *(Shimizu et al., 2011) Under LiNGAM and no confounding, O_i is independent of the residuals $R_{O_jO_i}$ for all $O_j \in (\boldsymbol{O} \setminus O_i)$ if and only if $O_i = E_i$. Moreover, partialing out $O_i = E_i$ from $\boldsymbol{O} \setminus O_i$ generates another LiNGAM model.*

All error terms correspond to root vertices, or vertices without ancestors. The algorithm therefore extracts an error term in each iteration by performing a series of univariate regressions to identify a root vertex. Partialing out the root vertex then recovers another LiNGAM model with a new set of root vertices, so DL repeats the process until it recovers all error terms.

We summarize DL in more detail in Algorithm 2. The algorithm calls FindRoot in Line 3, which we in turn summarize in Algorithm 3. FindRoot regresses each variable $O_i \in \boldsymbol{O}$ onto each variable $O_j \in (\boldsymbol{O} \setminus O_i)$. The algorithm then determines whether the residuals $R_{O_jO_i}$ and O_i are independent in Line 5, then vice versa in Line 6, using the independence measure $\mathcal{I}_{ij}$; the non-negative measure equals zero if and only if independence holds (Hyvärinen and Smith, 2013). FindRoot finally identifies the variable O_i *most* independent of its residuals in Line 9, thus bypassing the need for formal hypothesis testing with a predetermined Type I error rate. DL then removes O_i from consideration in Line 4 and replaces each O_j by its residuals $R_{O_jO_i}$ in Line 5. The algorithm iterates through this process until all variables in $\boldsymbol{O}$ have been replaced by their error terms. DL therefore ultimately outputs $\boldsymbol{E}$ as desired.

Algorithm 2 DirectLiNGAM (DL)

Input: $\boldsymbol{O}$
Output: $\boldsymbol{E}$

1: $\boldsymbol{U} \leftarrow \boldsymbol{O}$
2: **repeat**
3: $\quad G \leftarrow \text{FindRoot}(\boldsymbol{O}, \boldsymbol{U})$
4: $\quad \boldsymbol{U} \leftarrow \boldsymbol{U} \setminus G$
5: $\quad (\boldsymbol{O} \setminus G) \leftarrow$ partial out G from $\boldsymbol{O} \setminus G$
6: **until** $\boldsymbol{U} = \emptyset$
7: $\boldsymbol{E} \leftarrow \boldsymbol{O}$

9.2. Integrating Hypothesis Testing

DL unfortunately carries two main shortcomings:

1. The algorithm finds the variable most independent of its residuals in Line 9 of FindRoot. This process eliminates the need for hypothesis testing but ultimately slows down the algorithm by requiring that it check all pairs of variables in $\boldsymbol{U}$ in each iteration.

Algorithm 3 FindRoot

Input: $\boldsymbol{O}, \boldsymbol{U}$
Output: root G

return $\boldsymbol{U}$ if $|\boldsymbol{U}| = 1$
$\boldsymbol{T} = \mathbf{0}_{|\boldsymbol{U}|}$
for $i \in [|\boldsymbol{U}| - 1]$ **do**
    for $j \in \{i + 1, \ldots, |\boldsymbol{U}|\}$ **do**
        $T_i = T_i + \mathcal{I}_{ij}$
        $T_j = T_j + \mathcal{I}_{ji}$
    end for
end for
$G \leftarrow \boldsymbol{U}[\arg\min_i T_i]$

Algorithm 4 ExtractErrors (EE)

Input: $\boldsymbol{O}$
Output: $\boldsymbol{E}$

1: $\mathcal{G} \leftarrow$ complete undirected graph over $\boldsymbol{O}$
2: **for all** O_j s.t. $|\text{Adj}_{\mathcal{G}}(O_j)| > 0$ **do**
3: $Y \leftarrow \emptyset$
4: **repeat**
5: Choose a new $O_i \in \text{Adj}_{\mathcal{G}}(O_j)$
6: **if** $R_{O_j O_i} \perp\!\!\!\perp O_i$ **then**
7: $Y \leftarrow O_i$; **break**
8: **end if**
9: **until** all vertices in $\text{Adj}_{\mathcal{G}}(O_j)$ have been considered
10: **if** $Y \neq \emptyset$ **then**
11: Partial out Y from O_j
12: Remove Y from $\text{Adj}_{\mathcal{G}}(O_j)$
13: **end if**
14: **end for**
15: $\boldsymbol{E} \leftarrow \boldsymbol{O}$

2. DL partials out the effect of each root vertex from *all* remaining vertices because Equation (1) implies a linear relation from root vertices to their non-ancestors. We cannot apply this strategy to the confounded setting because $\boldsymbol{L}$ contains some of the root vertices.

We rectify both of these issues with a new method called Extract Errors (EE) that also assumes no confounding. EE capitalizes on Lemma 2, a simpler but analogous result to Proposition 4.

We can always find a variable $O_i \in \boldsymbol{O}$ where $O_i \perp\!\!\!\perp R_{O_j \boldsymbol{W}}$ with $|\boldsymbol{W}| = 1$ in the unconfounded setting because we observe all root vertices (for instance, $\boldsymbol{W} = O_i$ when O_i is a root vertex). We therefore set $|\boldsymbol{W}| = 1$ to focus on *univariate* regressions. We consider a new algorithm called Extract Errors (EE) in Algorithm 4. The algorithm first instantiates a complete undirected graph $\mathcal{G}$ over $\boldsymbol{O}$ in Line 1. EE then uses hypothesis testing to identify the variable $O_i \in \text{Adj}_{\mathcal{G}}(O_j)$ independent of the residuals $R_{O_j O_i}$ in Line 6. Subsequently, EE partials out O_i from O_j in Line

11 and eliminates the corresponding adjacency from $\mathcal{G}$ in Line 12. The algorithm terminates once $\mathcal{G}$ contains no adjacencies – i.e., once EE partials out all ancestral relationships. The correctness of EE follows as a corollary of Theorem 1:

Corollary 1. *Under LiNGAM and d-separation faithfulness, if no confounding exists, then EEL with $l \leq 1$ recovers the error terms $\boldsymbol{E}$.*

Line 6 represents the key step of EE because it allows the algorithm to select the first O_i inducing residuals independent of *some* $O_j \in (\boldsymbol{O} \setminus O_i)$, as opposed to *all* $O_j \in (\boldsymbol{O} \setminus O_i)$ like in DL. EE can therefore quickly partial out a variable even if it is *not* a root vertex in Line 11. This property will come in handy when we introduce confounders because we may not observe a root vertex when confounders exist.

9.3. Related Work

EEL is closely related to several lines of work. Lasko and Mesa (2019) proposed a methodology of extracting the error terms of an SEM using ICA, although the authors did not connect the approach to causality. We previously proposed methods for identifying sample-specific root causes in the linear and non-linear settings (Strobl and Lasko, 2022a,b). All of these methods however assume no confounding, whereas EEL accounts for unobserved variables by recovering inducing terms.

Our work is more broadly related to a suite of root causal analysis methodologies that identify sample-specific root causes in industrial or healthcare applications (Andersen and Fagerhaug, 2006; Wu et al., 2008). However, these methods take a painstaking manual approach that either implicitly or explicitly generates the underlying causal graph. Most strategies also focus on identifying human errors in *man-made* systems with well-understood causal processes. We on the other hand focus on identifying biological errors in nature or, more generally, errors where the underlying causal process is largely unknown and difficult to understand.

Other computational approaches also exist for identifying root causes, but they again assume a known or estimated causal graph. For example, a recent paper introduced a method called Root Causal Analysis of Outliers (RCAO) that considers the root causes of outlier events (Budhathoki et al., 2022b). RCAO however assumes that the user can recover the error terms even in the confounded setting. The algorithm also redraws the values of the error terms and therefore identifies root causes at the population rather than at the sample-specific level. Finally, RCAO assumes that the label corresponds to an outlier event with a noiseless cut-off, even though the cut-off score for a diagnosis is noisy because it depends on the diagnostician in practice. The Model Substitution (MS) method proposed in (Budhathoki et al., 2021) makes similar assumptions. Another strategy allows a noisy cut-off score but requires paired rather than more widely available case control data (Budhathoki et al., 2022a). EEL instead (1) utilizes case control data, (2) discovers the error terms directly without recovering or accessing the underlying causal graph, (3) identifies root causes at the sample-specific level and (4) allows a noisy diagnostic label with the logistic link. EEL is therefore more suitable for the biomedical setting with complex disease.

9.4. Proofs

Lemma 2. *(Darmois-Skitovitch) Suppose we can represent two random variables O_1 and O_2 as linear combinations of the mutually independent terms in $\boldsymbol{T}$:*

$$O_1 = \sum_{i=1}^{p} T_i \theta_{i1} \text{ and } O_2 = \sum_{i=1}^{p} T_i \theta_{i2}.$$

If some T_j for which $\theta_{j1}\theta_{j2} \neq 0$ is non-Gaussian, then O_1 and O_2 are dependent.

Let $\boldsymbol{W}$ denote a set of arbitrary linear combinations of error terms and latent variables in a minimal set $\boldsymbol{S} \subseteq \boldsymbol{T}$ for the proposition below.

Proposition 2. *Under LiNGAM, W_i is independent of the residuals $R_{O_j \boldsymbol{W}}$ for all $W_i \in \boldsymbol{W}$ if and only if O_j can be written as a linear function of $\boldsymbol{W}$ plus a linear function of $\boldsymbol{T} \setminus \boldsymbol{S}$. Thus, the residuals are a linear function of $\boldsymbol{T} \setminus \boldsymbol{S}$.*

Proof. For the forward direction, if $W_i \perp\!\!\!\perp R_{O_j \boldsymbol{W}}$ for all $W_i \in \boldsymbol{W}$, then W_i and $R_{O_j \boldsymbol{W}}$ are linear combinations of non-overlapping subsets of $\boldsymbol{T}$ for all $W_i \in \boldsymbol{W}$ by Lemma 2 under LiNGAM. This implies that $R_{O_j \boldsymbol{W}}$ is a linear function of $\boldsymbol{T} \setminus \boldsymbol{S}$, so O_j must be a linear function of $\boldsymbol{W}$ plus a linear function of $\boldsymbol{T} \setminus \boldsymbol{S}$. For the backward direction, if O_j can be written as a linear function of $\boldsymbol{W}$ plus a linear function of $\boldsymbol{T} \setminus \boldsymbol{S}$, then $R_{O_j \boldsymbol{W}}$ is a linear function of $\boldsymbol{T} \setminus \boldsymbol{S}$ only under LiNGAM. Hence $W_i \perp\!\!\!\perp R_{O_j \boldsymbol{W}}$ for all $W_i \in \boldsymbol{W}$. □

Lemma 3. *(Strobl, 2019) Under d-separation faithfulness, there exists an inducing path between X_i and X_j if and only if $X_i \not\perp\!\!\!\perp X_j | \boldsymbol{W}$ for all $\boldsymbol{W} \subseteq \boldsymbol{O} \setminus \{X_i, X_j\}$.*

Lemma 1. *Under LiNGAM and d-separation faithfulness, if some entry in $\boldsymbol{W} \subseteq \boldsymbol{F} \setminus F_i$ corresponds to an observed vertex lying on a directed inducing path to O_i, then $R_{O_i \boldsymbol{W}} \not\perp\!\!\!\perp F_j$ for some $F_j \in \boldsymbol{W}$.*

Proof. Let $F_k \in \boldsymbol{W}$ denote an entry lying on a directed inducing path to O_i. The directed inducing path Π must contain at least one non-collider in $\boldsymbol{L}$, lest Π induce a cycle in the DAG. Let L_r denote one such non-collider that is also a latent parent of O_k. Let $\boldsymbol{A} \subseteq \boldsymbol{O}$ contain the observed ancestors of O_i also corresponding to entries in $\boldsymbol{W}$. For example, $\boldsymbol{A}$ includes O_k because O_k is an ancestor of O_i, and $F_k \in \boldsymbol{W}$. Note that $E_k + L_r \gamma_{rk}$ is an additive component of F_k and $E_k \theta_{E_k i} + L_r \theta_{L_r i} = E_k \theta_{E_k i} + L_r(\gamma_{r\boldsymbol{A}} \theta_{E_{\boldsymbol{A}} i} + \delta)$ is an additive component of O_i by d-separation faithfulness. Assume $\delta = 0$ so that $\theta_{L_r i} = \gamma_{L_r \boldsymbol{A}} \theta_{E_{\boldsymbol{A}} i}$. But then $L_r \perp\!\!\!\perp O_i | \boldsymbol{A}$ which contradicts the fact that $L_r \not\perp\!\!\!\perp O_i | \boldsymbol{A}$ by the existence of an inducing path between L_r and O_i according to Lemma 3 under d-separation faithfulness. We thus have $\delta \neq 0$. But then we cannot partial out all of the entries in $\boldsymbol{W}$ corresponding to $\boldsymbol{A}$ from O_i, so $R_{O_i \boldsymbol{W}} \not\perp\!\!\!\perp F_j$ for some $F_j \in \boldsymbol{W}$. □

Theorem 1. *Under LiNGAM and d-separation faithfulness, if at most d observed variables lie on a confounding path of any member of $\boldsymbol{O}$, then EEL with $l \leq d$ recovers the inducing terms $\boldsymbol{E}^*$.*

Proof. We prove the statement by induction. Base: suppose that only one vertex exists in $\boldsymbol{O}$. Then $E_i^* = E_i = O_i$, so EEL terminates with $E_i^* = O_i$.

Step: suppose that EEL recovers $\boldsymbol{E}^*$, when there are q variables in $\boldsymbol{O}$. We need to prove the statement when there are $q+1$ variables in $\boldsymbol{O}$. Without loss of generality, choose O_{q+1} such that it is either an observed root vertex or a child of only an error term and latent variables so that $O_{q+1} = E_{q+1}^* = F_{q+1}$. We have two cases for any descendant O_l of O_{q+1}:

- E_{q+1} lies on a directed inducing path to O_l. EEL cannot partial out O_{q+1} from O_l by Line 9 and Lemma 1.

- E_{q+1} does not lie on a directed inducing path to O_l. Consider the largest set $\boldsymbol{U} \subseteq \boldsymbol{L}$ lying on a confounding path of O_l from E_{q+1}, where every collider is an ancestor of O_l and every non-collider is in $\boldsymbol{L}$ (the path may however not end at O_l). The at most d children of $\boldsymbol{U} \cup E_{q+1}$ on the path are all ancestors of O_l. Now place these observed children in $\boldsymbol{Y}$. Then EEL partials out $\boldsymbol{Y}$ from O_l in Line 14 when $l \leq d$.

We chose O_l as an arbitrary descendant of O_{q+1}, so we may repeat the above process for all descendants of O_{q+1}.

Next, for each O_m that is a child of O_{q+1}, set $E_m \leftarrow E_m + E_{q+1}\beta_{(q+1)m}$ and set $\gamma_{km} \leftarrow \gamma_{km} + \gamma_{k(q+1)}\beta_{(q+1)m}$ for each $L_k \in \text{Pa}(O_{q+1}) \cap \boldsymbol{L}$. We finally eliminate O_{q+1}. The conclusion follows by the inductive hypothesis. □

Corollary 1. *Under LiNGAM and d-separation faithfulness, if no confounding exists, then EEL with $l \leq 1$ recovers the error terms $\boldsymbol{E}$.*

Proof. At most one observed variable lies on a confounding path of O_i – that is, only O_i itself – so invoke Theorem 1 with $d = 1$. Observe further that E_i is the only member of $\boldsymbol{T}$ that lies on a directed inducing path to any $O_i \in \boldsymbol{O}$. As a result, EEL with $l = 1$ recovers $E_i^* = E_i$ for any $O_i \in \boldsymbol{O}$. □

Theorem 2. *The following relation holds under a linear model: $\gamma_{E_i^*\boldsymbol{W}} = E_i^*\delta_i - \mathbb{E}(E_i^*|\boldsymbol{V})\delta_i$, where $\boldsymbol{W} \subseteq (\boldsymbol{E}^* \setminus E_i^*)$ and $\boldsymbol{V} = (\boldsymbol{B}_i^* \setminus E_i^*) \cap \boldsymbol{W}$, so that:*

$$S_i^* = E_i^*\delta_i - \frac{\delta_i}{q} \sum_{\boldsymbol{V} \subseteq (\boldsymbol{B}_i^* \setminus E_i^*)} \psi_{|\boldsymbol{V}|}\mathbb{E}(E_i^*|\boldsymbol{V}).$$

Proof. We may write the following sequence for $\gamma_{E_i^*\boldsymbol{W}}$, where $\overline{\boldsymbol{W}} = (\boldsymbol{E}^* \setminus E_i^*) \setminus \boldsymbol{W}$ and $\boldsymbol{V} = (\boldsymbol{B}_i^* \setminus E_i^*) \cap \boldsymbol{W}$:

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{E}^*)|\boldsymbol{W}] &= \int f(\boldsymbol{W}, \overline{\boldsymbol{w}}) p(\overline{\boldsymbol{w}}|\boldsymbol{W})\, d\overline{\boldsymbol{w}} \\
&= \int \Big(\sum_{e_i^* \in \overline{\boldsymbol{W}}} e_i^*\delta_i + \sum_{E_i^* \in \boldsymbol{W}} E_i^*\delta_i \Big) p(\overline{\boldsymbol{w}}|\boldsymbol{W})\, d\overline{\boldsymbol{w}} \\
&= \sum_{E_i^* \in \overline{\boldsymbol{W}}} \delta_i \int e_i^* p(e_i^*|\boldsymbol{W})\, de_i^* + \sum_{E_i^* \in \boldsymbol{W}} E_i^*\delta_i \int p(\boldsymbol{W})\, d\boldsymbol{W} \\
&= \sum_{E_i^* \in \overline{\boldsymbol{W}}} \delta_i \mathbb{E}(E_i^*|\boldsymbol{V}) + \sum_{E_i^* \in \boldsymbol{W}} E_i^*\delta_i.
\end{aligned}$$

We finally arrive at $\gamma_{E_i^*\boldsymbol{W}}$ by subtraction:

$$\mathbb{E}[f(\boldsymbol{E}^*)|E_i^*, \boldsymbol{W}] - \mathbb{E}[f(\boldsymbol{E}^*)|\boldsymbol{W}] = E_i^*\delta_i - \mathbb{E}(E_i^*|\boldsymbol{V})\delta_i.$$

For S_i^*, we multiply (1) the number of sets $\boldsymbol{V} \subseteq (\boldsymbol{B}_i^* \setminus E_i^*)$ with $|\boldsymbol{V}| = k$ by (2) the Shapley weights to obtain:

$$\underbrace{\binom{|\boldsymbol{B}_i^*| - 1}{k} \sum_{j=0}^{q-|\boldsymbol{B}_i^*|} \binom{q - |\boldsymbol{B}_i^*|}{j}}_{(1)} \underbrace{\frac{1}{\binom{q-1}{j+k}}}_{(2)} = \frac{q}{|\boldsymbol{B}_i^*|}.$$

We identify (1) by choosing k elements from $\boldsymbol{B}_i^* \setminus E_i^*$ and then choosing the remaining elements from $(\boldsymbol{E}^* \setminus E_i^*) \cup (\boldsymbol{B}_i^* \setminus E_i^*)$. The equality follows by applying the creative telescoping algorithm (Zeilberger, 1991). We then have:

$$S_i^* = E_i^* \delta_i - \frac{\delta_i}{q} \sum_{k=0}^{|\boldsymbol{B}_i^*|-1} \sum_{\substack{\boldsymbol{V} \subseteq (\boldsymbol{B}_i^* \setminus E_i^*) \\ |\boldsymbol{V}| = k}} \psi_k \mathbb{E}(E_i^* | \boldsymbol{V}),$$

whence the conclusion follows. □

9.5. Additional Results

We refer to Table 2. EEL never came in first or last in terms of timing. The mean time increased most notably with sample size but remained within $O(n\log(n))$ as expected per the complexity analysis in Section 5. EEL only suffered a modest increase in time with higher degrees of confounding. We conclude that sample size drove most of the runtime of EEL in our experiments.

9.6. Diabetes Graph

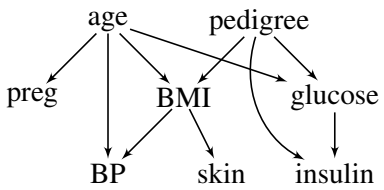


Figure 4: Insulin also causes glucose, but this introduces a cycle. We hypothesize that the direction insulin → glucose is better modeled as a mixture distribution for patients with type II diabetes versus healthy controls.

l	n	EEL	RCI	GRCI	ICA	RCAO	MS
0%	1,000	0.027	**0.008**	0.013	0.199	0.566	0.206
	10,000	0.002	**0.001**	**0.001**	0.193	0.250	0.209
	100,000	**0.006**	**0.000**	**0.000**	0.189	0.191	0.209
10%	1,000	0.075	**0.050**	0.063	0.268	0.686	0.280
	10,000	**0.010**	0.032	0.033	0.260	0.323	0.284
	100,000	**0.005**	0.032	0.035	0.256	0.266	0.286
20%	1,000	**0.147**	**0.159**	0.180	0.407	1.015	0.422
	10,000	**0.047**	0.137	0.147	0.399	0.491	0.427
	100,000	**0.023**	0.140	0.142	0.394	0.408	0.432

(a) MSE

l	n	EEL	RCI	GRCI	ICA	RCAO	MS
0%	1,000	2.686	0.031	0.228	0.872	2.396	0.611
	10,000	14.29	0.238	22.90	13.60	15.12	6.206
	100,000	89.21	2.382	160.84	521.0	480.7	48.17
10%	1,000	2.294	0.033	0.193	0.268	2.580	0.588
	10,000	14.45	0.280	16.34	0.260	13.56	5.506
	100,000	96.68	3.143	93.77	213.7	424.4	41.52
20%	1,000	1.798	0.035	0.131	0.407	2.264	0.521
	10,000	13.71	0.306	11.86	0.399	11.91	5.158
	100,000	114.2	3.242	79.38	177.7	392.0	36.96

(b) Time in seconds

Table 2: Results with the synthetic datasets in terms of mean (a) MSE and (b) time in seconds. EEL achieved the lowest MSE mean values with enough samples as highlighted in gray. However, EEL took more time to complete than RCI.