Efficient Identification of Direct Causal Parents via Invariance and Minimum Error Testing

Anonymous authors
Paper under double-blind review

Abstract

Invariant causal prediction (ICP) is a popular technique for finding causal parents (direct causes) of a target via exploiting distribution shifts and invariance testing (Peters et al., 2016). However, since ICP needs to run an exponential number of tests and fails to identify parents when distribution shifts only affect a few variables, applying ICP to practical large scale problems is challenging. We propose MMSE-ICP and fastICP, two approaches which employ an error inequality to address the identifiability problem of ICP. The inequality states that the minimum prediction error of the predictor using causal parents is the smallest among all predictors which do not use descendants. fastICP is an efficient approximation tailored for large problems as it exploits the inequality and a heuristic to run fewer tests. MMSE-ICP and fastICP not only outperform competitive baselines in many simulations but also achieve state-of-the-art result on a large scale real data benchmark.

1 Introduction

Causal discovery (CD) can offer insights into systems' dynamics (Pearl, 2018) which are helpful for influencing some outcomes (e.g. having diseases or not) or creating robust ML models. For example, a model based on a target's causal parents can robustly predict the target despite various distribution shifts. CD can be global or local: global CD searches for the complete causal graph while local CD only searches for causal relations surrounding a specific target Y. Thus, global CD is often intractable as the search domain grows exponentially with the number of variables. Local CD is more tractable and is sufficient if the goal is to change Y through interventions or to build domain-invariant ML models of Y. Recently, many works have built on local CD to improve ML models' out-of-distribution generalization (Rojas-Carulla et al., 2018; Magliacane et al., 2018; Arjovsky et al., 2019; Christiansen et al., 2021).

Invariant causal prediction (ICP) is a local CD method that finds the causal parents of Y using the invariance property: the distribution of Y conditioned on all of its causal parents will be invariant under interventions on variables other than Y (Peters et al., 2016). Specifically, ICP tests all subsets of variables for invariance and outputs the intersection (denoted as $\hat{\mathbf{S}}_{\text{ICP}}$) of all invariant subsets. As the number of subsets grows exponentially, applying ICP to large problems (> 20 variables) is difficult. Besides, ICP is guaranteed to identify all causal parents only when every variable in the system except Y is perturbed (Peters et al., 2016). However, in many practical problems (e.g. problems in cell biology with 20,000 variables), it is almost impossible to perturb all variables (Uhler, 2024). With limited perturbation/intervention, ICP may fail to identify causal parents (Rosenfeld et al., 2021; Mogensen et al., 2022), as disjoint subsets can be invariant, yielding an empty intersection (see Figure 1).

Invariant ancestry search (IAS) addresses ICP's failure to identify causal parents by outputting the causal ancestors of Y instead (Mogensen et al., 2022). The causal ancestors identified is the union (denoted as $\hat{\mathbf{S}}_{\text{IAS}}$) of all minimally invariant subsets. A set is minimally invariant if none of its proper subsets is invariant (Mogensen et al., 2022). Like ICP, $\hat{\mathbf{S}}_{\text{IAS}}$ can also be used to construct a stable predictor of Y across environments even though it is not the most compact stable predictor. Besides, intervening on an ancestor may not be as effective in influencing the value of Y as intervening on a parent. IAS can be sped

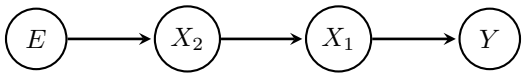


Figure 1: Consider this directed acyclic graph (DAG) representing a noisy causal structural model as a motivating example, where no causal mechanism is noise-free and thus there are no duplicate variables. Let E denote the variable capturing the environment (or context) (Mooij et al., 2020). Its children are directly affected by intervention. This reflect distribution shifts between contexts. Y is the target variable. $\hat{\mathbf{S}}_{\text{ICP}} = \emptyset$ because the invariant sets are $\{X_1\}$, $\{X_2\}$, and $\{X_1, X_2\}$. In contrast, $\hat{\mathbf{S}}_{\text{IAS}} = \{X_1, X_2\}$. Our methods output $\{X_1\}$ since the prediction error of $\hat{Y}_M(X_1)$ is less than $\hat{Y}_M(X_2)$'s.

up by only testing subsets up to a certain size at the cost of potentially finding fewer causal ancestors, thus trading accuracy for speed.

We propose two approaches that employ an error inequality to address the identifiability and scalability problem of ICP. The error inequality implies that the prediction error of the predictor using Y's causal parents is the smallest among all predictors which do not use Y's descendants. Thus, the first approach, MMSE-ICP (short for *minimum mean squared error ICP*), finds invariant sets that do not contain Y's descendants and outputs the invariant set with the smallest error for predicting Y. The second approach is fastICP, an efficient approximation that relies on a heuristic and constrained search scheme (see Section 3.4). fastICP also exploits the error inequality to reduce the number of tests for invariance to identify the causal parents. Consequently, fastICP can handle large-scale problems with numerous variables. Experiments on simulated and real data show MMSE-ICP and fastICP outperforming competitive baselines.

2 Background

In different environments/settings, data are generated using different generative mechanisms. An intervention refers to perturbation of generative mechanism of a specific variable (Pearl, 2009). Observational setting indicates no perturbation. Interventional setting indicates one or more variables are perturbed. Experimental data belong to interventional setting in which the interventions are known (which variables are perturbed and how are they perturbed). Most naturally occurring distribution-shift data also fall under interventional setting, however the interventions are unknown.

Many CD methods use solely observational data. Constraint-based (Spirtes et al., 2000) and score-based (Chickering, 2002) methods can identify up to the Markov equivalent class (MEC) of the true graph, leaving some edges with unresolved direction. With additional assumptions (e.g. non-Gaussianity (Shimizu et al., 2006), nonlinearity (Hyvärinen & Pajunen, 1999; Hoyer et al., 2008; Zhang & Hyvärinen, 2009; Peters et al., 2014; Zhang et al., 2015; Rolland et al., 2022), or independent causal mechanisms (Janzing et al., 2012)), methods are able to resolve more edges (Glymour et al., 2019). Optimization-base methods (Zheng et al., 2018; Geffner et al.) have become popular recently despite the lack of identifiability guarantees (Mogensen et al., 2022).

Having data from multiple settings can improve identifiability. When interventions are known, undirected edges in the MEC first estimated using observational data can be resolved using interventional data (He & Geng, 2008). Other approaches jointly model interventional and observational data (Hauser & Bühlmann, 2012; 2015; Wang et al., 2017). There are also optimization-based methods that leverage both types of data (Lorch et al., 2021; Lippe et al., 2022). In contrast, UT-IGSP (Squires et al., 2020), SDI (Ke et al., 2023), and DCDI (Brouillard et al., 2020) are methods that can work with unknown interventions. ICP (Peters et al., 2016; Pfister et al., 2019; Gamella & Heinze-Deml, 2020; Martinet et al., 2022) and IAS (Mogensen et al., 2022) both assume unknown interventions and only depends on the invariance property to identify causal structures. Our approaches also assume unknown interventions.

Prototypical global CD methods such as PC (Colombo et al., 2014; Li et al., 2019) and GES (Chickering, 2002) can be slow, so faster global CD ones (Cheng et al., 2002; Tsamardinos et al., 2006; Ramsey et al., 2017) have been proposed, although they make strong assumptions that may be inappropriate in some problems. In contrast, local CD methods only identify local causal structures surrounding a target variable. For example,

several algorithms only search for the Markov Blanket (MB) of the target (i.e. its parents, children, and spouses) (Koller et al., 1996; Margaritis & Thrun, 1999; Tsamardinos et al., 2003; Gao & Ji, 2015; Yang et al., 2021). Consequently, local CD methods often have faster runtime or can afford to make milder assumptions leading to more accurate identification. ICP and IAS only search for parents and ancestors of the target respectively, omitting all descendants of the target. They both make very mild assumptions and offer guarantees in terms of false positive rates. Another local CD approach is CORTH (Soleymani et al., 2022) which is based on double ML (Chernozhukov et al., 2018). CORTH assumes that Y is not a parent of any other variable (a strong assumption which can be unrealistic) and that Y is a linear combination of other variables, and achieves linear runtime finding parents of Y.

3 Method

3.1 Definitions and assumptions

We represent causal relations between variables using a Directed Acyclic Graph (DAG), where each node is a variable and directed edges between nodes represent direct causal influence (Pearl, 2009). We use the usual notations from graphical models (Lauritzen, 1996). Specifically, $\mathsf{PA}(Y)$, $\mathsf{CH}(Y)$, $\mathsf{AN}(Y)$, $\mathsf{DE}(Y)$, and $\mathsf{ND}(Y)$ denote the parents, children, ancestors, descendants, and non-descendants of Y respectively. Thus, $\mathsf{PA}(Y) \subset \mathsf{AN}(Y) \subset \mathsf{ND}(Y)$ and $\mathsf{CH}(Y) \subset \mathsf{DE}(Y)$. There is an additional node E in the graph to denote the different environments (or contexts) (Mooij et al., 2020). A set of variables/predictors $\mathbf{S}$ is invariant if $Y \perp\!\!\!\perp E|\mathbf{S}$. As is common in the causal inference literature, we assume (1) no hidden confounder, (2) no intervention on Y, (3) no feedback between variables, (4) independence of mechanisms, and (5) independent additive noise structural causal models (SCMs), with no noise-free mechanisms and duplicate variables. We further assume that (6) the additive noise variables in the SCM have unbounded support (e.g. Gaussian noise). Like prior work (Mogensen et al., 2022; Mooij et al., 2020), we assume that (7) E is exogenous (i.e., E has no parent).

3.2 ICP and IAS

ICP exploits distribution shifts between different environments E to learn a subset of $\mathsf{PA}(Y)$ (Peters et al., 2016). Given a test for the hypothesis $H_{0,\mathbf{S}}$ that $\mathbf{S}$ is invariant, ICP outputs the intersection of all invariant subsets. Specifically, $\hat{\mathbf{S}}_{\text{ICP}} := \bigcap_{\mathbf{S}:H_{0,\mathbf{S}} \text{ not rejected}} \mathbf{S}$. In contrast, IAS tries to identify a subset of $\mathsf{AN}(Y)$ (Mogensen et al., 2022). Let $\widehat{\mathcal{I}}$ be the set of all sets $\mathbf{S}$ for which $H_{0,\mathbf{S}}$ is not rejected. Define $\widehat{\mathcal{MI}} := \left\{\mathbf{S} \in \widehat{\mathcal{I}} \mid \forall \mathbf{S}' \subsetneq \mathbf{S} : \mathbf{S}' \notin \widehat{\mathcal{I}}\right\}$. Thus, $\hat{\mathbf{S}}_{\text{IAS}} := \bigcup_{\mathbf{S} \in \widehat{\mathcal{MI}}} \mathbf{S}$. Although Peters et al. (2016) proposed two different tests for invariance, $H_{0,\mathbf{S}}$, the approximate test based on residuals of a predictor (Method II) is usually used because of its speed. Specifically, this statistical test checks whether the distribution (means and variances) of the predictor's errors across environments E are the same. The predictor is fitted using the data from all environments (see Appendix A for more details).

3.3 Finding causal parents by minimizing error

Instead of taking the intersection or union of invariant subsets, we rely on the concept of minimum mean squared error predictor to find the causal parents. The mean squared error (MSE) of a predictior $\hat{Y}(\mathbf{X})$ with input variables $\mathbf{X}$ of target Y is $\mathbb{E}_{\mathbf{X}Y}(Y - \hat{Y}(\mathbf{X}))^2$, where $\mathbb{E}$ denotes expectation. Let the minimum MSE achieved by an optimal predictor $\hat{Y}_M(\mathbf{X})$ with input variables $\mathbf{X}$ be $\mathsf{MMSE}(\mathbf{X})$. Note that this is always with respect to predicting a target variable Y. It is well-known that, if the minimization is unconstrained, the $\mathsf{MMSE}(\mathbf{X})$ is the conditional expectation $\mathbb{E}_{Y|\mathbf{X}}(Y|X)$, and the corresponding MMSE is equal to $\mathbb{E}_{\mathbf{X}}\text{Var}_{Y|\mathbf{X}}(Y|X)$ (Leon-Garcia, 1994), where Var denotes variance. In practice, we will assume that a large-capacity model, trained on enough data with least square loss, can approximate this MMSE estimator. The following provides the theoretical grounding for MMSE-ICP.

3.3.1 Minimum mean squared error inequality

Lemma 3.1 (Error Inequality)**.** *Let* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *denote two sets of variables, not necessarily mutually exclusive. Then:* $\mathsf{MMSE}(\mathbf{X}_1 \cup \mathbf{X}_2) \leq \mathsf{MMSE}(\mathbf{X}_1)$*. Equality holds if* $Y \perp\!\!\!\perp \mathbf{X}_2 | \mathbf{X}_1$*.*

Proof. $\mathsf{MMSE}(\mathbf{X}_1 \cup \mathbf{X}_2) = \mathbb{E}_{\mathbf{X}_1,\mathbf{X}_2}\left[\min_{\hat{Y}} \mathbb{E}_{Y|\mathbf{X}_1,\mathbf{X}_2}(Y-\hat{Y})^2\right] = \mathbb{E}_{\mathbf{X}_1}\mathbb{E}_{\mathbf{X}_2|\mathbf{X}_1}\left[\min_{\hat{Y}} \mathbb{E}_{Y|\mathbf{X}_1,\mathbf{X}_2}(Y-\hat{Y})^2\right]$.

Let $\hat{Y}_M(\mathbf{X}_1,\mathbf{X}_2) := \arg\min_{\hat{Y}} \mathbb{E}_{Y|\mathbf{X}_1,\mathbf{X}_2}(Y-\hat{Y})^2$, then:

$$\mathbb{E}_{Y|\mathbf{X}_1,\mathbf{X}_2}(Y-\hat{Y}_M(\mathbf{X}_1,\mathbf{X}_2))^2 \leq \mathbb{E}_{Y|\mathbf{X}_1,\mathbf{X}_2}(Y-\hat{Y})^2, \;\; \forall \hat{Y} \tag{1}$$

Since expectation is monotonic,

$$\Rightarrow \mathbb{E}_{\mathbf{X}_2|\mathbf{X}_1}\left[\mathbb{E}_{Y|\mathbf{X}_1,\mathbf{X}_2}(Y-\hat{Y}_M(\mathbf{X}_1,\mathbf{X}_2))^2\right] \leq \mathbb{E}_{\mathbf{X}_2|\mathbf{X}_1}\left[\mathbb{E}_{Y|\mathbf{X}_1,\mathbf{X}_2}(Y-\hat{Y})^2\right] = \mathbb{E}_{Y|\mathbf{X}_1}(Y-\hat{Y})^2, \;\; \forall \hat{Y} \tag{2}$$

Since Equation 2 holds for every $\hat{Y}$, it is also true for $\hat{Y}_M(\mathbf{X}_1) = \arg\min_{\hat{Y}} \mathbb{E}_{Y|\mathbf{X}_1}(Y-\hat{Y})^2$,

$$\Rightarrow \mathbb{E}_{\mathbf{X}_1}\mathbb{E}_{\mathbf{X}_2|\mathbf{X}_1}\left[\mathbb{E}_{Y|\mathbf{X}_1,\mathbf{X}_2}(Y-\hat{Y}_M(\mathbf{X}_1,\mathbf{X}_2))^2\right] \leq \mathbb{E}_{\mathbf{X}_1}\left[\mathbb{E}_{Y|\mathbf{X}_1}(Y-\hat{Y}_M(\mathbf{X}_1))^2\right] = \mathsf{MMSE}(\mathbf{X}_1) \tag{3}$$

$$\Rightarrow \mathsf{MMSE}(\mathbf{X}_1 \cup \mathbf{X}_2) \leq \mathsf{MMSE}(\mathbf{X}_1) \tag{4}$$

Equality occurs when $\hat{Y}_M(\mathbf{X}_1,\mathbf{X}_2) = \hat{Y}_M(\mathbf{X}_1)$, i.e.,

$$\arg\min_{\hat{Y}} \mathbb{E}_{Y|\mathbf{X}_1,\mathbf{X}_2}(Y-\hat{Y})^2 = \arg\min_{\hat{Y}} \mathbb{E}_{Y|\mathbf{X}_1}(Y-\hat{Y})^2, \;\; \forall \mathbf{X}_1, \mathbf{X}_2 \tag{5}$$

If $Y \perp\!\!\!\perp \mathbf{X}_2 | \mathbf{X}_1$ then one can easily show that Equation 5 holds. □

Corollary 3.2. *In the causal DAG that satisfies the assumptions stated in Section 3.1, for all subset* $\mathbf{S}$ *of* $\mathsf{ND}(Y)$*:*

- $\mathsf{MMSE}(\mathbf{S}) \geq \mathsf{MMSE}(\mathsf{PA}(Y))$*,*
- $\mathsf{MMSE}(\mathbf{S}) = \mathsf{MMSE}(\mathsf{PA}(Y))$*, if and only if* $\mathsf{PA}(Y) \subset \mathbf{S}$*, and*
- $\mathsf{MMSE}(\mathbf{S}) > \mathsf{MMSE}(\mathsf{PA}(Y))$*, if and only if* $\mathsf{PA}(Y) \not\subset \mathbf{S}$*.*

Proof. The first inequality is an immediate corollary of Lemma 3.1. It is easy to show that, in the assumed DAG, $\mathsf{MMSE}(\mathsf{PA}(Y)) = \sigma_Y^2$, where σ_Y^2 is the variance of the independent additive noise for Y. Furthermore, for any $\mathbf{S}$ which is a subset of $\mathsf{ND}(Y)$, $\mathsf{MMSE}(\mathbf{S}) = \mathbb{E}_{\mathbf{S}}\mathrm{Var}_{Y|\mathbf{S}}(f_Y(\mathsf{PA}(Y))|\mathbf{S}) + \sigma_Y^2$, where $f_Y(\mathsf{PA}(Y))$ is the functional mapping from Y's parents to Y in the SCM. It can be shown that in the assumed noisy SCM where the noise distribution has unbounded support (e.g. Gaussian noise), $\mathrm{Var}_{Y|\mathbf{S}}(f_Y(\mathsf{PA}(Y))|\mathbf{S}) > 0$ if and only if $\mathsf{PA}(Y) \not\subset \mathbf{S}$ and $\mathrm{Var}_{Y|\mathbf{S}}(f_Y(\mathsf{PA}(Y))|\mathbf{S}) = 0$ if and only if $\mathsf{PA}(Y) \subset \mathbf{S}$. □

For example, in Figure 1, $\mathsf{MMSE}(\{X_2\}) > \mathsf{MMSE}(\{X_1, X_2\}) = \mathsf{MMSE}(\{X_1\})$. Furthermore, since $\{X_1, X_2\}$ has 2 elements while $\{X_1\}$ has only 1 element, we can deduce that $\{X_1\}$ is the causal parent and not $\{X_2\}$.

3.3.2 Combining MMSE and invariance test

Algorithm 1 combines the error inequality with the invariance test to find causal parents. Theorem 3.4 shows that Algorithm 1 can identify all causal parents of Y as long as $\mathsf{PA}(Y) \subset \mathsf{DE}(E)$. This implies that Algorithm 1 usually requires fewer interventions than the number of variables for complete identification. For example, a single intervention at an upstream variable that affects all causal parents of Y is sufficient. This is much more favorable than ICP, which may need as many interventions at as the number of variables (Peters et al., 2016).

Algorithm 1: MMSE-ICP

Input: $E, X_1, X_2, \ldots, X_d, Y$
Output: Potential parents of Y

1 **if** $is_invariant(E, \mathbf{X}, Y, \emptyset)$ **then** **return** $\emptyset$
2 $\mathsf{candidates} = []$
3 **for** $\mathbf{S} \subset \{X_1, X_2, \ldots, X_d\}$ *in increasing cardinality order* **do**
4 **if** $\exists \mathbf{S}' \in \mathsf{candidates}$ *such that* $\mathbf{S}' \subset \mathbf{S}$ **then** **continue**
5 **if** $is_invariant(E, \mathbf{X}, Y, \mathbf{S})$ **then** $\mathsf{candidates} = \mathsf{candidates} + [\mathbf{S}]$
6 **return** $\arg\min_{\mathbf{S} \in \mathsf{candidates}} \mathsf{MMSE}(\mathbf{S})$

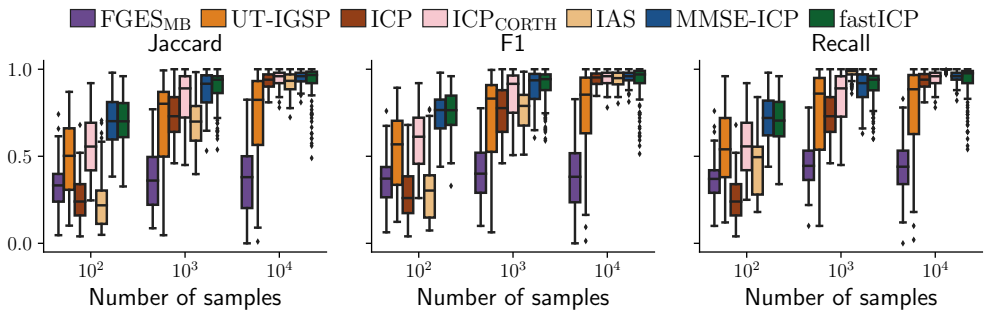


Figure 2: Performance when $N_{\text{int}} = d = 6$ (Table 1, No. 1). Linear simulation. Reference set: $\mathsf{PA}(Y)$.

Lemma 3.3. *If* $\mathbf{S}_1$ *is invariant, then* $\mathbf{S}_2 = \mathbf{S}_1 \cap (\mathsf{DE}(E) \cap \mathsf{ND}(Y))$ *is also invariant. In other words, removing all nodes in* $\mathbf{S}_1$ *that are not in* $\mathsf{DE}(E) \cap \mathsf{ND}(Y)$ *from* $\mathbf{S}_1$ *results in another invariant set.*

Proof. Since $\mathbf{S}_1$ is invariant, $Y \perp\!\!\!\perp E | \mathbf{S}_1$ so $\mathbf{S}_1$ blocks all paths between E and Y by d-separation (Pearl, 2009). If a node X is in $\mathbf{S}_1$ but not in $\mathsf{DE}(E) \cap \mathsf{ND}(Y)$, then one of these two cases must hold.

1. X is not on any blocking path so $\mathbf{S}_1 \setminus \{X\}$ still blocks all paths between E and Y.
2. X is on a blocking path between E and Y. Since E has no parent (see Section 3.1), $X \in \mathsf{DE}(E) \cap \mathsf{DE}(Y)$. Thus, this blocking path is out-going from Y. Removing all nodes on this path will keep the path blocked. The removed nodes are descendants of Y so they are not in $\mathsf{DE}(E) \cap \mathsf{ND}(Y)$.

In both cases, invariance is not affected by excluding X. Thus, excluding nodes in $\mathbf{S}_1$ that are not in $\mathsf{DE}(E) \cap \mathsf{ND}(Y)$ results in $\mathbf{S}_2$ that is also invariant. □

Theorem 3.4 (Identifiability)**.** *Given that all tests of invariance return the correct results and* $\mathsf{PA}(Y) \subset \mathsf{DE}(E)$*, Algorithm 1 will always find the set of causal parents* $\mathsf{PA}(Y)$*.*

Theorem 3.4 follows from Corollary 3.2 and Lemma 3.3. The loop in Algorithm 1 finds invariant subsets of $\mathsf{DE}(E) \cap \mathsf{ND}(Y)$ which include $\mathsf{PA}(Y)$ when $\mathsf{PA}(Y) \subset \mathsf{DE}(E)$. Amongst these subsets, $\mathsf{PA}(Y)$ has minimum MSE and is the most compact. The solution is unique (no other subset has the same MSE and cardinality) because other subsets with minimum MSE are all strict supersets of $\mathsf{PA}(Y)$ so they have larger cardinality than $\mathsf{PA}(Y)$.

Lemma 3.3 implies that for any invariant set $\mathbf{S}_1$ that is not a subset of $\mathsf{DE}(E) \cap \mathsf{ND}(Y)$, there exists an invariant set $\mathbf{S}_2$ that is a subset of $\mathbf{S}_1$ (having smaller cardinality than $\mathbf{S}_1$'s) and of $\mathsf{DE}(E) \cap \mathsf{ND}(Y)$. Since Algorithm 1 iterates through the sets of variables in increasing cardinality order, it will first find $\mathbf{S}_2$ and add $\mathbf{S}_2$ to the list. It will later exclude $\mathbf{S}_1$ because $\mathbf{S}_2$ is a subset of $\mathbf{S}_1$. Thus, all invariant sets that are not subsets of $\mathsf{DE}(E) \cap \mathsf{ND}(Y)$ will be excluded. $\mathsf{PA}(Y)$ will always be in the list because no subset of $\mathsf{PA}(Y)$ is invariant.

Algorithm 2: fastICP

Input: $E, X_1, X_2, \ldots, X_d, Y, \mathsf{MaxDepth}$
Output: Potential parents of Y

if $is_invariant(E, \mathbf{X}, Y, \emptyset)$ **then return** $\emptyset$
`// Find invariant set`
$\mathbf{S} = \{X_1, \ldots, X_d\}$
while *not* $is_invariant(E, \mathbf{X}, Y, \mathbf{S})$ **do**
  $\mathsf{candidates} = \{\mathbf{S}' \subset \mathbf{S} : |\mathbf{S}'| \geq |\mathbf{S}| - \mathsf{MaxDepth}\}$
  if *no* $\mathsf{candidates}$ **then return** $\emptyset$
  $\mathbf{S} = \arg\min_{\mathbf{S}' \in \mathsf{candidates}} \mathsf{statDependency}(Y, \mathbf{S}')$
`// Prune invariant set`
while *True* **do**
  $\mathsf{candidates} = []$
  for $Z \in \mathbf{S}$ **do**
    if $is_invariant(E, \mathbf{X}, Y, \mathbf{S} \setminus \{Z\})$ **then**
      $\mathsf{candidates} = \mathsf{candidates} + [\mathbf{S} \setminus \{Z\}]$
  if *no* $\mathsf{candidates}$ **then break**
  else $\mathbf{S} = \arg\min_{\mathbf{S}' \in \mathsf{candidates}} \mathsf{MMSE}(\mathbf{S}')$
return $\mathbf{S}$

3.4 Faster search sequence

Although Algorithm 1 should be more accurate than ICP, it still needs to run an exponential number of invariance tests. We propose Algorithm 2, which is faster and consists of two stages.

Stage 1 aims to find the largest invariant set $\mathbf{S}$ that includes $\mathsf{PA}(Y)$. Starting with the set of all variables, it removes potential colliders and their descendants, which might be unblocking paths from E to Y, using the heuristic presented in Cheng et al. (1998). The heuristic assumes that blocking a path by removing some nodes from the conditioning set decreases the statistical dependency between E and Y (Cheng et al., 1998). In Algorithm 2, $\mathsf{MaxDepth}$ is a hyper-parameter for the maximum number of nodes to be removed at one time. The statistical dependency can be measured using the invariance test of Peters et al. (2016) where the lower the statistical dependency, the higher the probability of being invariant. Stage 2, based on Proposition 3.5, removes variables from $\mathbf{S}$ one-by-one while maintaining invariance. The algorithm terminates when $\mathbf{S} = \mathsf{PA}(Y)$.

Proposition 3.5. *We are given that* $\mathsf{PA}(Y) \subset \mathsf{DE}(E)$*,* $\mathbf{S}$ *is invariant and has the lowest MMSE amongst subsets with the same cardinality, and* $\mathsf{PA}(Y) \subset \mathbf{S}$*. For each* $Z \in \mathbf{S}$*, let* $\mathbf{S}_{-Z} := \mathbf{S} \setminus \{Z\}$*. If no subset* $\mathbf{S}_{-Z}$ *of* S *is invariant, then* $\mathbf{S} = \mathsf{PA}(Y)$*. Otherwise,* $\mathsf{PA}(Y) \subset \arg\min_{invariant\ \mathbf{S}_{-Z}} \mathsf{MMSE}(\mathbf{S}_{-Z})$*.*

Proof. If no subset $\mathbf{S}_{-Z}$ of S is invariant, all nodes of $\mathbf{S}$ must be on blocking paths because removing any node Z from $\mathbf{S}$ leads to loss of invariance (unblocks a path from E to Y). If multiple nodes are on the same blocking path, one node (e.g. W) can be removed and $\mathbf{S}_{-W}$ is still invariant which is contradictory. Thus, $Z \in \mathsf{PA}(Y)$ because otherwise, exchanging Z with the parent in the same blocking path would result in a subset with the same cardinality but with smaller MMSE, which is contradictory. Since $Z \in \mathsf{PA}(Y), \forall Z \in \mathbf{S}$, then $\mathbf{S} = \mathsf{PA}(Y)$.

Suppose there exists a $Z \in \mathbf{S}$ where $\mathbf{S}_{-Z}$ is invariant. Let's assume that $\mathsf{PA}(Y) \nsubseteq \mathbf{S}_{-Z'} := \arg\min_{\text{invariant } \mathbf{S}_{-Z}} \mathsf{MMSE}(\mathbf{S}_{-Z})$.
Since $\mathsf{PA}(Y) \nsubseteq \mathbf{S}_{-Z'}$ but $\mathsf{PA}(Y) \subset \mathbf{S}$ so $Z' \in \mathsf{PA}(Y)$. Besides, $\mathbf{S}_{-Z'}$ is invariant but $Z' \notin \mathbf{S}_{-Z'}$, so there must be another node $T \in \mathbf{S}_{-Z'}$ that is on the same blocking path as Z'.
Let $\mathbf{S}' := \mathbf{S} \setminus \{Z', T\}$. Consequently, $\mathbf{S}_{-Z'} = \mathbf{S} \setminus \{Z'\} = \mathbf{S}' \cup \{T\}$ and $\mathbf{S}_{-T} = \mathbf{S} \setminus \{T\} = \mathbf{S}' \cup \{Z'\}$.
Since T and Z' are on the same blocking path, $\mathbf{S}_{-T}$ is also invariant. In addition, because $Z' \in \mathsf{PA}(Y)$, T

Table 1: Different setups with different number of nodes $(X_1, \ldots, X_d, Y, E)$, graph densities, number of interventions (N_{int}), and type of interventions. MB: Markov Blanket

No.	d	Density	MB size	N_{int}	Type
1.	6	0.240	2.98	6	Perfect
2.	6	0.145	2.58	1,2,3	Perfect
3.	6	0.158	2.89	1	Imperfect
4.	6	0.153	2.86	1	Noise
5.	100	0.010	3.30	1–5	Perfect
6.	100	0.050	26.99	1–5	Perfect

must precede Z', i.e. $Y \perp\!\!\!\perp T | \mathbf{S}' \cup \{Z'\} \Leftrightarrow Y \perp\!\!\!\perp T | \mathbf{S}_{-T}$. As a result, by Lemma 3.1 and Corollary 3.2,

$$\mathsf{MMSE}(\mathbf{S}_{-T}) = \mathsf{MMSE}(\mathbf{S}_{-T} \cup \{T\}) = \mathsf{MMSE}(\mathbf{S}) = \mathsf{MMSE}(\mathbf{S}_{-Z'} \cup \{Z'\}) < \mathsf{MMSE}(\mathbf{S}_{-Z'}).$$

This is contradictory because $\mathsf{MMSE}(\mathbf{S}_{-Z'}) < \mathsf{MMSE}(\mathbf{S}_{-T})$ by definition. Hence, $\mathsf{PA}(Y) \subset \mathbf{S}_{-Z'}$. □

3.5 Complexity Analysis

For a problem with V variables, the complexity of ICP and MMSE-ICP is $O(2^V)$ as the number of invariance tests run is exponential. Stage 1 of fastICP is $O(V * 2^{\mathsf{MaxDepth}})$ because of the number of candidates checked grows exponentially with $\mathsf{MaxDepth}$. Stage 2 is $O(V^2)$. Thus, the complexity of fastICP is $O(V * 2^{\mathsf{MaxDepth}} + V^2)$. When $\mathsf{MaxDepth}$ is less than V, fastICP is faster than ICP.

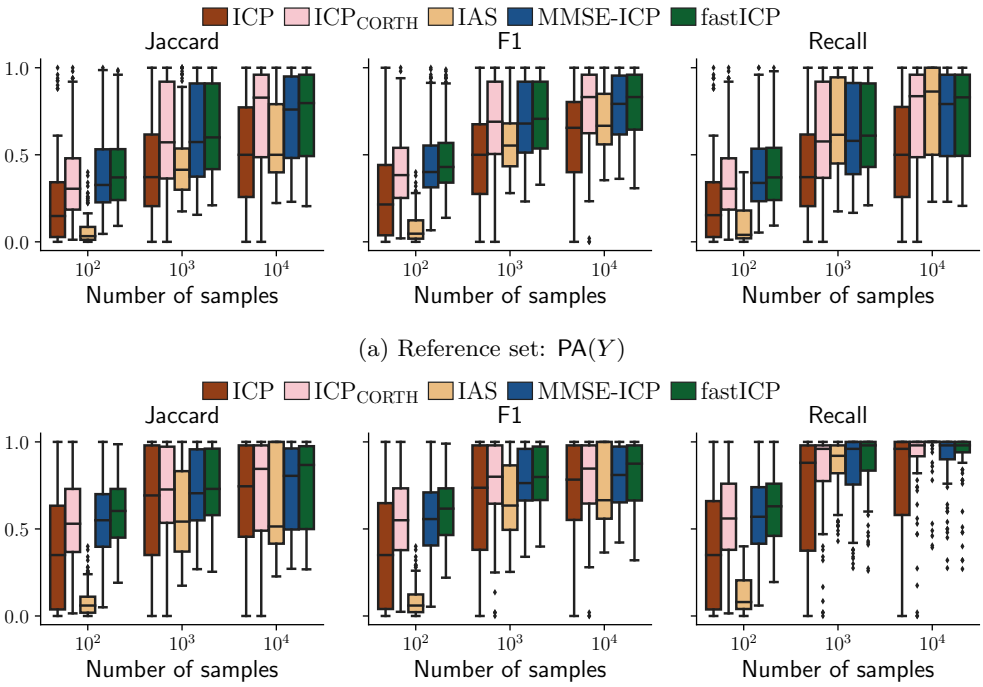
4 Simulation Experiments

4.1 Data

We generate synthetic datasets according to the assumptions in Section 3.1 (additive noise model with i.i.d. Gaussian noise). Similar to Peters et al. (2016); Mogensen et al. (2022), we consider 2-environment setups (observational and interventional). In each setup, 100 graphs are randomly generated. In a graph, beside the target node Y and the environment indicator node E, there are d additional nodes $\{X_1, \ldots, X_d\}$. The edges from E to a subset of $\{X_1, \ldots, X_d\}$ specify the nodes that may be intervened on. The number of interventions (N_{int}) varies between different setups. We do not enforce the constraint that $\mathsf{PA}(Y) \subset \mathsf{DE}(E)$ to test the robustness of the proposed methods when this assumption is invalid. Values of a node is generated based on its parents' values, through linear or nonlinear functions. Beside linear simulations whereby all functions are linear, we also include nonlinear simulations (see Appendix C). For each graph, 50 sets of coefficients of the generative functions are drawn randomly. For each set of coefficients, we sample 4 datasets with different sample sizes (10^2, 10^3, 10^4, 10^5). The data are standardized along the causal order to prevent shortcut learning (Reisach et al., 2021). In the interventional environment, interventions are applied to a random subset of children of E. We consider 3 types of interventions: (1) perfect intervention which severs all causal dependencies from parents, (2) imperfect intervention which modifies causal relations between a node and its parents, and (3) noise intervention in which the intervened variable's noise variance changes (Cooper & Yoo, 1999; Peters et al., 2016). The different setups are summarized in Table 1.

4.2 Baselines and implementation details

We compare MMSE-ICP and fastICP against ICP (Peters et al., 2016),IAS (Mogensen et al., 2022),fGES-MB (Ramsey et al., 2017),UT-IGSP (Squires et al., 2020),and ICP-CORTH. ICP-CORTH uses ICP to remove false positives in CORTH's output (Soleymani et al., 2022). Since CORTH assumes Y is a linear combination of its parents (similar to the simulation setup), ICP-CORTH is a competitive baseline. To test for invariance, MMSE-ICP and fastICP use the same test as ICP and IAS. Specifically, the test checks whether the mean and variance of the prediction residuals (of a regression model) is equal across environments. Linear regression



(a) Reference set: $\mathsf{PA}(Y)$

(b) Reference set: $\mathbf{S}^*{=}\mathsf{DE}(E) \cap \mathsf{PA}(Y)$

Figure 3: Performance when $N_{\text{int}}{=}1; d{=}6$ (Table 1, No. 2). Linear simulation. See Appendix D.1 for results when $N_{\text{int}}{=}2$ or $N_{\text{int}}{=}3$.

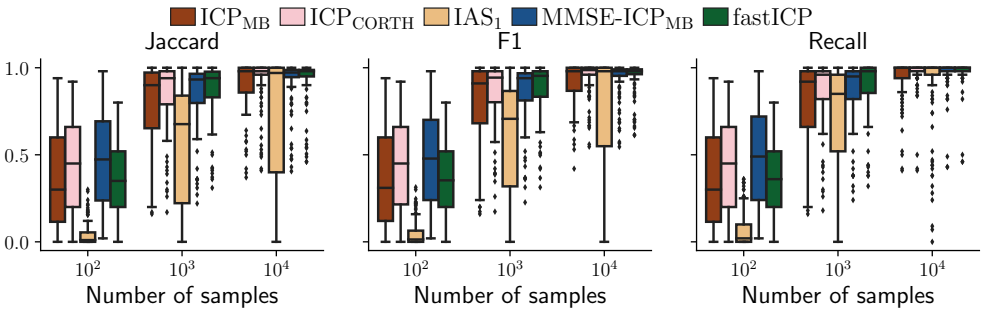
is used for linear simulations while gradient boosted tree is used for nonlinear simulations (see Appendix A). To keep the runtime of the invariance test comparable to ICP and IAS, MMSE is estimated by averaging the prediction residuals (see Appendix D.2 for experiments to test the robustness of MMSE estimation).

The MaxDepth hyper-parameter is set at 2. For ICP, IAS, UT-IGSP, MMSE-ICP, and fastICP, the significance level α is set at 0.05. The parameters of IAS (C and α_0) are set according to the original paper (Mogensen et al., 2022). Since it is intractable to search exhaustively using ICP, IAS and MMSE-ICP for large d, their search scopes are restricted in these cases. In particular, ICP and MMSE-ICP only search within an estimated MB of size 10. The 10 variables that are considered for further analysis are determined via L2-boosting (Friedman, 2001; Bühlmann & Yu, 2003; Hothorn et al., 2010). IAS only tests for sets with a single element.

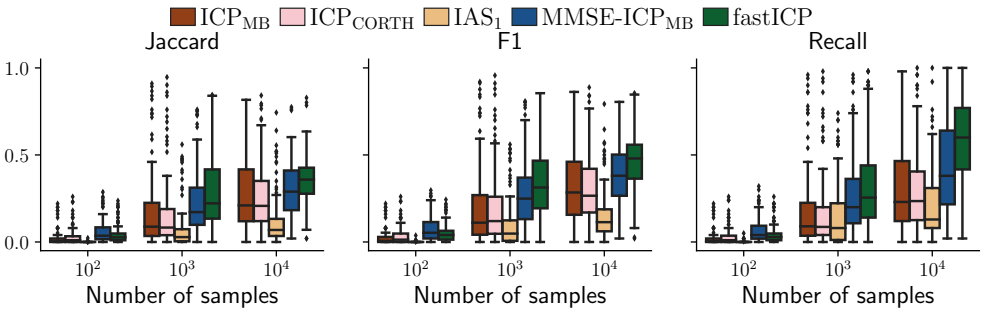
Jaccard similarity and *F1-score* are used as metrics. The prediction is compared against a reference set, which could be either (1) the full set parents $\mathsf{PA}(Y)$ and (2) the set of perturbed parents $\mathbf{S}^*{=}\mathsf{DE}(E) \cap \mathsf{PA}(Y)$. Since IAS finds ancestors instead of parents, we also look at the parents' *Recall* rate.

4.3 Results from linear simulations

When, $N_{\text{int}} = d = 6$, invariance-based algorithms should be able to discover all parents. Figure 2 shows their results in this setting. Their average runtimes are reported in Appendix B. With sufficient samples, invariance-based algorithms outperform fGES-MB (observational constraint-based CD) and UT-IGSP. When the number of interventions is limited (i.e. $N_{\text{int}} < d$, Table 1, No. 2–6), identifying all direct parents using invariance is more challenging since some invariance subsets may be strict subset of the parents, i.e. $|\mathbf{S}^*| < |\mathsf{PA}(Y)|$. Due to the strict inequality, we only benchmark our approaches against invariance-based algorithms. For perfect interventions, MMSE-ICP and fastICP achieve similar performance and outperform the baselines in both Jaccard similarity and F1-score (Figure 3). The recall of our approaches is the same as IAS's and higher than ICP's. Figure 3 shows that our approaches often found most of the perturbed parents



(a) Large sparse graphs (Table 1, No. 5)



(b) Large dense graphs (Table 1, No. 6)

Figure 4: Performance for large graphs. Reference set: $\mathbf{S}^*$. Also see Figure 11 and 12 in the Appendix.

($\mathbf{S}^*$). The same trends are observed for imperfect interventions (Figure 16), noise interventions (Figure 17), and large sparse graphs (Figure 4a). They hold for large sparse graphs since the estimated MB of size 10 is sufficient to cover the true MB (see Table 1).

For large dense graph (Figure 4b), MMSE-ICP and fastICP still outperform the baselines. Since the estimated MB is much smaller than the true MB, methods that cannot search exhaustively will miss out many parents. ICP-CORTH is worse than ICP probably because CORTH's assumption that the target Y has no children is more likely to be wrong. When the graph is large and dense, MMSE-ICP cannot search exhaustively so the ability to search through all nodes gives fastICP an edge over MMSE-ICP.

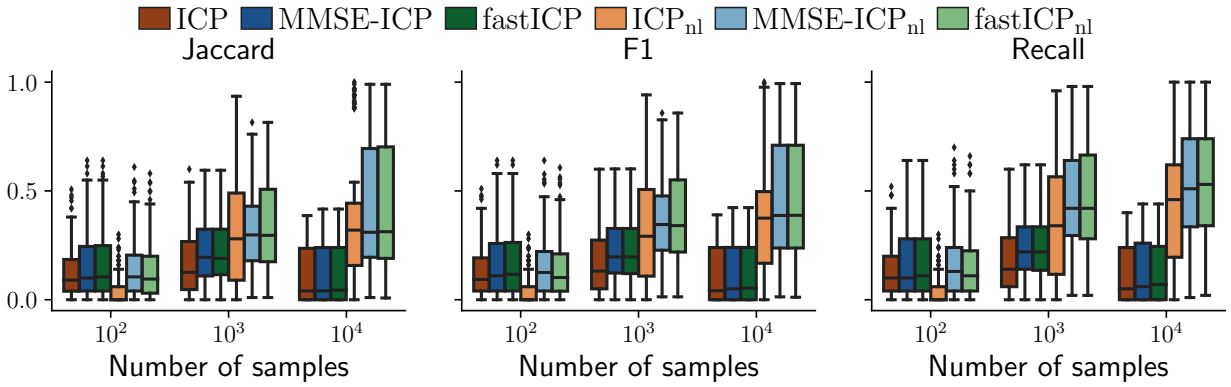


Figure 5: Performance for nonlinear simulation. N_{int}=1; d=6. Reference set: $\mathbf{S}^*$. *nl*: using nonlinear regression in invariance test. Also see Figure 18 in the Appendix.

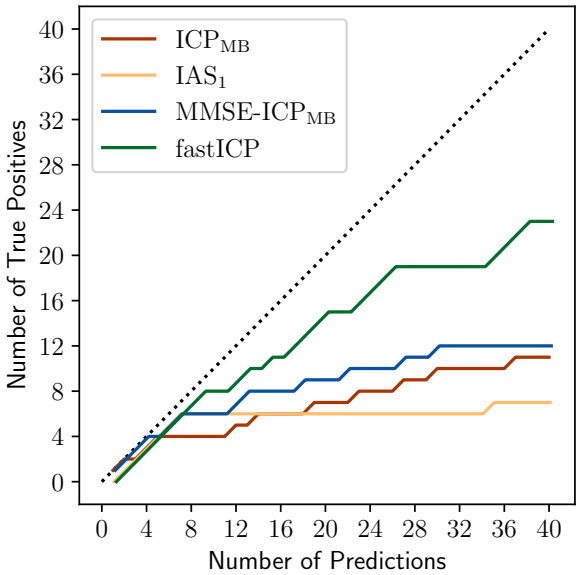


Figure 6: Performance on genes' parents prediction. The dotted line delineates perfect accuracy.

4.4 Results from nonlinear simulations

For nonlinear simulation, our approaches are still better than ICP (Figure 5 and Appendix D.4). As predicted by Heinze-Deml et al. (2018), it is sometime possible to use the linear invariance test (invariance test with linear regression) to find causal parents in nonlinear simulations. However, it is preferable to use a nonlinear invariance test to obtain more accurate results if runtime is not a constraint.

5 Experiment on gene expression data

5.1 Data

We apply our methods to a real-world large-scale yeast gene expression dataset with 6170 variables (Kemmeren et al., 2014). There are 160 observational samples and 1479 interventional samples. Each interventional sample corresponds to an experiment where single X_k has been perturbed. Following Peters et al. (2016); Mogensen et al. (2022), we assume that a direct causal effect $X \to Y$ exists (true positive) if the expression level of gene Y after intervening on gene X lies in the 1% lower or upper tail of the observational distribution of gene Y.

Since neither the ground-truth graph nor a separate set of validation data is available, we must use the same data for validation. Hence, we employ the same cross-validation scheme used in the original ICP paper to prevent information leakage (Peters et al., 2016). Specifically, when predicting whether X_k is a parent of Y, we do not include the sample when intervening on X_k (if the sample exists) in the data used for inference. The interventional samples are split into 3 folds. In each fold, two thirds of the interventional samples not containing X_k are used as interventional data, and remaining interventional data are used for validation. Thus, for each target, we need to run the algorithms the same number of times s the number of folds. Additionally, when looking for potential causes of Y, we exclude samples corresponding to intervention on Y (if it exists).

5.2 Baselines and implementation details

ICP and IAS are used as baselines as they give confidence estimates for individual predicted parents/ancestors. The same linear invariance test is used although the threshold is set at $\alpha = 0.01$, fol-

lowing Peters et al. (2016). L2-boosting (Friedman, 2001; Bühlmann & Yu, 2003; Hothorn et al., 2010) is used to estimate the MB of size 10 for ICP and MMSE-ICP. Even though fastICP can search exhaustively in simulations of 100 variables, it would be too slow for this problem. Hence, we restrict the fastICP search scope to an estimated MB of 100 variables. ICP-CORTH is excluded because it takes more than 3 days for one gene so obtaining the result for 6170 genes would take more than 18000 days. Instead scoring the methods' output at fixed thresholds, we ranked the set of predicted causal relations and score the most confident predicted relations.

5.3 Results

Figure 6 shows how the number of true positives vary as the methods are allowed to make more predictions. When the number of allowed predictions is lower than 8, the performance of the methods are very similar. However, when the number of predictions is more than 10, MMSE-ICP is generally more accurate than ICP. Moreover, fastICP is the best approach for this task as it can afford to search more widely for the parents of each gene. fastICP obtained 20 true positives when making 30 predictions (66% success probability), setting a new state-of-the-art (SOTA) for this benchmark.

6 Discussion

While using only observational data for causal discovery (CD) is common, interventional data with mechanism changes are very valuable for inferring causal relations. In fact, changes (e.g. different equipments, locations, demographics, weathers) often arise naturally in distribution-shift data. Given their abundance, harnessing changes for CD would reveal insights useful for understanding and manipulating complex systems. However, exploiting changes in general is difficult as they are often imperfect interventions (merely altering the mechanism generating a variable) affecting some unknown variables. In contrast, controlled experiments are perfect interventions that fix designated variables to predetermined values (Tian & Pearl, 2001; Eberhardt & Scheines, 2007). As such, invariance-based CD methods are appealing because they can work with unknown interventions.

Like other invariance-based CD methods, ours also make no assumption about where interventions are and what precisely the effect of interventions may be. We only assume that there is no intervention on the target. This make our approaches appealing in many problems whereby specifying what an intervention or change of environment actually means is difficult. In addition, unlike ICP which needs a sufficient number of interventions to identify all parents (Rosenfeld et al., 2021; Mogensen et al., 2022), our approaches can identify all parents even if they are only indirectly effected by interventions. Thus, our methods can be applied to problems where it is logistically challenging to exhaustively intervene on many variables (Uhler, 2024).

Besides, our work may have implications for robust representation learning. Building ML models with invariant representations has become popular recently since they can generalize better to distribution-shift data (Arjovsky et al., 2019; Rosenfeld et al., 2021). It may be possible to build even more accurate invariant representations by identifying causal variables (Subbaswamy et al., 2019). Thus, as an efficient way to find causal variables from distribution-shift data, fastICP can be a key to more adaptive ML models.

In this work, we proposed two algorithms: MMSE-ICP which has similar runtime but better recall than ICP; and fastICP which is more scalable. Both approaches outperform multiple baselines in simulations. fastICP also achieves SOTA result on the large-scale Kemmeren et al. (2014)'s benchmark. MMSE-ICP and fastICP are based on general theoretical results are orthogonal to the implementations of invariance test and MSE estimation. However, more efficient and accurate tests could lead to more robust and accurate results (e.g. using neural nets or double ML (Chernozhukov et al., 2018)). Finally, despite its speed, fastICP still has worst-case exponential complexity and whether there is a general polynomial-complexity algorithm remains an open question.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. Eprint arXiv:1907.02893, 2019.

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. volume 33, pp. 21865–21877, 2020.

Peter Bühlmann and Bin Yu. Boosting with the l 2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

Jie Cheng, David Bell, and Weiru Liu. Learning bayesian networks from data: An efficient approach based on information theory. *On World Wide Web at http://www. cs. ualberta. ca/˜ jcheng/bnpc. htm*, 1998.

Jie Cheng, Russell Greiner, Jonathan Kelly, David Bell, and Weiru Liu. Learning bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137(1):43–90, 2002. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(02)00191-1. URL `https://www.sciencedirect.com/science/article/pii/S0004370202001911`.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

David Maxwell Chickering. Optimal structure identification with greedy search. *JMLR*, 3(Nov):507–554, 2002.

Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE TPAMI*, 44(10):6614–6630, 2021.

Diego Colombo, Marloes H Maathuis, et al. Order-independent constraint-based causal structure learning. *JMLR*, 15(1):3741–3782, 2014.

Gregory F Cooper and Changwon Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of UAI*, pp. 116–125, 1999.

Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of science*, 74(5): 981–995, 2007.

Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Juan L Gamella and Christina Heinze-Deml. Active invariant causal prediction: Experiment selection through stability. In *Processing of NeurIPS*, volume 33, pp. 15464–15475, 2020.

Tian Gao and Qiang Ji. Local causal discovery of direct causes and effects. volume 28, 2015.

Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Agrin Hilmkil, et al. Deep end-to-end causal inference. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*.

C Glymour, K Zhang, and P Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in Genetics*, 10:524, 2019.

Alain Hauser and Peter Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *JMLR*, 13(1):2409–2464, 2012.

Alain Hauser and Peter Bühlmann. Jointly interventional and observational data: estimation of interventional markov equivalence classes of directed acyclic graphs. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pp. 291–318, 2015.

Yang-Bo He and Zhi Geng. Active learning of causal networks with intervention experiments and optimal designs. *JMLR*, 9(Nov):2523–2547, 2008.

Christina Heinze-Deml, Jonas Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

Torsten Hothorn, Peter Bühlmann, Thomas Kneib, Matthias Schmid, and Benjamin Hofner. Model-based boosting 2.0. *JMLR*, 11(71):2109–2113, 2010.

P Hoyer, D Janzing, J Mooij, J Peters, and B Schölkopf. Nonlinear causal discovery with additive noise models. *Proceedings of NIPS*, 21, 2008.

A Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Dominik Janzing, Joris Mooij, Kun Zhang, Jan Lemeire, Jakob Zscheischler, Povilas Daniušis, Bastian Steudel, and Bernhard Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Neural Causal Structure Discovery from Interventions. *TMLR*, 2023.

Patrick Kemmeren, Katrin Sameith, Loes AL Van De Pasch, Joris J Benschop, Tineke L Lenstra, Thanasis Margaritis, Eoghan O'Duibhir, Eva Apweiler, Sake van Wageningen, Cheuk W Ko, et al. Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, 157 (3):740–752, 2014.

Daphne Koller, Mehran Sahami, et al. Toward optimal feature selection. In *Proceedings of ICML*, volume 96, pp. 292, 1996.

Steffen L Lauritzen. *Graphical models*, volume 17. Clarendon Press, 1996.

Alberto Leon-Garcia. *Probability and random processes for electrical engineering*. Pearson Education India, 1994.

Howard Levene. Robust tests for equality of variances. *Contributions to probability and statistics*, pp. 278–292, 1960.

Honghao Li, Vincent Cabeli, Nadir Sella, and Hervé Isambert. Constraint-based causal structure learning with consistent separating sets. *Proceedings of NIPS*, 32, 2019.

Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints. In *Proceedings of ICLR*, 2022.

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. Dibs: Differentiable bayesian structure learning. volume 34, pp. 24111–24123, 2021.

Sara Magliacane, Thijs Van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. volume 31, 2018.

Dimitris Margaritis and Sebastian Thrun. Bayesian network induction via local neighborhoods. *Proceedings of NIPS*, 12, 1999.

Guillaume G Martinet, Alexander Strzalkowski, and Barbara Engelhardt. Variance minimization in the wasserstein space for invariant causal prediction. In *Proceedings of AISTATS*, pp. 8803–8851. PMLR, 2022.

Phillip B Mogensen, Nikolaj Thams, and Jonas Peters. Invariant ancestry search. In *Proceedings of ICML*, pp. 15832–15857. PMLR, 2022.

Joris M Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *JMLR*, 21(1):3919–4026, 2020.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge university press, 2009.

Judea Pearl. Theoretical impediments to machine learning with seven sparks from the causal revolution. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 3–3, 2018.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *JMLR*, 15(1):2009–2053, 2014.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

Joseph Ramsey, Madelyn Glymour, Ruben Sanchez-Romero, and Clark Glymour. A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. *International journal of data science and analytics*, 3:121–129, 2017.

Alexander G. 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.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *JMLR*, 19(1):1309–1342, 2018.

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russell, Dominik Janzing, Bernhard Schölkopf, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. In *ICML*, pp. 18741–18753. PMLR, 2022.

Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *Proceedings of ICLR*, 2021.

S Shimizu, P Hoyer, A Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *JMLR*, 7(10), 2006.

Ashkan Soleymani, Anant Raj, Stefan Bauer, Bernhard Schölkopf, and Michel Besserve. Causal feature selection via orthogonal search. *TMLR*, 2022.

P Spirtes, C Glymour, R Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

Chandler Squires, Yuhao Wang, and Caroline Uhler. Permutation-based causal structure learning with unknown intervention targets. In *Proceedings of UAI*, pp. 1039–1048. PMLR, 2020.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *Proceedings of AISTATS*, pp. 3118–3127. PMLR, 2019.

Jin Tian and Judea Pearl. Causal discovery from changes. In *Proceedings of UAI*, pp. 512–521, 2001.

Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of KDD*, pp. 673–678, 2003.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

Caroline Uhler. Building a two-way street between cell biology and machine learning. *Nature Cell Biology*, 26(1):13–14, 2024.

Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. volume 30, 2017.

Shuai Yang, Hao Wang, Kui Yu, Fuyuan Cao, and Xindong Wu. Towards efficient local causal structure learning. *IEEE Transactions on Big Data*, 8(6):1592–1609, 2021.

K Zhang and A Hyvärinen. Causality discovery with additive disturbances: An information-theoretical perspective. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 570–585, 2009.

K Zhang, Zhikun Wang, Jiji Zhang, and B Schölkopf. On estimation of functional causal models: general results and application to the post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):1–22, 2015.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In *Proceedings of NeurIPS*, 2018.