



Model-Based Causal Feature Selection for General Response Types

Lucas Kook, Sorawit Saengkyongam, Anton Rask Lundborg, Torsten Hothorn & Jonas Peters

To cite this article: Lucas Kook, Sorawit Saengkyongam, Anton Rask Lundborg, Torsten Hothorn & Jonas Peters (28 Oct 2024): Model-Based Causal Feature Selection for General Response Types, Journal of the American Statistical Association, DOI: [10.1080/01621459.2024.2395588](https://doi.org/10.1080/01621459.2024.2395588)

To link to this article: <https://doi.org/10.1080/01621459.2024.2395588>



© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.



[View supplementary material](#)



Published online: 28 Oct 2024.



[Submit your article to this journal](#)



Article views: 607



[View related articles](#)






[View Crossmark data](#)



Citing articles: 1 [View citing articles](#)

Model-Based Causal Feature Selection for General Response Types

Lucas Kook^a , Sorawit Saengkyongam^b , Anton Rask Lundborg^c , Torsten Hothorn^d , and Jonas Peters^b 

^aInstitute for Statistics and Mathematics, WU Vienna, Vienna, Austria; ^bSeminar for Statistics, ETH Zurich, Zurich, Switzerland; ^cDepartment of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark; ^dEpidemiology, Biostatistics & Prevention Institute, University of Zurich, Zurich, Switzerland

ABSTRACT

Discovering causal relationships from observational data is a fundamental yet challenging task. Invariant causal prediction (ICP, Peters, Bühlmann, and Meinshausen) is a method for causal feature selection which requires data from heterogeneous settings and exploits that causal models are invariant. ICP has been extended to general additive noise models and to nonparametric settings using conditional independence tests. However, the latter often suffer from low power (or poor Type I error control) and additive noise models are not suitable for applications in which the response is not measured on a continuous scale, but reflects categories or counts. Here, we develop transformation-model (TRAM) based ICP, allowing for continuous, categorical, count-type, and uninformatively censored responses (these model classes, generally, do not allow for identifiability when there is no exogenous heterogeneity). As an invariance test, we propose TRAM-GCM based on the expected conditional covariance between environments and score residuals with uniform asymptotic level guarantees. For the special case of linear shift TRAMS, we also consider TRAM-Wald, which tests invariance based on the Wald statistic. We provide an open-source R package **tramicp** and evaluate our approach on simulated data and in a case study investigating causal features of survival in critically ill patients. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

ARTICLE HISTORY

Received July 2023
Accepted August 2024

KEYWORDS

Invariant causal prediction;
Lifetime and survival
analysis; Transformation
model

1. Introduction



1.1. Motivation

Establishing causal relationships from observational data is a common goal in several scientific disciplines. However, systems are often too complex to allow for recovery of the full causal structure underlying the data-generating process. In this work, we consider the easier task of uncovering the causal drivers of a particular response variable of interest. We present methods, theoretical results and user-friendly software for model-based causal feature selection, where the response may represent a binary, ordered, count, or continuous outcome and may additionally be uninformatively censored. We propose TRAMICP for causal feature selection, which is based on invariant causal prediction (ICP, Peters, Bühlmann, and Meinshausen 2016) and a flexible class of regression models, called transformation models (TRAMS, Hothorn, Möst, and Bühlmann 2018). TRAMICP relies on data from heterogeneous environments and the assumption, that the causal mechanism of the response given its direct causes (direct w.r.t. the considered sets of covariates) is correctly specified by a TRAM and does not change across those environments (Haavelmo 1943; Frisch et al. 1948; Aldrich 1989; Pearl 2009; Schölkopf et al. 2012). The causal TRAM will then produce score residuals (residuals defined specifically for TRAMS and potentially censored observations) that are invariant across the environments. If this assumption is violated (for instance, if

the environment, which is not included as a covariate, directly impacts the response) but faithfulness (Spirtes et al. 2000, p. 56) holds, TRAMICP is conservative and will produce an uninformative output. We propose an invariance test based on the expected conditional covariance between the score residuals and the environments given a subset S of the covariates, called TRAM-GCM. With this invariance test, TRAMICP recovers a subset of the direct causes with high probability, by fitting a TRAM for all subsets of covariates, computing score residuals, testing whether those score residuals are uncorrelated with the residualized environments and lastly, intersecting all subsets for which the null hypothesis of invariance was not rejected. For the special case of additive linear TRAMS, we propose another invariance test, TRAM-Wald, based on the Wald statistic for testing whether main and interaction effects involving the environments are zero.

We illustrate the core ideas of TRAMICP in the following example with a binary response and the logistic regression model (McCullagh and Nelder 2019), which is a TRAM. We defer all details on how TRAMS and score residuals are defined to Section 2 and describe the TRAM-GCM and TRAM-Wald invariance tests in Section 3 and Appendix A1, respectively.

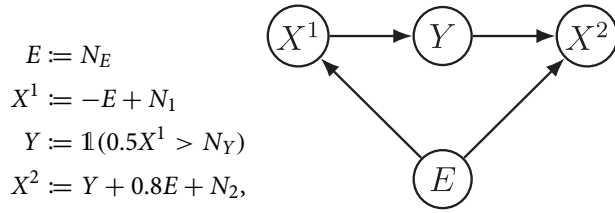
Example 1 (Invariance in binary generalized linear models). Consider the following structural causal model (Pearl 2009) over (Y, X^1, X^2, E) :

CONTACT Lucas Kook  lucasheinrich.kook@gmail.com  Institute for Statistics and Mathematics, WU Vienna, Welthandelsplatz 1, Building D4, AT-1020 Vienna, Austria.

 Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.



where $N_E \sim \text{Bernoulli}(0.5)$, $N_1 \sim N(0, 1)$, $N_2 \sim N(0, 1)$, N_Y are jointly independent noise variables and N_Y follows a standard logistic distribution. Here, E encodes two environments in which the distribution of X^1 and X^2 differ, but the causal mechanism of Y given its direct cause X^1 does not change.

Let us assume that both the above structural causal model and its implied structure are unknown and that we observe an iid sample $\{(e_i, x_i^1, x_i^2, y_i)\}_{i=1}^n$ from the joint distribution of (E, X^1, X^2, Y) . We further know that Y given its direct causes is correctly specified by a logistic regression. All remaining conditionals do not need to satisfy any model assumptions. Our task is now to infer (a subset of) the direct causes of Y .

To do so, for each subset of the covariates X^S , $S \subseteq \{1, 2\}$ (i.e., for \emptyset , $\{1\}$, $\{2\}$, and $\{1, 2\}$), we now (i) fit a binary logistic regression model, (ii) compute the score residuals $y_i - \widehat{\mathbb{P}}(Y = 1 \mid X^S = x_i^S)$ (from the logistic regression) and residualized environments $e_i - \widehat{\mathbb{P}}(E = 1 \mid X^S = x_i^S)$ (via a random forest), and (iii) test whether the two residuals are correlated. Figure 1 shows the residuals obtained in step (iii) for each non-empty subset of the covariates.

In this example, even though the model using $\{X^1, X^2\}$ achieves higher predictive accuracy than the model using the causal parent $\{X^1\}$, only the model $Y \mid X^1$ is stable across the environments. If more than one set is invariant, one can take the intersection of the invariant sets to obtain a subset of the direct causes of Y (Peters, Bühlmann, and Meinshausen 2016).

With our openly available R package **tramicp** (<https://CRAN.R-project.org/package=tramicp>), the analysis in this example can be reproduced with the following code, where `df` is a data frame with 500 independent observations from the structural causal model above.

```

R> library("tramicp")
R> icp <- glmICP(Y ~ X1 + X2, data = df,
  env = ~ E, family = "binomial")
R> pvalues(icp, which = "set")
      Empty      X1      X2    X1+X2
1.82e-02  5.10e-01  4.54e-09  2.22e-03

```

1.2. Related Work

Several algorithms exist to tackle the problem of causal discovery, that is learning the causal graph from data, including constraint-based and score-based methods (Spirtes et al. 2000; Chickering 2002; Pearl 2009; Glymour, Zhang, and Spirtes 2019). Assuming faithfulness, one can hope to recover the causal graph up to the Markov equivalence class (Verma and Pearl 1990; Andersson, Madigan, and Perlman 1997; Tian and Pearl 2001), for which several algorithms have been proposed based on observational data, interventional data, or a combination of both (Spirtes et al. 2000; Chickering 2002; Castelo and Kocka 2003; He and Geng 2008; Hauser and Bühlmann 2015). However, in many real-world applications learning the full causal graph may be too ambitious or unnecessary for tackling the problem at hand. As opposed to causal discovery, causal feature selection aims to identify the direct causes of a given variable of interest (the response) from potentially many measured covariates, instead of the full graph (Guyon, Aliferis, and Elisseeff 2007).

Invariant causal prediction (ICP) is an approach to causal feature selection which exploits invariance of the conditional distribution of a response given its direct causes under perturbations of the covariates (ICP, Peters, Bühlmann, and Meinshausen 2016). ICP can be formulated from a structural causal modeling, as well as potential outcome perspective (Hernán and Robins 2010). In contrast to constraint- and score-based algorithms, ICP requires a specific response variable and data from heterogeneous environments.

ICP builds on the concept of invariance and can generally be formulated as conditional independence between the response and the environments given a candidate set (Heinze-Deml,

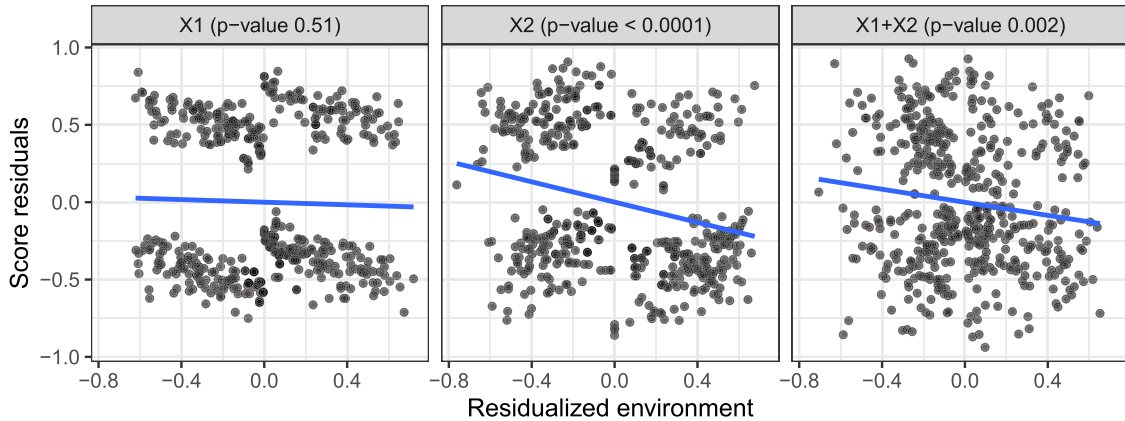


Figure 1. Invariance in binary generalized linear models. By the data generating mechanism in Example 1, we know that the conditional distribution of Y given its direct cause X^1 does not change across the two environments $E = 0$ and $E = 1$. When predicting both Y and E from the three sets of covariates $\{1\}$, $\{2\}$ and $\{1, 2\}$, the resulting residuals are uncorrelated only when conditioning on the invariant set $\{1\}$. The p -values of the invariance test we introduce in Section 3.1.1 are shown in the panel strips for the corresponding subset of covariates (we have also added linear model fits, see blue lines). The empty set is omitted, since the score residuals and residualized environments only take two values.

Peters, and Meinshausen 2018). Thus, nonparametric conditional independence tests (Fukumizu et al. 2007; Zhang et al. 2011; Candès et al. 2018; Strobl, Zhang, and Visweswaran 2019; Berrett et al. 2019) can, in principle, always be applied. However, if one of the conditioning variables is continuous, conditional independence testing is not feasible without further assumptions in the sense that there is no test that simultaneously is level and has nontrivial power (Shah and Peters 2020). This holds also if the environments are discrete (Shah and Peters 2020, Remark 4).

As an alternative to conditional independence testing, model-based formulations of ICP have been formulated for linear (Peters, Bühlmann, and Meinshausen 2016) and non-linear additive noise models (“invariant residual distribution test” proposed in Heinze-Deml, Peters, and Meinshausen 2018). Diaz et al. (2022) use an “invariant target prediction” test from Heinze-Deml, Peters, and Meinshausen (2018) for testing invariance with a binary response by nonparametrically comparing out-of-sample area under the receiver operating characteristic (ROC) curve (AUC). Under correct model specification, model-based ICP can have considerably higher power than its nonparametric alternative. Model-based ICP has been extended to generalized linear models (GLMs, see discussion in Peters, Bühlmann, and Meinshausen 2016) and sequential data (Pfister, Bühlmann, and Peters 2019). ICP for GLMs and additive and multiplicative hazard models has been investigated in Laksafoss (2020). For real-world applications of ICP with exogenous environments, see, for example, Meinshausen et al. (2016), Heinze-Deml, Peters, and Meinshausen (2018), Christiansen and Peters (2020), and Migliavacca et al. (2021).

Many applications feature complex response types, such as ordinal scales, survival times, or counts and the data-generating mechanism can seldomly be assumed to be additive in the noise. This is reflected in the most common model choices for these responses, namely proportional odds logistic (McCullagh 1980; Tutz 2011), Cox proportional hazards (Cox 1972), and generalized linear models (McCullagh and Nelder 2019), which do not assume additive noise in general. Together, noncontinuous responses and nonadditive noise render many causal feature selection algorithms inapplicable. Moreover, proposed extensions to GLMs and hazard-based models rely on case-specific definitions of invariance and thus a unified view on linear, generalized linear, hazards, and general distributional regression is yet to be established.

In practice, a model-based approach can be desirable, because it leads to interpretable effect estimates, such as odds or hazard ratios. However, there is a tradeoff between model intelligibility and misspecification. Many commonly applied regression models are not closed under marginalization or the inclusion or exclusion of covariates that are associated with the response (collapsibility, Greenland 1996; Greenland, Pearl, and Robins 1999; Didelez and Stensrud 2022, see also Appendix A2).

1.3. Summary

Formally, we are interested in discovering the direct causes of a response $Y \in \mathcal{Y} \subseteq \mathbb{R}$ among a potentially large number of covariates $\mathbf{X} \in \mathcal{X}_1 \times \dots \times \mathcal{X}_d \subseteq \mathbb{R}^d$. Consider a set

$S_* \subseteq \{1, \dots, d\}$ (the reader may think about the “direct causes” of Y) and assume that $Y \mid X^{S_*}$ is correctly specified by a TRAM while all other conditionals remain unspecified. In Section 2.2, we define structural causal TRAMS and there, S_* will be the set of causal parents of Y . TRAMS characterize the relationship between features and response via the conditional cumulative distribution function (CDF) $F_{Y \mid X^{S_*} = \mathbf{x}^{S_*}}(y) := \mathbb{P}(Y \leq y \mid X^{S_*} = \mathbf{x}^{S_*})$ on the quantile-scale of a user-specified CDF F_Z . More specifically, when using TRAMS, one models the increasing function $h(\cdot \mid \mathbf{x}^{S_*}) := F_Z^{-1} \circ F_{Y \mid X^{S_*} = \mathbf{x}^{S_*}}(\cdot)$, called a transformation function. The name stems from the fact that for all \mathbf{x}^{S_*} its (generalized) inverse transforms samples of $Z \sim F_Z$ to samples from the conditional distribution $Y \mid X^{S_*} = \mathbf{x}^{S_*}$. Specific choices of F_Z and further modeling assumptions on the functional form of h give rise to many well-known models (examples below). Throughout the article we illustrate TRAMICP with a binary response (Example 2) and give additional examples with a count and survival response in Appendix A3. None of the examples can be phrased as additive noise models of the form $Y = f(X) + \varepsilon$ with $X \perp\!\!\!\perp \varepsilon$. Together with the hardness of conditional independence testing (Shah and Peters 2020, see also above), this motivates the need for causal feature selection algorithms in more flexible nonadditive noise models.

Example 2 (Binary logistic regression). The binary logistic regression model (binomial GLM) with $\mathcal{Y} := \{0, 1\}$ can be phrased in terms of the conditional distribution $F_{Y \mid X^{S_*} = \mathbf{x}^{S_*}}(0) = \text{expit}(\vartheta - (\mathbf{x}^{S_*})^\top \boldsymbol{\beta})$, where $\text{expit}(\cdot) = \text{logit}^{-1}(\cdot) = (1 + \exp(-\cdot))^{-1}$ denotes the standard logistic CDF, and ϑ denotes the baseline ($\mathbf{x}^{S_*} = 0$) log-odds for belonging to class 0 rather than 1. Here, $\boldsymbol{\beta}$ is interpretable as a vector of log odds-ratios. The model can informally be written as $F_{Y \mid X^{S_*} = \mathbf{x}^{S_*}}(y) = \text{expit}(h_Y(y) + (\mathbf{x}^{S_*})^\top \boldsymbol{\beta})$, where $h_Y(0) := \vartheta$ and $h_Y(1) := +\infty$. The latter way of writing the model extends to ordered responses with more than two levels $\mathcal{Y} := \{y_1, y_2, \dots, y_K\}$ with $y_1 < y_2 < \dots < y_K$, $h_Y(y_k) := \vartheta_k$, for all k and for $k = 2, \dots, K$, $\vartheta_k > \vartheta_{k-1}$, using the convention $\vartheta_K = +\infty$ (see McCullagh 1980, proportional odds logistic regression).

In Example 2, we have assumed that the response given its causal parents is correctly specified by a linear shift TRAM (see Definition 5 for more details). If conditioning on a set that is not S_* always yielded a model misspecification, one could attempt to identify the set of causal parents by testing, for different sets X^S of covariates, whether the model for Y given X^S is correctly specified. However, in Proposition 10, we prove that, in general, such a procedure does not work. More precisely, there exists a pair of structural causal models such that both induce the same observational distribution, and in both, the response given its causal parents is correctly specified by an (linear shift) TRAM but the parental sets differ.

In this work, following a line of work in causal discovery (Peters, Bühlmann, and Meinshausen 2016; Meinshausen et al. 2016; Heinze-Deml, Peters, and Meinshausen 2018; Christiansen and Peters 2020), we instead assume to have access to data from heterogeneous environments. Given such data, we define invariance in TRAMS and propose invariance tests based

on the expected conditional covariance between the environments and score residuals (TRAM-GCM) and an invariance test based on the Wald statistic for linear shift TRAMS in particular (TRAM-Wald). We prove that the TRAM-GCM test is uniformly asymptotically level α for any $\alpha \in (0, 1)$ (Theorem 15) and demonstrate empirically that it has power comparable to or higher than nonparametric conditional independence testing. In the context of the result on the hardness of assumption-free conditional independence testing assumptions for continuous distributions (Shah and Peters 2020), our theoretical results show that, under mild assumptions on the relationship between E and X , the model class of TRAMS can be sufficiently restrictive to allow for useful conditional independence tests.

The rest of this article is structured as follows. Section 2.1 gives a technical introduction to transformation models which can be skipped at first reading. We introduce structural causal TRAMS in Section 2.2 and show that in this class, the set of causal parents is, in general, not identified (Section 2.3). In Section 3, we present the proposed TRAM-GCM invariance test and its theoretical guarantees. We apply TRAMICP to discover causal features of survival in critically ill hospitalized patients in Section 4.

2. Using Transformation Models for Causal Inference

Transformation models, as introduced by Box and Cox (1964) in their earliest form, are models for the conditional cumulative distribution function of a response given covariates (Doksum 1974; Bickel and Doksum 1981; Cheng, Wei, and Ying 1995; Hothorn, Kneib, and Bühlmann 2014). TRAMS transform the response conditional on covariates such that the transformed response can be modeled on a fixed, continuous latent scale. Given data and a finite parameterization, the transformation can be estimated via maximum likelihood (Hothorn, Möst, and Bühlmann 2018). We formally define TRAMS as a class of non-linear nonadditive noise models depending on the sample space of both response and covariates. Our treatment of TRAMS may appear overly mathematical; however, the formalism is needed to formulate and prove the identification result (see Proposition 10 in Section 2.3) and the uniform asymptotic level guarantee for the TRAM-GCM invariance test (Theorem 15). A more intuitive introduction to TRAMS can be found in Hothorn, Möst, and Bühlmann (2018), for example. We then embed TRAMS into a causal modeling framework, using structural causal models (SCMs, Pearl 2009; Bongers et al. 2021). We adapt standard results from parametric (Hothorn, Möst, and Bühlmann 2018) and semi-parametric (McLain and Ghosh 2013) maximum likelihood estimation to obtain results on consistency and asymptotic normality, which are exploited by the proposed invariance tests.

2.1. Transformation Models

Let $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}$ denote the extended real line. Throughout the article, let \mathcal{Z} denote the set of functions $F_Z : \overline{\mathbb{R}} \rightarrow [0, 1]$ that are (i) strictly increasing with $\lim_{x \rightarrow -\infty} F_Z(x) = 0$, $\lim_{x \rightarrow \infty} F_Z(x) = 1$, (ii) three-times

differentiable and have a log-concave derivative $f_Z = F'_Z$ when restricted to \mathbb{R} , and (iii) satisfy $F_Z(-\infty) = 0$ and $F_Z(+\infty) = 1$. We call \mathcal{Z} the set of *extended differentiable cumulative distribution functions*. Given that a CDF $F : \mathbb{R} \rightarrow \mathbb{R}$ satisfies (i) and (ii), we may add (iii) and refer to the resulting function as an *extended CDF*. For instance, the extended standard logistic CDF is given by $F_{\text{SL}}(z) = (1 + \exp(-z))^{-1}$ for all $z \in \mathbb{R}$ and $F_{\text{SL}}(-\infty) = 0$ and $F_{\text{SL}}(+\infty) = 1$. Besides F_{SL} , in our applications, we consider the extended versions of the standard normal CDF Φ , and the standard minimum extreme value CDF $F_{\text{minEV}} : z \mapsto 1 - \exp(-\exp(z))$. By slight abuse of notation, we use the same letters $\Phi, F_{\text{SL}}, F_{\text{minEV}}$, for the extended CDFs. In general, specification of a transformation model requires choosing a particular $F_Z \in \mathcal{Z}$. Further, for a symmetric positive semidefinite matrix A , let $\lambda_{\min}(A)$ denote its smallest eigenvalue and $\|A\|_{\text{op}}$ denote its operator norm. For all $n \in \mathbb{N}$, we write $[n]$ as shorthand for $\{1, \dots, n\}$.

We call a function $h : \mathbb{R} \rightarrow \overline{\mathbb{R}}$ *extended right-continuous and increasing* (ERCI) on $\mathcal{Y} \subseteq \mathbb{R}$ if (i) it is right-continuous and strictly increasing on \mathcal{Y} and fulfills $h(\min \mathcal{Y}) > -\infty$ (if $\min \mathcal{Y}$ exists), (ii) for all $y < \inf \mathcal{Y}$, we have $h(y) = -\infty$, (iii) for all $y > \sup \mathcal{Y}$, we have $h(y) = +\infty$, (iv) for all $t \in (\inf \mathcal{Y}, \sup \mathcal{Y}) \setminus \mathcal{Y}$, we have $h(t) = h(\underline{t})$, where $\underline{t} := \sup\{v \in \mathcal{Y} : v < t\}$ and (v) $\lim_{v \rightarrow -\infty} h(v) = -\infty$ and $\lim_{v \rightarrow \infty} h(v) = \infty$. Condition (iv) is needed to ensure that h is piece-wise constant outside of \mathcal{Y} . Finally, for a function $f : \overline{\mathbb{R}} \rightarrow \mathbb{R}$, we denote the derivative $f' : \mathbb{R} \rightarrow \mathbb{R}$ s.t. for all $x \in \mathbb{R}$, $f'(x) = \frac{d}{du}f(u)|_{u=x}$. We are now ready to define the class of transformation models.

Definition 3 (Transformation model). Let $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{X} := \mathcal{X}_1 \times \dots \times \mathcal{X}_d \subseteq \mathbb{R}^d$, where for all i , $\mathcal{X}_i \subseteq \mathbb{R}$. The set of all *transformation functions* on $\mathcal{Y} \times \mathcal{X}$ is defined as

$$\mathcal{H}_{\mathcal{Y}, \mathcal{X}}^* := \left\{ h : \mathbb{R} \times \mathcal{X} \rightarrow \overline{\mathbb{R}} \mid \forall \mathbf{x} \in \mathcal{X}, h(\cdot \mid \mathbf{x}) \text{ is ERCI on } \mathcal{Y} \right\}.$$

Then, for a fixed *error distribution* $F_Z \in \mathcal{Z}$ and a set of transformation functions $\mathcal{H}_{\mathcal{Y}, \mathcal{X}} \subseteq \mathcal{H}_{\mathcal{Y}, \mathcal{X}}^*$, the *family of TRAMS* $\mathcal{M}(F_Z, \mathcal{Y}, \mathcal{X}, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ is defined as the following set of conditional cumulative distribution functions¹ (see also Definition 2 in Hothorn, Möst, and Bühlmann 2018):

$$\begin{aligned} \mathcal{M}(F_Z, \mathcal{Y}, \mathcal{X}, \mathcal{H}_{\mathcal{Y}, \mathcal{X}}) := & \left\{ F_{Y|X=} : \mathbb{R} \times \mathcal{X} \rightarrow [0, 1] \mid \right. \\ & \exists h \in \mathcal{H}_{\mathcal{Y}, \mathcal{X}} : \forall \mathbf{x} \in \mathcal{X} \forall y \in \mathbb{R}, \\ & \left. F_{Y|X=\mathbf{x}}(y) = F_Z(h(y \mid \mathbf{x})) \right\}. \end{aligned}$$

As such, a single TRAM is fully specified by (F_Z, h) , $F_Z \in \mathcal{Z}$, $h \in \mathcal{H}_{\mathcal{Y}, \mathcal{X}}$. The condition that for all $\mathbf{x} \in \mathcal{X}$, $h(\cdot \mid \mathbf{x})$ is ERCI on \mathcal{Y} ensures that the support of the induced conditional distribution specified by $F_{Y|X=\mathbf{x}}$ is \mathcal{Y} . Further, for all $\mathbf{x} \in \mathcal{X}$ and $z \in \overline{\mathbb{R}}$, we write $h^{-1}(z \mid \mathbf{x}) := \inf\{y \in \mathcal{Y} : z \leq h(y \mid \mathbf{x})\}$ for the inverse transformation function.

¹In Proposition 27 in Appendix E2, we show that \mathcal{M} indeed only contains CDFs.

The inverse transformation function $h^{-1}(\cdot | \mathbf{x})$ at a given \mathbf{x} can be interpreted analogously to a quantile function: Given some $X = x$, we can obtain an observation from $F_{Y|X=x}$ by sampling an observation from F_Z and passing it through $h^{-1}(\cdot | \mathbf{x})$.

In statistical modeling, it is common to additionally assume additivity of the effects of X on a specific scale. For instance, in linear regression the covariates enter as a linear predictor on the scale of the conditional mean. In this work, we restrict ourselves to the class of shift TRAMS in which additivity is assumed on the scale of the transformation function.

Definition 4 (Shift TRAMS). Let \mathcal{Y} , \mathcal{X} , and $F_Z \in \mathcal{Z}$ be as in Definition 3. Further, let $\mathcal{F} := \{f : \mathcal{X} \rightarrow \mathbb{R} \mid f \text{ measurable}\}$ and $\mathcal{H}_{\mathcal{Y}} := \{h_Y : \mathbb{R} \rightarrow \overline{\mathbb{R}} \mid h_Y \text{ is ERCI on } \mathcal{Y}\}$. Let the set of *shift transformation functions* be defined as

$$\mathcal{H}_{\mathcal{Y}, \mathcal{X}}^{\text{shift}} := \left\{ h \in \mathcal{H}_{\mathcal{Y}, \mathcal{X}}^* \mid \exists h_Y \in \mathcal{H}_{\mathcal{Y}}, \right. \\ \left. f \in \mathcal{F} : \forall \mathbf{x} \in \mathcal{X}, h(\cdot | \mathbf{x}) = h_Y(\cdot) - f(\mathbf{x}) \right\}.$$

Then, $\mathcal{M}(F_Z, \mathcal{Y}, \mathcal{X}, \mathcal{H}_{\mathcal{Y}, \mathcal{X}}^{\text{shift}})$ denotes the family of *shift TRAMS* and a TRAM $F_Z \circ h$ is called *shift TRAM* iff $h \in \mathcal{H}_{\mathcal{Y}, \mathcal{X}}^{\text{shift}}$. Further, any $h_Y \in \mathcal{H}_{\mathcal{Y}}$ is referred to as a *baseline transformation*.

We next introduce the subset of linear shift TRAMS in which the covariates enter as a linear predictor.

Definition 5 (Linear shift TRAMS). Consider shift TRAMS specified by $F_Z, \mathcal{Y}, \mathcal{X}, \mathcal{F}, \mathcal{H}_{\mathcal{Y}, \mathcal{X}}^{\text{shift}}$, as in Definition 4. Let $\mathbf{b} : \mathcal{X} \rightarrow \mathbb{R}^b$ be a finite collection of basis transformations and define $\mathcal{F}_{\mathbf{b}} := \{f \in \mathcal{F} \mid \exists \boldsymbol{\beta} \in \mathbb{R}^b \text{ s.t. } f(\cdot) = \mathbf{b}(\cdot)^\top \boldsymbol{\beta}\}$. The set of *linear shift transformation functions w.r.t. \mathbf{b}* is defined as

$$\mathcal{H}_{\mathcal{Y}, \mathcal{X}}^{\text{linear}}(\mathbf{b}) := \left\{ h \in \mathcal{H}_{\mathcal{Y}, \mathcal{X}}^{\text{shift}} \mid \exists h_Y \in \mathcal{H}_{\mathcal{Y}}, \right. \\ \left. f \in \mathcal{F}_{\mathbf{b}} : \forall \mathbf{x} \in \mathcal{X} : h(\cdot | \mathbf{x}) = h_Y(\cdot) - f(\mathbf{x}) \right\}.$$

Then, $\mathcal{M}(F_Z, \mathcal{Y}, \mathcal{X}, \mathcal{H}_{\mathcal{Y}, \mathcal{X}}^{\text{linear}}(\mathbf{b}))$ denotes the family of *linear shift TRAMS w.r.t. \mathbf{b}* . Further, a TRAM $F_Z \circ h$ is called *linear shift TRAM w.r.t. \mathbf{b}* iff $h \in \mathcal{H}_{\mathcal{Y}, \mathcal{X}}^{\text{linear}}(\mathbf{b})$. For the special case of $\mathbf{b} : \mathbf{x} \mapsto \mathbf{x}$, we write $\mathcal{H}_{\mathcal{Y}, \mathcal{X}}^{\text{linear}}$ and refer to the class and its members as *linear shift TRAMS*.

Estimation and inference in TRAMS can be based on the log-likelihood function—if it exists. The following assumption ensures that this is the case.

Assumption 1. We have $\mathcal{H}_{\mathcal{Y}, \mathcal{X}} \subseteq \mathcal{H}_{\mathcal{Y}, \mathcal{X}}^{\text{shift}}$. Furthermore, if \mathcal{Y} is uncountable, $F_Z, \mathcal{X}, \mathcal{H}_{\mathcal{Y}, \mathcal{X}}$ are such that for all $\mathbf{x} \in \mathcal{X}$ and $h \in \mathcal{H}_{\mathcal{Y}, \mathcal{X}}$,

$$f_{Y|X=x}(\cdot; h) := F'_Z(h(\cdot | \mathbf{x}))h'(\cdot | \mathbf{x}), \quad (1)$$

where $h'(y | \mathbf{x}) := \frac{d}{dv} h(v | \mathbf{x})|_{v=y}$, is well-defined and a density (w.r.t. Lebesgue measure) of the conditional CDF induced by the TRAM.

Assumption 1 allows us to define (strictly positive) canonical conditional densities with respect to a fixed measure that we

denote by μ : If \mathcal{Y} is countable, we let μ denote the counting measure on \mathcal{Y} and for all $y \in \mathcal{Y}$, define the canonical conditional density by $f_{Y|X=x}(y; h) := F_Z(h(y | \mathbf{x})) - F_Z(h(y | \mathbf{x}))$, where $y := \sup\{v \in \mathcal{Y} : v < y\}$.² If \mathcal{Y} is uncountable, we let μ denote the Lebesgue measure restricted to \mathcal{Y} and the canonical conditional density is then defined by (1). In either case, $\mathcal{H}_{\mathcal{Y}, \mathcal{X}} \subseteq \mathcal{H}_{\mathcal{Y}, \mathcal{X}}^{\text{shift}}$ ensures that for all \mathbf{x} and $y \in \mathcal{Y}$, $f_{Y|X=x}(y; h) > 0$. Thus, for $(F_Z, \mathcal{Y}, \mathcal{X}, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ satisfying **Assumption 1**, we can define the TRAM log-likelihood as $\ell : \mathcal{H}_{\mathcal{Y}, \mathcal{X}} \times \mathcal{Y} \times \mathcal{X} \rightarrow \mathbb{R}$ with $\ell(h; y, \mathbf{x}) := \log f_{Y|X=x}(y; h)$.

When applying ICP to linear additive noise models, invariance can be formulated as uncorrelatedness between residuals and environments. In TRAMS, however, the response can be categorical, reducing the usefulness of classical residuals. Instead, score residuals (Lagakos 1981; Korepanova et al. 2020; Kook, Sick, and Bühlmann 2022) are a natural choice for testing invariance of TRAMS. Score residuals were first introduced by Lagakos (1981) for multiplicative hazard models (see also Korepanova et al. 2020, for non-multiplicative hazard models) and extended to linear shift TRAMS by Kook, Sick, and Bühlmann (2022, Definition 2). Score residuals coincide with scaled least-squares residuals in linear regression with normal errors and martingale residuals in the Cox proportional hazards model (Barlow and Prentice 1988) and directly extend to censored responses (Lagakos 1981; Farrington 2000). In this work, score residuals play a major role in formulating invariance tests (Section 3) and have been used for causal regularization in a distributional version of anchor regression (Rothenhäusler et al. 2021; Kook, Sick, and Bühlmann 2022). For defining score residuals, we require the following assumption (which, by definition, is satisfied for $\mathcal{H}_{\mathcal{Y}, \mathcal{X}}^{\text{shift}}$ and $\mathcal{H}_{\mathcal{Y}, \mathcal{X}}^{\text{linear}}$).

Assumption 2. $\mathcal{H}_{\mathcal{Y}, \mathcal{X}}$ is closed under scalar addition, that is, for all $h \in \mathcal{H}_{\mathcal{Y}, \mathcal{X}}$ and $\alpha \in \mathbb{R}$, we have³ $h + \alpha \in \mathcal{H}_{\mathcal{Y}, \mathcal{X}}$.

Definition 6 (Score residuals, Lagakos, 1981; Kook, Sick, and Bühlmann, 2022). Let $\mathcal{Y}, \mathcal{X}, F_Z \in \mathcal{Z}$ and $\mathcal{H}_{\mathcal{Y}, \mathcal{X}} \subseteq \mathcal{H}_{\mathcal{Y}, \mathcal{X}}^*$ be as in Definition 3. Impose **Assumptions 1** and **2**. Then, considering $\alpha \in \mathbb{R}$, the *score residual* $R : \mathcal{H}_{\mathcal{Y}, \mathcal{X}} \times \mathcal{Y} \times \mathcal{X} \rightarrow \mathbb{R}$ is defined as

$$R : (h; y, \mathbf{x}) \mapsto \frac{\partial}{\partial \alpha} \ell(h + \alpha; y, \mathbf{x}) \Big|_{\alpha=0}.$$

Example 7 (Binary logistic regression, cont'd). The family of binary linear shift logistic regression models is given by $\mathcal{M}(F_{\text{SL}}, \{0, 1\}, \mathcal{X}, \mathcal{H}_{\mathcal{Y}, \mathcal{X}}^{\text{linear}})$. We can thus write for all $\mathbf{x} \in \mathcal{X}$, $h(\cdot | \mathbf{x}) := h_Y(\cdot) - \mathbf{x}^\top \boldsymbol{\beta}$ with $h_Y(0) := \vartheta$ and, by convention, $h_Y(1) := +\infty$. The likelihood contribution for a given observation (y, \mathbf{x}) is $F_{\text{SL}}(h(0 | \mathbf{x}))^{1-y} (1 - F_{\text{SL}}(h(0 | \mathbf{x})))^y$. The score residual is given by $R(h; y, \mathbf{x}) = 1 - y - F_{\text{SL}}(h(0 | \mathbf{x}))$. Further, the inverse transformation function is given by $h^{-1} : (z, \mathbf{x}) \mapsto \mathbb{1}(z \geq \vartheta - \mathbf{x}^\top \boldsymbol{\beta})$.

2.2. Structural Causal Transformation Models

Next, we cast TRAMS into a structural causal modeling framework (Pearl 2009) and return to our examples from Section 1.

²We adopt the convention that the supremum of the empty set is $-\infty$.

³We adopt the convention that for all $\alpha \in \mathbb{R}$, $-\infty + \alpha = -\infty$ and $\infty + \alpha = \infty$.

For all subsets $S \subseteq [d]$, define \mathcal{X}^S to be the projection of \mathcal{X} onto the ordered coordinates in S . For the rest of this article, we restrict ourselves to shift TRAMS. In this case, any “global” model class $\mathcal{H}_{\mathcal{Y},\mathcal{X}}$ naturally induces submodel classes $\mathcal{H}_{\mathcal{Y},\mathcal{X}^S} \subseteq \mathcal{H}_{\mathcal{Y},\mathcal{X}^S}^*$ for all $S \subseteq [d]$ by the following construction: $\mathcal{H}_{\mathcal{Y},\mathcal{X}^S} := \{h \in \mathcal{H}_{\mathcal{Y},\mathcal{X}^S}^* \mid \exists h^{\text{global}} \in \mathcal{H}_{\mathcal{Y},\mathcal{X}} \text{ s.t. } \forall \mathbf{x} \in \mathcal{X}, h^{\text{global}}(\cdot \mid \mathbf{x}) = h(\cdot \mid \mathbf{x}^S)\}$. If $(F_Z, \mathcal{Y}, \mathcal{X}, \mathcal{H}_{\mathcal{Y},\mathcal{X}})$ satisfies [Assumption 1](#), then $(F_Z, \mathcal{Y}, \mathcal{X}^S, \mathcal{H}_{\mathcal{Y},\mathcal{X}^S})$ does too. We are now ready to define structural causal TRAMS.

Definition 8 (Structural causal TRAM). Let $\mathcal{Y}, \mathcal{X}, F_Z \in \mathcal{Z}$ be as in [Definition 3](#). Let $\mathcal{H}_{\mathcal{Y},\mathcal{X}} \subseteq \mathcal{H}_{\mathcal{Y},\mathcal{X}}^*$ be a class of transformation functions such that [Assumption 1](#) holds. Let (Z, N_X) be jointly independent with $Z \sim F_Z$. Then, a *structural causal TRAM* C over (Y, X) is defined as

$$C := \begin{cases} X^j := g_j(X, Y, N_{X^j}), & \forall j \in [d] \\ Y := h^{-1}(Z \mid X^{S_*}), \end{cases} \quad (2)$$

where $S_* \subseteq [d]$, $h \in \mathcal{H}_{\mathcal{Y},\mathcal{X}^{S_*}}$ is the *causal transformation function* and $\text{pa}_C(Y) := S_*$ denotes the set of causal parents of Y in C and $g_j, j \in [d]$, are arbitrary measurable functions. By $\mathbb{P}_{(Y,X)}^C$ we denote the observational distribution induced by C . We assume that the induced graph (obtained by drawing directed edges from the observed variables on the right-hand side to variables on the left-hand side) is acyclic. (This, in particular, implies that the function g_j does not depend on X^j .) We denote by $\mathcal{C}(F_Z, \mathcal{Y}, \mathcal{X}, \mathcal{H}_{\mathcal{Y},\mathcal{X}})$ the collection of all structural causal TRAMS with error distribution F_Z and causal transformation function $h \in \mathcal{H}_{\mathcal{Y},\mathcal{X}}$.

2.3. Non-Identifiability of the Causal Parents in Transformation Models

We now show that performing causal feature selection in structural causal transformation models requires further assumptions. We consider a response variable Y and a set of covariates X and assume that (Y, X) are generated from an (unknown) structural causal TRAM (defined in (2)) with (known) $\mathcal{H}_{\mathcal{Y},\mathcal{X}} \subsetneq \mathcal{H}_{\mathcal{Y},\mathcal{X}}^*$. In our work, the problem of causal feature selection concerns learning the causal parents $\text{pa}(Y)$ given a sample of (Y, X) and knowledge of $F_Z, \mathcal{Y}, \mathcal{X}, \mathcal{H}_{\mathcal{Y},\mathcal{X}}$ (which specifies the model class $\mathcal{M}(F_Z, \mathcal{Y}, \mathcal{X}, \mathcal{H}_{\mathcal{Y},\mathcal{X}})$).

In this work, we specify the model class for the conditional of the response, given its causal parents, $Y \mid X^{\text{pa}(Y)}$ by a TRAM; the remaining conditionals are unconstrained. Identifiability of causal structure has been studied for several model classes that constrain the joint distribution (Y, X) . When considering the class of linear Gaussian SCMs, for example, the causal parents are in general not identifiable from the observational distribution (as there are linear Gaussian SCMs with a different structure inducing the same distribution). This is different for other model classes: When considering linear Gaussian SCMs with equal noise variances (Peters and Bühlmann 2013), linear non-Gaussian SCMs (Shimizu 2014) or nonlinear Gaussian SCMs (Hoyer et al. 2008; Peters et al. 2014), for example, the graph

structure (and thus the set of causal parents of Y) is identifiable under weak assumptions (identification then becomes possible by using goodness-of-fit procedures). To the best of our knowledge, identifiability in such model classes (i.e., recovering the causal parents of Y , not the entire graph or equivalence classes) has not been studied when constraining only the conditional distribution of Y given $X^{\text{pa}(Y)}$.

TRAMS are generally not closed under marginalization (see [Appendix A2](#) for a detailed discussion on non-collapsability) and one may hypothesize that this model class allows for identifiability of the parents (e.g., by considering different subsets of covariates and testing for goodness of fit). We now prove that this is not the case: In general, for TRAMS (and even for linear shift TRAMS), the causal parents are not identifiable from the observed distribution. Instead, additional assumptions are needed to facilitate causal feature selection in TRAMS.

[Definition 9](#) formally introduces the notion of identifiability of the causal parents and [Proposition 10](#) provides the non-identifiability result.

Definition 9 (Subset-identifiability of the causal parents). Let \mathcal{C} denote a collection of structural causal models. The set of causal parents is said to be *\mathcal{C} -subset-identifiable* if for all pairs $C_1, C_2 \in \mathcal{C}$ it holds that

$$\mathbb{P}_{(Y,X)}^{C_1} = \mathbb{P}_{(Y,X)}^{C_2} \implies \text{pa}_{C_1}(Y) \subseteq \text{pa}_{C_2}(Y) \vee \text{pa}_{C_2}(Y) \subseteq \text{pa}_{C_1}(Y).$$

Proposition 10 (Non-subset-identifiability). For all $A \subseteq \mathbb{R}$ that are either an interval or countable, $F_Z \in \mathcal{Z}, \mathcal{Y} \subseteq \mathbb{R}$, there exists a class of transformation functions $\mathcal{H}_{\mathcal{Y},A \times A} \subseteq \mathcal{H}_{\mathcal{Y},A \times A}^{\text{shift}} \subsetneq \mathcal{H}_{\mathcal{Y},A \times A}^*$, such that the set of causal parents is not $\mathcal{C}(F_Z, \mathcal{Y}, A \times A, \mathcal{H}_{\mathcal{Y},A \times A})$ -subset identifiable.

A proof is given in [Appendix E1.1](#), where we construct a joint distribution over three random variables (Y, X^1, X^2) , in which the two conditionals $Y \mid X^1$ and $Y \mid X^2$ are TRAMS. This implies that there are two structural causal TRAMS that have identical observational distributions, while Y has two different (non-empty) sets of causal parents that do not overlap. The proof in [Appendix E1.1](#) characterizes how to construct such a joint distribution for shift TRAMS. For illustrative purposes, we present a concrete example in [Appendix E1.1](#) in which $\mathcal{Y} = \mathcal{X}^1 = \mathcal{X}^2 = \{1, 2, 3\}$ and $Y \mid X^1$ and $Y \mid X^2$ are proportional odds logistic regression models. We then sample from the induced distribution. We sample from the induced distributions of the two structural causal TRAMS constructed in the proof and apply the naive method described above of performing goodness-of-fit tests to identify the parents. We see that this method indeed fails to identify a non-empty subset of the parents in this example.

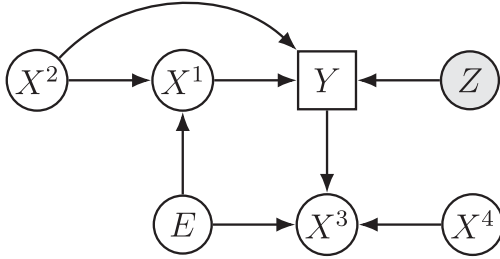
Instead of subset-identifiability, one can also consider a stronger notion of *full identifiability*, which states that the set of causal parents can be uniquely determined by the observed distribution (formally defined in [Appendix A4](#)). [Proposition 10](#) immediately implies that the set of causal parents is not fully identifiable either.

3. Transformation Model Invariant Causal Prediction

Even if the observational distribution is insufficient to identify causal parents, identifiability can become possible if we have access to data from multiple, heterogeneous environments. Invariant causal prediction (ICP, Peters, Bühlmann, and Meinshausen 2016) exploits the invariance of causal mechanisms (Haavelmo 1943; Frisch et al. 1948; Aldrich 1989; Pearl 2009; Schölkopf et al. 2012) under interventions on variables other than the response. Depending on the response variable, multi-center clinical trials, data collected from different countries or different points in time may fall into this category. We then show that under Setting 1, the set of causal parents is subset-identifiable (Proposition 12) and fully identifiable if the environments are sufficiently heterogeneous (Proposition 13).

Setting 1 (Data from multiple environments). Let $\mathcal{Y}, \mathcal{X}, F_Z \in \mathcal{Z}$ be as in Definition 3 and let $\mathcal{H}_{\mathcal{Y}, \mathcal{X}} \subseteq \mathcal{H}_{\mathcal{Y}, \mathcal{X}}^*$ be a class of transformation functions such that Assumptions 1 and 2 hold. Let C_* be a structural causal TRAM (Definition 8) over (Y, \mathbf{X}, E) such that

$$C_* := \begin{cases} E^k := m_k(\mathbf{X}, N_{E^k}), & \forall k \in [q] \\ X^j := g_j(\mathbf{X}, \mathbf{E}, Y, N_{X^j}), & \forall j \in [d] \\ Y := h_*^{-1}(Z | \mathbf{X}^{S_*}), \end{cases}$$



where $h_* \in \mathcal{H}_{\mathcal{Y}, \mathcal{X}^{S_*}}$ with $S_* \subseteq [d]$ denoting the parents of Y and (Z, N_X, N_E) denoting the jointly independent noise variables. By definition, the induced graph \mathcal{G}_* (containing the variables $E^1, \dots, E^q, X^1, \dots, X^d, Y$) is acyclic. In this setup, the random vector \mathbf{E} encodes the environments and takes values in $\mathcal{E} \subseteq \mathbb{R}^q$. We further assume that the parents of \mathbf{E} can only be non-descendants⁴ of Y in \mathcal{G}_* , which is satisfied, for example, if \mathbf{E} is exogenous (that is, each E^k is a function of N_{E^k} only); \mathbf{E} may be discrete or continuous. An example of a DAG contained in this setup is depicted above. By $\mathcal{D}_n := \{(y_i, \mathbf{x}_i, \mathbf{e}_i)\}_{i=1}^n$, we denote an iid sample from $\mathbb{P}_{(Y, \mathbf{X}, E)}^{C_*}$.

As for ICP, invariance plays a key role for TRAMICP. We say a subset of covariates is invariant if the corresponding transformation model correctly describes the conditional distribution across the environments \mathbf{E} . More formally, we have the following definition.

Definition 11 ($(F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ -invariance). Assume Setting 1. A subset of covariates $S \subseteq [d]$ is $(F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ -invariant if there

exists $h^S \in \mathcal{H}_{\mathcal{Y}, \mathcal{X}^S}$, such that for $\mathbb{P}_{(X^S, E)}$ -almost all $(\mathbf{x}^S, \mathbf{e})$,

$(Y | \mathbf{X}^S = \mathbf{x}^S, \mathbf{E} = \mathbf{e})$ and $(Y | \mathbf{X}^S = \mathbf{x}^S)$ are identical with conditional CDF $F_Z(h^S(\cdot | \mathbf{x}^S))$.

If an *invariant transformation function* h^S according to Definition 11 exists, it is \mathbb{P}_{X^S} -almost surely unique (see Lemma 31 in Appendix E2). Proposition 12 shows that the parental set fulfills $(F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ -invariance, which is sufficient to establish coverage guarantees for invariant causal prediction in TRAMS. A proof is given in Appendix E1.2.

Proposition 12. Assuming Setting 1, the set of causal parents S_* is $(F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ -invariant.

The set of causal parents S_* together with the causal transformation function h_* in Setting 1 may not be the only set satisfying $(F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ -invariance. In this vein, we define the set of *identifiable causal predictors* as

$$S_I := \bigcap_{S \subseteq [d]: S \text{ is } (F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})\text{-invariant}} S.$$

Since $(F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ -invariance is a property of the observed distribution, S_I is identifiable from the observed distribution, too. By Proposition 12, $S_I \subseteq S_*$. Thus, the causal parents S_* are subset-identifiable. In a modified version of Setting 1 in which \mathbf{E} is among the causal parents of Y (see Setting 2 in Appendix A5) and the induced distribution is faithful w.r.t. the induced graph (see Spirtes et al. 2000, p. 56), the set of identifiable causal predictors is empty (because there exists no $(F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ -invariant set; we prove this statement as Proposition 20 in Appendix A5).

Furthermore, if the environments induce a sufficient amount of heterogeneity in the data, in the sense that $S_* \subseteq \text{ch}(\mathbf{E})$, then $S_I = S_*$, so the causal parents are fully identified (this result assumes faithfulness).⁵

Proposition 13. Assume Setting 1. Let \mathcal{G} be the DAG induced by C_* and assume that $\mathbb{P}_{(Y, \mathbf{X}, E)}^{C_*}$ is faithful w.r.t. to \mathcal{G} . If $S_* \subseteq \text{ch}(\mathbf{E})$, where $\text{ch}(\mathbf{E})$ denotes the children of \mathbf{E} , we have $S_I = S_*$.

A proof is given in Appendix E1.3. For simple model classes such as linear Gaussian SCMs, sufficient conditions for faithfulness are known (Spirtes et al. 2000). In our setting, analyzing the faithfulness assumption is particularly challenging due to non-collapsibility and non-closure under marginalization of TRAMS (see Appendix A2). Nonetheless, we empirically show in our simulations (see Appendix B3) that faithfulness is not violated, for example, if the coefficients in linear shift TRAMS are sampled from a continuous distribution.

3.1. Testing for Invariance

We now translate $(F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ -invariance into testable conditions which are applicable to general TRAMS and thus general

⁴A node is called a non-descendant of Y in \mathcal{G}_* if there is no directed path from Y to that node in \mathcal{G}_* .

⁵Strictly speaking, assuming faithfulness for the whole graph when proving Proposition 13 is too strong. As can be seen from the proof, it suffices to assume that for all $S \subseteq [d]$, we have \mathbf{E} is not d -separated from Y given \mathbf{X}^S implies \mathbf{E} is not independent of Y given \mathbf{X}^S .

response types. Here, we propose an invariance condition based on score residuals (Definition 6). The following proposition shows that the score residuals are uncorrelated with the environments (in Setting 1) when conditioning on an invariant set.

Proposition 14 (Score-residual-invariance). Assume Setting 1 and that (8) in Appendix E2 holds. Then, we have the following implication:

$$S \text{ is } (F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})\text{-invariant} \implies \mathbb{E}[R(h^S; Y, \mathbf{X}^S) \mid \mathbf{X}^S] = 0, \text{ and} \\ \mathbb{E}[\text{cov}[\mathbf{E}, R(h^S; Y, \mathbf{X}^S) \mid \mathbf{X}^S]] = 0, \quad (3)$$

where $\mathbb{E}[\text{cov}[\mathbf{E}, R(h^S; Y, \mathbf{X}^S) \mid \mathbf{X}^S]] := \mathbb{E}[\mathbf{E}R(h^S; Y, \mathbf{X}^S) \mid \mathbf{X}^S] - \mathbb{E}[\mathbf{E} \mid \mathbf{X}^S]\mathbb{E}[R(h^S; Y, \mathbf{X}^S) \mid \mathbf{X}^S]$ denotes the expected conditional covariance between the residuals and environments.

A proof is given in Appendix E1.4. In Appendix A6, we extend TRAMICP (in particular, Proposition 14) to uninformatively censored observations, where Y itself is unobserved.

We now turn to the problem of testing invariance from finite data. Section 3.1.1 develops a test, based on similar ideas as the Generalised Covariance Measure (GCM, Shah and Peters 2020), on how to test the implication in (3). As a second, alternative, invariance test, we also propose a Wald test for the existence of main and interaction terms involving the environments in Appendix A1; we show in Proposition 16 that for linear shift TRAMs, such a test is closely related to the implication in Proposition 14.

For all $S \subseteq [d]$, and sample sizes n , let $p_{S,n} : (\mathbb{R} \times \mathcal{X}^S \times \mathcal{E})^n \rightarrow [0, 1]$ be the p -value for the null hypothesis that S is $(F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ -invariant. All proposed invariance tests are embedded in a subset-search over the set of covariates, in which we return the intersection of all non-rejected sets at a given level $\alpha \in (0, 1)$ (ICP; Algorithm 1).

Algorithm 1 Invariant causal prediction (Peters, Bühlmann, and Meinshausen 2016)

Require: Data \mathcal{D}_n from Setting 1, significance level $\alpha \in (0, 1)$, and a family of invariance tests $(p_{S,n})_{S \subseteq \{1, \dots, d\}}$ (outputting a p -value; see Algorithms 2, and A1 and the comparators in Appendix B1)

- 1: For all $S \subseteq [d]$, compute $p_{S,n}(\mathcal{D}_n)$ \triangleright Compute p -value of invariance test
 - 2: **return** $S_n := \bigcap_{S: p_{S,n}(\mathcal{D}_n) > \alpha} S$ \triangleright Intersection over all non-rejected sets
-

It directly follows from Proposition 12 that if the tests are level α , then the output of Algorithm 1 is contained in the causal parents with large probability (see Peters, Bühlmann, and Meinshausen 2016, Theorem 1), that is, $\mathbb{P}(S_n \subseteq \text{pa}_{C_*}(Y)) \geq 1 - \alpha$.⁶ This coverage guarantee does not require faithfulness or sufficiently heterogeneous environments as assumed in Proposition 13.⁷ It only requires that the environment is a measurable function of non-descendants of Y and is not a causal parent

⁶The coverage guarantee holds by $\mathbb{P}(S_n \subseteq \text{pa}_{C_*}(Y)) \geq \mathbb{P}(p_{S^*,n}(\mathcal{D}_n) > \alpha) = 1 - \alpha$.

⁷If the test had perfect power, then under the conditions assumed in Proposition 13, the procedure would output C^* . In practice, even under the

Algorithm 2 TRAM-GCM invariance test

Require: Data \mathcal{D}_n from Setting 1, $S \subseteq [d]$, estimator $\hat{\boldsymbol{\mu}}$ for $\boldsymbol{\mu}(\mathbf{X}^S) := \mathbb{E}[\mathbf{E} \mid \mathbf{X}^S]$.

- 1: Fit the TRAM: $\hat{h} \leftarrow \arg \max_{h \in \mathcal{H}_{\mathcal{Y}, \mathcal{X}^S}} \ell(h; \mathcal{D}_n)$
 - 2: Obtain $\hat{\boldsymbol{\mu}}$ using data \mathcal{D}_n
 - 3: Compute residual product terms: $\mathbf{L}_i \leftarrow R(\hat{h}; y_i, \mathbf{x}_i^S) \{\mathbf{e}_i - \hat{\boldsymbol{\mu}}(\mathbf{x}_i^S)\}, i = 1, \dots, n$
 - 4: Compute residual covariance: $\hat{\Sigma} \leftarrow n^{-1} \sum_{i=1}^n \mathbf{L}_i \mathbf{L}_i^\top - (n^{-1} \sum_{i=1}^n \mathbf{L}_i) (n^{-1} \sum_{i=1}^n \mathbf{L}_i)^\top$
 - 5: Compute test statistic: $\mathbf{T}_n \leftarrow \hat{\Sigma}^{-1/2} (n^{-1/2} \sum_{i=1}^n \mathbf{L}_i)$
 - 6: Compute p -value: $p_{S,n}(\mathcal{D}_n) \leftarrow 1 - F_{\chi_q^2}(\|\mathbf{T}_n\|_2^2)$
 - 7: **return** $p_{S,n}(\mathcal{D}_n)$
-

of Y . Assuming the induced distribution is faithful w.r.t. the induced graph and oracle tests for $(F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ -invariance, the coverage guarantee even holds if \mathbf{E} is a causal parent of Y (this, in particular, includes cases of linear shift TRAMs in which \mathbf{E} only interacts with \mathbf{X}^{S^*} to cause Y , that is, effect modification). In this case, there exists no $S \in [d]$ that is $(F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ -invariant and TRAMICP returns the empty set with high probability (see Proposition 20 in Appendix A5).

We refer to the combination of ICP (Algorithm 1) with the proposed TRAM-GCM invariance test (Algorithm 2) as TRAM-ICP-GCM, with the proposed TRAM-Wald invariance test (Algorithm A1) as TRAMICP-Wald and using a nonparametric conditional independence test (see Appendix B1) as nonparametric ICP.

3.1.1. Invariance Tests based on Score Residuals

We can test the null hypothesis of $(F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ -invariance by testing the implication in (3), that is uncorrelatedness between score residuals and residualized environments in a GCM-type invariance test (Algorithm 2). This requires that the maximum likelihood estimator exists and is unique.

Assumption 3. Under Setting 1 and for all $S \subseteq [d]$, the maximum likelihood estimator, given by $\arg \max_{h \in \mathcal{H}_{\mathcal{Y}, \mathcal{X}^S}} \ell(h; \mathcal{D}_n)$, exists and is unique.

See also the regularity conditions in McLain and Ghosh (2013, Assumptions I–V). Theorem 15 shows that the proposed test is uniformly asymptotically level α for any $\alpha \in (0, 1)$.

Theorem 15 (Uniform asymptotic level of the invariance test in Algorithm 2). Assume Setting 1 and Assumption 3 and for a fixed $S \subseteq [d]$ let $\mathcal{P} := \{\mathbb{P}_{(Y, \mathbf{X}^S, \mathbf{E})} \mid S \text{ is } (F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})\text{-invariant}\}$ denote the set of null distributions for the hypothesis $H_0(S) : S$ is $(F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ -invariant (Definition 11). For all P in \mathcal{P} , we denote by h_P the $h^S \in \mathcal{H}_{\mathcal{Y}, \mathcal{X}^S}$ in the definition of $(F_Z, \mathcal{H}_{\mathcal{Y}, \mathcal{X}})$ -invariance and $\boldsymbol{\mu}(\mathbf{X}^S) := \mathbb{E}_P[\mathbf{E} \mid \mathbf{X}^S]$. Let $\boldsymbol{\xi} := \mathbf{E} - \boldsymbol{\mu}(\mathbf{X}^S)$. Assume that

- (a) $\inf_{P \in \mathcal{P}} \lambda_{\min}(\mathbb{E}_P[R(h_P; Y, \mathbf{X}^S)^2 \boldsymbol{\xi} \boldsymbol{\xi}^\top]) > 0$,

conditions assumed in Proposition 13, we may not correctly reject all non-invariant sets, but the coverage guarantee still holds. In this sense, the method adapts automatically to settings, in which the heterogeneity is sufficiently strong.

- (b) There exists $\delta > 0$, s.t. $\sup_{P \in \mathcal{P}} \mathbb{E}_P[\|R(h_P; Y, \mathbf{X}^S)\xi\|_2^{2+\delta}] < \infty$,
- (c) $\sup_{P \in \mathcal{P}} \max\{\mathbb{E}_P[\|\xi\|_2^2 | \mathbf{X}^S], \mathbb{E}_P[R(h_P; Y, \mathbf{X}^S)^2 | \mathbf{X}^S]\} < \infty$.

Further, we require the following rate conditions on $M := n^{-1} \sum_{i=1}^n \|\widehat{\boldsymbol{\mu}}(\mathbf{X}_i^S) - \boldsymbol{\mu}(\mathbf{X}_i^S)\|_2^2$ and $W := n^{-1} \sum_{i=1}^n \{R(\widehat{h}; Y_i, \mathbf{X}_i^S) - R(h_P; Y_i, \mathbf{X}_i^S)\}^2$:

- (i) $M = o_{\mathcal{P}}(1)$,
- (ii) $W = o_{\mathcal{P}}(1)$,
- (iii) $MW = o_{\mathcal{P}}(n^{-1})$.

Then T_n converges to a standard q -variate normal distribution uniformly over \mathcal{P} . As a consequence, for all $\alpha \in (0, 1)$,

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P(p_{S,n}(\mathcal{D}_n) \leq \alpha) = \alpha,$$

where $p_{S,n}(\mathcal{D}_n)$ is the p -value computed by Algorithm 2.

A proof is given in Appendix E1.5. Conditions (a)–(c) are mild regularity conditions on the distributions of $(Y, \mathbf{X}^S, \mathbf{E})$. Of the remaining conditions it is usually (iii) that is the strictest. In the case of a parametric linear shift TRAM, we would expect $W = O_{\mathcal{P}}(n^{-1})$ and therefore would only need the regression of \mathbf{E} on \mathbf{X}^S to be consistent. However, the TRAM-GCM invariance test can still be correctly calibrated even if the score residuals are learned at a slower-than-parametric rate. Slower rates occur, for instance, in mixed-effects (Tamási and Hothorn 2021), penalized linear shift (Kook and Hothorn 2021), or conditional TRAMS (Hothorn, Kneib, and Bühlmann 2014).

In Appendix B3, we demonstrate in a simulation study that TRAMICP-GCM and TRAMICP-Wald are level at the nominal α and have nontrivial power (at least as high as nonparametric ICP) against the considered alternatives in several model classes including binary logistic, Weibull and Cox regression. However, the TRAM-Wald invariance test hinges critically on correct model specification. Despite its high power in the simulation study (Appendix B3), the TRAM-Wald invariance test has size greater than its nominal level under slight model misspecification (for instance, presence of a nonlinear effect, see Appendix B6). The TRAM-GCM test, however, directly extends to more flexible shift TRAMS which can incorporate the non-linearity, comes with theoretical guarantees, and does not lead to anti-conservative behavior under the null when testing invariance. The robustness property of the TRAM-GCM invariance test does not necessarily hold without residualization of the environments (Chernozhukov et al. 2017; Shah and Peters 2020). We illustrate empirically how also the naive correlation test may not be level, in case of shift and penalized linear shift TRAMS, in Appendix B6.

3.2. Practical Aspects

Plausible Causal Predictors. The procedure in Algorithm 1 can be used to compute p -values for all $S \in \{1, \dots, d\}$. Based on Peters, Bühlmann, and Meinshausen (2016) and as implemented in **InvariantCausalPrediction** (Meinshausen 2019), we can transform the set-specific p -values into predictor-specific p -values: For all $j \in [d]$, $\widehat{p}_j := 1$ if $\max_{S \subseteq [d]} p_{S,n}(\mathcal{D}_n) < \alpha$ and $\widehat{p}_j := \max_{S \subseteq [d]: j \notin S} p_{S,n}(\mathcal{D}_n)$ otherwise. Now, for $j \in [d]$, \widehat{p}_j is

a valid p -value for the null hypothesis $H_0(j) : X^j \notin \text{pa}(Y)$ (assuming that the true parents satisfy $(F_Z, \mathcal{H}_{Y, \mathcal{X}})$ -invariance). We then refer to X^j with $\widehat{p}_j \leq \alpha$, $j \in [d]$ as *plausible causal predictors*.

Unmeasured Confounding. In Setting 1, we assume that all confounding variables between covariates and response and all parents of the response have been measured. This assumption can be weakened by instead assuming that there exists a subset of observed ancestors $A \subseteq \text{an}(Y)$, such that $\mathbf{E} \perp_{\mathcal{G}^*} Y | \mathbf{X}^A$ (where $\perp_{\mathcal{G}^*}$ denotes d -separation in \mathcal{G}^*) and the model for Y given \mathbf{X}^A is correctly specified by a TRAM. Such transformation models can be constructed in special cases (Wienke 2010; Barbanti and Hothorn 2019), but a characterization of this assumption is, to the best of our knowledge, an open problem. As in ICP in the presence of hidden confounders (Peters, Bühlmann, and Meinshausen 2016, Proposition 5), TRAMICP, under this assumption, returns a subset of the ancestors of Y with large probability.

Nonparametric Extension. If the assumption that the response given its parents is correctly specified by a TRAM is violated, we can still apply nonparametric approaches to estimate the conditional CDF of Y given \mathbf{X}^S , $S \in [d]$. Appendix B7 shows empirically that the TRAM-GCM test based on score residuals obtained via survival random forests (Ishwaran et al. 2008) is level in a data-generating process with right-censored responses where nonparametric ICP, ignoring the censoring, is not. We leave a theoretical extension of our results for shift TRAMS to the nonparametric case for future work.

4. Causal Drivers of Survival in Critically Ill Adults

We apply TRAMICP to the SUPPORT2 dataset (Knaus et al. 1995) with time-to-death in a population of critically ill hospitalized adults being the response variable. SUPPORT2 contains data from 9105 patients of whom 68.1% died after a maximum follow-up of 5.55 years and the remaining 31.9% of observations were right-censored due to loss of follow-up. We consider the following predictors measured at baseline (determined at most three days after hospital admission): Sex (male/female), race (white, black, asian, hispanic, other), number of comorbidities (0–9; num.co), coma score (0–100, scoma), cancer (no cancer, cancer, metastatic cancer; ca), age (years), diabetes (yes/no), dementia (yes/no), disease group (nine groups, including colon and lung cancer; dzgroup).⁸ For our analysis, we treat num.co (0, 1, ..., 5, 6 or more) and scoma (11 levels) as factors, square-root transform age and omit 43 patients with missing values in any of the predictors listed above. We apply TRAMICP using both TRAM-GCM and TRAM-Wald. For TRAM-Wald, we only test the presence of main effects of the environments (without additional first-order interaction effects) due to non-convergence when fitting the models with interaction effects.

⁸ca is not a deterministic function of dzgroup.

Table 1. TRAMICP applied to the SUPPORT2 dataset in the different settings described in Section 4.

| Invariance test | Predictor-specific p -values | | | | | | | | Environment |
|--|--------------------------------|---------|--------------|--------------|----------|----------|-------|-------|-------------|
| | scoma | dzgroup | ca | age | diabetes | dementia | sex | race | |
| <i>Evidence of age and cancer being direct causes of time-to-death</i> | | | | | | | | | |
| TRAM-GCM | 0.239 | 0.239 | 0.000 | 0.003 | 0.157 | 0.176 | 0.162 | 0.220 | num.co |
| TRAM-Wald | 0.127 | 0.127 | 0.000 | 0.001 | 0.080 | 0.077 | 0.089 | 0.127 | |
| <i>Incorporating prior knowledge about direct causes</i> | | | | | | | | | |
| TRAM-GCM | 0.273 | 0.273 | 0.000 | – | – | – | 0.163 | 0.216 | num.co |
| TRAM-Wald | 0.127 | 0.127 | 0.000 | – | – | – | 0.089 | 0.127 | |

NOTE: Predictor-specific p -values (see Section 3.2) are reported for the TRAM-GCM and TRAM-Wald invariant test, together with the environment variable used. p -values in bold are significant at the 5% level; in each row, the set of predictors with bold numbers corresponds to the output of TRAMICP.

4.1. Choice of Environments

When applying oracle tests and assuming faithfulness, TRAMICP maintains the coverage guarantee as long as the environment variables are non-descendant of the response (Peters, Bühlmann, and Meinshausen 2016, sec. 3.3). In our study, all measured predictors precede the response chronologically, so, if all model assumptions are satisfied and faithfulness holds, all choices of environments come with the correct coverage but may differ in power. We choose num.co as the environment as, we believe, it is associated with several other predictors and subsequently creates enough heterogeneity. In addition, because num.co is constructed from the presence/absence of other (recorded and unrecorded) comorbidities, it is a sink node in the corresponding graph. If an unrecorded comorbidity or num.co were to directly cause (time to) death, the population output of TRAMICP (assuming faithfulness) would be empty since the path from num.co to (time to) death cannot be blocked without conditioning on the presence/absence of this comorbidity itself. In Appendix C2, we apply TRAMICP when additionally using race as an environment. (For a single choice of a valid environment, no multiple testing correction is needed; however, when applying TRAMICP to several choices of environments, in order to obtain a family-wise coverage guarantee, one would need to apply a multiple testing correction, such as Bonferroni with the number of choices of environments.)

4.2. Results

The Set of All Predictors is Not Invariant. In the model including all predictors the standard Wald test rejects the null hypothesis of no effect for all predictors except race. A Wald test for the main effect of num.co yields a p -value < 0.0001 . This provides strong evidence that the purely predictive model using all predictors is not invariant across num.co and thus uses a set of features that is different from the set of causal parents.

Evidence of Age and Cancer Being Direct Causes of Time-to-Death. We now apply TRAMICP-GCM and TRAMICP-Wald to the SUPPORT2 dataset specifying the survival time as the response in a Cox proportional hazard model, using num.co as the environment and including all other predictors. Both algorithms output ca and age as plausible causal predictors (i.e., the intersection of all sets for which the invariance test was not rejected equals {ca, age}). This can be seen in Figure C1 in Appendix C1, where all non-rejected sets include both ca and age. The predictor-specific p -values (see Section 3.2) are

given in Table 1 (“Evidence of age and cancer being direct causes of time-to-death”). In their original analysis of the SUPPORT2 dataset, Knaus et al. (1995) have assumed that the censoring is uninformative. In a sensitivity analysis in Appendix C3, we show that while TRAMICP is somewhat robust when inducing (potentially) additional informative censoring, it eventually returns the empty set.

Incorporating Prior Knowledge About Direct Causes. If a set of predictors is known to cause the outcome, this set can always be included in the conditioning set (which reduces computational complexity, because fewer invariance tests have to be performed). We illustrate this by including age, dementia, and diabetes as “mandatory” covariates when running TRAMICP (see Appendix D). In this case, both TRAMICP-GCM and TRAMICP-Wald still output ca as a causal predictor of survival. The predictor p -values are given in Table 1 (“Incorporating prior knowledge about direct causes”).

5. Discussion

In this article, we generalize invariant causal prediction to transformation models, which encompass many classical regression models and different types of responses including categorical and discrete variables. We show that, despite most of these models being neither closed under marginalization nor collapsible, TRAMICP retains the same theoretical guarantees in terms of identifying a subset the causal parents of a response with high probability. We generalize the notion of invariance to discrete and categorical responses by considering score residuals which are uncorrelated with the environment under the null hypothesis. Since score residuals remain well-defined for categorical responses, our proposal is one way to phrase invariance in classification settings.

We have applied TRAMICP to roughly ten real world datasets (which technically would require a multiple testing correction), and have often observed that, depending on the choice of environment, either no subset of covariates is invariant (i.e., all invariance tests are rejected) or all subsets of covariates are invariant. In both cases, TRAMICP outputs the empty set—an output that is not incorrect but uninformative.

Supplementary Materials

The online supplement contains a pdf file with Appendices A to E and another pdf file with information on how to reproduce the results in the

paper. The code for reproducing the results can be found at <https://github.com/LucasKook/tramicp.git>.

Acknowledgments

LK carried out part of this work at the University of Copenhagen and University of Zurich. We thank Niklas Pfister and Alexander Mangulad Christgau for insightful discussions. We would also like to thank Juraj Bodik for helpful comments.

Disclosure Statement

The authors report that there are no competing interests to declare.

Funding

The research of LK was supported by the Swiss National Science Foundation (SNF; grant no. 214457). During parts of this research project, SS and JP worked at the University Copenhagen and were supported by a research grant (18968) from the VILLUM Foundation. ARL is supported by a research grant (0069071) from Novo Nordisk Fonden. TH is supported by the Swiss National Science Foundation (grant 200021_219384).

ORCID

Lucas Kook  <http://orcid.org/0000-0002-7546-7356>
 Sorawit Saengkyongam  <http://orcid.org/0000-0003-1581-259X>
 Anton Rask Lundborg  <http://orcid.org/0000-0001-5565-5678>
 Torsten Hothorn  <http://orcid.org/0000-0001-8301-0471>
 Jonas Peters  <http://orcid.org/0000-0002-1487-7511>

References

- Aldrich, J. (1989), "Autonomy," *Oxford Economic Papers*, 41, 15–34. [1,7]
- Andersson, S. A., Madigan, D., and Perlman, M. D. (1997), "On the Markov Equivalence of Chain Graphs, Undirected Graphs, and Acyclic Digraphs," *Scandinavian Journal of Statistics*, 24, 81–102. [2]
- Barbanti, L., and Hothorn, T. (2019), "A Transformation Perspective on Marginal and Conditional Models," arXiv:1910.09219. [9]
- Barlow, W. E., and Prentice, R. L. (1988), "Residuals for Relative Risk Regression," *Biometrika*, 75, 65–74. [5]
- Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2019), "The Conditional Permutation Test for Independence While Controlling for Confounders," *Journal of the Royal Statistical Society, Series B*, 82, 175–197. [3]
- Bickel, P. J., and Doksum, K. A. (1981), "An Analysis of Transformations Revisited," *Journal of the American Statistical Association*, 76, 296–311. [4]
- Bongers, S., Forré, P., Peters, J., and Mooij, J. M. (2021), "Foundations of Structural Causal Models with Cycles and Latent Variables," *The Annals of Statistics*, 49, 2885–2915. [4]
- Box, G. E., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, Series B*, 26, 211–243. [4]
- Candès, E., Fan, Y., Janson, L., and Lv, J. (2018), "Panning for Gold: 'Model-X' Knockoffs for High Dimensional Controlled Variable Selection," *Journal of the Royal Statistical Society, Series B*, 80, 551–577. [3]
- Castelo, R., and Kocka, T. (2003), "On Inclusion-Driven Learning of Bayesian Networks," *Journal of Machine Learning Research*, 4, 527–574. [2]
- Cheng, S., Wei, L., and Ying, Z. (1995), "Analysis of Transformation Models with Censored Data," *Biometrika*, 82, 835–845. [4]
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017), "Double/Debiased/Neyman Machine Learning of Treatment Effects," *American Economic Review*, 107, 261–65. [9]
- Chickering, D. M., (2002), "Optimal Structure Identification with Greedy Search," *Journal of Machine Learning Research*, 3, 507–554. [2]
- Christiansen, R., and Peters, J. (2020), "Switching Regression Models and Causal Inference in the Presence of Discrete Latent Variables," *Journal of Machine Learning Research*, 21, 1–46. [3]
- Cox, D. R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, Series B*, 34, 187–202. [3]
- Diaz, E., Tazi, K., Braude, A. S., Okoh, D., Lamb, K., Watson-Parris, D., Harder, P., and Meinert, N. (2022), "Identifying causes of Pyrocumulonimbus (PyroCb)," in *NeurIPS 2022 Workshop on Causality for Real-world Impact*. [3]
- Didelez, V., and Stensrud, M. J. (2022), "On the Logic of Collapsibility for Causal Effect Measures," *Biometrical Journal*, 64, 235–242. [3]
- Doksum, K. (1974), "Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case," *The Annals of Statistics*, 2, 267–277. [4]
- Farrington, C. P. (2000), "Residuals for Proportional Hazards Models with Interval-Censored Survival Data," *Biometrics*, 56, 473–482. [5]
- Frisch, R., Haavelmo, T., Koopmans, T., and Tinbergen, J. (1948), "Autonomy of Economic Relations," Technical report, Universitets Sosialøkonomiske Institutt, Oslo, Norway. [1,7]
- Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2007), "Kernel Measures of Conditional Dependence," in *Advances in Neural Information Processing Systems* (Vol. 20), Curran Associates, Inc. [3]
- Glymour, C., Zhang, K., and Spirtes, P. (2019), "Review of Causal Discovery Methods Based on Graphical Models," *Frontiers in Genetics*, 10, 524. [2]
- Greenland, S. (1996), "Absence of Confounding Does Not Correspond to Collapsibility of the Rate Ratio or Rate Difference," *Epidemiology*, 7, 498–501. [3]
- Greenland, S., Pearl, J., and Robins, J. M. (1999), "Confounding and Collapsibility in Causal Inference," *Statistical Science*, 14, 29–46. [3]
- Guyon, I., Aliferis, C., and Elisseeff, A. (2007), "Causal Feature Selection," in *Computational Methods of Feature Selection* (1st ed.), eds. H. Liu and H. Motoda, pp. 79–102, Boca Raton, FL: Chapman and Hall/CRC. [2]
- Haavelmo, T. (1943), "The Statistical Implications of a System of Simultaneous Equations," *Econometrica*, 11, 1–12. [1,7]
- Hauser, A., and Bühlmann, P. (2015), "Jointly Interventional and Observational Data: Estimation of Interventional Markov Equivalence Classes of Directed Acyclic Graphs," *Journal of the Royal Statistical Society, Series B*, 77, 291–318. [2]
- He, Y.-B., and Geng, Z. (2008), "Active Learning of Causal Networks with Intervention Experiments and Optimal Designs," *Journal of Machine Learning Research*, 9, 2523–2547. [2]
- Heinze-Deml, C., Peters, J., and Meinshausen, N. (2018), "Invariant Causal Prediction for Nonlinear Models," *Journal of Causal Inference*, 6, 20170016. [3]
- Hernán, M. A., and Robins, J. M. (2010), "Causal Inference: What If." Boca Raton: Chapman & Hall/CRC. [2]
- Hothorn, T., Kneib, T., and Bühlmann, P. (2014), "Conditional Transformation Models," *Journal of the Royal Statistical Society, Series B*, 76, 3–27. [4,9]
- Hothorn, T., Möst, L., and Bühlmann, P. (2018), "Most Likely Transformations," *Scandinavian Journal of Statistics*, 45, 110–134. [1,4]
- Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. (2008), "Nonlinear Causal Discovery With Additive Noise Models," in *Advances in Neural Information Processing Systems* Volume 21 of *NeurIPS*, Curran Associates, Inc., pp. 689–696. [6]
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008), "Random Survival Forests," *The Annals of Applied Statistics*, 2, 841–860. [9]
- Knaus, W. A., Harrell, F. E., Lynn, J., Goldman, L., Phillips, R. S., Connors, A. F., Dawson, N. V., Fulkerson, W. J., Califf, R. M., Desbiens, N. et al. (1995), "The Support Prognostic Model: Objective Estimates of Survival for Seriously Ill Hospitalized Adults," *Annals of Internal Medicine*, 122, 191–203. [9]
- Kook, L., and Hothorn, T. (2021), "Regularized Transformation Models: The **tramnet** Package," *The R Journal*, 13, 581–594. [9]
- Kook, L., Sick, B., and Bühlmann, P. (2022), "Distributional Anchor Regression," *Statistics and Computing*, 32, 39. [5]
- Korepanova, N., Seibold, H., Steffen, V., and Hothorn, T. (2020), "Survival Forests Under Test: Impact of the Proportional Hazards Assumption

- on Prognostic and Predictive Forests for Amyotrophic Lateral Sclerosis Survival,” *Statistical Methods in Medical Research*, 29, 1403–1419. [5]
- Lagakos, S. W. (1981), “The Graphical Evaluation of Explanatory Variables in Proportional Hazard Regression Models,” *Biometrika*, 68, 93–98. [5]
- Laksafoss, A. D. (2020), “Invariant Causal Prediction for Event and Time to Event Data,” Master’s Thesis, University of Copenhagen, Department of Mathematical Sciences. [3]
- McCullagh, P. (1980), “Regression Models for Ordinal Data,” *Journal of the Royal Statistical Society, Series B*, 42, 109–127. [3]
- McCullagh, P. and Nelder, J. A. (2019), *Generalized Linear Models*, London: Routledge. [1,3]
- McLain, A. C., and Ghosh, S. K. (2013), “Efficient Sieve Maximum Likelihood Estimation of Time-transformation Models,” *Journal of Statistical Theory and Practice*, 7, 285–303. [4,8]
- Meinshausen, N. (2019), *InvariantCausalPrediction: Invariant Causal Prediction*, R package version 0.8. [9]
- Meinshausen, N., Hauser, A., Mooij, J. M., Peters, J., Versteeg, P., and Bühlmann, P. (2016), “Methods for Causal Inference from Gene Perturbation Experiments and Validation,” *Proceedings of the National Academy of Sciences of the United States of America*, 113, 7361–7368. [3]
- Migliavacca, M., Musavi, T., Mahecha, M. D., Nelson, J. A., Knauer, J., Baldocchi, D. D., Perez-Priego, O., Christiansen, R., Peters, J., Anderson, K., et al. (2021), “The Three Major Axes of Terrestrial Ecosystem Function,” *Nature*, 598, 468–472. [3]
- Pearl, J. (2009), *Causality*, Cambridge, MA: Cambridge University Press. [1,2,4,5,7]
- Peters, J., and Bühlmann, P. (2013), “Identifiability of Gaussian Structural Equation Models with Equal Error Variances,” *Biometrika*, 101, 219–228. [6]
- Peters, J., Bühlmann, P., and Meinshausen, N. (2016), “Causal Inference by Using Invariant Prediction: Identification and Confidence Intervals,” *Journal of the Royal Statistical Society, Series B*, 78, 947–1012. [1,2,3,7,8,9,10]
- Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014), “Causal Discovery with Continuous Additive Noise Models,” *Journal of Machine Learning Research*, 15, 2009–2053. [6]
- Pfister, N., Bühlmann, P., and Peters, J. (2019), “Invariant Causal Prediction for Sequential Data,” *Journal of the American Statistical Association*, 114, 1264–1276. [3]
- Rothenhäusler, D., Meinshausen, N., Bühlmann, P., and Peters, J. (2021), “Anchor Regression: Heterogeneous Data Meet Causality,” *Journal of the Royal Statistical Society, Series B*, 83, 215–246. [5]
- Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. (2012), “On Causal and Anticausal Learning,” in *Proceedings of the 29th International Conference on International Conference on Machine Learning*, ICML, Omnipress, pp. 459–466. [1,7]
- Shah, R. D., and Peters, J. (2020), “The Hardness of Conditional Independence Testing and the Generalised Covariance Measure,” *The Annals of Statistics*, 48, 1514–1538. [3,4,8,9]
- Shimizu, S. (2014), “LiNGAM: Non-Gaussian Methods for Estimating Causal Structures,” *Behaviormetrika*, 41, 65–98. [6]
- Spirites, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000), *Causation, Prediction, and Search*, Cambridge, MA: MIT Press. [1,2,7]
- Strobl, E. V., Zhang, K., and Visweswaran, S. (2019), “Approximate Kernel-Based Conditional Independence Tests for Fast Non-Parametric Causal Discovery,” *Journal of Causal Inference*, 7, 20180017. [3]
- Tamási, B., and Hothorn, T. (2021), “**tramME**: Mixed-Effects Transformation Models Using Template Model Builder,” *The R Journal*, 13, 398–418. [9]
- Tian, J., and Pearl, J. (2001), “Causal Discovery from Changes,” in *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann. [2]
- Tutz, G. (2011), *Regression for Categorical Data* (Vol. 34), Cambridge, UK: Cambridge University Press. [3]
- Verma, T., and Pearl, J. (1990), “Causal Networks: Semantics and Expressiveness,” in *Machine Intelligence and Pattern Recognition* (Vol. 9), pp. 69–76, Amsterdam: Elsevier. [2]
- Wienke, A. (2010), *Frailty Models in Survival Analysis*, Boca Raton, FL: CRC Press. [9]
- Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011), “Kernel-Based Conditional Independence Test and Application in Causal Discovery,” in *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, AUAI Press, pp. 804–813. [3]