Towards causality-aware predictions in static anticausal machine learning tasks: the linear structural causal model case

Elias Chaibub Neto
Sage Bionetworks, Seattle, WA 98121
elias.chaibub.neto@sagebase.org

Abstract

We propose a counterfactual approach to train "causality-aware" predictive models that are able to leverage causal information in static anticausal machine learning tasks (i.e., prediction tasks where the outcome influences the features). In applications plagued by confounding, the approach can be used to generate predictions that are free from the influence of observed confounders. In applications involving observed mediators, the approach can be used to generate predictions that only capture the direct or the indirect causal influences. Mechanistically, we train supervised learners on (counterfactually) simulated features which retain only the associations generated by the causal relations of interest. We focus on linear models, where analytical results connecting covariances, causal effects, and prediction mean squared errors are readily available. Quite importantly, we show that our approach does not require knowledge of the full causal graph. It suffices to know which variables represent potential confounders and/or mediators. We discuss the stability of the method with respect to dataset shifts generated by selection biases and validate the approach using synthetic data experiments.

1 Introduction

Causal modeling has been recognized as a potential solution to many challenging problems in machine learning (ML) [42]. Current approaches operating at the intersection between causality and ML can be roughly split into three different classes. The first, focus on the prediction of the consequences of different actions, policies, and interventions, aiming to improve decision making. These approaches attempt to answer "what if" counterfactual questions such as "What if I had treated a patient differently?". The second class focus on the generation of invariant/stable predictions aiming to improve model generalization under dataset shifts [39], while the third class is largely concerned with the estimation of causal effects and only uses ML techniques as a tool to improve the estimation of causal effects. (These approaches will be reviewed in more detail in the Related work section.)

In this paper, our goal is to generate causality-inspired predictions that only leverage associations generated by the causal mechanisms that we are interested in modeling. To this end, we propose a simple counterfactual approach to train "causality-aware" predictive models, where we train and evaluate ML algorithms on (counterfactually) simulated features which retain only the associations of interest. For instance, in anticausal prediction tasks influenced by mediators and/or confounders where we are interested in the direct effects of the outcome on the features, we simulate counterfactual features containing only the associations generated by the direct causal effects. This ability to generate learners that only leverage associations generated by the causal relations of interest is important in practice. For instance, in situations where confounding is unstable across the training and target populations (while direct causal effects are stable), the approach can be used to generate more stable

Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems, 2020.

predictions. Furthermore, in situations where the confounders and/or mediators represent sensitive variables, the approach can also be used to generate predictions that are free from the direct influence of the sensitive variables[1]. (In this paper, however, we present synthetic data illustrations focusing on stable prediction applications, rather than on the analysis of sensitive variables.)

We focus on linear models, where analytical results connecting covariances, causal effects, and prediction mean squared error (MSE) are readily available. At first sight, the proposed approach appears to require the strong assumption that one needs to know the full causal graph describing the data generation process. We point out, however, that this is not the case. The approach only requires partial domain knowledge about which variables represent potential confounders and/or mediators. Noteworthy, we will describe how we can always reparameterize the model in a way that the covariance generated by the causal relations among the features is pushed towards the feature error terms (and similarly for the covariances among the mediators and the covariances among the confounders) so that we can safely generate counterfactual data without even knowing how these variables are causally related. In practice, this is an important advantage in applications involving high-dimensional feature spaces and metadata, where it is unlikely that domain knowledge about these causal relationships will be available.

We also investigate the stability of the proposed approach with respect to (w.r.t.) dataset shifts [39]. A standard assumption in supervised ML is that the training and test sets are independent and identically distributed. In practice, however, this assumption is often violated, and dataset shifts are commonly observed in the real world. At the same time, ML models are often capable of leveraging subtle statistical associations between the input ($\boldsymbol{X}$) and outcome (Y) variables in the training data, including spurious associations generated by confounders ($\boldsymbol{C}$) and other sources of biases in the data. As a consequence, predictions from confounded learners are often unstable across shifted test sets, and can fail to generalize.

We focus on dataset shifts generated by selection biases [21, 23, 1] affecting the joint distribution of the confounders and outcome variable, $P(\boldsymbol{C}, Y)$. In real word applications, selection biases often lead to the collection of non-representative training sets and represent an important challenge for ML. While simple approaches such as matching and inverse probability weighting can be used to neutralize these issues in situations where the joint distribution of $\boldsymbol{C}$ and Y in the target population is known, here we focus on the case where the test set can be shifted in unknown ways w.r.t. $P(\boldsymbol{C}, Y)$. This more challenging setting requires more sophisticated adjustment methods, which are sometimes applied to the training data alone with the hope that training an unconfounded model will be enough to generate stable predictions in shifted test sets. Here, we show that this is insufficient, and that deconfounding both the training and test set features can produce more stable predictions.

2 Related work

Causal approaches based on counterfactual thinking have been used in the context of ML applications to predict the outcomes of different actions, policies, and interventions using non-experimental data [7, 57, 27, 49]. The goal is to make "what if" predictions of the consequences of different actions in order to guide decisions. These approaches, however, are only applicable in situations where the "treatment" variables correspond to features of the ML model, so that prediction goes in the same direction of the causal effect (i.e., the features influence the response variable). Our approach, on the other hand, focus on static anticausal ML tasks where the response influences the features.

Our work is similar in spirit to invariant prediction approaches [43, 17, 22, 47, 36, 3] or stable prediction approaches [31, 52, 53, 32] in the sense that it can also be used to generate predictions based on the stable properties of the data, without absorbing unstable spurious associations. Invariant prediction approaches, however, rely on multiple training sets to learn invariances while the causality-aware (and stable prediction) approaches only requires a single training set. Some stable prediction approaches require, nonetheless, full knowledge of the causal graph [52], or can only be directly used in causal prediction tasks [31, 32], while the causality-aware method only requires partial knowledge of the causal graph, and is suited to anticausal tasks. (Supplementary Section 1 provides more detailed discussions on these more closely related approaches.)

[1]The approach can also be used to generate predictions that are exclusively driven by associations generated by sensitive variables. Such models could be used, for example, to demonstrate how the sensitive variables can still impact the predictive performance of a learner, even when they are not included as inputs in the model.

Supervised ML has also been extensively used to aid the estimation of causal effects, where it can potentially attenuate model mispecification issues [30]. In particular, supervised ML has been used to: (i) improve the calculation of propensity scores [38, 58, 33, 60, 44, 61]; (ii) fit regression approaches to estimate outcome models [24, 4, 20]; and (iii) also for the development of double-robust approaches that combine propensity score and outcome regression approaches together [19, 12]. In this paper, however, we take an opposite strategy where instead of using ML to improve causal inference we leverage (partial) causal knowledge to improve the explainability and robustness of ML predictions.

3 Preliminaries

Throughout the text we let $\boldsymbol{X} = (X_1, X_2, \ldots, X_{n_X})^T$, $\boldsymbol{C} = (C_1, C_2, \ldots, C_{n_C})^T$, and $\boldsymbol{M} = (M_1, M_2, \ldots, M_{n_M})^T$ represent, respectively, sets of features, confounders, and mediators, while Y represents the response (outcome) variable. The causality-aware counterfactual versions of $\boldsymbol{X}$ and $\boldsymbol{M}$ are represented, respectively, by $\boldsymbol{X}^*$ and $\boldsymbol{M}^*$. Following [40, 56], we adopt a mechanism-based approach to causation, where the statistical information encoded in the joint probability distribution of a set of variables is supplemented by a *directed acyclic graph* (DAG) describing our qualitative assumptions about the causal relation between the variables. Following [48] we denote prediction tasks where the response influences the features as *anticausal prediction tasks*, whereas tasks where the features influence the response are denoted as *causal prediction tasks*. Figure 1 presents the DAG of a general anticausal predictive task, where $\boldsymbol{X}$, $\boldsymbol{C}$, and $\boldsymbol{M}$ are organized into arbitrary DAG subdiagrams (see Supplementary Figure S6 for an illustrative example).

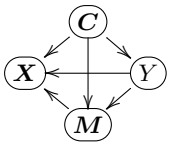


Figure 1: Anticausal prediction task.

4 The proposed approach

4.1 The univariate case

For the sake of clarity, we first describe our approach in the special case where $\boldsymbol{X}$, $\boldsymbol{C}$, and $\boldsymbol{M}$ are composed of a single variable. We describe how to use counterfactual reasoning to simulate features where the association between the response and the features is due exclusively to the causal effects of interest. For simplicity, we assume that the data is generated from a standardized linear model[2], so that the variances of X, C, M, and Y are equal to 1, and the direct causal effect of a variable Z_j on another variable Z_k is represented by the path coefficient [59], $\theta_{Z_k Z_j}$.

The anticausal task presented in Figure 1 is represented by the set of structural equations, $C = U_C$, $Y = \theta_{YC}\, C + U_Y$, $M = \theta_{MC}\, C + \theta_{MY} Y + U_M$, and $X = \theta_{XC}\, C + \theta_{XM}\, M + \theta_{XY}\, Y + U_X$, where U_C, U_Y, U_M, and U_X are independent background (residual) variables. Using Wright's method of path analysis [59], we have that the total covariance (correlation) between X and Y,

$$Cov(X, Y) = \underbrace{\theta_{XY}}_{X \leftarrow Y} + \underbrace{\theta_{XM}\, \theta_{MY}}_{X \leftarrow M \leftarrow Y} + \underbrace{\theta_{XC}\, \theta_{YC}}_{X \leftarrow C \rightarrow Y} .$$

can be decomposed into the contribution of the direct causal path, $Y \rightarrow X$, the indirect causal path $Y \rightarrow M \rightarrow X$, and the spurious association generated by the backdoor path $X \leftarrow C \rightarrow Y$. Clearly, the predictive performance of any ML model trained with data generated by this model will be biased by the influence of the confounder C since the learner will leverage the total association between X and Y during training.

Now, suppose that our goal is to build a ML model whose predictive performance is only informed by the direct influence of Y on X and is free from the influence of C, as well as, from the indirect influence of Y that is mediated by M. To this end, we need to simulate counterfactual data where the association between X and Y is due exclusively to the direct causal effect of Y on X. In other words, we want to simulate counterfactual feature data, X^*, such that $Cov(X^*, Y) = \theta_{XY}$. In theory, this

[2]Note that any linear model $Z_k^o = \mu_k + \Sigma_j \beta_{kj} Z_j^o + U_k^o$, where Z_k^o represents the original data, can be reparameterized into its equivalent standardized form $Z_k = \sum_j \theta_{kj} Z_j + U_k$, where $Z_k = (Z_k^o - E(Z_k^o))/\sqrt{Var(Z_k^o)}$ represent standardized variables with $E(Z_k) = 0$ and $Var(Z_k) = 1$; $\theta_{Z_k Z_j} = \beta_{Z_k Z_j} \sqrt{Var(Z_j^o)/Var(Z_k^o)}$ represent the path coefficients; and $U_k = U_k^o/\sqrt{Var(Z_k^o)}$ represent the standardized error terms.

could be done by simulating data according to the twin network[3] [5, 40] in Figure 2, where the new counterfactual feature data, X^*, is generated from the model $X^* = \theta_{XY}Y + U_X$. (In practice, we can estimate θ_{XY} and U_X by regressing X on C, M and Y, and simulate the counterfactual feature data using $\hat{X}^* = X - \hat{\theta}_{XC}C = \hat{\theta}_{XY}Y + \hat{U}_X$. In other words, we can employ a variation of Pearl's "abduction, action, prediction" approach to simulate deterministic counterfactuals [40, 41]. In the next subsection we explain in detail how the proposed approach differs from Pearl's approach at the "action" step.) Direct calculation of the covariance between X^* and Y shows that,

$$Cov(X^*, Y) = Cov(\theta_{XY}Y + U_X, Y) = \theta_{XY}\, Var(Y) + Cov(U_X, Y) = \theta_{XY}\ . \tag{1}$$

Supplementary Section 2 describes the cases where the goal is to build a ML model whose predictive performance is only informed by the indirect causal effect of Y on X, as well as, when the goal is to capture the predictive performance informed by the spurious associations generated by the confounder alone. At this point, a natural question is whether alternative interventions would also work. In Supplementary Section 3, we show that a requirement for the intervention to work is that Y is not altered by the intervention. Furthermore, in Supplementary Section 4 we also show that node-splitting transformations in SWIGs [46] can also be used as alternative interventions.

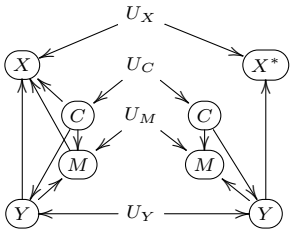


Figure 2: Twin network approach in the case where the direct effect represents the causal effect of interest.

Remarks It is important to highlight that our proposed interventions are different from Pearl's atomic $do(Z = z)$ interventions, and that our counterfactual approach is implemented using a modification of Pearl's "abduction, action, prediction" procedure for the computation of deterministic counterfactuals. While in Pearl's approach the action step is enforced by a $do(Z = z)$ intervention, where the causal structural model $Z = f(pa(Z), U_Z)$ is replaced by $Z = z$, our interventions are different. For instance, in the case where the direct effect represents the causal effect of interest, our intervention corresponds to replacing $X = f_X(pa(X), U_X) = f_X(C, M, Y, U_X)$ by $X = f_X(pa(X) \setminus \{C \cup M\}, U_X) = f_X(Y, U_X)$. (Note that while our interventions at the action step differs from Pearl's approach, the abduction and prediction steps are still the same.) Also, from a more "philosophical" point of view, note that even though our proposed interventions represent a different type of microsurgery on the structural causal models, they are still consistent with Lewis' framework of possible worlds [34]. Instead of considering counterfactual worlds that develop from different actions than the actions taken in the factual world, our approach considers counterfactual worlds where the data generation mechanisms/laws are different from the mechanisms/laws of the factual world[4]. Observe, as well, that our interventions operate at the population level, rather than at the individual level.

4.2 The multivariate case

Next, we extend our results to the multivariate case, where the nodes $\boldsymbol{X}$, $\boldsymbol{C}$, and $\boldsymbol{M}$ in Figure 1 represent arbitrary DAG subdiagrams. But first, we describe how we can always reparameterize linear structural causal models in a way that, in practice, we do not need to know how the DAG subdiagrams are organized in order to estimate the causal effects and the residuals employed in the computation of the counterfactual data.

[3]The twin network approach provides a graphical method for evaluating conditional independence relations between counterfactual and factual variables. The basic idea is to use two networks, one representing the factual world and the other the counterfactual world, which share the same background (residual) variables. The factual network (shown to the left of the residual terms) represents the data generation process for the original data, while the counterfactual network (show to the right of the residual terms) shows the modified causal model.

[4]As an example, consider an anticausal prediction task described by the DAG $C \rightarrow X \leftarrow Y$, where Y represents the severity score of a disease, X represents a symptom, C represents age, and where the goal is to predict Y using X, after removing the spurious association generated by C. In our proposed approach, we consider a counterfactual world, $C \quad X^* \leftarrow Y$, where age no longer influences the symptom X. Note that this intervention can be seen as a type of soft or stochastic intervention where the data generation process differs from the natural system only in the mechanism associated with the feature X. Related types of soft/stochastic interventions have been studied in [13, 29, 15, 37].

4.2.1 Reparameterization in linear models

For linear structural causal models, we can always reparameterize any arbitrary DAG model to a simpler model where the covariance structure between the observed variables is "pushed" to the unobserved error terms. Figure 3 provides an illustrative example of this well-known fact in the structural equations modelling literature [55, 2]. The DAG in panel a represents the actual data generation process for the variables $\boldsymbol{X} = (X_1, X_2, X_3)^T$, where the error terms $\boldsymbol{U}_X = (U_{X_1}, U_{X_2}, U_{X_3})^T$ are independent, whereas the DAG in panel b shows the reparameterized model with correlated error terms $\boldsymbol{W}_X = (W_{X_1}, W_{X_2}, W_{X_3})^T$. The set of linear structural causal models describing the DAG in Figure 3a is given by, $\boldsymbol{X} = \boldsymbol{\Theta}_{XX}\,\boldsymbol{X} + \boldsymbol{U}_X$, which can be reparameterized as $\boldsymbol{X} = \boldsymbol{W}_X$, where $\boldsymbol{W}_X = (\boldsymbol{I} - \boldsymbol{\Theta}_{XX})^{-1}\boldsymbol{U}_X$[5].

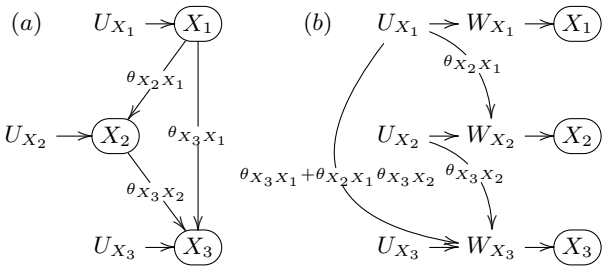


Figure 3: Original (a) and reparameterized (b) models.

Next, we describe the above reparameterization for the arbitrary anticausal predictive task. From the DAG in Figure 1, we have that the joint distribution of the anticausal prediction tasks is factorized as,

$$P(\boldsymbol{C}, Y, \boldsymbol{M}, \boldsymbol{X}) \;=\; P(\boldsymbol{C})\,P(Y \mid \boldsymbol{C})\,P(\boldsymbol{M} \mid \boldsymbol{C}, Y)\,P(\boldsymbol{X} \mid \boldsymbol{C}, \boldsymbol{M}, Y)\;,$$

where the components of this factorization are described, respectively, by the structural causal models,

$$\begin{aligned}
\boldsymbol{C} &= \boldsymbol{\Theta}_{CC}\,\boldsymbol{C} + \boldsymbol{U}_C\;, \qquad\qquad Y = \boldsymbol{\Theta}_{YC}\,\boldsymbol{C} + U_Y\;,\\
\boldsymbol{M} &= \boldsymbol{\Theta}_{MM}\,\boldsymbol{M} + \boldsymbol{\Theta}_{MC}\,\boldsymbol{C} + \boldsymbol{\Theta}_{MY}\,Y + \boldsymbol{U}_M\;,\\
\boldsymbol{X} &= \boldsymbol{\Theta}_{XX}\,\boldsymbol{X} + \boldsymbol{\Theta}_{XC}\,\boldsymbol{C} + \boldsymbol{\Theta}_{XM}\,\boldsymbol{M} + \boldsymbol{\Theta}_{XY}\,Y + \boldsymbol{U}_X\;,
\end{aligned}$$

where $\boldsymbol{U}_C$, U_Y, $\boldsymbol{U}_M$, and $\boldsymbol{U}_X$ are vectors of independent error terms with zero mean and finite variance; $\boldsymbol{\Theta}_{CC}$, $\boldsymbol{\Theta}_{MM}$, and $\boldsymbol{\Theta}_{XX}$ represent, respectively, square matrices of dimension $n_C \times n_C$, $n_M \times n_M$, and $n_X \times n_X$, containing the path coefficients connecting the confounders among themselves, the mediators among themselves and the features among themselves; and $\boldsymbol{\Theta}_{YC}$, $\boldsymbol{\Theta}_{MC}$, $\boldsymbol{\Theta}_{MY}$, $\boldsymbol{\Theta}_{XC}$, $\boldsymbol{\Theta}_{XM}$, and $\boldsymbol{\Theta}_{XY}$, represent retangular matrices of path coefficients connecting variables from separate sets. (For instance, $\boldsymbol{\Theta}_{MC}$, corresponds to a $n_M \times n_C$ matrix of path coefficients connecting confounder variables to mediator variables, whereas $\boldsymbol{\Theta}_{XY}$, corresponds to a $n_X \times 1$ matrix of path coefficients connecting the response to the features.)

Using simple algebraic manipulations, we can re-write the above linear structural models as,

$$\begin{aligned}
\boldsymbol{C} &= \boldsymbol{W}_C\;, \qquad\qquad Y = \boldsymbol{\Gamma}_{YC}\,\boldsymbol{C} + W_Y\;,\\
\boldsymbol{M} &= \boldsymbol{\Gamma}_{MC}\,\boldsymbol{C} + \boldsymbol{\Gamma}_{MY}\,Y + \boldsymbol{W}_M\;,\\
\boldsymbol{X} &= \boldsymbol{\Gamma}_{XC}\,\boldsymbol{C} + \boldsymbol{\Gamma}_{XM}\,\boldsymbol{M} + \boldsymbol{\Gamma}_{XY}\,Y + \boldsymbol{W}_X\;,
\end{aligned}$$

where $W_Y = U_Y$, and $\boldsymbol{W}_V = (\boldsymbol{I} - \boldsymbol{\Theta}_{VV})^{-1}\boldsymbol{U}_V$ for V equal to C, M, or X, and $\boldsymbol{\Gamma}_{YC} = \boldsymbol{\Theta}_{YC}$, and $\boldsymbol{\Gamma}_{ZV} = (\boldsymbol{I} - \boldsymbol{\Theta}_{ZZ})^{-1}\boldsymbol{\Theta}_{ZV}$ for $\{Z, V\}$ pairs equal to $\{M, C\}$, $\{M, Y\}$, $\{X, C\}$, $\{X, M\}$, and $\{X, Y\}$. Supplementary Section 5 presents a concrete illustrative example of the above reparameterization.

[5]Explicitly, we have that,

$$\underbrace{\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}}_{\boldsymbol{X}} = \underbrace{\begin{pmatrix} 0 & 0 & 0 \\ \theta_{X_2X_1} & 0 & 0 \\ \theta_{X_3X_1} & \theta_{X_3X_2} & 0 \end{pmatrix}}_{\boldsymbol{\Theta}_{XX}} \underbrace{\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}}_{\boldsymbol{X}} + \underbrace{\begin{pmatrix} U_{X_1} \\ U_{X_2} \\ U_{X_3} \end{pmatrix}}_{\boldsymbol{U}_X}\;, \quad \underbrace{\begin{pmatrix} W_{X_1} \\ W_{X_2} \\ W_{X_3} \end{pmatrix}}_{\boldsymbol{W}_X} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ \theta_{X_2X_1} & 1 & 0 \\ \theta_{X_3X_1} + \theta_{X_2X_1}\theta_{X_3X_2} & \theta_{X_3X_2} & 1 \end{pmatrix}}_{(\boldsymbol{I} - \boldsymbol{\Theta}_{XX})^{-1}} \underbrace{\begin{pmatrix} U_{X_1} \\ U_{X_2} \\ U_{X_3} \end{pmatrix}}_{\boldsymbol{U}_X}\;.$$

Note that because model $\boldsymbol{X} = \boldsymbol{W}_X$ is just a reparameterization of model $\boldsymbol{X} = \boldsymbol{\Theta}_{XX}\,\boldsymbol{X} + \boldsymbol{U}_X$, we have that the association structure between the X_j variables is still the same after the model reparameterization. Observe, as well, that for any arbitrary DAG, the matrix $(\boldsymbol{I} - \boldsymbol{\Theta}_{XX})$ is always invertible (as fully explained in Supplementary Section 5.1).

4.2.2 Estimation of causal effects and residuals in the reparameterized model

In practice, our counterfactual approach requires the estimation of causal effects and residuals using regression models. For an anticausal task, we regress each feature X_j, $j = 1, \ldots, n_X$, on the set of observed confounders and mediators using the regression equations, $X_j = \sum_{k=1}^{n_C} \gamma_{X_j C_k} C_k + \sum_{k=1}^{n_M} \gamma_{X_j M_k} M_k + \gamma_{X_j Y}\, Y + W_{X_j}$, to estimate the causal effects $\hat{\gamma}_{X_j C_k}$, $\hat{\gamma}_{X_j M_k}$, $\hat{\gamma}_{X_j Y}$, and residuals $\hat{W}_{X_j}$ using least squares[6], and then generate counterfactual features by adding back the estimated residuals to a linear predictor containing only the causal effects of interest. That is, in order to estimate the predictive performance that is separately due to direct causal effects, indirect causal effects, or confounding, we generate counterfactual features using, respectively, $\hat{\boldsymbol{X}}^* = \hat{\boldsymbol{\Gamma}}_{XY} Y + \hat{\boldsymbol{W}}_X$, $\hat{\boldsymbol{X}}^* = \hat{\boldsymbol{\Gamma}}_{XM}\, \hat{\boldsymbol{M}}^* + \hat{\boldsymbol{W}}_X$[7], or $\hat{\boldsymbol{X}}^* = \hat{\boldsymbol{\Gamma}}_{XC}\, \boldsymbol{C} + \hat{\boldsymbol{W}}_X$. Importantly, note that when we regress X_j on $\boldsymbol{C}$, $\boldsymbol{M}$, and Y only the coefficients associated with the parents of X_j in the reparameterized model will be statistically different from zero (for large enough sample sizes). Therefore, in practice, we don't need to know before hand which variables are the parents of X_j in the reparameterized model. The parent set will be learned automatically from the data by the regression model fit. (This, of course, assumes the absence of unmeasured confounders. Supplementary Section 6 provides further remarks on potential identification issues.)

4.2.3 The connection between covariances and causal effects in the multivariate general case

Here, we extend the univariate results of Section 4.1 to the multivariate case (see Supplementary Section 7 for the proofs).

Theorem 1. *Consider an anticausal prediction task:*

1. *For causal effects generated by the paths in $Y \rightarrow \boldsymbol{X}$, if $\boldsymbol{X}^*$ is given by $\boldsymbol{X}^* = \boldsymbol{\Gamma}_{XY}\, Y + \boldsymbol{W}_X$, then $Cov(\boldsymbol{X}^*, Y) = \boldsymbol{\Gamma}_{XY}$.*
2. *For causal effects generated by the paths in $Y \rightarrow \boldsymbol{M} \rightarrow \boldsymbol{X}$, if $\boldsymbol{X}^*$ is given by $\boldsymbol{X}^* = \boldsymbol{\Gamma}_{XM}\, \boldsymbol{M}^* + \boldsymbol{W}_X$, and $\boldsymbol{M}^* = \boldsymbol{\Gamma}_{MY}\, Y + \boldsymbol{W}_M$, then $Cov(\boldsymbol{X}^*, Y) = \boldsymbol{\Gamma}_{XM}\, \boldsymbol{\Gamma}_{MY}$.*
3. *For the spurious associations generated by the paths in $\boldsymbol{X} \leftarrow \boldsymbol{C} \rightarrow Y$, if $\boldsymbol{X}^*$ is given by $\boldsymbol{X}^* = \boldsymbol{\Gamma}_{XC}\, \boldsymbol{C} + \boldsymbol{W}_X$, then $Cov(\boldsymbol{X}^*, Y) = \boldsymbol{\Gamma}_{XC}\, Cov(\boldsymbol{C})\, \boldsymbol{\Gamma}_{YC}^T$.*

The above result, together with the estimation approach described In Section 4.2.2, show that by generating causality-aware counterfactual features, $\boldsymbol{X}^*$, and then training and evaluating ML learners on this counterfactual data, we are able to leverage only the associations generated by the causal mechanisms of interest. Quite importantly, because the counterfactual data is estimated from the reparameterized model, the approach does not require full knowledge of the causal graph. It suffices to know which variables are confounders and which are mediators.

5 Confounding adjustment in anticausal tasks

5.1 An algorithmic description for confounding adjustment

When the goal is confounding adjustment, the causality-aware features are generated according to Algorithm 1. Observe that the algorithm requires test set confounding data (but not the test set labels). Note that for large sample sizes, and under the assumption that the causal effects are stable between the training and test sets, we have that $\hat{\beta}^{tr}_{X_j C_i} \approx \hat{\beta}^{ts}_{X_j C_i}$ so that we can estimate the test set counterfactual features without using test set labels since,

$$X^*_{j,ts} = X_{j,ts} - \sum \hat{\beta}^{tr}_{X_j C_i}\, C_{i,ts} \approx X_{j,ts} - \sum \hat{\beta}^{ts}_{X_j C_i}\, C_{i,ts} = \hat{\mu}^{ts}_j + \hat{\beta}^{ts}_{X_j Y}\, Y_{ts} + \hat{W}^{ts}_{X_j}\ .$$

[6]Here, we assume that the number of samples is larger than the number of covariates in the regression fits, and that multicolinearity is not an issue too. Note that we do not need to assume Gaussian error terms.

[7]Where, $\hat{\boldsymbol{M}}^* = M - \hat{\boldsymbol{\Gamma}}_{MC}\, C = \hat{\boldsymbol{\Gamma}}_{MY}\, Y + \hat{\boldsymbol{W}}_M$ is calculated by first fitting the regressing models $M_j = \sum_{k=1}^{n_C} \gamma_{M_j C_k} C_k + \gamma_{M_j Y}\, Y + W_{M_j}$, to estimate the causal effects $\hat{\gamma}_{M_j C_k}$, $\hat{\gamma}_{M_j Y}$ and error terms $\hat{W}_{M_j}$.

Algorithm 1: Causality-aware feature computation in anticausal prediction tasks

Data: Training data, $\{\boldsymbol{X}_{tr}, \boldsymbol{C}_{tr}, Y_{tr}\}$; test set features and confounders, $\{\boldsymbol{X}_{ts}, \boldsymbol{C}_{ts}\}$.

1 **for** *each feature* X_j **do**

2 • Using the training set, estimate regression coefficients and residuals from, $X_{j,tr} = \mu_j^{tr} + \beta_{X_jY}^{tr} Y_{tr} + \sum_i \beta_{X_jC_i}^{tr} C_{i,tr} + W_{X_j}^{tr}$, and then compute the respective counterfactual feature as, $\hat{X}_{j,tr}^{*} = \hat{\mu}_j^{tr} + \hat{\beta}_{X_jY}^{tr} Y_{tr} + \hat{W}_{X_j}^{tr}$.

3 • Using the test set, compute the counterfactual feature, $\hat{X}_{j,ts}^{*} = X_{j,ts} - \sum_i \hat{\beta}_{X_jC_i}^{tr} C_{i,ts}$.

Result: Counterfactual features, $\hat{\boldsymbol{X}}_{tr}^{*}$ and $\hat{\boldsymbol{X}}_{ts}^{*}$.

5.2 Dataset shifts generated by selection biases

In anticausal prediction tasks, dataset shifts in the joint distribution of the confounders and outcome variable, $P(\boldsymbol{C}, Y)$, are often caused by selection biases. The confounded anticausal prediction task influenced by selection bias is described by the causal graph in Figure 4, where the auxiliary variable S indicates the presence of a selection mechanism contributing to the association between $\boldsymbol{C}$ and Y. (Here, S represents a binary variable which indicates whether the sample was included or not in the dataset, and the square frame around S indicates that our dataset is generated conditional on S being set to 1. Note that, the application of the d-separation criterion [40] to the causal graph shows that because S is a collider, we have that, conditional on $S = 1$, the additional path $\boldsymbol{C} \rightarrow S \leftarrow Y$ is open and, therefore, contributes to the association between $\boldsymbol{C}$ and Y.) In the stability analysis that we present in the next subsection, we assume that the causal effects $\boldsymbol{\beta}_{XY}$ and $\boldsymbol{\beta}_{XC}$ and the residual covariance, $Cov(\boldsymbol{U}_X)$, are the same across the training and test sets, so that $P(\boldsymbol{X} \mid \boldsymbol{C}, Y)$ is stable. We also assume that the causal effect β_{YC} is stable, and that the dataset shifts in $P(\boldsymbol{C}, Y)$ are generated by selection biases.

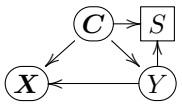


Figure 4:

5.3 Stability under dataset shifts of $P(\boldsymbol{C}, Y)$ generated by selection biases

While it might seen intuitive that training a learner on unconfounded data will prevent it from learning the confounding signal and, therefore, will lead to more stable predictions in shifted target populations[8], here we show that adjusting the training data alone is insufficient, and that better stability can be achieved by deconfounding the test set features as well.

Next, we present an analysis of this issue using a toy linear model example (the result, nonetheless, holds for more general linear models, as described in Supplementary Section 8). Consider the causal graph in $C \rightrightarrows X \leftleftarrows Y$ where $C = U_C$, $Y = \beta_{YC}\, C + U_Y$, and $X = \beta_{XY}\, Y + \beta_{XC}\, C + U_X$, with $E[U_V] = 0$, $Var(U_V) = \sigma_V^2$, for $V = \{C, Y, X\}$. The goal is to predict the outcome Y using the feature X. Assume without loss of generality that the data has been centered. Let $\hat{Y} = X_{ts}\hat{\beta}_{tr}$ represent the test set prediction from a linear regression model, where $\hat{\beta}_{tr}$ represents the coefficient estimated with the training data, and X_{ts} represents the test set feature. By definition the expected MSE is given by,

$$\begin{aligned} E[(Y_{ts} - \hat{Y})^2] &= E[Y_{ts}^2] + E[\hat{Y}^2] - 2E[\hat{Y}Y_{ts}] = Var[Y_{ts}] + E[\hat{Y}^2] - 2Cov(\hat{Y}, Y_{ts}) \\ &= Var[Y_{ts}] + \hat{\beta}_{tr}^2 Var[X_{ts}] - 2\hat{\beta}_{tr} Cov(X_{ts}, Y_{ts})\ , \end{aligned} \tag{2}$$

where the expectation is w.r.t. the test set (so that $\hat{\beta}_{tr}$ is a fixed constant w.r.t. the expectation).

For any approach which does not process the test set features we have that,

$$\begin{aligned} Var(X_{ts}) &= Var(\beta_{XY}Y_{ts} + \beta_{XC}C + U_X) \\ &= \sigma_X^2 + \beta_{XY}^2 Var(Y_{ts}) + \beta_{XC}^2 Var(C_{ts}) + 2\beta_{XY}\beta_{XC} Cov(Y_{ts}, C_{ts})\ , \end{aligned}$$

$$Cov(X_{ts}, Y_{ts}) = Cov(\beta_{XY}Y_{ts} + \beta_{XC}C_{ts} + U_X, Y_{ts}) = \beta_{XY} Var(Y_{ts}) + \beta_{XC} Cov(Y_{ts}, C_{ts})$$

[8] Examples of approaches that only adjust the training data include pre-processing techniques to reduce discrimination in ML [8, 28].

showing that both $Var(X_{ts})$ and $Cov(X_{ts}, Y_{ts})$ depend on $Cov(Y_{ts}, C_{ts})$ (so that the $E[MSE]$ will be unstable under dataset shifts of the association between the confounder and the outcome variable). On the other hand, we have that for the causality-aware approach,

$$Var(X^*_{ts}) = Var(\beta_{XY} Y_{ts} + U_X) = \beta^2_{XY} Var(Y_{ts}) + \sigma^2_X \ ,$$

$$Cov(X^*_{ts}, Y_{ts}) = Cov(\beta_{XY} Y_{ts} + U_X, Y_{ts}) = \beta_{XY} Var(Y_{ts}) \ ,$$

do not depend on $Cov(Y_{ts}, C_{ts})$, so that the $E[MSE]$ will be stable w.r.t. this particular type of dataset shift (although, as shown by eq. (2) it will be still influenced by dataset shifts on $Var(Y_{ts})$). Note that this is true even when we apply a confounding adjustment to the training set (a situation where the $\hat{\beta}_{tr}$ estimate is not influenced by the spurious associations generated by the confounder). This explains why it is not enough to deconfound the training features alone. While training a regression model using deconfounded features allows us to estimate deconfounded model weights, $\hat{\boldsymbol{\beta}}_{tr}$, the prediction $\hat{Y} = \boldsymbol{X}_{ts} \hat{\boldsymbol{\beta}}_{tr}$ is a function of both the trained model $\hat{\boldsymbol{\beta}}_{tr}$ and the test set feature, $\boldsymbol{X}_{ts}$. As a consequence, if we do not deconfound the test set features, the expected MSE will still be influenced by the confounders (since, in anticausal prediction tasks, the original test set features, $X_{j,ts} = \beta_{X_j Y} Y_{ts} + \beta_{XC} C_{ts} + U_{X_j}$ are still functions of the confounder variable). This point is described in more general terms in Supplementary Section 9, where we show that the expected value of an arbitrary performance metric is still a function of C_{ts} when the features $X_{j,ts}$ are generated by arbitrary structural causal model $X_{j,ts} = f(Y_{ts}, C_{ts}, U^{ts}_X)$, even when we train the ML model using deconfounded training set features, $X^*_{j,tr} = f^*(Y_{tr}, U^{tr}_X)$.

5.4 Synthetic data experiments

We illustrate the above points in synthetic data experiments investigating the influence of dataset shifts in $P(\boldsymbol{C}, Y)$ on the predictive performance (measured by MSE). In order to investigate the influence of shifts in $Var(Y_{ts})$ on the prediction stability, we performed two experiments, where $Var(Y_{ts})$ was kept constant in the first, but was allowed to vary in the second. In both experiments, we compared the causality-aware adjustment against two alternative approaches denoted as *baseline 1* and *baseline 2* adjustments. The *baseline 1* adjustment represents approaches that remove the causal effect of the confounders on the features in the training set alone, while *baseline 2* represents approaches that remove the association between the confounders and the output in the training set alone (see Supplementary Section 10 for further details). For completeness we also report results based on the "*no adjustment*" approach, where no adjustments are applied to the training or test sets.

Each experiment was based on 1,000 replications where, for each replication, we generated training sets with $Var(Y_{ts}) = 1$, $Var(C_{ts}) = 1$, and $Cov(C_{ts}, Y_{ts}) = 0.8$, and 9 distinct test sets showing increasing amounts of dataset shifts in the $P(\boldsymbol{C}, Y)$ relative to the training data. In the first experiment (the fixed $Var(Y_{ts})$ case), this was accomplished by varying $Cov(Y_{ts}, C_{ts})$ according to $\{0.8, 0.6, 0.4, 0.2, 0.0, -0.2, -0.4, -0.6, -0.8\}$ across the 9 test sets, and by varying $Var(C_{ts})$ according to $\{1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00\}$, while keeping $Var(Y_{ts})$ fixed at 1. In the second experiment (the varying $Var(Y_{ts})$ case), we varied $Cov(Y_{ts}, C_{ts})$ as before, but kept $Var(C_{ts})$ fixed at 1, while increasing $Var(Y_{ts})$ according to $\{1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00\}$ across the test sets.

Our experiments were based on linear models containing 10 features and 1 confounder, and on training and test sets containing 1,000 samples. (See Supplementary Section 10 for further details on the synthetic data generation and simulation parameter choices.) The causal effects $\boldsymbol{\beta}_{XY}$, $\boldsymbol{\beta}_{XC}$, and β_{YC} and $Cov(\boldsymbol{U}_X)$ were kept constant across the training and test sets in order to guarantee that $P(\boldsymbol{X} \mid \boldsymbol{C}, Y)$ was stable.

Figures 5 and 6 report the results for the fixed and varying $Var(Y_{ts})$ cases, respectively. In both figures, panels a to d report boxplots of the MSE scores (y-axis) across 1,000 simulation replications for the 9 test sets (x-axis), while panel e presents a comparison of the stability-errors, defined as the standard deviation of the MSE scores across the 9 test sets in each simulation replication.

Figure 5 reports the results for the first experiment. Note that because we kept $Var(Y_{ts})$ constant across the test sets we see perfect stability for the causality-aware approach (panel a). (Observe that varying $Cov(Y_{ts}, C_{ts})$ and $Var(C_{ts})$ has no influence on the stability of the results, since the expected MSE for the causality-aware approach only depends on $Var(Y_{ts})$.)

Figure 6 reports results for the second experiment based on increasing $Var(Y_{ts})$ values. As expected, we now observe instability in the causality-aware approach too. The causality-aware predictions, however, are still more stable than the predictions from the other approaches.

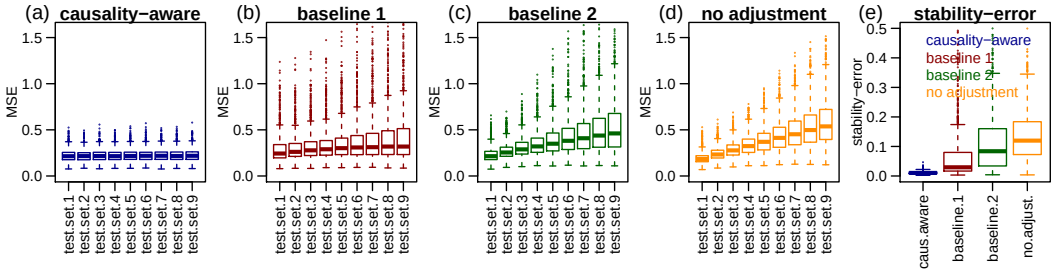


Figure 5: Synthetic data experiment results for the fixed $Var(Y_{ts})$ case.

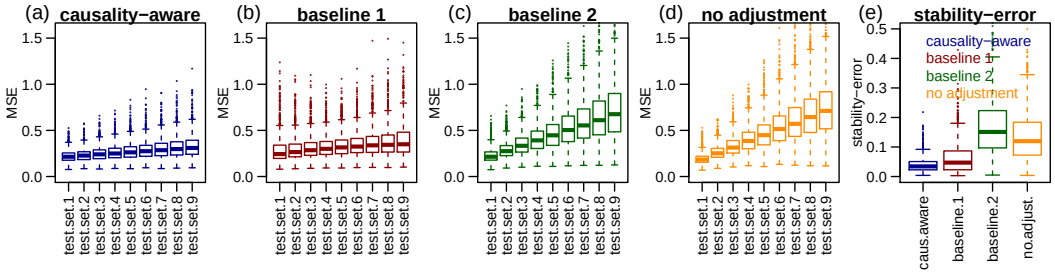


Figure 6: Synthetic data experiment results for the increasing $Var(Y_{ts})$ case.

6 Final remarks

This paper has three main contributions. First, we describe a novel counterfactual approach to train "causality-aware" predictive models, which leverages only the associations generated by the causal mechanisms of interest. Second, by leveraging a reparameterization of the linear structural causal models (described in Section 4.2.1), we show that the approach does not require full knowledge of the data generation process. It suffices to know which variables are confounders and mediators, without knowing how these variables are causally related. This represent an important practical advantage of the method relative to alternative approaches such as counterfactual normalization [52], which requires knowledge of the full causal graph. Third, we investigate the stability properties of the method w.r.t. dataset shifts generated by selection biases. We show that the $E[MSE]$ for adjustment approaches that fail to deconfound the test set features will be unstable w.r.t. shifts in $Cov(\boldsymbol{C}, Y)$, even when the ML models are trained with unconfounded data (and there are no shifts in $Var(Y_{ts})$). This is an important observation that (we feel) is not well appreciated in the ML community.

One important drawback of the approach is its reliance on the linearity assumption. The present work, however, represents a first step that, we believe, will serve as inspiration for more flexible approaches. Along these lines, in a separate contribution [10] (where we compare the causality-aware approach against the residualization confounding adjustment - an ad-hoc approach, widely used in applied fields such as neuroimaging), we describe an extension of the causality-aware approach to additive models. Furthermore, in another separate contribution [11], we also describe how the causality-aware approach (based on linear models) can still be used to deconfound the feature representations learned by deep neural network models in classification tasks. The key idea is that by training a highly accurate DNN using softmax activation at the classification layer, we have that, by construction, the feature representation learned by the last layer prior to the output layer will fit well a logistic regression model (since the softmax activation used to classify the outputs of the DNN is essentially performing logistic regression classification). This reference illustrates the practicality of the causality-aware approach in real world applications. (Finally, while this work has focused on anticausal tasks, we present some analogous results for causal prediction tasks in Supplementary Section 11.)

References

[1] Bareinboim, E. and Pearl, J. (2012) Controlling selection bias in causal inference. AISTATS 2012.

[2] Bollen, K. A. (1989) *Structural equations with latent variables.* First edition, John Wiley and Sons.

[3] Arjovsky M., Bottou L., Gulrajani I., Lopez-Paz D. (2019) Invariant risk minimization. *arXiv:1907.02893v3*.

[4] Austin, P. C. (2012). Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based g-computation. *Multivariate Behavioral Research*, **47**, 115-135.

[5] Balke, A., Pearl, J. (1994) Probabilistic evaluation of counterfactual queries. *Proceedings of the 12th National Conference on Artificial Intelligence*, pp 230-237.

[6] Bickel, S., Bruckner, M., and Scheffer, T. (2009) Discriminative learning under covariate shift. *Journal of Machine Learning Research*, **10**, 2137-2155.

[7] Bottou, J., et al (2013) Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, **14**, 3207–3260.

[8] Calders T., Kamiran, F., Pechenizkiy, M. (2009) Building classifiers with independency constraints. ICDM Workshop on Domain Driven Data Mining.

[9] Chaibub Neto, E., et al. (2019) Causality-based tests to detect the influence of confounders on mobile health diagnostic applications: a comparison with restricted permutations. In Machine Learning for Health (ML4H) Workshop at NeurIPS 2019 - Extended Abstract. arXiv:1911.05139.

[10] Chaibub Neto, E. (2020) Causality-aware counterfactual confounding adjustment as an alternative to linear residualization in anticausal prediction tasks based on linear learners. *arXiv:2011.04605*

[11] Chaibub Neto, E. (2020) Causality-aware counterfactual confounding adjustment for feature representations learned by deep models. *arXiv:2004.09466*

[12] Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, **107**, 261-265.

[13] Correia, D. J. and Bareinboim, E. (2020) A calculus for stochastic interventions: causal effect identification and surrogate experiments. In AAAI 2020.

[14] Dudik, M., Phillips, S. J., and Schapire, R. E. (2006) Correcting sample selection bias in maximum entropy density estimation. NeurIPS 2006.

[15] Eberhardt, F. and Scheines, R. (2007) Interventions and causal inference. *Philosophy of Science*, **74**, 981-995.

[16] Fong, C., Hazlett, C., and Imai, K. (2018). Covariate balancing propensity score for a continuoustreatment: Application to the efficacy of political advertisements. The Annals of Applied Statistics,12(1), 156-177.

[17] Ghassami, A. E., Salehkaleybar, S., Kiyavash, N., Zhang, K. (2017) Learning causal structures using regression invariance. In *NIPS 2017*.

[18] Gretton, A., Smola, A. J., Huang, J., Schmittfull, M.,Borgwardt, K. M., and Scholkopf, B. (2009). Covariate shift by kernel mean matching. In Quinonero-Candela, et al., editors, *Dataset Shift in Machine Learning*, 131-160. The MIT Press.

[19] Gruber, S. and M. J. van der Laan (2010). A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *The International Journal of Biostatistics* **6** (1).

[20] Hahn, P. R., J. S. Murray, and C. M. Carvalho (2017). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects.

[21] Heckman, J. J. (1979) Sample selection bias as a specification error. *Econometrica*, **47**, 153-161.

[22] Heinze-Deml, C., Peters, J., Meinshausen, N. (2018) Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 20170016.

[23] Hernan, M., Hernandez-Diaz, S. and Robins, J. (2004). A structural approach to selection bias. *Epidemiology*, **15**, 615-625.

[24] Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* **20**, 217-240.

[25] Hirano, K. and Imbems, G. W. (2004). The propensity score with continuous treatments. In Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family 73–84. Wiley, New York.

[26] Huang, J., et al (2007) Correcting sample selection bias by unlabeled data. In NeurIPS 2007.

[27] Johansson, F. D., Shalit, U., and Sontag, D. (2016) Learning representations for counterfactual inference. *International Conference on Machine Learning (ICML)*, 2017.

[28] Kamiran, F. and Calders, T. (2012) Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, **33**, 1-33.

[29] Kocaoglu, M., Jaber, A., Shanmugam, K., and Bareinboim, E. (2019) Characterization and learning of causal graphs with latent variables from soft interventions. In NeurIPS 2019.

[30] Kreif, N. and DiazOrdaz, K. (2019) Machine learning in policy evaluation: new tools for causal inference. arXiv:1903.00402.

[31] Kuang, K., Cui, C., Athey, S., Xiong, R., Li, B. (2018) Stable prediction across unknown environments. In *SIGKDD 2018*.

[32] Kuang, K., Xiong, R., Cui, C., Athey, S., Li, B. (2020) Stable prediction with model misspecification and agnostic distribution shift. *arXiv:2001.11713*.

[33] Lee, B. K., J. Lessler, and E. A. Stuart (2010) Improving propensity score weighting using machine learning. *Statistics in Medicine*, **29**, 337-346.

[34] Lewis D. (2013) *Counterfactuals*. John Wiley & Sons.

[35] Liu, A. and Ziebart, B. (2014) Robust classification under sample selection bias. NeurIPS 2014.

[36] Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. (2018). Domain adaptation by using causal inference to predict invariant conditional distributions. *NeuRIPS 2018*.

[37] Malinsky D. (2018). Intervening on structure. *Synthese* **195**, 2295-2312.

[38] McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, **9**, 403.

[39] Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. (2009). *Dataset shift in machine learning.* MIT Press.

[40] Pearl, J. (2009) *Causality: models, reasoning, and inference.* Cambridge University Press New York, NY, 2nd edition.

[41] Pearl, J., Glymour, M., Jewell, N. P. (2016) *Causal inference in statistics: a primer.* Wiley.

[42] Pearl, J. (2019) The seven tools of causal inference with reflections on machine learning. *Communications of ACM*, **62**, 54-60.

[43] Peters, J., Buhlmann, P., Meinshausen, N. (2016) Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, series B*, **78**, 947-1012.

[44] Pirracchio, R., M. L. Petersen, and M. van der Laan (2015) Improving propensity score estimators' robustness to model misspecification using super learner. *American Journal of Epidemiology*, **181**, 108-119.

[45] R Core Team. (2019) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

[46] Richardson T. S., and Robins J. M. (2013) Single world intervention graphs (SIWGs): a unification of the counterfactual and graphical approaches to causality. *Working Paper Number 128 Center for Statistics and the Social Sciences, University of Washington*.

[47] Rojas-Carulla, M., Scholkopf, B., Turner, R., Peters, J. (2018) Invariant models for causal transfer learning. In *JMLR 2018*.

[48] Schölkopf B, Janzing D, Peters J, et al. (2012) On causal and anticausal learning. ICML 2012, 1255-1262.

[49] Schulam, P., Saria, S. (2017) Reliable Decision Support Using Counterfactual Models. In *NIPS 2017*.

[50] Shimodaira H. (2000) Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, **90**, 227-244.

[51] Sugiyama, M., Krauledat, M., and MAzller, K. R. (2007). Covariate shift adaptation by importance weighted cross-validation. *Journal of Machine Learning Research*, **8**, 985-1005.

[52] Subbaswamy A., Saria, S. (2018) Counterfactual normalization: proactively addressing dataset shift and improving reliability using causal mechanisms. *UAI 2018*.

[53] Subbaswamy, A., Schulam, P., Saria, S. (2019) Learning Predictive Models that Transport. *AISTATS 2019*.

[54] Subbaswamy A., Saria, S. (2020) From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, **2**, 345-352.

[55] Sobel, M. E. (1987) Direct and indirect effects in linear structural equation models. *Sociological Methods and Research*, **16**, 155-176.

[56] Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction and Search.* MIT Press, Cambridge, MA, 2nd edition.

[57] Swaminathan, A., and Joachims, T. (2015) Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research*, **16**, 1731-1755.

[58] Westreich, D., J. Lessler, and M. J. Funk (2010). Propensity score estimation: neural networks, support vector machines, decision trees (cart), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, **63**, 826-833.

[59] Wright, S. (1934) The method of path coefficients. *The Annals of Mathematical Statistics*, **5**:161-215.

[60] Wyss, R., A. R. Ellis, M. A. Brookhart, C. J. Girman, M. Jonsson Funk, R. LoCasale, and T. Sturmer (2014). The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bcart, and the covariate-balancing propensity score. *American Journal of Epidemiology*, **180**, 645-655.

[61] Zhu, Y., D. L. Coffman, and D. Ghosh (2015). A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of Causal Inference*, **3**, 25-40.

Supplement

1 Further related work

As clearly articulated by [54] there are, broadly speaking, two types of stable prediction approaches: (i) *reactive* methods, that use data (or knowledge) from the intended deployment/target population to correct for shifts; and (ii) *proactive* methods, that do not require data from the deployment/target populations, and are able to learn models that are stable with respect to unknown dataset shifts.

Many reactive approaches in the literature [50, 51, 14, 26, 18, 6, 35] deal with dataset shift by reweighting the training data to make it more closely aligned it with the target test distribution. In this paper, however, we focus on anticausal prediction tasks [48] and address only dataset shifts in the joint distribution of the confounders and outcome variable, $P(\boldsymbol{C}, Y)$, caused by selection biases [21, 23, 1]. In our particular context, we can still use simple reactive approaches when the target (test set) joint distribution, $P(\boldsymbol{C}_{ts}, Y_{ts})$ is known. For instance, if we know, a priori, the prevalence of a disease with respect to a given demographic risk factor in the target population, then we can either subsample or oversample the training data in order to make the training set distribution $P(\boldsymbol{C}_{tr}, Y_{tr})$ match the test set set distribution $P(\boldsymbol{C}_{ts}, Y_{ts})$. In classification tasks, simple balancing approaches, such as matching or approximate inverse probability weighting, can be used to subsample or oversample the training data. In regression tasks, approaches such as propensity scores for continuous variables [25], covariate balancing propensity score methods for continuous variables [16], or standard propensity score matching applied to dichotomized outcome data, can be used.

The more challenging case where we face unknown shifts in $P(\boldsymbol{C}_{ts}, Y_{ts})$ (the case we address in this paper) requires more sophisticated adjustment approaches. Several proactive methods have been proposed in the literature. For instance, invariant learning approaches [43, 47, 36, 3] employ multiple training datasets in order to learn invariant predictions. The causality-aware approach (adopted in this paper), on the other hand, only requires a single training set.

Another proactive approach, which can be applied to anticausal tasks based on a single training set, is the counterfactual normalization method proposed by [52]. The approach requires full knowledge of the causal graph describing the data generation process and is implemented in several steps. First, it identifies a set vulnerable variables that make the ML model susceptible to learning unstable relationships that might lead to poor generalization across shifted dataset. Second, the approach performs a node-splitting operation in order to augment the causal graph with counterfactual variables which isolate unstable paths of statistical associations and allow the retention of some stable paths involving vulnerable variables. Third, the approach determines a stable set of input variables that can be used to train a more stable ML model. In practice, the approach is implemented with linear (or additive) models.

Similarly to counterfactual normalization, the causality-aware approach also leverages counterfactual features to improve stability and is also implemented with linear models[9]. There are, nonetheless, important differences. The key idea (in the context of anticausal prediction tasks) is to train and evaluate supervised ML algorithms on counterfactually simulated data which retains only the associations generated by the causal influences of the output variable on the inputs. Noteworthy, as described in the main text, it is always possible to reparameterize the model in a way that the covariance among the features and among the confounders is pushed towards the respective error terms. This allows the generation of counterfactual features without even knowing the causal relations among features and the causal relations among the confounders. As a consequence, the causality-aware approach does not require knowledge of the full data generation process (at least for linear models). Contrary to counterfactual normalization, where the full causal diagram needs to be specified, the causality-aware approach only requires knowledge of which variables are confounders.

Finally, the methods proposed by [31, 32] represent another set of stable prediction approaches. The key idea behind these methods is to find a set of covariates for which the expected value of the outcome is stable across distinct test set environments. These covariates fall into two classes: stable variables ($\boldsymbol{S}$) that have an structural relationship with the outcome, and unstable variables ($\boldsymbol{V}$) that can be associated with both the outcome and the stable variables but do not have a causal relation with the outcome. Assuming that there exists a stable function $f(\boldsymbol{s})$ such that for all testing environments

[9] In reference [10] we describe how to causality-aware approach can be extended to additive models.

$E(Y \mid \boldsymbol{S} = \boldsymbol{s}, \boldsymbol{V} = \boldsymbol{v}) = E(Y \mid \boldsymbol{S} = \boldsymbol{s}) = f(\boldsymbol{s})$ - a condition which is fulfilled when $Y \perp\!\!\!\perp \boldsymbol{V} \mid \boldsymbol{S}$ - the approach is able to learn the stable function $f(\boldsymbol{s})$ without prior knowledge about which variables are stable or unstable. These methods, however, are tailored to causal prediction tasks (i.e., where the inputs have a causal effect on the outcome), and cannot be directly applied in anticausal tasks[10].

2 Additional univariate examples

Consider an anticausal prediction task, and suppose that our goal is to build a ML model whose predictive performance is only informed by the indirect causal effect of Y on X. In this case, we simulate data according to the twin network in Figure S1a, so that,

$$\begin{aligned} Cov(X^*, Y) &= Cov(\theta_{XM} M^* + U_X, Y) = \theta_{XM}\, Cov(M^*, Y) = \theta_{XM}\, Cov(\theta_{MY} Y + U_M, Y) \\ &= \theta_{XM}\, \theta_{MY}\, Cov(Y, Y) = \theta_{XM}\, \theta_{MY}\ . \end{aligned} \tag{3}$$

Now, suppose that the goal is to build a ML model whose predictive performance is only informed by the spurious associations generated by the confounder, we can simulate data according to the twin network in Figure S1b, so that,

$$Cov(X^*, Y) = Cov(\theta_{XC}\, C + U_X, \theta_{YC}\, C + U_Y) = \theta_{XC}\, \theta_{YC}\, Cov(C, C) = \theta_{XC}\, \theta_{YC}\ . \tag{4}$$

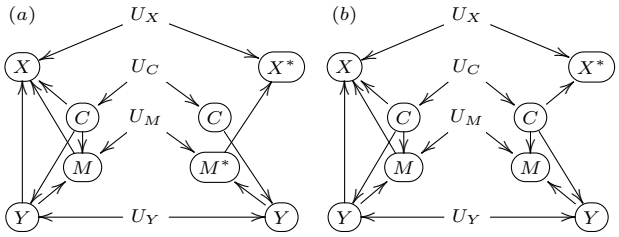


Figure S1: Twin network approach in the case where the indirect effect represents the causal effect of interest (panel a), and in the case where we are interested in estimating predictive performance that is due to confounding effects.

3 On alternative model modifications for simulating counterfactual data

In the main text (as well as, in the above section) we showed how to generate counterfactual data that contains only associations generated by the causal effects of interest. A natural question is whether alternative modifications of the causal diagram (other than the ones presented in the main text and in Supplementary Section 2) would also lead to counterfactual datasets containing only the associations due to the causal effects of interest. Here, we show that this is sometimes possible, and clarify that, for anticausal prediction tasks, the requirement for the intervention to work is that Y is not altered by the intervention.

We start with the case where the interest focus on the direct causal effects in anticausal predictive tasks. Here, the goal is to simulate counterfactual data where $Cov(X^*, Y^*) = \theta_{XY}$. Starting with examples involving confounding alone, consider first an alternative modification where we simulate data with the confounder variable C set to a fixed value c, as described in the twin network in Figure S12a. Direct calculation shows that,

$$\begin{aligned} Cov(X^*, Y^*) &= Cov(\theta_{XC}\, c + \theta_{XY}\, Y^* + U_X, Y^*) \\ &= \theta_{XY}\, Cov(Y^*, Y^*) = \theta_{XY}\, Var(Y^*) \\ &= \theta_{XY}\, Var(\theta_{YC}\, c + U_Y) = \theta_{XY}\, Var(U_Y) \\ &= \theta_{XY}(1 - \theta_{YC}^2)\ , \end{aligned}$$

for any chosen c value. (Note that $Var(U_Y) = 1 - \theta_{YC}^2$ since $1 = Var(Y) = Var(\theta_{YC}\, C + U_Y) = \theta_{YC}^2 Var(C) + Var(U_Y) = \theta_{YC}^2 + Var(U_Y)$.) Now, consider another alternative modification

[10]Note that in anticausal prediction tasks $\boldsymbol{S}$ might be a collider. Hence, if $\boldsymbol{S}$ is a collider, it follows that conditional on $\boldsymbol{S}$, Y cannot be independent of $\boldsymbol{V}$, and the assumption $Y \perp\!\!\!\perp \boldsymbol{V} \mid \boldsymbol{S}$ cannot hold.

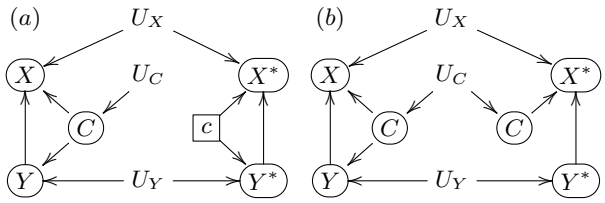


Figure S2: Alternative model modifications for the confounding only examples.

where we drop the causal link $C \rightarrow Y$ (rather than $C \rightarrow X$) as shown in Figure S12b. Note that direct calculation of $Cov(X^*, Y^*)$ shows again that,

$$\begin{aligned} Cov(X^*, Y^*) &= Cov(\theta_{XY}\, Y^* + \theta_{XC}\, C + U_X, Y^*) \\ &= \theta_{XY}\, Cov(Y^*, Y^*) \\ &= \theta_{XY}\, Var(Y^*) = \theta_{XY}\, Var(U_Y) \\ &= \theta_{XY}(1 - \theta_{YC}^2)\ . \end{aligned}$$

Hence, we see that for both alternative modifications presented in Figure S12 the covariance between the response and the feature does not equal θ_{XY}, the association due to the causal effect of Y on X. (Note that in both examples the intervention altered the original variable Y.)

Now, we show that for the mediation only example, these alternative modifications still capture the correct covariance because, in this case, these modifications do not alter Y. For instance, by setting the mediator M to the fixed value m, as described in Figure S3a, we still have that,

$$\begin{aligned} Cov(X^*, Y) &= Cov(\theta_{XY}\, Y + \theta_{XM}\, m + U_X, Y) \\ &= \theta_{XY}\, Cov(Y, Y) + \theta_{XM}\, Cov(m, Y) + Cov(U_X, Y) \\ &= \theta_{XY}\, Var(Y) = \theta_{XY}\ . \end{aligned}$$

Similarly, note that by dropping the causal link $Y \rightarrow M$ (rather than $M \rightarrow X$), as described in Figure S3b, we still have that,

$$\begin{aligned} Cov(X^*, Y) &= Cov(\theta_{XY}\, Y + \theta_{XM}\, M^* + U_X, Y) \\ &= \theta_{XY}\, Cov(Y, Y) + \theta_{XM}\, Cov(M^*, Y) + Cov(U_X, Y) \\ &= \theta_{XY}\, Var(Y) = \theta_{XY}\ . \end{aligned}$$

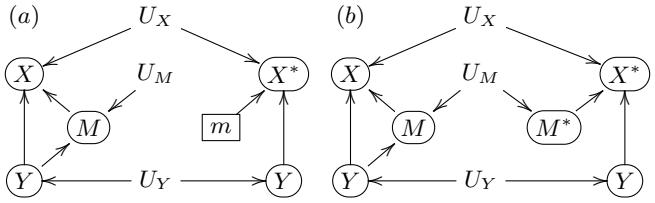


Figure S3: Alternative model modifications for the for the mediation only examples.

These examples show that for the mediation problem we don't necessarily need to simulate counterfactual features by dropping M from the parent set of X. From a practical point of view, however, it is still more advantageous to simulate counterfactual features by dropping the causal link $M \rightarrow X$ since this approach only requires the simulation of the counterfactual features, whereas the approach described in Figure S3a requires us to set M to m, and the approach in Figure S3b requires the simulation of counterfactual mediator data, M^*, in addition to the simulation of counterfactual feature data, X^*.

Now, let's consider indirect causal effects in anticausal prediction tasks. Here, the goal is to simulate counterfactual data where $Cov(X^*, Y^*) = \theta_{XM}\, \theta_{MY}$. Consider first the alternative intervention where we remove the link $C \rightarrow Y$ (rather than $C \rightarrow X$, as we did in Figure 2 in the main text) in

addition to removing $Y \rightarrow X$ and $C \rightarrow M$, as shown in Figure S4a. Note that, in this case, the intervention altered Y and we have that,

$$\begin{aligned} Cov(X^*, Y^*) &= Cov(\theta_{XC} C + \theta_{XM} M^* + U_X, U_Y) \\ &= \theta_{XM}\, Cov(M^*, U_Y) \\ &= \theta_{XM}\, Cov(\theta_{MY} Y^* + U_M, U_Y) \\ &= \theta_{XM}\, \theta_{MY}\, Cov(Y^*, U_Y) = \theta_{XM}\, \theta_{MY}\, Var(U_Y) \\ &= \theta_{XM}\, \theta_{MY}\, (1 - \theta_{YC}^2)\ . \end{aligned}$$

Similarly, for the intervention where we set C to c we also alter Y and we have that,

$$\begin{aligned} Cov(X^*, Y^*) &= Cov(\theta_{XC} c + \theta_{XM} M^* + U_X, Y^*) \\ &= \theta_{XM}\, Cov(M^*, Y^*) \\ &= \theta_{XM}\, Cov(\theta_{MY} Y^* + \theta_{MC}\, c + U_M, Y^*) \\ &= \theta_{XM}\, \theta_{MY}\, Var(Y^*) \\ &= \theta_{XM}\, \theta_{MY}\, Var(\theta_{YC}\, c + U_Y) \\ &= \theta_{XM}\, \theta_{MY}\, Var(U_Y) \\ &= \theta_{XM}\, \theta_{MY}\, (1 - \theta_{YC}^2)\ . \end{aligned}$$

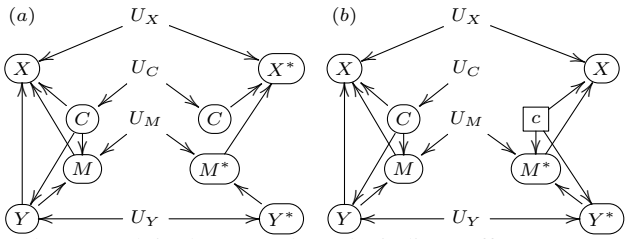


Figure S4: Twin network approach in the case where the indirect effect represents the causal effect of interest.

These examples once again illustrate that we are unable to recover the associations generated by the indirect effects (namely, $\theta_{XM}\, \theta_{MY}$) when we alter Y in anticausal tasks.

4 Node-splitting transformations as alternative interventions

In this section we show that the adoption of node-splitting transformations [46] encoded in single world intervention graphs (SWIGs) can also be used as an alternative intervention for the generation of counterfactual data that contains only the associations generated by the causal mechanisms of interest. Here, we present SWIGs that capture exactly the same marginal associations between the counterfactual features and responses, as the twin-networks presented in Figure 2 in the main text, and in Supplementary Figures S1a and b.

Figure S5 presents the SWIGs for the generation of counterfactual features in the anticausal prediction tasks. Here, a node-split operation associated with the intervention $do(Z = z)$ is represented by splitting the node $\textcircled{Z}$ into two elements: $\boxed{z}$ representing the instantiation of Z to the fixed value z; and $\textcircled{Z}$ representing the random variable Z.

In Figure S5a we split the C and M nodes in order to obtain a counterfactual feature $X_{c,m,Y}$, whose association with Y is generated by the direct causal effect θ_{XY}, since for any fixed values of c and m we have that,

$$\begin{aligned} Cov(X_{c,m,Y}, Y) &= Cov(\theta_{XC}\, c + \theta_{XM}\, m + \theta_{XY}\, Y + U_X, Y) \\ &= \theta_{XY}\, Cov(Y, Y) = \theta_{XY}\ . \end{aligned}$$

In Figure S5b we split the C and Y nodes in order to obtain a counterfactual feature $X_{y,c,M_{c,Y}}$, whose association with Y is generated by the indirect causal effect $\theta_{XM}\, \theta_{MY}$, since for any fixed

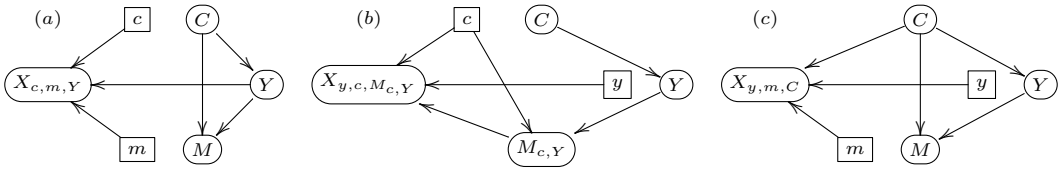


Figure S5: SWIGs for the anticausal predictive tasks.

values of c and y we have that,

$$\begin{aligned} Cov(X_{y,c,M_{c,Y}}, Y) &= Cov(\theta_{XC}\, c + \theta_{XY}\, y + \theta_{XM}\, M_{c,Y} + U_X, Y) \\ &= \theta_{XM}\, Cov(M_{c,Y}, Y) \\ &= \theta_{XM}\, Cov(\theta_{MC}\, c + \theta_{MY}\, Y + U_M, Y) \\ &= \theta_{XM}\, \theta_{MY}\, Cov(Y, Y) = \theta_{XM}\, \theta_{MY}\ . \end{aligned}$$

Finally, in Figure S5c we split the Y and M nodes in order to obtain a counterfactual feature $X_{y,m,C}$, whose association with Y, measured by $\theta_{XC}\, \theta_{YC}$, is generated by the confounder C. Note that for any fixed values of y and m we have that,

$$\begin{aligned} Cov(X_{y,m,C}, Y) &= Cov(\theta_{XC}\, C + \theta_{XM}\, m + \theta_{XY}\, y + U_X, Y) \\ &= \theta_{XC}\, Cov(C, Y) \\ &= \theta_{XC}\, Cov(C, \theta_{YC}\, C + U_Y) \\ &= \theta_{XC}\, \theta_{YC}\, Cov(C, C) = \theta_{XC}\, \theta_{YC}\ . \end{aligned}$$

Note that in the SWIG framework, even when we split the Y node into $\boxed{y}$ and $\textcircled{Y}$ in anticausal prediction tasks (e.g., Figure S5b and c), we have that the component $\textcircled{Y}$ still represents the un-altered random variable Y. (This observation is again consistent with the point made in the previous section that for anticausal prediction tasks, the requirement for the intervention to work is that Y is not altered by the intervention.)

5 Anticausal reparameterization example

Here, we present an illustrative example of the reparameterization for the anticausal prediction task. The goal is to provide a concrete example to help out readers that are not familiar with the notation used in the linear structural equations models. Figure S6a presents an illustrative example of the actual data generation process, whereas Figure S7 represents the reparameterized model.

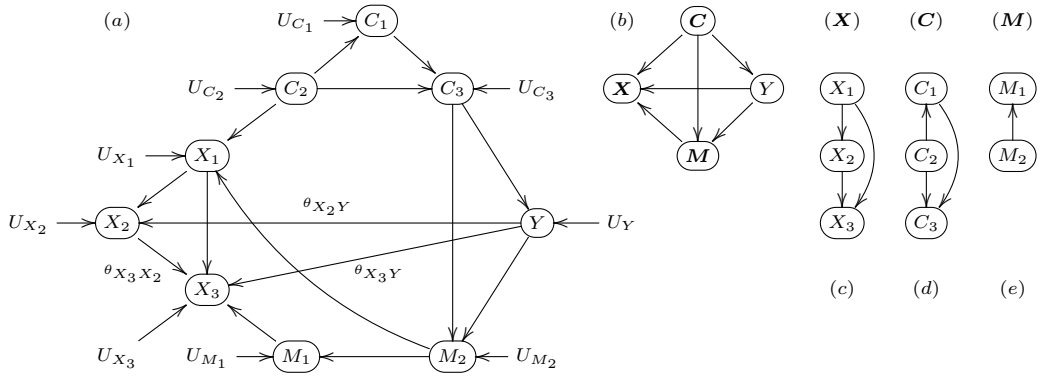


Figure S6: Original anticausal prediction task example. Panel a shows the actual data generation process. Panel b shows the multivariate representation of the DAG in panel a. Panels c, d, and e show, respectively, the DAG subdiagrams represented by the $\boldsymbol{X}$, $\boldsymbol{C}$, and $\boldsymbol{M}$ nodes in panel b.

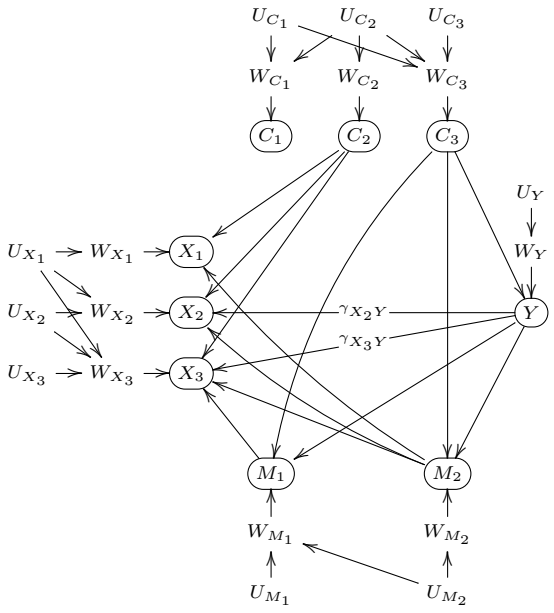


Figure S7: Reparameterized model for the anticausal prediction task example in Figure S6a.

For the anticausal prediction task DAG in Figure S6, we have that the structural equations,

$$
\begin{aligned}
\boldsymbol{C} &= \boldsymbol{\Theta}_{CC}\,\boldsymbol{C} + \boldsymbol{U}_C\ , \\
Y &= \boldsymbol{\Theta}_{YC}\,\boldsymbol{C} + U_Y\ , \\
\boldsymbol{M} &= \boldsymbol{\Theta}_{MM}\,\boldsymbol{M} + \boldsymbol{\Theta}_{MC}\,\boldsymbol{C} + \boldsymbol{\Theta}_{MY}\,Y + \boldsymbol{U}_M\ , \\
\boldsymbol{X} &= \boldsymbol{\Theta}_{XX}\,\boldsymbol{X} + \boldsymbol{\Theta}_{XC}\,\boldsymbol{C} + \boldsymbol{\Theta}_{XM}\,\boldsymbol{M} + \boldsymbol{\Theta}_{XY}\,Y + \boldsymbol{U}_X\ ,
\end{aligned}
$$

are explicitly given by,

$$
\underbrace{\begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}}_{\boldsymbol{C}} = \underbrace{\begin{pmatrix} 0 & \theta_{C_1C_2} & 0 \\ 0 & 0 & 0 \\ \theta_{C_3C_1} & \theta_{C_3C_2} & 0 \end{pmatrix}}_{\boldsymbol{\Theta}_{CC}} \underbrace{\begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}}_{\boldsymbol{C}} + \underbrace{\begin{pmatrix} U_{C_1} \\ U_{C_2} \\ U_{C_3} \end{pmatrix}}_{\boldsymbol{U}_C}\ ,
$$

$$
Y = \underbrace{\begin{pmatrix} 0 & 0 & \theta_{YC_3} \end{pmatrix}}_{\boldsymbol{\Theta}_{YC}} \underbrace{\begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}}_{\boldsymbol{C}} + U_Y\ ,
$$

$$
\underbrace{\begin{pmatrix} M_1 \\ M_2 \end{pmatrix}}_{\boldsymbol{M}} = \underbrace{\begin{pmatrix} 0 & \theta_{M_1M_2} \\ 0 & 0 \end{pmatrix}}_{\boldsymbol{\Theta}_{MM}} \underbrace{\begin{pmatrix} M_1 \\ M_2 \end{pmatrix}}_{\boldsymbol{M}} + \underbrace{\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \theta_{M_2C_3} \end{pmatrix}}_{\boldsymbol{\Theta}_{MC}} \underbrace{\begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}}_{\boldsymbol{C}} + \underbrace{\begin{pmatrix} 0 \\ \theta_{M_2Y} \end{pmatrix}}_{\boldsymbol{\Theta}_{MY}} Y + \underbrace{\begin{pmatrix} U_{M_1} \\ U_{M_2} \end{pmatrix}}_{\boldsymbol{U}_M}\ ,
$$

$$
\begin{aligned}
\underbrace{\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}}_{\boldsymbol{X}} = {} & \underbrace{\begin{pmatrix} 0 & 0 & 0 \\ \theta_{X_2X_1} & 0 & 0 \\ \theta_{X_3X_1} & \theta_{X_3X_2} & 0 \end{pmatrix}}_{\boldsymbol{\Theta}_{XX}} \underbrace{\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}}_{\boldsymbol{X}} + \underbrace{\begin{pmatrix} 0 & \theta_{X_1C_2} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}}_{\boldsymbol{\Theta}_{XC}} \underbrace{\begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}}_{\boldsymbol{C}} + \\
& + \underbrace{\begin{pmatrix} 0 & \theta_{X_1M_2} \\ 0 & 0 \\ \theta_{X_3M_1} & 0 \end{pmatrix}}_{\boldsymbol{\Theta}_{XM}} \underbrace{\begin{pmatrix} M_1 \\ M_2 \end{pmatrix}}_{\boldsymbol{M}} + \underbrace{\begin{pmatrix} 0 \\ \theta_{X_2Y} \\ \theta_{X_3Y} \end{pmatrix}}_{\boldsymbol{\Theta}_{XY}} Y + \underbrace{\begin{pmatrix} U_{X_1} \\ U_{X_2} \\ U_{X_3} \end{pmatrix}}_{\boldsymbol{U}_X}\ .
\end{aligned}
$$

Using simple algebraic manipulations, we can re-write the above linear structural models as,

$$\begin{aligned}
\boldsymbol{C} &= \boldsymbol{W}_C \ , \\
Y &= \boldsymbol{\Gamma}_{YC}\, \boldsymbol{C} + W_Y \ , \\
\boldsymbol{M} &= \boldsymbol{\Gamma}_{MC}\, \boldsymbol{C} + \boldsymbol{\Gamma}_{MY}\, Y + \boldsymbol{W}_M \ , \\
\boldsymbol{X} &= \boldsymbol{\Gamma}_{XC}\, \boldsymbol{C} + \boldsymbol{\Gamma}_{XM}\, \boldsymbol{M} + \boldsymbol{\Gamma}_{XY}\, Y + \boldsymbol{W}_X \ ,
\end{aligned}$$

where,

$$\boldsymbol{W}_C = (\boldsymbol{I} - \boldsymbol{\Theta}_{CC})^{-1}\boldsymbol{U}_C \ , \ W_Y = U_Y \ , \ \boldsymbol{W}_M = (\boldsymbol{I} - \boldsymbol{\Theta}_{MM})^{-1}\boldsymbol{U}_M \ , \ \boldsymbol{W}_X = (\boldsymbol{I} - \boldsymbol{\Theta}_{XX})^{-1}\boldsymbol{U}_X \ ,$$

and,

$$\begin{aligned}
&\boldsymbol{\Gamma}_{YC} = \boldsymbol{\Theta}_{YC} \ , \ \ \boldsymbol{\Gamma}_{MC} = (\boldsymbol{I} - \boldsymbol{\Theta}_{MM})^{-1}\boldsymbol{\Theta}_{MC} \ , \ \ \boldsymbol{\Gamma}_{MY} = (\boldsymbol{I} - \boldsymbol{\Theta}_{MM})^{-1}\boldsymbol{\Theta}_{MY} \ , \\
&\boldsymbol{\Gamma}_{XC} = (\boldsymbol{I} - \boldsymbol{\Theta}_{XX})^{-1}\boldsymbol{\Theta}_{XC} \ , \ \ \boldsymbol{\Gamma}_{XM} = (\boldsymbol{I} - \boldsymbol{\Theta}_{XX})^{-1}\boldsymbol{\Theta}_{XM} \ , \ \ \boldsymbol{\Gamma}_{XY} = (\boldsymbol{I} - \boldsymbol{\Theta}_{XX})^{-1}\boldsymbol{\Theta}_{XY} \ .
\end{aligned}$$

Next, the present the explicit form of parameters and error terms for the particular example in Figure S6. Starting with model $\boldsymbol{C} = \boldsymbol{W}_C$, we have that,

$$(\boldsymbol{I} - \boldsymbol{\Theta}_{CC})^{-1} = \begin{pmatrix} 1 & \theta_{C_1C_2} & 0 \\ 0 & 1 & 0 \\ \theta_{C_3C_1} & \theta_{C_3C_2} + \theta_{C_3C_1}\,\theta_{C_1C_2} & 1 \end{pmatrix} \ ,$$

so that,

$$\begin{aligned}
\boldsymbol{W}_C &= (\boldsymbol{I} - \boldsymbol{\Theta}_{CC})^{-1}\boldsymbol{U}_C \\
&= \begin{pmatrix} 1 & \theta_{C_1C_2} & 0 \\ 0 & 1 & 0 \\ \theta_{C_3C_1} & \theta_{C_3C_2} + \theta_{C_3C_1}\,\theta_{C_1C_2} & 1 \end{pmatrix} \begin{pmatrix} U_{C_1} \\ U_{C_2} \\ U_{C_3} \end{pmatrix} \\
&= \begin{pmatrix} U_{C_1} + \theta_{C_1C_2}\, U_{C_2} \\ U_{C_2} \\ U_{C_3} + U_{C_2}(\theta_{C_3C_2} + \theta_{C_3C_1}\,\theta_{C_1C_2}) + U_{C_1}\,\theta_{C_3C_1} \end{pmatrix} = \begin{pmatrix} W_{C_1} \\ W_{C_2} \\ W_{C_3} \end{pmatrix} \ ,
\end{aligned}$$

For the model $Y = \boldsymbol{\Gamma}_{YC}\, \boldsymbol{C} + W_Y$, we have that,

$$\begin{aligned}
\boldsymbol{\Gamma}_{YC} &= \boldsymbol{\Theta}_{YC} \\
&= (0 \quad 0 \quad \theta_{YC_3}) \\
&= (\gamma_{YC_1} \quad \gamma_{YC_2} \quad \gamma_{YC_3}) \ , \\
W_Y &= U_Y \ .
\end{aligned}$$

For the model $\boldsymbol{M} = \boldsymbol{\Gamma}_{MC}\, \boldsymbol{C} + \boldsymbol{\Gamma}_{MY}\, Y + \boldsymbol{W}_M$, we have that,

$$(\boldsymbol{I} - \boldsymbol{\Theta}_{MM})^{-1} = \begin{pmatrix} 1 & \theta_{M_1M_2} \\ 0 & 1 \end{pmatrix} \ ,$$

so that,

$$\begin{aligned}
\boldsymbol{\Gamma}_{MC} &= (\boldsymbol{I} - \boldsymbol{\Theta}_{MM})^{-1}\boldsymbol{\Theta}_{MC} \\
&= \begin{pmatrix} 1 & \theta_{M_1M_2} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \theta_{M_2C_3} \end{pmatrix} \\
&= \begin{pmatrix} 0 & 0 & \theta_{M_1M_2}\,\theta_{M_2C_3} \\ 0 & 0 & \theta_{M_2C_3} \end{pmatrix} = \begin{pmatrix} \gamma_{M_1C_1} & \gamma_{M_1C_2} & \gamma_{M_1C_3} \\ \gamma_{M_2C_1} & \gamma_{M_2C_2} & \gamma_{M_2C_3} \end{pmatrix} \ ,
\end{aligned}$$

and,

$$\begin{aligned}
\boldsymbol{\Gamma}_{MY} &= (\boldsymbol{I} - \boldsymbol{\Theta}_{MM})^{-1}\boldsymbol{\Theta}_{MY} \\
&= \begin{pmatrix} 1 & \theta_{M_1M_2} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ \theta_{M_2Y} \end{pmatrix} = \begin{pmatrix} \theta_{M_1M_2}\,\theta_{M_2Y} \\ \theta_{M_2Y} \end{pmatrix} = \begin{pmatrix} \gamma_{M_1Y} \\ \gamma_{M_2Y} \end{pmatrix} \ ,
\end{aligned}$$

and,

$$\begin{aligned}\boldsymbol{W}_M &= (\boldsymbol{I} - \boldsymbol{\Theta}_{MM})^{-1}\boldsymbol{U}_M \\ &= \begin{pmatrix} 1 & \theta_{M_1M_2} \\ 0 & 1 \end{pmatrix}\begin{pmatrix} U_{M_1} \\ U_{M_2} \end{pmatrix} = \begin{pmatrix} U_{M_1} + \theta_{M_1M_2}\, U_{M_2} \\ U_{M_2} \end{pmatrix} = \begin{pmatrix} W_{M_1} \\ W_{M_2} \end{pmatrix}.\end{aligned}$$

Finally, for the model $\boldsymbol{X} = \boldsymbol{\Gamma}_{XC}\,\boldsymbol{C} + \boldsymbol{\Gamma}_{XM}\,\boldsymbol{M} + \boldsymbol{\Gamma}_{XY}\,Y + \boldsymbol{W}_X$, we have that,

$$(\boldsymbol{I} - \boldsymbol{\Theta}_{XX})^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ \theta_{X_2X_1} & 1 & 0 \\ \theta_{X_3X_1} + \theta_{X_2X_1}\theta_{X_3X_2} & \theta_{X_3X_2} & 1 \end{pmatrix},$$

so that,

$$\begin{aligned}\boldsymbol{\Gamma}_{XC} &= (\boldsymbol{I} - \boldsymbol{\Theta}_{XX})^{-1}\boldsymbol{\Theta}_{XC} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ \theta_{X_2X_1} & 1 & 0 \\ \theta_{X_3X_1} + \theta_{X_2X_1}\,\theta_{X_3X_2} & \theta_{X_3X_2} & 1 \end{pmatrix}\begin{pmatrix} 0 & \theta_{X_1C_2} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & \theta_{X_1C_2} & 0 \\ 0 & \theta_{X_1C_2}\,\theta_{X_2X_1} & 0 \\ 0 & \theta_{X_1C_2}\,\theta_{X_3X_1} + \theta_{X_1C_2}\,\theta_{X_2X_1}\,\theta_{X_3X_2} & 0 \end{pmatrix} = \begin{pmatrix} \gamma_{X_1C_1} & \gamma_{X_1C_2} & \gamma_{X_1C_3} \\ \gamma_{X_2C_1} & \gamma_{X_2C_2} & \gamma_{X_2C_3} \\ \gamma_{X_3C_1} & \gamma_{X_3C_2} & \gamma_{X_3C_3} \end{pmatrix},\end{aligned}$$

and,

$$\begin{aligned}\boldsymbol{\Gamma}_{XM} &= (\boldsymbol{I} - \boldsymbol{\Theta}_{XX})^{-1}\boldsymbol{\Theta}_{XM} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ \theta_{X_2X_1} & 1 & 0 \\ \theta_{X_3X_1} + \theta_{X_2X_1}\,\theta_{X_3X_2} & \theta_{X_3X_2} & 1 \end{pmatrix}\begin{pmatrix} 0 & \theta_{X_1M_2} \\ 0 & 0 \\ \theta_{X_3M_1} & 0 \end{pmatrix} \\ &= \begin{pmatrix} 0 & \theta_{X_1M_2} \\ 0 & \theta_{X_1M_2}\,\theta_{X_2X_1} \\ \theta_{X_3M_1} & \theta_{X_1M_2}\,\theta_{X_3X_1} + \theta_{X_1M_2}\,\theta_{X_2X_1}\,\theta_{X_3X_2} \end{pmatrix} = \begin{pmatrix} \gamma_{X_1M_1} & \gamma_{X_1M_2} \\ \gamma_{X_2M_1} & \gamma_{X_2M_2} \\ \gamma_{X_3M_1} & \gamma_{X_3M_2} \end{pmatrix},\end{aligned}$$

and,

$$\begin{aligned}\boldsymbol{\Gamma}_{XY} &= (\boldsymbol{I} - \boldsymbol{\Theta}_{XX})^{-1}\boldsymbol{\Theta}_{XY} \\ &= \begin{pmatrix} 1 & 0 & 0 \\ \theta_{X_2X_1} & 1 & 0 \\ \theta_{X_3X_1} + \theta_{X_2X_1}\theta_{X_3X_2} & \theta_{X_3X_2} & 1 \end{pmatrix}\begin{pmatrix} 0 \\ \theta_{X_2Y} \\ \theta_{X_3Y} \end{pmatrix} \\ &= \begin{pmatrix} 0 \\ \theta_{X_2Y} \\ \theta_{X_3Y} + \theta_{X_3X_2}\,\theta_{X_2Y} \end{pmatrix} = \begin{pmatrix} \gamma_{X_1Y} \\ \gamma_{X_2Y} \\ \gamma_{X_3Y} \end{pmatrix},\end{aligned}$$

and,

$$\begin{aligned}\boldsymbol{W}_X &= (\boldsymbol{I} - \boldsymbol{\Theta}_{XX})^{-1}\boldsymbol{U}_X \\ &= \begin{pmatrix} 1 & 0 & 0 \\ \theta_{X_2X_1} & 1 & 0 \\ \theta_{X_3X_1} + \theta_{X_2X_1}\theta_{X_3X_2} & \theta_{X_3X_2} & 1 \end{pmatrix}\begin{pmatrix} U_{X_1} \\ U_{X_2} \\ U_{X_3} \end{pmatrix} \\ &= \begin{pmatrix} U_{X_1} \\ U_{X_1}\theta_{X_2X_1} + U_{X_2} \\ U_{X_1}(\theta_{X_3X_1} + \theta_{X_2X_1}\theta_{X_3X_2}) + U_{X_2}\theta_{X_3X_2} + U_{X_3} \end{pmatrix} = \begin{pmatrix} W_{U_1} \\ W_{U_2} \\ W_{U_3} \end{pmatrix}.\end{aligned}$$

Table S1 compiles all the elements of $\boldsymbol{\Gamma}_{YC}$, $\boldsymbol{\Gamma}_{MC}$, $\boldsymbol{\Gamma}_{MY}$, $\boldsymbol{\Gamma}_{XC}$, $\boldsymbol{\Gamma}_{XM}$, and $\boldsymbol{\Gamma}_{XY}$. It presents the causal effects in the reparameterized model (represented by the γs) in terms of the original causal effects (represented by the θs). Note that the arrows in Figure S7 correspond to the non-zero γ causal effects in Table S1.

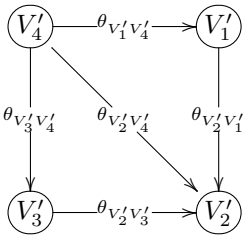
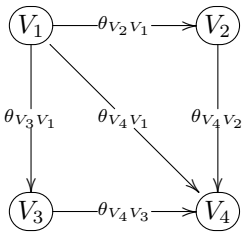
5.1 On the invertibility of $(\boldsymbol{I} - \boldsymbol{\Theta}_{VV})$

Here, it is important to point out that for any arbitrary DAG, we have that the matrix $(\boldsymbol{I} - \boldsymbol{\Theta}_{VV})$ is always invertible. To see why this is the case, note that for any arbitrary DAG we can always

$\gamma \in \mathbf{\Gamma}_{YC}$	$\gamma_{YC_3} = \theta_{YC_3}$ $\gamma_{YC_1} = \gamma_{YC_2} = 0$
$\gamma \in \mathbf{\Gamma}_{MC}$	$\gamma_{M_1C_3} = \theta_{M_1M_2}\,\theta_{M_2C_3}$ $\gamma_{M_2C_3} = \theta_{M_2C_3}$ $\gamma_{M_1C_1} = \gamma_{M_1C_2} = \gamma_{M_2C_1} = \gamma_{M_2C_2} = 0$
$\gamma \in \mathbf{\Gamma}_{MY}$	$\gamma_{M_1Y} = \theta_{M_1M_2}\,\theta_{M_2Y}$ $\gamma_{M_2Y} = \theta_{M_2Y}$
$\gamma \in \mathbf{\Gamma}_{XC}$	$\gamma_{X_1C_2} = \theta_{X_1C_2}$ $\gamma_{X_2C_2} = \theta_{X_2X_1}\,\theta_{X_1C_2}$ $\gamma_{X_3C_2} = \theta_{X_1C_2}\,\theta_{X_3X_1} + \theta_{X_1C_2}\,\theta_{X_2X_1}\,\theta_{X_3X_2}$ $\gamma_{X_1C_1} = \gamma_{X_1C_3} = \gamma_{X_2C_1} = \gamma_{X_2C_3} =$ $= \gamma_{X_3C_1} = \gamma_{X_3C_3} = 0$
$\gamma \in \mathbf{\Gamma}_{XM}$	$\gamma_{X_3M_1} = \theta_{X_3M_1}$ $\gamma_{X_1M_2} = \theta_{X_1M_2}$ $\gamma_{X_2M_2} = \theta_{X_1M_2}\,\theta_{X_2X_1}$ $\gamma_{X_3M_2} = \theta_{X_1M_2}\,\theta_{X_3X_1} + \theta_{X_1M_2}\,\theta_{X_2X_1}\,\theta_{X_3X_2}$ $\gamma_{X_1M_1} = \gamma_{X_2M_1} = 0$
$\gamma \in \mathbf{\Gamma}_{XY}$	$\gamma_{X_2Y} = \theta_{X_2Y}$ $\gamma_{X_3Y} = \theta_{X_3Y} + \theta_{X_3X_2}\,\theta_{X_2Y}$ $\gamma_{X_1Y} = 0$

Table S1: Causal effects in the reparameterized model.

rearrange the order of the variables so that $\mathbf{\Theta}_{VV}$ is a lower triangular matrix. For instance, we can rename and rearrange the order of the variables in the DAG M_a in (5) as $V_4' = V_1$, $V_1' = V_2$, $V_3' = V_3$, and $V_2' = V_4$, in order to obtain the rearranged DAG M_b in (5).

(M_a)  (M_b)  (5)

In this way, the original matrix $\mathbf{\Theta}_{V'V'}$,

$$\mathbf{\Theta}_{V'V'} = \begin{pmatrix} 0 & 0 & 0 & \theta_{V_1'V_4'} \\ \theta_{V_2'V_1'} & 0 & \theta_{V_2'V_3'} & \theta_{V_2'V_4'} \\ 0 & 0 & 0 & \theta_{V_3'V_4'} \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{6}$$

is rearranged as the lower triangular matrix,

$$\mathbf{\Theta}_{VV} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \theta_{V_2V_1} & 0 & 0 & 0 \\ \theta_{V_3V_1} & 0 & 0 & 0 \\ \theta_{V_4V_1} & \theta_{V_4V_2} & \theta_{V_4V_3} & 0 \end{pmatrix}. \tag{7}$$

Now, recalling that the determinant of a (lower or upper) triangular matrix is given by the product of its diagonal elements and that a triangular matrix is invertible if and only if none of its diagonal elements is zero, we see that $(\boldsymbol{I} - \mathbf{\Theta}_{VV})$ is always invertible because all diagonal elements are always equal to 1.

6 Remarks on identification issues

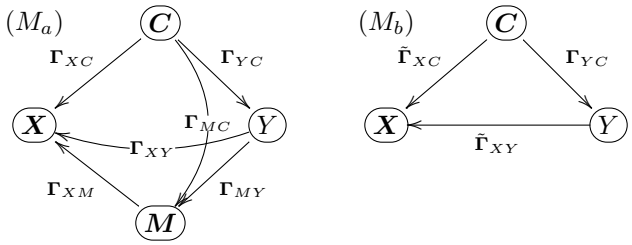
Under the assumption that all the confounders and mediators are observed, we can identify the direct and indirect causal effects of response on the features. In particular, a simple least squares

estimation procedure provides consistent estimates of these causal effects[11]. To see why, note that for the reparameterized model, if all confounders and mediators are observed, it follows from the Markov property of DAGs that $X_j = f_{X_j}(\boldsymbol{C}, \boldsymbol{M}, Y, W_{X_j}) = f_{X_j}(pa(X_j), W_{X_j})$. (Here, f_{X_j} represents linear structural causal models). Hence, for the anticausal task, it follows that, when we regress X_j on the elements of $\boldsymbol{C}$, $\boldsymbol{M}$, and Y only the coefficients associated with the parents of X_j in the reparameterized model will be statistically different from zero (for large enough sample sizes). Therefore, in practice, we don't need to know before hand which variables are the parents of X_j in the reparameterized model. The parent set will be learned automatically from the data by the regression model fit.

Observe, as well, that even if the mediators are unobserved, but the confounders are still observed, we can still identify total causal effects. For instance, we have that,

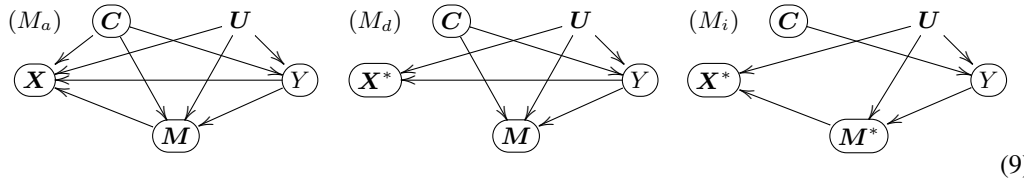
$$\begin{aligned}
\boldsymbol{X} &= \boldsymbol{\Gamma}_{XC}\,\boldsymbol{C} + \boldsymbol{\Gamma}_{XM}\,\boldsymbol{M} + \boldsymbol{\Gamma}_{XY}\,Y + \boldsymbol{W}_X\ , \\
&= \boldsymbol{\Gamma}_{XC}\,\boldsymbol{C} + \boldsymbol{\Gamma}_{XM}\,(\boldsymbol{\Gamma}_{MC}\,\boldsymbol{C} + \boldsymbol{\Gamma}_{MY}\,Y + \boldsymbol{W}_M) + \boldsymbol{\Gamma}_{XY}\,Y + \boldsymbol{W}_X\ , \\
&= \underbrace{(\boldsymbol{\Gamma}_{XC} + \boldsymbol{\Gamma}_{XM}\,\boldsymbol{\Gamma}_{MC})}_{\tilde{\boldsymbol{\Gamma}}_{XC}}\,\boldsymbol{C} + \underbrace{(\boldsymbol{\Gamma}_{XY} + \boldsymbol{\Gamma}_{XM}\,\boldsymbol{\Gamma}_{MY})}_{\tilde{\boldsymbol{\Gamma}}_{XY}}\,Y + \underbrace{\boldsymbol{\Gamma}_{XM}\,\boldsymbol{W}_M + \boldsymbol{W}_X}_{\tilde{\boldsymbol{W}}_X}\ , \\
&= \tilde{\boldsymbol{\Gamma}}_{XC}\,\boldsymbol{C} + \tilde{\boldsymbol{\Gamma}}_{XY}\,Y + \tilde{\boldsymbol{W}}_X\ ,
\end{aligned}$$

where $\tilde{\boldsymbol{\Gamma}}_{XY} = \boldsymbol{\Gamma}_{XY} + \boldsymbol{\Gamma}_{XM}\,\boldsymbol{\Gamma}_{MY}$ represents the total causal effect of Y on $\boldsymbol{X}$, as represented in the DAG M_b in in the causal task model (8).



(8)

On the other hand, if the mediators are observed, but some the confounders are unobserved, then neither the direct, the indirect, or the total causal effects are identifiable, and the predictions generated by the causality-aware approach will still be confounded. For instance, for the anticausal prediction tasks in model (9) the unmeasured confounders of the feature/response relationship will still confound the predictions.



(9)

Finally, observe that while so far we have discussed confounding of the feature/response relationship, it is also possible that the causal relations between features and mediators or between mediators and response are also influenced by confounders. If these confounders are unobserved, then we cannot identify the causal effects $\boldsymbol{\Gamma}_{XM}$ and $\boldsymbol{\Gamma}_{MY}$. Clearly, in the presence of unobserved confounding the causality-aware predictions will be biased, whenever the causal effects of interest are not identifiable.

7 Proof of Theorem 1

Before we present the proof, we first clarify that, in the multivariate case, the covariance between two vectors of random variables, $\boldsymbol{A} = (A_1, \ldots, A_{N_A})^T$ and $\boldsymbol{B} = (B_1, \ldots, B_{N_B})^T$, is given by the cross-covariance operator, $Cov(\boldsymbol{A}, \boldsymbol{B})$, defined and the $N_A \times N_B$ matrix with elements $Cov(A_i, B_j)$.

For the proof we will use the following properties of the cross-covariance operator:

[11]Here, we assume that the number of samples is larger than the number of covariates in the regression fits, and that multicolinearity is not an issue too.

1. $Cov(\boldsymbol{Z}_1 + \boldsymbol{Z}_2, \boldsymbol{Z}_3) = Cov(\boldsymbol{Z}_1, \boldsymbol{Z}_3) + Cov(\boldsymbol{Z}_2, \boldsymbol{Z}_3)$,
2. $Cov(\boldsymbol{A}\,\boldsymbol{Z}_1, \boldsymbol{B}\,\boldsymbol{Z}_2) = \boldsymbol{A}\,Cov(\boldsymbol{Z}_1, \boldsymbol{Z}_2)\,\boldsymbol{B}^T$, where $\boldsymbol{A}$ and $\boldsymbol{B}$ are constant matrices
3. $Cov(\boldsymbol{Z}, \boldsymbol{Z}) = Cov(\boldsymbol{Z})$, where $Cov(\boldsymbol{Z})$ is the variance covariance matrix of $\boldsymbol{Z}$.

The proof is straight forward, and follow directly from the above three properties. For completeness we restate the Theorem.

Theorem 1. *Consider an anticausal prediction task:*

(i) When the interest focus on the causal effects generated by the paths in $Y \rightarrow \boldsymbol{X}$. If $\boldsymbol{X}^*$ is given by $\boldsymbol{X}^* = \boldsymbol{\Gamma}_{XY}\,Y + \boldsymbol{W}_X$, then $Cov(\boldsymbol{X}^*, Y) = \boldsymbol{\Gamma}_{XY}$.

(ii) When the interest focus on the causal effects generated by the paths in $Y \rightarrow \boldsymbol{M} \rightarrow \boldsymbol{X}$. If $\boldsymbol{X}^*$ is given by $\boldsymbol{X}^* = \boldsymbol{\Gamma}_{XM}\,\boldsymbol{M}^* + \boldsymbol{W}_X$, and $\boldsymbol{M}^* = \boldsymbol{\Gamma}_{MY}\,Y + \boldsymbol{W}_M$, then $Cov(\boldsymbol{X}^*, Y) = \boldsymbol{\Gamma}_{XM}\,\boldsymbol{\Gamma}_{MY}$.

(iii) When the interest focus on the spurious associations generated by the paths in $\boldsymbol{X} \leftarrow \boldsymbol{C} \rightarrow Y$. If $\boldsymbol{X}^*$ is given by $\boldsymbol{X}^* = \boldsymbol{\Gamma}_{XC}\,\boldsymbol{C} + \boldsymbol{W}_X$, then $Cov(\boldsymbol{X}^*, Y) = \boldsymbol{\Gamma}_{XC}\,Cov(\boldsymbol{C})\,\boldsymbol{\Gamma}_{YC}^T$.

Proof.

Result *i*: If $\boldsymbol{X}^* = \boldsymbol{\Gamma}_{XY}\,Y + \boldsymbol{W}_X$, then,

$$\begin{aligned} Cov(\boldsymbol{X}^*, Y) &= Cov(\boldsymbol{\Gamma}_{XY}\,Y + \boldsymbol{W}_X, Y) \\ &= \boldsymbol{\Gamma}_{XY}\,Cov(Y, Y) \\ &= \boldsymbol{\Gamma}_{XY} \end{aligned}$$

Result *ii*: If $\boldsymbol{X}^* = \boldsymbol{\Gamma}_{XM}\,\boldsymbol{M}^* + \boldsymbol{W}_X$ and $\boldsymbol{M}^* = \boldsymbol{\Gamma}_{MY}\,Y + \boldsymbol{W}_M$, then,

$$\begin{aligned} Cov(\boldsymbol{X}^*, Y) &= Cov(\boldsymbol{\Gamma}_{XM}\,\boldsymbol{M}^* + \boldsymbol{W}_X, Y) \\ &= \boldsymbol{\Gamma}_{XM}\,Cov(M^*, Y) \\ &= \boldsymbol{\Gamma}_{XM}\,Cov(\boldsymbol{\Gamma}_{MY}\,Y + \boldsymbol{W}_M, Y) \\ &= \boldsymbol{\Gamma}_{XM}\,\boldsymbol{\Gamma}_{MY}\,Var(Y) \\ &= \boldsymbol{\Gamma}_{XM}\,\boldsymbol{\Gamma}_{MY} \end{aligned}$$

Result *iii*: If $\boldsymbol{X}^* = \boldsymbol{\Gamma}_{XC}\,\boldsymbol{C} + \boldsymbol{W}_X$, then,

$$\begin{aligned} Cov(\boldsymbol{X}^*, Y) &= Cov(\boldsymbol{\Gamma}_{XC}\,\boldsymbol{C} + \boldsymbol{W}_X, Y) \\ &= \boldsymbol{\Gamma}_{XC}\,Cov(\boldsymbol{C}, Y) \\ &= \boldsymbol{\Gamma}_{XC}\,Cov(\boldsymbol{C}, \boldsymbol{\Gamma}_{YC}\,\boldsymbol{C} + W_Y) \\ &= \boldsymbol{\Gamma}_{XC}\,Cov(\boldsymbol{C}, \boldsymbol{C})\,\boldsymbol{\Gamma}_{YC}^T \\ &= \boldsymbol{\Gamma}_{XC}\,Cov(\boldsymbol{C})\,\boldsymbol{\Gamma}_{YC}^T \end{aligned}$$

□

8 Expected MSE for arbitrary anticausal prediction tasks based on linear models

Consider the arbitrary anticausal prediction task model in Figure S8, where the double arrows connecting the variables $\{U_{X_1}, \ldots, U_{X_p}\}$ (and $\{U_{C_1}, \ldots, U_{C_m}\}$) represent the fact that these error terms are correlated[12]. Without loss of generality assume that the data has been centered, so that the

[12]Note that the above model might represent a reparameterization of a model with uncorrelated error terms and unknown causal relations among the $\boldsymbol{X}$ input variables, as well as, among the $\boldsymbol{C}$ confounder variables. As described in detail in the main text, for linear structural equation models, we can always reparameterize the original model in a way where the covariance structure among the input variables, as well as, the covariance structure among the confounder variables is pushed towards the respective error terms as illustrated in Figure S8.

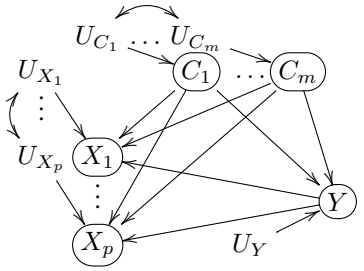


Figure S8: Confounded anticausal prediction task example.

linear structural causal models describing the data generation process are given by,

$$C_j = U_{C_j} \ , \tag{10}$$

$$Y = \sum_i \beta_{YC_i} C_i + U_Y \ , \tag{11}$$

$$X_j = \beta_{X_jY} Y + \sum_i \beta_{X_jC_i} C_i + U_{X_j} \ , \tag{12}$$

for $j = 1, \ldots, p$ and $i = 1, \ldots, m$. The causality-aware features are estimated as,

$$\hat{X}_j^* = X_j - \sum_i \hat{\beta}_{X_jC_i} C_i \ , \tag{13}$$

and converge asymptotically to,

$$X_j^* \ = \ X_j - \sum_i \beta_{X_jC_i} C_i \ = \ \beta_{X_jY} Y + U_{X_j} \ . \tag{14}$$

Now, let $\hat{Y} = \boldsymbol{X}_{ts}\hat{\boldsymbol{\beta}}^{tr}$ represent the prediction of a linear regression model, where $\boldsymbol{X}_{ts}$ represents the test set features, and $\hat{\boldsymbol{\beta}}^{tr}$ represents the regression coefficients estimated from the training set. By definition, the expected mean squared error of the prediction is given by,

$$\begin{aligned} E[MSE] &= E[(Y_{ts} - \hat{Y})^2] = E[Y_{ts}^2] + E[\hat{Y}^2] - 2E[\hat{Y}Y_{ts}] \\ &= Var(Y_{ts}) + E[\hat{Y}^2] - 2Cov(\hat{Y}, Y_{ts}) \ , \end{aligned}$$

since $E[Y_{ts}] = 0$. Direct computation shows that,

$$E[\hat{Y}^2] = E[(\sum_{j=1}^{p} X_{j,ts}\hat{\beta}_j^{tr})^2] = \sum_{j=1}^{p} (\hat{\beta}_j^{tr})^2 Var(X_{j,ts}) \ + 2\sum_{j<k} \hat{\beta}_j^{tr}\hat{\beta}_k^{tr} Cov(X_{j,ts}, X_{k,ts}) \ ,$$

and,

$$Cov(\hat{Y}, Y_{ts}) = \sum_{j=1}^{p} \hat{\beta}_j^{tr} Cov(X_{j,ts}, Y_{ts}) \ ,$$

so that,

$$\begin{aligned} E[MSE] \ = \ Var(Y_{ts}) \ + \sum_{j=1}^{p} (\hat{\beta}_j^{tr})^2 Var(X_{j,ts}) + \\ + 2\sum_{j<k} \hat{\beta}_j^{tr}\hat{\beta}_k^{tr} Cov(X_{j,ts}, X_{k,ts}) - 2\sum_{j=1}^{p} \hat{\beta}_j^{tr} Cov(X_{j,ts}, Y_{ts}) \ . \end{aligned}$$

Next, we derive the expressions for $Var(X_{j,ts})$, $Cov(X_{j,ts}, X_{k,ts})$, and $Cov(X_{j,ts}, Y_{ts})$ and show that they still depend on $Cov(Y_{ts}, C_{i,ts})$. From equation (12) we have that,

$$\begin{aligned} Var(X_{j,ts}) &= Var(\beta_{X_jY}\, Y_{ts} + \sum_i \beta_{X_jC_i}\, C_{i,ts} + U^{ts}_{X_j}) \\ &= \sigma^2_{X_j} + \beta^2_{X_jY}\, Var(Y_{ts}) + \sum_i \beta^2_{X_jC_i}\, Var(C_{i,ts}) + \\ &\quad + 2 \sum_{i<i'} \beta_{X_jC_i}\, \beta_{X_jC_{i'}}\, Cov(C_{i,ts}, C_{i',ts}) + 2\, \beta_{X_jY} \sum_i \beta_{X_jC_i}\, Cov(Y_{ts}, C_{i,ts})\ , \end{aligned}$$

$$\begin{aligned} Cov(X_{j,ts}, X_{k,ts}) &= Cov(\beta_{X_jY}\, Y_{ts} + \sum_i \beta_{X_jC_i}\, C_{i,ts} + U^{ts}_{X_j}\ ,\ \beta_{X_kY}\, Y_{ts} + \sum_i \beta_{X_kC_i}\, C_{i,ts} + U^{ts}_{X_k}) \\ &= \beta_{X_jY}\, \beta_{X_kY}\, Var(Y_{ts}) + \beta_{X_jY} \sum_i \beta_{X_kC_i}\, Cov(Y_{ts}, C_{i,ts}) + \\ &\quad + \beta_{X_kY} \sum_i \beta_{X_jC_i}\, Cov(Y_{ts}, C_{i,ts}) + \sum_i \sum_{i'} \beta_{X_jC_i}\, \beta_{X_kC_{i'}}\, Cov(C_{i,ts}, C_{i',ts}) + \\ &\quad + Cov(U^{ts}_{X_j}, U^{ts}_{X_k}) \end{aligned}$$

$$\begin{aligned} Cov(X_{j,ts}, Y_{ts}) &= Cov(\beta_{X_jY}\, Y_{ts} + \sum_i \beta_{X_jC_i}\, C_{i,ts} + U_{X_j}, Y_{ts}) \\ &= \beta_{X_jY}\, Var(Y_{ts}) + \sum_i \beta_{X_jC_i}\, Cov(Y_{ts}, C_{i,ts})\ , \end{aligned}$$

showing that these three quantities still depend on $Cov(Y_{ts}, C_{i,ts})$ (in addition to depending on $Var(Y_{ts})$, $Var(C_{ts})$, and $Cov(C_{i,ts}, C_{i',ts})$). This observation implies that the $E[MSE]$ will still be unstable w.r.t. shifts in these quantities, even when the regression model is trained in unconfounded data (a situation where the estimates $\hat{\beta}^{tr}_j$ are not influenced by spurious associations generated by the confounders). This explains why it is not enough to deconfound the training features alone. While training a regression model using deconfounded features allows us to estimate deconfounded model weights[13], $\hat{\boldsymbol{\beta}}^{tr}$, the prediction $\hat{Y} = \boldsymbol{X}_{ts}\hat{\boldsymbol{\beta}}^{tr}$ is a function of both the trained model $\hat{\boldsymbol{\beta}}^{tr}$ and the test set features, $\boldsymbol{X}_{ts}$. As a consequence, if we do not deconfound the test set features, the expected MSE will still be influenced by the confounders (since, in anticausal prediction tasks, the features are functions of both the confounder and outcome variables).

The expected MSE of models trained with test set features processed according to the causality-aware approach, on the other hand, do not depend on $Cov(Y_{ts}, C_{i,ts})$, $Var(C_{ts})$, or $Cov(C_{i,ts}, C_{i',ts})$, since the approach also deconfounds the test set features. Note that direct computation of $Var(X^*_{j,ts})$, $Cov(X^*_{j,ts}, X^*_{k,ts})$, and $Cov(X^*_{j,ts}, Y_{ts})$ based on the causality-aware features, $X^*_{j,ts} = \beta_{X_jY}\, Y_{ts} + U^{ts}_{X_j}$, shows that,

$$\begin{aligned} Var(X^*_{j,ts}) &= Var(\beta_{X_jY}\, Y_{ts} + U^{ts}_{X_j}) = \sigma^2_{X_j} + \beta^2_{X_jY}\, Var(Y_{ts})\ , \\ Cov(X^*_{j,ts}, X^*_{k,ts}) &= Cov(\beta_{X_jY}\, Y_{ts} + U^{ts}_{X_j}, \beta_{X_kY}\, Y_{ts} + U^{ts}_{X_k}) \\ &= \beta_{X_jY}\, \beta_{X_kY}\, Var(Y_{ts}) + Cov(U^{ts}_{X_j}, U^{ts}_{X_k})\ , \\ Cov(X^*_{j,ts}, Y_{ts}) &= Cov(\beta_{X_jY}\, Y_{ts} + U^{ts}_{X_j}, Y_{ts}) = \beta_{X_jY}\, Var(Y_{ts})\ , \end{aligned}$$

no longer depend on $Cov(Y_{ts}, C_{i,ts})$, $Var(C_{ts})$, or $Cov(C_{i,ts}, C_{i',ts})$, so that the approach will be stable against shifts in these quantities. Observe, nonetheless, that it will still be influenced by shifts in $Var(Y_{ts})$. (We point out, however, that the dependence of $E[MSE]$ on $Var(Y_{ts})$ is, in general, unavoidable since, by definition, $E[MSE] = Var(Y_{ts}) + E[\hat{Y}^2] - 2Cov(\hat{Y}, Y_{ts})$.)

[13]Note that the weights $\hat{\boldsymbol{\beta}}_{tr}$ are not causal effects, since they represent the coefficients of the regression of Y_{tr} on $\boldsymbol{X}_{tr}$, while in the true data generation process Y_{tr} is the independent variable and $\boldsymbol{X}_{tr}$ represents the dependent variables. Still, the estimate $\hat{\boldsymbol{\beta}}_{tr}$ will not absorb spurious associations when the model is trained with unconfounded data.

9 Extensions to arbitrary performance metrics and arbitrary structural causal models

Here, we extend the argument presented in the previous section to arbitrary performance metrics and arbitrary structural causal models.

Let $M = h_1(Y_{ts}, \hat{Y})$ represent an arbitrary performance metric, and let $\hat{Y} = h_2(\boldsymbol{\omega}_{tr}, \boldsymbol{X}_{ts}) = h_2(\boldsymbol{\omega}_{tr}, X_{1,ts}, \ldots, X_{p,ts})$ represent a prediction generated with an arbitrary ML model $\boldsymbol{\omega}_{tr}$. Note that $\boldsymbol{\omega}_{tr} = h_3(\boldsymbol{X}_{tr}, Y_{tr})$ is a function of the training data. Assume the features X_j are generated by an arbitrary structural causal model $X_j = f(Y, \boldsymbol{C}, U_{X_j})$. Then the expected value of M, with respect to the test set data distribution is given by,

$$\begin{aligned} E[M] &= E[h_1(Y_{ts}, \hat{Y})] \\ &= E[h_1(Y_{ts}, h_2(\boldsymbol{\omega}_{tr}, X_{1,ts}, \ldots, X_{p,ts}))] \\ &= E[h_1(Y_{ts}, h_2(\boldsymbol{\omega}_{tr}, f(Y_{ts}, \boldsymbol{C}_{ts}, U^{ts}_{X_1}), \ldots, f(Y_{ts}, \boldsymbol{C}_{ts}, U^{ts}_{X_p})))] \,, \end{aligned}$$

showing that even when we train the model $\boldsymbol{\omega}_{tr}$ using deconfounded training data, we have that $E[M]$ is still a function of $\boldsymbol{C}_{ts}$, and will be unstable with respect to shifts in $P(\boldsymbol{C}_{ts}, Y_{ts})$.

Observe, however, that if we are able to deconfound the test set features, so that the counterfactual features $X^*_{j,ts} = f^*(Y_{ts}, U^{ts}_{X_j})$ are no longer a function of the confounders, then we have that,

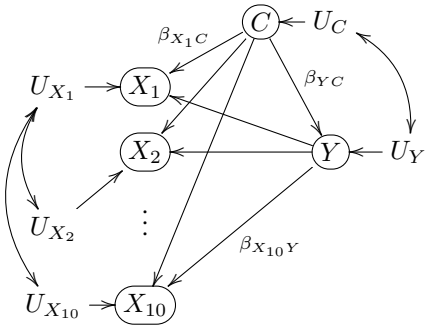
$$\begin{aligned} E[M] &= E[h_1(Y_{ts}, \hat{Y}^*)] \\ &= E[h_1(Y_{ts}, h_2(\boldsymbol{\omega}_{tr}, X^*_{1,ts}, \ldots, X^*_{p,ts}))] \\ &= E[h_1(Y_{ts}, h_2(\boldsymbol{\omega}_{tr}, f^*(Y_{ts}, U^{ts}_{X_1}), \ldots, f^*(Y_{ts}, U^{ts}_{X_p})))] \,, \end{aligned}$$

will no longer depend on $\boldsymbol{C}_{ts}$. Note that while the predictive performance will still depend on the distribution of Y_{ts} and, therefore, will still be unstable with respect to shifts in the marginal distribution $P(Y_{ts})$, the approach will still be stable with respect to shifts in the conditional distribution $P(\boldsymbol{C}_{ts} \mid Y_{ts})$.

10 Additional details - synthetic data experiments

In our experiments, we compare the causality-aware approach against two "archetypical" baselines: (1) one representing adjustment approaches that remove the causal effect of the confounders from the features, denoted *baseline 1*; and (2) another representing approaches that remove the association between the confounders and the output, denoted *baseline 2*. In both baseline approaches we adjust the training data but not the test set. Note that for both of these baselines, while the training data is unconfounded, the test data is still confounded. For the causality-aware approach, on the other hand, we generated confounded training and test sets and then apply our adjustment for both the training and test sets.

The confounded data is generated from the model,



where we change the covariance of the error terms U_C and U_Y in order to simulate the effects of selection biases in the joint distribution $P(C, Y)$.

The model is described by the following set of linear structural causal equations,

$$C = U_C \,, \tag{15}$$
$$Y = \beta_{YC}\, C + U_Y \,, \tag{16}$$
$$X_j = \beta_{X_jY}\, Y + \beta_{X_jC}\, C + U_{X_j} \,, \tag{17}$$

for $j = 1, \ldots, 10$, and where the error terms U_C and U_X are distributed according to,

$$\begin{pmatrix} U_C \\ U_Y \end{pmatrix} \sim \mathrm{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \phi_{CC} & \phi_{CY} \\ \phi_{CY} & \phi_{YY} \end{pmatrix} \right) \,, \tag{18}$$

and $\boldsymbol{U}_X = (U_{X_1}, \ldots, U_{X_{10}})^T$ is distributed according to a multivariate normal distribution,

$$\boldsymbol{U}_X \;\sim\; N_{10}(\mathbf{0}\,,\; \boldsymbol{\Sigma}_{\boldsymbol{U}_X}) \,, \tag{19}$$

where the ijth entry of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{U}_X}$ is given by 1 for $i = j$, and by $\rho^{|i-j|}$ for $i \neq j$.

Note that, for the above model, we have that,

$$Var(C) = \phi_{CC} \,, \tag{20}$$
$$\begin{aligned} Cov(Y, C) &= Cov(\beta_{YC}\, C + U_Y, C) = \beta_{YC}\, Var(C) + Cov(U_Y, C) \\ &= \beta_{YC}\, \phi_{CC} + \phi_{CY} \,, \end{aligned} \tag{21}$$
$$\begin{aligned} Var(Y) &= Var(\beta_{YC}\, C + U_Y) = \beta_{YC}^2\, Var(C) + Var(U_Y) + 2\, \beta_{YC}\, Cov(C, U_Y) \\ &= \beta_{YC}^2\, \phi_{CC} + \phi_{YY} + 2\, \beta_{YC}\, \phi_{CY} \,, \end{aligned} \tag{22}$$

so that for fixed values of $Var(C)$, $Cov(Y, C)$, $Var(Y)$, and β_{YC} we can determine the values of ϕ_{CC}, ϕ_{CY}, and ϕ_{YY} as follows,

$$\phi_{CC} = Var(C) \,, \tag{23}$$
$$\phi_{CY} = Cov(Y, C) - \beta_{YC}\, Var(C) \,, \tag{24}$$
$$\phi_{YY} = Var(Y) - \beta_{YC}^2\, Var(C) - 2\, \beta_{YC}\, Cov(Y, C) \,. \tag{25}$$

In our experiments, we simulate training and test set data as follows:

1. Sample the simulation parameters β_{X_jY}, β_{X_jC}, and β_{YC} from a $U(-1, 1)$ distribution, and ρ from a $U(-0.5, 0.5)$ distribution.
2. Given the fixed values for $Var(C_{tr})$, $Cov(Y_{tr}, C_{tr})$, and $Var(Y_{tr})$, and the sampled value for β_{YC}, we compute ϕ_{CC}, ϕ_{CY}, and ϕ_{YY} as described in equations (23), (24), and (25).
3. Sample the error terms U_C^{tr} and U_Y^{tr} according to (18), and the error terms $\boldsymbol{U}_X^{tr}$ according to (19).
4. Simulate 3 separate training sets, the confounded one (where we apply the causality-aware adjustment), and the baseline 1 and baseline 2 training sets (using the exact same error terms sampled in the previous step). The confounded training set was generated according to the following model,

$$C_{tr} = U_C^{tr} \,, \tag{26}$$
$$Y_{tr} = \beta_{YC} C_{tr} + U_Y^{tr} \,, \tag{27}$$
$$X_{j,tr} = \beta_{X_jY} Y_{tr} + \beta_{X_jC} C_{tr} + U_{X_j}^{tr} \,. \tag{28}$$

The baseline 1 training data was generated according to the model,

$$C_{tr} = U_C^{tr} \,, \tag{29}$$
$$Y_{tr} = \beta_{YC} C_{tr} + U_Y^{tr} \,, \tag{30}$$
$$X_{j,tr} = \beta_{X_jY} Y_{tr} + U_{X_j}^{tr} \,, \tag{31}$$

while the baseline 2 training data was generated according to the model,

$$C_{tr} = U_C^{tr} \,, \tag{32}$$
$$Y_{tr} = U_Y^{tr} \,, \tag{33}$$
$$X_{j,tr} = \beta_{X_jY} Y_{tr} + \beta_{X_jC} C_{tr} + U_{X_j}^{tr} \,. \tag{34}$$

5. Simulate 9 distinct confounded test sets (indexed by ts_k, for $k = 1, \ldots, 9$). Each test set is simulated as follows:
   (a) Given the fixed values for $Var(C_{ts_k})$, $Cov(Y_{ts_k}, C_{ts_k})$, and $Var(Y_{ts_k})$, and the sampled value for β_{YC}, we compute ϕ_{CC}, ϕ_{CY}, and ϕ_{YY} as described in equations (23), (24), and (25).
   (b) Sample the error terms $U_C^{ts_k}$ and $U_Y^{ts_k}$ according to (18), and the error terms $\boldsymbol{U}_X^{ts_k}$ according to (19).
   (c) Simulate the test set data according to the model,

$$C_{ts_k} = U_C^{ts_k} \,, \tag{35}$$

$$Y_{ts_k} = \beta_{YC} C_{ts_k} + U_Y^{ts_k} \,, \tag{36}$$

$$X_{j,ts_k} = \beta_{X_jY} Y_{ts_k} + \beta_{X_jC} C_{ts_k} + U_{X_j}^{ts_k} \,, \tag{37}$$

Note that, in order to generate dataset shifts in $P(C, Y)$, we allow $Var(C)$, $Cov(Y, C)$, and $Var(Y)$ to vary between the training and test sets. However, in order to maintain the stability of $P(\boldsymbol{X} \mid C, Y)$ we use the same sampled values of β_{X_jY}, β_{X_jC}, β_{YC} and ρ in the generation of the training and test sets.

In order to illustrate the influence of $Var(Y_{ts})$ in the stability of the predictions, we performed two experiments. In the first, we kept the $Var(Y_{ts})$ constant across the test sets. In the second, we increased $Var(Y_{ts})$ across the test sets. Each of these experiments were based on 1000 simulations. For each simulation replication we:

1. Generate the 3 training sets ($n = 1,000$) by setting $Var(C_{tr}) = 1$, $Cov(Y_{tr}, C_{tr}) = 0.8$, and $Var(Y_{tr}) = 1$ and then simulating the data as described above.
2. Generate 9 distinct test sets (each containing $n = 1,000$ samples). Each test set was generated with an increasing amount of shift in the $P(C, Y)$ distribution. In the first experiment this was accomplished by varying $Cov(Y_{ts_k}, C_{ts_k})$ according to $\{0.8$, 0.6, 0.4, 0.2, 0.0, -0.2, -0.4, -0.6, $-0.8\}$ across the 9 test sets, and by varying $Var(C_{ts_k})$ according to $\{1.00$, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, $3.00\}$, while keeping $Var(Y_{ts_k})$ fixed at 1 for all k. In the second experiment, we varied $Cov(Y_{ts_k}, C_{ts_k})$ as before, but kept $Var(C_{ts_k})$ fixed at 1, while varying $Var(Y_{ts_k})$ according to $\{1.00$, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, $3.00\}$ across the test sets.
3. For the causality-aware approach we: adjust the confounded training set, and each of the 9 confounded test sets; fit a regression model on the adjusted training set data; use the same trained model to predict on the 9 adjusted test sets; and evaluate the test set performances using MSE.
4. For the baseline 1 approach we: fit a regression model to the (unconfounded) baseline 1 training set; use the trained model to predict on the 9 confounded test sets; and evaluate the test set performances using MSE.
5. For the baseline 2 approach we: fit a regression model to the (unconfounded) baseline 2 training set; use the trained model to predict on the 9 confounded test sets; and evaluate the test set performances using MSE.
6. For the "no adjustment" approach we: fit a regression model to the confounded training data; use the trained model to predict on the 9 confounded test sets; and evaluate the test set performances using MSE.

Note that the first test set is generated using the same values of $Cov(Y, C)$, $Var(C)$, and $Var(Y)$ as the training set, so that it illustrates the case where the training and test sets are independent and identically distributed. (Observe that in this setting, performing confounding adjustment may decrease the predictive performance of the learner in situations where the confounder strengths the association between the features and the outcome variable.)

Next, we present a few important remarks.

1. Note that the ("archetypical") baseline 1 approach is meant to represent methods that attempt to remove the causal effects of the confounders from the features in the training set alone.

This includes a poor man's version of the causality-aware approach where we do not process the test set features. Our goal here is to illustrate that while it might seen intuitive that training a learner on unconfounded data will prevent it from learning the confounding signal and, therefore, will lead to more stable predictions in shifted target populations, the unconfounded trained model, $\hat{\boldsymbol{\beta}}^{tr}$ is only one component of the prediction, $\hat{Y} = \boldsymbol{X}_{ts}\hat{\boldsymbol{\beta}}^{tr}$, so that better stability can be achieved by deconfounding the test set features, $\boldsymbol{X}_{ts}$, as well.

2. Second, note that the baseline 2 approach is meant to represent methods that attempt to remove the association between the confounders and the outcome. Those include approaches such as propensity scores for continuous variables [25], covariate balancing propensity score methods for continuous variables [16], or standard propensity score matching applied to dichotomized outcome data[14]. As described before, rather than implementing these methods, we simulate unconfounded training data where the output is statistically independent from the confounders, which mimics the case where these adjustments worked perfectly. (Observe that, in the particular context of classification tasks, removing the association between labels and confounders represents a common strategy to combat discrimination in fairness research, where data pre-processing techniques such as re-weighting and (under-) over-sampling are applied to the training data alone, in order remove the association between sensitive variables (i.e., confounders) and the classifier labels [8, 28])

3. Third, it is important to point out that several approaches proposed in the stable prediction literature are not applicable in our illustrations. For instance, in the context of classification tasks, approaches such as invariant risk minimization [3] or invariant causal prediction [43] rely on training data from multiple training sets while our approach focus on a single training set. Furthermore, stable prediction approaches [31, 32], which only require a single training set, can only be applied in causal prediction tasks, while our illustrations focus on anticausal tasks.

4. Fourth, observe that our approach assumes that $P(\boldsymbol{X} \mid \boldsymbol{C}, Y)$ is stable across the test set domains. This assumption is reasonable in several application domains. For instance, in health diagnostic applications, where the goal is to classify (for example) mild vs severe cases of a given disease, using the disease symptoms as inputs, we have that $P(\boldsymbol{X} \mid \boldsymbol{C}, Y)$ tends to be stable for demographic confounders such as age and gender. Note that this distribution would be unstable in the less likely scenario where the individuals in the training set have different symptom severities (caused by age, gender and disease status) than individuals in distinct test sets, pointing to biological/physiological differences between the individuals in training and testing populations. Dataset shifts on $P(\boldsymbol{C}, Y)$, on the other hand, are much more commonly observed in health applications, because selection biases during data collection often mean that the $P(\boldsymbol{C}, Y)$ distribution in the target/test populations are shifted relative to the training data.

5. Finally, note that application of counterfactual normalization [52] approach to the particular causal graph used in our experiments would augment the causal graph with the counterfactual variables $X_j(C = \emptyset)$ (representing the values of X_j we would have seen, had C not being a parent of X_j), and would return the counterfactual variables $X_j(C = \emptyset)$ as the stable set for predicting Y. Because (for this particular example) the counterfactual features, $X_j(C = \emptyset)$, are computed in exactly the same way as the causality-aware training features, $X_j^* = X_j - \hat{\beta}_{X_jC}\, C$, it follows that application of counterfactual normalization would produce the same results as the causality-aware approach.

11 The causal prediction task case

In this paper, we have focused in anticausal prediction tasks. A few analogous results are, nonetheless, available for causal prediction tasks (i.e., prediction tasks where the inputs influence the outcome). In the next subsections, we present these results.

[14] For classification tasks these methods include standard matching and IPW by propensity score methods.

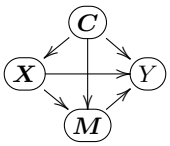


Figure S9: Causal prediction task.

11.1 Reparameterization in causal prediction tasks

For the causal prediction task presented in Figure S9 we have that the joint distribution factorizes as,

$$P(\boldsymbol{C})P(\boldsymbol{X} \mid \boldsymbol{C})P(\boldsymbol{M} \mid \boldsymbol{C}, \boldsymbol{X})P(Y \mid \boldsymbol{C}, \boldsymbol{M}, \boldsymbol{X}) \ ,$$

where each component is described by the structural model,

$$\begin{aligned}
\boldsymbol{C} &= \boldsymbol{\Theta}_{CC}\, \boldsymbol{C} + \boldsymbol{U}_C \ , \\
\boldsymbol{X} &= \boldsymbol{\Theta}_{XX}\, \boldsymbol{X} + \boldsymbol{\Theta}_{XC}\, \boldsymbol{C} + \boldsymbol{U}_X \ , \\
\boldsymbol{M} &= \boldsymbol{\Theta}_{MM}\, \boldsymbol{M} + \boldsymbol{\Theta}_{MC}\, \boldsymbol{C} + \boldsymbol{\Theta}_{MX}\, \boldsymbol{X} + \boldsymbol{U}_M \ , \\
Y &= \boldsymbol{\Theta}_{YC}\, \boldsymbol{C} + \boldsymbol{\Theta}_{YM}\, \boldsymbol{M} + \boldsymbol{\Theta}_{YX}\, \boldsymbol{X} + U_Y \ ,
\end{aligned}$$

which can also be reparameterized as,

$$\begin{aligned}
\boldsymbol{C} &= \boldsymbol{W}_C \ , \\
\boldsymbol{X} &= \boldsymbol{\Gamma}_{XC}\, \boldsymbol{C} + \boldsymbol{W}_X \ , \\
\boldsymbol{M} &= \boldsymbol{\Gamma}_{MC}\, \boldsymbol{C} + \boldsymbol{\Gamma}_{MX}\, \boldsymbol{X} + \boldsymbol{W}_M \ , \\
Y &= \boldsymbol{\Gamma}_{YC}\, \boldsymbol{C} + \boldsymbol{\Gamma}_{YM}\, \boldsymbol{M} + \boldsymbol{\Gamma}_{YX}\, \boldsymbol{X} + W_Y \ ,
\end{aligned}$$

where $\boldsymbol{\Gamma}_{MX} = (\boldsymbol{I} - \boldsymbol{\Theta}_{MM})^{-1}\boldsymbol{\Theta}_{MX}$, $\boldsymbol{\Gamma}_{YC} = \boldsymbol{\Theta}_{YC}$, $\boldsymbol{\Gamma}_{YM} = \boldsymbol{\Theta}_{YM}$, $\boldsymbol{\Gamma}_{YX} = \boldsymbol{\Theta}_{YX}$, $W_Y = U_Y$, and the other parameters and error terms are given as before.

12 Estimation of the causal effects in the causal task, and remarks on identification issues

For the causal prediction task, we regress the response on the confounders, mediators, and features,

$$Y = \sum_{k=1}^{n_C} \gamma_{Y C_k}\, C_k + \sum_{k=1}^{n_M} \gamma_{Y M_k}\, M_k + \sum_{k=1}^{n_X} \gamma_{Y X_k}\, X_k + W_Y \ ,$$

and then generate the counterfactual response by adding back $\hat{W}_Y$ to a linear predictor containing only the causal effects of interest. In particular, we can generate counterfactual response data that captures the predictive performance due to direct causal effects, indirect causal effects, or to confounding, using, respectively,

$$\begin{aligned}
\hat{Y}^* &= \hat{\boldsymbol{\Gamma}}_{YX} \boldsymbol{X} + \hat{W}_Y \ , \\
\hat{Y}^* &= \hat{\boldsymbol{\Gamma}}_{YM}\, \hat{\boldsymbol{M}}^* + \hat{W}_Y \ , \\
\hat{Y}^* &= \hat{\boldsymbol{\Gamma}}_{YC}\, \boldsymbol{C} + \hat{W}_Y \ .
\end{aligned}$$

Under the assumption that all the confounders and mediators are observed, we can identify the direct and indirect causal effects of the features on the response. In particular, a simple least squares estimation procedure provides consistent estimates of these causal effects[15]. To see why, note that for the reparameterized model, if all confounders and mediators are observed, it follows from the Markov property of DAGs that $Y = f_Y(\boldsymbol{C}, \boldsymbol{M}, \boldsymbol{X}, W_Y) = f_Y(pa(Y), W_Y)$. (Here, f_Y represent a linear structural causal model). Hence, it follows that, when we regress Y on the elements of $\boldsymbol{C}$, $\boldsymbol{M}$, and $\boldsymbol{X}$ only the coefficients associated with the parents of Y will be statistically different from

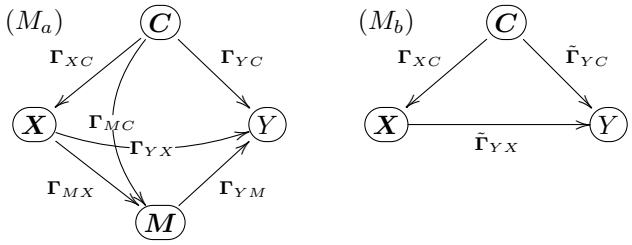
[15]Here, we assume that the number of samples is larger than the number of covariates in the regression fits, and that multicolinearity is not an issue too.

zero (for large enough sample sizes). Therefore, in practice, we don't need to know before hand which variables are the parents of Y. The parent set will be learned automatically from the data by the regression model fit.

Observe, as well, that even if the mediators are unobserved, but the confounders are still observed, we can still identify total causal effects. For instance, in causal tasks we have that,

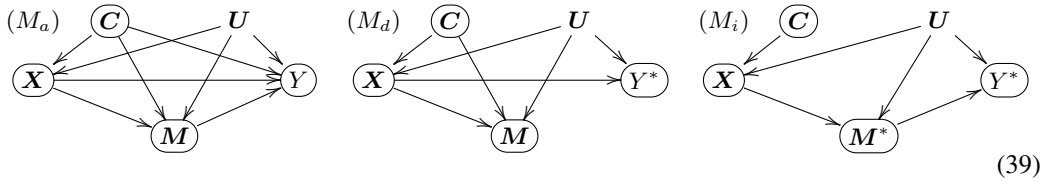
$$\begin{aligned} Y &= \boldsymbol{\Gamma}_{YC}\,\boldsymbol{C} + \boldsymbol{\Gamma}_{YM}\,\boldsymbol{M} + \boldsymbol{\Gamma}_{YX}\,\boldsymbol{X} + W_Y \;, \\ &= \boldsymbol{\Gamma}_{YC}\,\boldsymbol{C} + \boldsymbol{\Gamma}_{YM}\,(\boldsymbol{\Gamma}_{MC}\,\boldsymbol{C} + \boldsymbol{\Gamma}_{MX}\,\boldsymbol{X} + \boldsymbol{W}_M) + \boldsymbol{\Gamma}_{YX}\,\boldsymbol{X} + W_Y \;, \\ &= \underbrace{(\boldsymbol{\Gamma}_{YC} + \boldsymbol{\Gamma}_{YM}\,\boldsymbol{\Gamma}_{MC})}_{\tilde{\boldsymbol{\Gamma}}_{YC}}\,\boldsymbol{C} + \underbrace{(\boldsymbol{\Gamma}_{YX} + \boldsymbol{\Gamma}_{YM}\,\boldsymbol{\Gamma}_{MX})}_{\tilde{\boldsymbol{\Gamma}}_{YX}}\,\boldsymbol{X} + \underbrace{\boldsymbol{\Gamma}_{YM}\,\boldsymbol{W}_M + W_Y}_{\tilde{W}_Y} \;, \\ &= \tilde{\boldsymbol{\Gamma}}_{YC}\,\boldsymbol{C} + \tilde{\boldsymbol{\Gamma}}_{YX}\,\boldsymbol{X} + \tilde{W}_Y \;, \end{aligned}$$

where $\tilde{\boldsymbol{\Gamma}}_{YX} = \boldsymbol{\Gamma}_{YX} + \boldsymbol{\Gamma}_{YM}\,\boldsymbol{\Gamma}_{MX}$ represents the total causal effect of $\boldsymbol{X}$ on Y, as represented in the DAG M_b in in the causal task model (38).



(38)

On the other hand, if the mediators are observed, but some the confounders are unobserved, then neither the direct, the indirect, or the total causal effects are identifiable, and the predictions generated by the causality-aware approach will still be confounded. For instance, for the causal prediction tasks in model (39), we have that the unobserved confounders, $\boldsymbol{U}$, still confound the direct causal effect of $\boldsymbol{X}$ on Y^* in model M_d, and the indirect causal effect in model M_i. As a consequence, the spurious associations contributed by $\boldsymbol{U}$ will still bias the predictions from models trained with the counterfactual data.



(39)

Finally, observe that while so far we have discussed confounding of the feature/response relationship, it is also possible that the causal relations between features and mediators or between mediators and response are also influenced by confounders. If these confounders are unobserved, then we cannot identify the causal effects $\boldsymbol{\Gamma}_{MX}$ and $\boldsymbol{\Gamma}_{YM}$. Clearly, in the presence of unobserved confounding the causality-aware predictions will be biased, whenever the causal effects of interest are not identifiable.

12.1 Causality-aware predictions in causal prediction tasks - the univariate case

Consider a causal prediction task where the goal is to build a ML model whose predictive performance is only informed by the direct causal effect of X on Y. We can simulate counterfactual response data, Y^*, according to the twin network in Figure S10a so that,

$$Cov(X, Y^*) = Cov(X,\, \theta_{YX} X + U_Y) = \theta_{YX}\, Var(X) = \theta_{YX} \;, \tag{40}$$

Now, consider a causal prediction task where the goal is to build a ML model whose predictive performance is only informed by the indirect causal effect of X on Y. Now, we can simulate counterfactual response data, Y^*, according to the twin network in Figure S10b so that,

$$\begin{aligned} Cov(X, Y^*) &= Cov(X,\, \theta_{YM} M^* + U_Y) = \theta_{YM} Cov(X, M^*) \\ &= \theta_{YM} Cov(X, \theta_{MX} X + U_M) = \theta_{YM}\theta_{MX} Var(X) = \theta_{YM}\theta_{MX} \;, \end{aligned} \tag{41}$$

Finally, suppose that the goal is to build a ML model whose predictive performance is only informed by the spurious associations generated by the confounder. We can simulate data according to the twin network in Figure S10, so that,

$$\begin{aligned} Cov(X, Y^*) &= Cov(X, \theta_{YC}C + U_Y) = \theta_{YC} Cov(X, C) \\ &= \theta_{YC} Cov(\theta_{XC}C + U_X, C) = \theta_{YC}\theta_{XC} Var(C) = \theta_{YC}\theta_{XC} \, . \end{aligned} \tag{42}$$

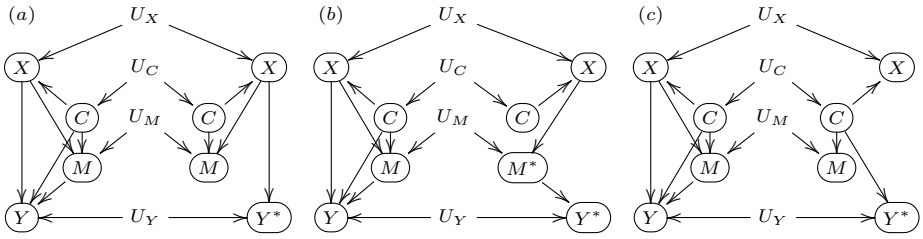


Figure S10: Twin network approach for the causal prediction tasks.

Similarly to the anticausal prediction task case, alternative interventions based on SWIGs can also be used. Figure S11 shows the respective SWIGs for the generation of counterfactual responses.

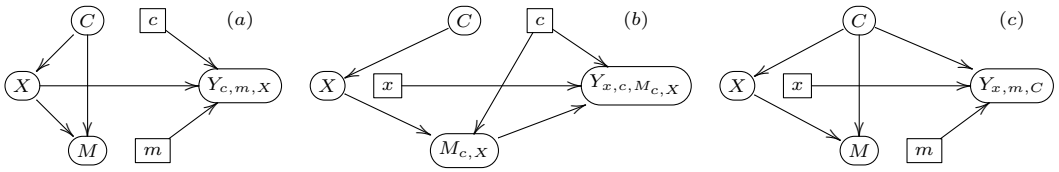


Figure S11: SWIGs for the causal predictive tasks.

Direct calculation of the covariances shows that, $Cov(X, Y_{c,m,X}) = \theta_{YX}$ for the SWIG in panel a, $Cov(X, Y_{x,c,M_{c,X}}) = \theta_{YM}\,\theta_{MX}$ for the SWIG in panel b, and $Cov(X, Y_{x,m,C}) = \theta_{XC}\,\theta_{YC}$ for the SWIG in panel c.

Observe, that alternative interventions where we intervene on the features will not recover the correct associations. To illustrate this point, consider the simplified situation where we are interested in the direct causal effect, θ_{YX}, in a model containing a confounder but no mediator. For the interventions presented in Figure S12a we have that,

$$\begin{aligned} Cov(X^*, Y^*) &= Cov(X^*,\, \theta_{YX}X^* + \theta_{YC}C + U_Y) \\ &= \theta_{YX}\, Var(X^*) + \theta_{YC} Cov(X^*, C) \\ &= \theta_{YX}\, Var(U_X) + \theta_{YC} Cov(U_X, C) \\ &= \theta_{YX}\, Var(U_X) \\ &= \theta_{YX}\,(1 - \theta_{XC}^2) \, , \end{aligned}$$

where the last equality follows from the fact that $Var(U_X) = (1 - \theta_{XC}^2)$ since $1 = Var(X) = Var(\theta_{XC}C + U_X) = \theta_{XC}^2 Var(C) + Var(U_X) = \theta_{XC}^2 + Var(U_X)$. Similarly, even for the intervention in Figure S12b we still have that,

$$\begin{aligned} Cov(X^*, Y^*) &= Cov(X^*,\, \theta_{YX}X^* + U_Y) \\ &= \theta_{YX}\, Var(X^*) = \theta_{YX}\, Var(U_X) = \theta_{YX}\,(1 - \theta_{XC}^2) \, . \end{aligned}$$

These examples illustrate that for causal prediction tasks, only interventions that do not modify X can generate associations that recover the causal effects of interest.

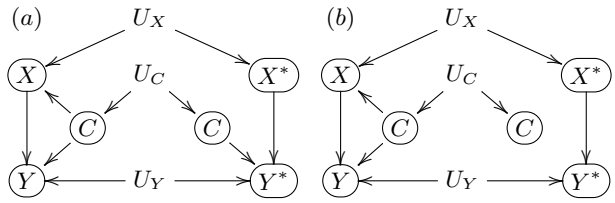


Figure S12: Alternative model modifications for the confounding only examples.

Remarks: The fact that the causality-aware approach requires the computation of counterfactual responses, Y^*, implies that, contrary to anticausal prediction tasks (which requires the computation of counterfactual features, X^*, and where it is possible to estimate counterfactual features for both the training and test sets without having access to the test set responses), causal prediction tasks require access to the test set responses, Y_{ts}, in order to estimate the causal effects and residuals needed for the computation of the counterfactual test set responses, Y^*_{ts}. Since, in practice, Y_{ts} is unavailable (as it is the quantity we want to predict) it follows that the approach cannot be used to generate, for example, stable predictions w.r.t. unknown shifts in target populations, as was done in the anticausal tasks. In causal prediction tasks, and under the assumption of no dataset shifts between the training and target populations, the causality-aware approach can still be used to estimate the predictive performance that is due to (or is free from) the influence of sensitive variables. For instance, we still can split our development data into independent and identically distributed training and validation sets and then compute counterfactual versions of the training and validation responses, in order to generate causality-aware predictions that can still be used to answer important questions such as, for example: "what would the predictive performance of the learner be, had the (in)direct path not contributed to the association between the features and the response?" or "what would the predictive performance of the learner be, had the observed confounders not biased the data?"

12.2 Causality-aware predictions in causal prediction tasks - the multivariate case

Theorem 2. *Consider a causal prediction task:*

(i) Suppose the interest focus on the causal effects generated by the paths in the path set $\boldsymbol{X} \rightarrow Y$. If Y^* is given by $Y^* = \boldsymbol{\Gamma}_{YX}\, \boldsymbol{X} + W_Y$, then $Cov(Y^*, \boldsymbol{X}) = \boldsymbol{\Gamma}_{YX}\, Cov(\boldsymbol{X})$.

(ii) Suppose the interest focus on the causal effects generated by the paths in the path set $\boldsymbol{X} \rightarrow \boldsymbol{M} \rightarrow Y$. If Y^* is given by $Y^* = \boldsymbol{\Gamma}_{YM}\, \boldsymbol{M}^* + W_Y$, and $\boldsymbol{M}^* = \boldsymbol{\Gamma}_{MX} \boldsymbol{X} + \boldsymbol{W}_M$, then $Cov(Y^*, \boldsymbol{X}) = \boldsymbol{\Gamma}_{YM}\, \boldsymbol{\Gamma}_{MX}\, Cov(\boldsymbol{X})$.

(iii) Suppose the interest focus on the spurious associations generated by the paths in the path set $\boldsymbol{X} \leftarrow \boldsymbol{C} \rightarrow Y$. If Y^* is given by $Y^* = \boldsymbol{\Gamma}_{YC}\, \boldsymbol{C} + W_Y$, then $Cov(Y^*, \boldsymbol{X}) = \boldsymbol{\Gamma}_{YC}\, Cov(\boldsymbol{C})\, \boldsymbol{\Gamma}_{XC}^T$.

Proof.

Result *i*: If $Y^* = \boldsymbol{\Gamma}_{YX}\, \boldsymbol{X} + W_Y$, then,

$$\begin{aligned} Cov(Y^*, \boldsymbol{X}) &= Cov(\boldsymbol{\Gamma}_{YX}\, \boldsymbol{X} + W_Y, \boldsymbol{X}) \\ &= \boldsymbol{\Gamma}_{YX}\, Cov(\boldsymbol{X}, \boldsymbol{X}) \\ &= \boldsymbol{\Gamma}_{YX}\, Cov(\boldsymbol{X}) \end{aligned}$$

Result *ii*: If $Y^* = \boldsymbol{\Gamma}_{YM}\, \boldsymbol{M}^* + W_Y$ and $\boldsymbol{M}^* = \boldsymbol{\Gamma}_{MX} \boldsymbol{X} + \boldsymbol{W}_M$, then,

$$\begin{aligned} Cov(Y^*, \boldsymbol{X}) &= Cov(\boldsymbol{\Gamma}_{YM}\, \boldsymbol{M}^* + W_Y, \boldsymbol{X}) \\ &= \boldsymbol{\Gamma}_{YM}\, Cov(\boldsymbol{X}, \boldsymbol{M}^*) \\ &= \boldsymbol{\Gamma}_{YM}\, Cov(\boldsymbol{\Gamma}_{MX} \boldsymbol{X} + \boldsymbol{W}_M, \boldsymbol{X}) \\ &= \boldsymbol{\Gamma}_{YM}\, \boldsymbol{\Gamma}_{MX}\, Cov(\boldsymbol{X}, \boldsymbol{X}) \\ &= \boldsymbol{\Gamma}_{YM}\, \boldsymbol{\Gamma}_{MX}\, Cov(\boldsymbol{X}) \end{aligned}$$

Result *iii*: If $Y^* = \boldsymbol{\Gamma}_{YC}\,\boldsymbol{C} + W_Y$, then,

$$\begin{aligned} Cov(Y^*, \boldsymbol{X}) &= Cov(\boldsymbol{\Gamma}_{YC}\,\boldsymbol{C} + W_Y, \boldsymbol{X}) \\ &= \boldsymbol{\Gamma}_{YC}\,Cov(\boldsymbol{C}, \boldsymbol{X}) \\ &= \boldsymbol{\Gamma}_{YC}\,Cov(\boldsymbol{C}, \boldsymbol{\Gamma}_{XC}\,\boldsymbol{C} + \boldsymbol{W}_X) \\ &= \boldsymbol{\Gamma}_{YC}\,Cov(\boldsymbol{C}, \boldsymbol{C})\,\boldsymbol{\Gamma}_{XC}^T \\ &= \boldsymbol{\Gamma}_{YC}\,Cov(\boldsymbol{C})\,\boldsymbol{\Gamma}_{XC}^T \end{aligned}$$

□

Note that, in the univariate case, results (i), (ii), and (iii) in Theorem 2 reduce to the univariate results presented in equations (40), (41), and (42), respectively (note that $Cov(\boldsymbol{X})$ reduces to 1). Observe, as well, that results (*i*) and (*ii*) in Theorem 2 show that, in addition to the direct causal effect ($\boldsymbol{\Gamma}_{YX}$, in result *i*) and the indirect causal effect ($\boldsymbol{\Gamma}_{YM}\,\boldsymbol{\Gamma}_{MX}$, in result *ii*) the marginal covariances between the elements of $\boldsymbol{X}$ and Y^* also depend on $Cov(\boldsymbol{X})$. This makes sense, since $Cov(\boldsymbol{X})$ captures the associations between the elements of $\boldsymbol{X}$. Note that for each element X_j of $\boldsymbol{X}$, the operation $\boldsymbol{\Gamma}_{YX}\,Cov(\boldsymbol{X})$ captures not only the association generated by the direct causal path $X_j \to Y^*$, but also the association generated by indirect and backdoor paths that start at X_j and end at Y^*, but where the last node prior to Y^* is another element X_k of $\boldsymbol{X}$.

As an illustration, consider the DAG describing the causal prediction task in Figure S13a, where $Cov(\boldsymbol{X})$,

$$\begin{pmatrix} 1 & \theta_{X_2X_1} + \theta_{X_1C_1}\,\theta_{X_2C_1} \\ \theta_{X_2X_1} + \theta_{X_1C_1}\,\theta_{X_2C_1} & 1 \end{pmatrix}.$$

In this example, the association between X_1 and X_2,

$$Cov(X_1, X_2) = \underbrace{\theta_{X_2X_1}}_{X_1 \to X_2} + \underbrace{\theta_{X_1C_1}\,\theta_{X_2C}}_{X_1 \leftarrow C_1 \to X_2},$$

is generated by the paths $X_1 \to X_2$ and $X_1 \leftarrow C_1 \to X_2$. From result *i* in Theorem 2 we have that,

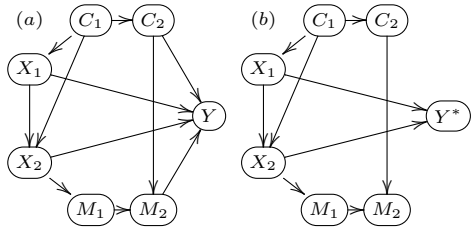


Figure S13: A causal prediction task illustrative example.

$$\begin{aligned} Cov(Y^*, \boldsymbol{X}) &= \boldsymbol{\Gamma}_{YX}\,Cov(\boldsymbol{X}) = (\theta_{YX_1}, \theta_{YX_2})\,Cov(\boldsymbol{X}) \\ &= \begin{pmatrix} \theta_{YX_1} + \theta_{X_2X_1}\,\theta_{YX_2} + \theta_{X_1C_1}\,\theta_{X_2C_1}\,\theta_{YX_2} \\ \theta_{YX_2} + \theta_{X_2X_1}\,\theta_{YX_1} + \theta_{X_2C_1}\,\theta_{X_1C_1}\,\theta_{YX_1} \end{pmatrix}^T, \\ &= \begin{pmatrix} Cov(Y^*, X_1) \\ Cov(Y^*, X_2) \end{pmatrix}^T. \end{aligned}$$

Note that the direct application of Wright's path analysis to the diagram in Figure S13b shows that we can decompose the covariance of X_1 and Y^*,

$$Cov(Y^*, X_1) = \underbrace{\theta_{YX_1}}_{X_1 \to Y^*} + \underbrace{\theta_{X_2X_1}\,\theta_{YX_2}}_{X_1 \to X_2 \to Y^*} + \underbrace{\theta_{X_1C_1}\,\theta_{X_2C_1}\,\theta_{YX_2}}_{X_1 \leftarrow C_1 \to X_2 \to Y^*},$$

in terms of the direct path $X_1 \to Y^*$, the indirect path $X_1 \to X_2 \to Y^*$, and the backdoor path $X_1 \leftarrow C_1 \to X_2 \to Y^*$. Similarly, the covariance of X_2 and Y^*,

$$Cov(Y^*, X_2) = \underbrace{\theta_{YX_2}}_{X_2 \to Y^*} + \underbrace{\theta_{X_2X_1}\,\theta_{YX_1}}_{X_2 \leftarrow X_1 \to Y^*} + \underbrace{\theta_{X_2C_1}\,\theta_{X_1C_1}\,\theta_{YX_1}}_{X_2 \leftarrow C_1 \to X_1 \to Y^*},$$

can be decomposed in terms of the direct path $X_2 \to Y^*$, and the backdoor paths $X_2 \leftarrow X_1 \to Y^*$ and $X_2 \leftarrow C_1 \to X_1 \to Y^*$. (Note that all the indirect and backdoor paths in this example either start at X_1 and end at X_2 before connecting to Y^*, or start at X_2 and end at X_1 before connecting to Y^*.)