Causality-aware counterfactual confounding adjustment as an alternative to linear residualization in anticausal prediction tasks based on linear learners

Elias Chaibub Neto [1]

Abstract

Linear residualization is a common practice for confounding adjustment in machine learning applications. Recently, causality-aware predictive modeling has been proposed as an alternative causality-inspired approach for adjusting for confounders. In this paper, we compare the linear residualization approach against the causality-aware confounding adjustment in anticausal prediction tasks. Our comparisons include both the settings where the training and test sets come from the same distributions, as well as, when the training and test sets are shifted due to selection biases. In the absence of dataset shifts, we show that the causality-aware approach tends to (asymptotically) outperform the residualization adjustment in terms of predictive performance in linear learners. Importantly, our results still holds even when the true model generating the data is not linear. We illustrate our results in both regression and classification tasks. Furthermore, in the presence of dataset shifts in the joint distribution of the confounders and outcome variables, we show that the causality-aware approach is more stable than linear residualization.

1. Introduction

Linear residualization is a common technique for confounding adjustment in applied machine learning (ML) work. The basic idea is to regress the input data on the observed confounders and use the residuals of the regression fits as the new inputs for ML algorithms. The technique is widely used in applied fields such as social sciences, bioinformatics/genomics, neuroimaging, and others. For instance, as pointed by Snoek et al. (2019), linear residualization is perhaps the most common confounding adjustment approach in neuroimage studies[1].

In this paper, we focus on anticausal prediction tasks (Schölkopf et al., 2012), namely, ML tasks where the output variable has a causal influence on the inputs, and we provide a "head-to-head" comparison between the ad hoc linear residualization technique against the recently proposed causality-aware adjustment (Chaibub Neto, 2020) - which is implemented by regressing each input on both the confounders and output, and then generating counterfactual inputs by adding back the estimated residuals to a linear predictor that no longer includes the confounder variables. The new counterfactual inputs are then used as the inputs for the ML algorithm.

We focus on anticausal prediction tasks because, from a causality perspective, the residualization procedure performs the wrong adjustment in this setting. By failing to include the outcome variable as a covariate in the regression models, the residualization approach removes not only the direct causal influence of the confounders from the inputs, but also the indirect influences that are mediated by the outcome variable. In a sense, it goes too far and removes too much. For example, consider the toy model in Figure 1a, where X, Y, and A represent, respectively, an input, the outcome, and a confounder variable, and γ_{XA}, γ_{YA}, and γ_{XY}, represent the causal effects among these variables. We see that by regressing X on A alone the residualized input $X_r = X - \omega A$ is obtained by removing the total effect $\omega = \gamma_{XA} + \gamma_{XY}\,\gamma_{YA}$ of A from X, whereas by regressing X on both A and Y the causality-aware input, $X_c = \gamma_{XY}\,Y + W_X = X - \gamma_{XA} A$ is obtained by remov-

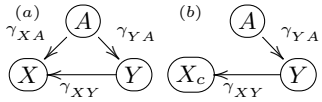

Figure 1. Anticausal prediction task.

[1]Sage Bionetworks, Seattle, United States. Correspondence to: Elias Chaibub Neto <elias.chaibub.neto@sagebase.org>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

[1]Examples of works that either apply, discuss, or evaluate linear residualization in neuroimage studies include Abdulkadir et al (2014), Dubois et al (2017) Dukart et al (2011), Kostro et al (2014), Rao et al (2017), Todd et al (2013), Greenstein et al (2012), Doan et al (2017), Friston et al (1994), and Maglanoc et al (2020). Also, note that the term "linear residualization" is similarly denoted as "confounding regression", "image correction", or as "regressing out" confounding effects in this applied literature.

ing only the direct causal effect of A on X, as described in Figure 1b[2]. (Note that we cannot even represent the residualization input, X_r, in a causal graph, because the conditional independence relations between X_r, A, and Y are not faithful (Spirtes, Glymour, and Scheines, 2000) to any causal graph representing an anticausal prediction task. See Supplementary Section 1 for further details.)

In causal prediction tasks, on the other hand, residualization performs the correct adjustment (from a causal perspective) and removes only the direct causal effect of A on X, as described in Figure 2. (Note that, now, we can represent the residualization input using the causal graph in Figure 2b.)

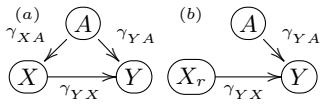

Figure 2. Causal prediction task.

In this paper, we show how a more principled, causality-inspired approach, can better handle confounding in anticausal prediction task and generate substantial gains in both predictive performance and stability, when compared to a widely used ad hoc statistical adjustment approach. While the subfield of causality-inspired machine learning has raised awareness regarding the potential advantages of using causality to inform learning approaches, there has been few clear demonstrations of the concrete advantages of leveraging the causal structure of the learning task. Our work represents an important step in this direction[3].

Moreover, the work has also important practical consequences. Anticausal tasks are frequently found in real world applications and notable examples include health related diagnostic applications, where the goal is to classify disease status using disease symptoms as the inputs of the ML model. These applications are clearly anticausal since the disease is a cause of the symptoms, and not the other way around. We further illustrate this point with a real data illustration comparing the causality-aware and residualization approaches using data from a Parkinson's disease (PD) mobile health application (where we build classifies of PD vs non-PD subjects using inputs extracted from accelerometer data collected by smartphones). Other examples of anticausal prediction tasks include computer vision applications where the goal is to classify different objects using their images. (These tasks are also anticausal since the real world objects cause the pixel patterns/intensities in their images, and not the other way around).

In this paper, we compare the causality-aware and residualization approaches in settings where the training and test sets come from the same distributions, as well as, in situations where the training and test sets are different due to dataset shifts (Quiñonero-Candela et al., 2009) generated by selection biases (Heckman, 1979; Hernán et al., 1979, Bareinboim & Pearl, 2012). In real word applications, selection biases often lead to the collection of non-representative training sets and represent an important challenge for ML. Furthermore, because in many applications the target populations where the ML model will be deployed can be shifted in unknown ways, researchers often prefer to perform confounding adjustment in the development data available to train and evaluate their ML models, even when the confounding adjustment decreases the predictive performance in the development data[4].

In situations where the training and test sets come from the same distribution (e.g., in development datasets where the researcher splits the data into independent and identically distributed (i.i.d.) training and test sets), we prove that the asymptotic expected mean squared error (MSE) of regression models trained with residualized inputs is always greater than or equal to the expected MSE of regression models trained with causality-aware inputs, even when the regression models are mispecified. We illustrate this result in synthetic data experiments based on both correct and mispecified models.

While we do not provide an analogous proof for classification tasks, we, nonetheless, prove that the strength of the covariance between the causality-aware counterfactual features and the output variable is always asymptotically stronger than the covariance between the residualized features and the output (again even when the true data generating process is not linear). Intuitively, this result suggests that, for linear classifiers such as logistic regression, the causality-aware approach will tend to (asymptotically) outperform the linear residualization technique. We provide empirical support for this conjecture using simulation experiments based again on correct and mispecified models, as well as, on real data illustrations.

In situations where the training and test sets are shifted due to selection biases, we prove that while the causality-aware approach is stable with respect to (w.r.t.) shifts in the association between the outcome and the confounders, the expected MSE of learners trained with residualized inputs is still a function of this association. As a consequence, the residualization approach is unstable w.r.t. dataset shifts generated by selection biases. We also illustrate this analytical result using synthetic data experiments.

[2]Here, we assume that sample size converges to infinity, so that the least squares estimates $\hat{\gamma}_{XA}$, $\hat{\gamma}_{YA}$, $\hat{\gamma}_{XY}$, and $\hat{\omega}$ converge to the true parameter values and γ_{XA}, γ_{YA}, γ_{XY}, and ω, and $\hat{X}_r = X - \hat{\omega}\, A$ and $\hat{X}_c = X - \hat{\gamma}_{XA}\, A$ converge to $X_r = X - \omega\, A$ and $X_c = X - \gamma_{XA}\, A$.

[3]We thank an anonymous reviewer for raising this point.

[4]This is also the reason why stable prediction approaches are attractive in safety critical applications where the trade-off between predictive performance and predictive stability tips in favor of stability (Subbaswamy et al., 2019b).

2. Related work

This work relies heavily on the causality-aware approach, recently proposed by Chaibub Neto (2020). The basic idea behind the causality-aware confounding adjustment is to generate counterfactual data which is free from the spurious associations generated by the observed confounders. The approach can be used to generate stable predictions and is similar in spirit to invariant prediction approaches (Peters et al., 2016; Ghassami et al., 2017; Heinze-Deml et al., 2018; Rojas-Carulla et al., 2018; Magliacane et al., 2018; Arjovsky et al., 2019) and other stable prediction approaches (Kuang et al., 2018; Subbaswamy et al., 2018; Subbaswamy et al., 2019a; Kuang et al., 2020) in the sense that it can 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 in the data while the causality-aware approach only requires a single training set, a characteristic that is also shared by the other stable prediction methods listed above. Observe, however, that some stable prediction approaches (Subbaswamy et al., 2018; Subbaswamy et al., 2019a) require, nonetheless, full knowledge about the data generation process, while the causality-aware approach only requires partial domain knowledge about which variables are confounders, without requiring knowledge about how the inputs are causally related (nor about how the confounders are related). Furthermore, the stable approaches proposed by Kuang et al. (2018; 2020) can only be applied to causal prediction tasks, while the causality-aware adjustment is suited to anticausal tasks. Finally, observe that while the causality-aware approach is based on linear models, other stable and invariant prediction approaches can handle non-linear data.

3. Background

3.1. Notation and causality definitions

Throughout the text, we let X, Y, and A represent, respectively, the input, output and confounder variables. Sets of random variables are represented in italic and boldface, and we use the superscripts tr and ts to represent the training and test sets, respectively. We adopt Pearl's mechanism-based approach to causation (Pearl, 2009) where the joint distribution of a set of variables is accompanied by a *directed acyclic graph* (DAG), also denoted as a *causal diagram/graph*, representing our prior knowledge (or assumptions) about the causal relation between the variables. The *nodes* on the causal graph represent the random variables, and the *directed edges* represent causal influences of one variable on another. In a DAG, a *path* corresponds to any unbroken, nonintersecting sequence of edges in the DAG, which may go along or against the direction of the arrows. A path is *d-separated* or *blocked* by a set of nodes $\boldsymbol{Z}$ if and only if: (i) the path contains a chain $V_j \rightarrow V_m \rightarrow V_k$ or a fork $V_j \leftarrow V_m \rightarrow V_k$ such that the middle node V_m is in $\boldsymbol{Z}$; or (ii) the path contains a collider $V_j \rightarrow V_m \leftarrow V_k$ such that V_m is not in $\boldsymbol{Z}$ and no descendant of V_m is in $\boldsymbol{Z}$. Otherwise, the path is *d-connected* or *open*. Following Pearl (2009), we adopt a causal definition of confounding where a variable A is a confounder of the relationship between variables X and Y, if there is an open path from A to X that does not go through Y, and an open path from A to Y that does not go through X. Throughout most of the text we assume that X, Y, and A have been standardized to have mean 0 and variance 1[5]. In the context of anticausal prediction tasks, the non-representativeness of the development data often arises due to selection mechanisms operating during the data collection phase. As illustrated in Figure 3a, confounding can be generated by selection mechanisms alone. Furthermore, even when the confounder has (stable) causal effects on X and on Y (Figure 3b), selection mechanisms can still contribute to the association between A and Y, making the data non-representative relative to target populations where this association is shifted.[6]

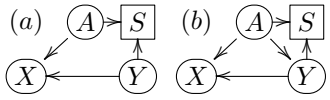

Figure 3. Confounding as a consequence of selection biases.

3.2. The confounded anticausal prediction task

Figure 4 describes the confounded anticausal prediction task. Note that the variables $\{W_{X_1}, \ldots, W_{X_p}\}$ and $\{W_{A_1}, \ldots, W_{A_k}\}$ represent sets of correlated error terms, and that the causal model in Figure 4 might represent a reparameterization of a

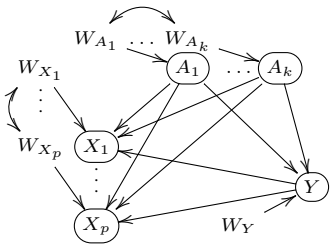

Figure 4. Confounded anticausal prediction task.

[5]Note that any linear model $V_s^o = \mu_s + \Sigma_{j \neq s} \beta_{sj} V_j^o + W_s^o$, where V_s^o represents the original data, can be reparameterized into its equivalent standardized form $V_s = \sum_{j \neq s} \gamma_{sj} V_j + W_s$, where $V_s = (V_s^o - E(V_s^o))/Var(V_s^o)^{\frac{1}{2}}$ represent standardized variables with $E(V_s) = 0$ and $Var(V_s) = 1$; $\gamma_{V_s V_j} = \beta_{V_s V_j}(Var(V_j^o)/Var(V_s^o))^{\frac{1}{2}}$ represent the path coefficients (Wright, 1934); and $W_s = W_s^o/Var(V_s^o)^{\frac{1}{2}}$ represent the standardized error terms.

[6]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 conditional on $S = 1$, we have that the path $A \rightarrow S \leftarrow Y$ is open, since S is a collider. This shows that A satisfies the definition of a confounder even in Figure 3a, where A is not a cause of Y.

model with uncorrelated error terms and unknown causal relations among the $\boldsymbol{X}$ inputs, as well as, among the $\boldsymbol{A}$ confounders.

This point has been described in detail in Chaibub Neto (2020), where it has been shown that, in the special case where the true data generation process corresponds to linear structural causal 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 4. However, it is important to clarify, that even when the true data generation process does not correspond to a set of linear structural equations, we can still model the data according to the diagram in Figure 4, with the understanding that we are working with a mispecified model. In either way, we model the input variables, X_j, $j = 1, \ldots, p$, according to the linear structural equations,

$$X_j = \sum_{i=1}^{k} \gamma_{X_j A_i} A_i + \gamma_{X_j Y} Y + W_{X_j} \ , \tag{1}$$

which can be represented in matrix form by,

$$\underbrace{\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}}_{\boldsymbol{X}} = \underbrace{\begin{pmatrix} \gamma_{X_1 A_1} & \cdots & \gamma_{X_1 A_k} \\ \vdots & \ddots & \vdots \\ \gamma_{X_p A_1} & \cdots & \gamma_{X_p A_k} \end{pmatrix}}_{\boldsymbol{\Gamma}_{XA}} \underbrace{\begin{pmatrix} A_1 \\ \vdots \\ A_k \end{pmatrix}}_{\boldsymbol{A}} + \underbrace{\begin{pmatrix} \gamma_{X_1 Y} \\ \vdots \\ \gamma_{X_p Y} \end{pmatrix}}_{\boldsymbol{\Gamma}_{XY}} Y + \underbrace{\begin{pmatrix} W_{X_1} \\ \vdots \\ W_{X_p} \end{pmatrix}}_{\boldsymbol{W}_X} .$$

Similarly, we model the output variable, Y, as,

$$Y = \sum_{j=1}^{k} \gamma_{Y A_j} A_j + W_Y = \underbrace{\begin{pmatrix} \gamma_{Y A_1} & \cdots & \gamma_{Y A_k} \end{pmatrix}}_{\boldsymbol{\Gamma}_{YA}} \underbrace{\begin{pmatrix} A_1 \\ \vdots \\ A_k \end{pmatrix}}_{\boldsymbol{A}} + W_Y \ ,$$

so that our inferences will be based on the potentially mispecified models,

$$\boldsymbol{X} = \boldsymbol{\Gamma}_{XA}\, \boldsymbol{A} + \boldsymbol{\Gamma}_{XY}\, Y + \boldsymbol{W}_X \ , \tag{2}$$

$$Y = \boldsymbol{\Gamma}_{YA}\, \boldsymbol{A} + W_Y \ , \tag{3}$$

where the variables $\boldsymbol{X}$, $\boldsymbol{A}$, and Y are scaled to have mean 0 and variance 1, and the error terms have mean 0 and finite variance (but are not assumed to be Gaussian.)

3.3. Linear residualization adjustment

The linear residualization approach is implemented by regressing each separate input variable X_j on the confounders, and then using the residuals of the linear regression fits as the new inputs for machine learning. Since the output variable is not included as a covariate in the regression fits, we have that the approach is actually based on the reduced model obtained by replacing eq. (3) on eq. (2),

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{\Gamma}_{XA}\, \boldsymbol{A} + \boldsymbol{\Gamma}_{XY}\, (\boldsymbol{\Gamma}_{YA}\, \boldsymbol{A} + W_Y) + \boldsymbol{W}_X \\ &= (\boldsymbol{\Gamma}_{XA} + \boldsymbol{\Gamma}_{XY}\, \boldsymbol{\Gamma}_{YA}) \boldsymbol{A} + \boldsymbol{\Gamma}_{XY}\, W_Y + \boldsymbol{W}_X \\ &= \boldsymbol{\Omega}_{XA}\, \boldsymbol{A} + \boldsymbol{W}^{\star}_X \end{aligned} \tag{4}$$

where $\boldsymbol{\Omega}_{XA} = \boldsymbol{\Gamma}_{XA} + \boldsymbol{\Gamma}_{XY}\, \boldsymbol{\Gamma}_{YA}$, and $\boldsymbol{W}^{\star}_X = \boldsymbol{W}_X + \boldsymbol{\Gamma}_{XY}\, W_Y$. In practice, the residualized inputs, $\hat{\boldsymbol{X}}_r$, are estimated as,

$$\hat{\boldsymbol{X}}_r = \boldsymbol{X} - \hat{\boldsymbol{\Omega}}_{XA}\, \boldsymbol{A} \ , \tag{5}$$

by regressing the train and test set inputs on the train and test confounder data in order to estimate $\boldsymbol{\Omega}_{XA}$ via least squares. Note that $\hat{\boldsymbol{X}}_r$ corresponds to the estimated residuals, $\hat{\boldsymbol{W}}^{\star}_X$.

3.4. Causality-aware counterfactual adjustment

The causality-aware counterfactual confounding adjustment is implemented using a modification of Pearl's three step approach for the computation of deterministic counterfactuals (Pearl, 2009; Pearl et al., 2016)[7], where we regress the inputs on the confounders and output variable in order to estimate the model residuals and regression coefficients, and then simulate counterfactual data by adding back the model residuals to a linear predictor that no longer contains the confounder variables. Mechanistically, the causality-aware inputs are calculated as follows:

1. Using the training set, estimate regression coefficients and residuals from the linear regression model, $\boldsymbol{X}^{tr} = \boldsymbol{\Gamma}^{tr}_{XA}\, \boldsymbol{A}^{tr} + \boldsymbol{\Gamma}^{tr}_{XY}\, Y^{tr} + \boldsymbol{W}^{tr}_X$, using least squares, and then estimate the counterfactual inputs,

$$\hat{\boldsymbol{X}}^{tr}_c = \hat{\boldsymbol{\Gamma}}^{tr}_{XY}\, Y^{tr} + \hat{\boldsymbol{W}}^{tr}_X \ , \tag{6}$$

where $\hat{\boldsymbol{W}}^{tr}_X = \boldsymbol{X}^{tr} - \hat{\boldsymbol{\Gamma}}^{tr}_{XA}\, \boldsymbol{A}^{tr} - \hat{\boldsymbol{\Gamma}}^{tr}_{XY}\, Y^{tr}$. (Note that $\hat{\boldsymbol{X}}^{tr}_c$ can also be computed as $\hat{\boldsymbol{X}}^{tr}_c = \boldsymbol{X}^{tr} - \hat{\boldsymbol{\Gamma}}^{tr}_{XA}\, \boldsymbol{A}^{tr}$.)

2. Using the test set, compute the counterfactual inputs,

$$\hat{\boldsymbol{X}}^{ts}_c = \boldsymbol{X}^{ts} - \hat{\boldsymbol{\Gamma}}^{tr}_{XA}\, \boldsymbol{A}^{ts} \ , \tag{7}$$

using the coefficients $\hat{\boldsymbol{\Gamma}}^{tr}_{XA}$ estimated in the training set.

Once the training and test set counterfactual inputs, $\hat{\boldsymbol{X}}^{tr}_c$ and $\hat{\boldsymbol{X}}^{ts}_c$, have been generated we can then use $\hat{\boldsymbol{X}}^{tr}_c$ and Y^{tr}

[7] Pearl's approach is comprised of the "abduction, action, prediction" steps. In the context of linear models, they are implemented as follows: (i) causal effects and residuals are estimated in the "abduction" step; (ii) the causal graph is modified using an atomic intervention of the type "$do(Z = z)$" in the "action" step; and (iii) using the intervened graph and the quantities computed in the abduction step, the new counterfactual variables are generated in the "prediction" step. While the action step in Pearl's approach is enforced by an atomic intervention, in our approach it is based on a soft intervention. The abduction and prediction steps are, nonetheless, still the same.

to train a linear learner, and then use $\hat{\boldsymbol{X}}_c^{ts}$ to generate predictions that are free from the influence, or at least impacted by a lesser degree, by the observed confounders. Observe that the calculation of the test set causality-aware inputs in eq. (7) does not uses the test set output, Y^{ts}. Observe, as well, that for large sample sizes, we have that the computation of the test set inputs using eq. (7) is equivalent to computing the test set inputs using $\hat{\boldsymbol{X}}_c^{ts} = \hat{\boldsymbol{\Gamma}}_{XY}^{ts} Y^{ts} + \hat{\boldsymbol{W}}_X^{ts}$ since for large enough sample sizes we have that $\hat{\boldsymbol{\Gamma}}_{XA}^{tr} \approx \hat{\boldsymbol{\Gamma}}_{XA}^{ts}$ (assuming that the effects are stable across the training and test data) so that,

$$\begin{aligned}\hat{\boldsymbol{X}}_c^{ts} &= \boldsymbol{X}^{ts} - \hat{\boldsymbol{\Gamma}}_{XA}^{tr}\, \boldsymbol{A}^{ts} \approx \boldsymbol{X}^{ts} - \hat{\boldsymbol{\Gamma}}_{XA}^{ts}\, \boldsymbol{A}^{ts} \\ &= \hat{\boldsymbol{\Gamma}}_{XY}^{ts}\, Y^{ts} + \hat{\boldsymbol{W}}_X^{ts}\,. \end{aligned} \tag{8}$$

4. Results

Before we present the main results of this section (namely, Theorems 2 and 3), we first present a few pre-requisite results in Result 1 and Theorem 1. (The proofs of all results are presented in Supplementary Section 2.)

Result 1. *For centered data, the asymptotic MSE of (potentially mispecified) regression models trained by least squares is given by,*

$$E[MSE] = Var(Y) - Cov(Y, \boldsymbol{X}) Cov(\boldsymbol{X})^{-1} Cov(\boldsymbol{X}, Y)\,.$$

Theorem 1. *For an anticausal prediction task influenced by a set of observed confounders $\boldsymbol{A}$, the asymptotic cross-covariances between Y and $\boldsymbol{X}_c$, $\boldsymbol{X}_r$, and $\boldsymbol{A}$ are given respectively by,*

$$Cov(\boldsymbol{X}_c, Y) = \boldsymbol{\Gamma}_{XY}\,, \tag{9}$$

$$Cov(\boldsymbol{X}_r, Y) = \boldsymbol{\Gamma}_{XY}(1 - \boldsymbol{\Gamma}_{YA}\, Cov(\boldsymbol{A})\, \boldsymbol{\Gamma}_{YA}^T)\,, \tag{10}$$

$$Cov(\boldsymbol{A}, Y) = \boldsymbol{\Gamma}_{YA}\, Cov(\boldsymbol{A})\,, \tag{11}$$

while the covariances of $\boldsymbol{X}_c$ and $\boldsymbol{X}_r$ are given by,

$$Cov(\boldsymbol{X}_c) = \boldsymbol{\Gamma}_{XY}\, \boldsymbol{\Gamma}_{XY}^T + Cov(\boldsymbol{W}_X)\,, \tag{12}$$

$$\begin{aligned} Cov(\boldsymbol{X}_r) = Cov(\boldsymbol{X}_c) &- \\ &- \boldsymbol{\Gamma}_{XY}\, \boldsymbol{\Gamma}_{YA}\, Cov(\boldsymbol{A})\, \boldsymbol{\Gamma}_{YA}^T\, \boldsymbol{\Gamma}_{XY}^T\,. \end{aligned} \tag{13}$$

Using Result 1 and Theorem 1 we can show the following.

Theorem 2. *Let $E[MSE_r]$ and $E[MSE_c]$ represent the expected MSE of (potentially mispecified) regression models trained with residualized and causality-aware adjusted inputs, respectively. For an anticausal prediction task, where training and test sets come from the same distribution, and sample sizes tend to infinity we have that $E[MSE_r] \geq E[MSE_c]$.*

The above result guarantees that the causality-aware approach dominates residualization under the particular conditions described in Theorem 2. While we do not provide analogous results for alternative metrics, models, or for classification tasks, the next result suggests that the causality-aware approach will tend to outperform residualization for linear learners outside the settings of Theorem 2.

Theorem 3. *Under the conditions of Theorem 1, for each element j of the vectors $Cov(\boldsymbol{X}_c, Y)$ and $Cov(\boldsymbol{X}_r, Y)$, we have that $|Cov(X_{c,j}, Y)| \geq |Cov(X_{r,j}, Y)|$.*

In the special case of a single confounder variable, A, equations (9) and (10) in Theorem 1 reduce to, $Cov(X_c, Y) = \gamma_{XY}$ and $Cov(X_r, Y) = \gamma_{XY}(1 - \gamma_{YA}^2)$, and the result in Theorem 3 follows from,

$$|Cov(X_c, Y)| = |\gamma_{XY}| \;\geq\; |\gamma_{XY}|(1 - \gamma_{YA}^2) = |Cov(X_r, Y)|\,,$$

since $(1 - \gamma_{YA}^2) \leq 1$ because γ_{YA} corresponds to the correlation between the Y and A variables[8] and can only assume values between -1 and 1.

Now, observe that under the assumption of the absence of dataset shift between training and test sets, we conjecture that the above result leads to a better performance by linear learners trained with causality-aware inputs since Theorem 3 guarantees that for each input X_j, the linear association between the counterfactual input, $X_{c,j}$, and Y is always stronger or equal than the linear association between the residual input, $X_{r,j}$, and Y. Since linear learners are only able to leverage linear associations between the inputs and the output for the predictions, it seems reasonable to expect that a linear learner trained with causality-aware counterfactual inputs will tend to outperform the respective learner trained on the residualized inputs.

5. Synthetic data illustrations

Here, we present synthetic data illustrations of the points in the previous section for both regression and classification tasks. We evaluate predictive performance using mean squared error (MSE) in the regression task experiments, and accuracy (ACC) in the classification task experiments. We first report in detail the results based on models containing two inputs and 2 confounders, before presenting simulations based on larger numbers of inputs. All experiments were run in R (R Core Team, 2019).

5.1. Regression task experiments

To illustrate the regression tasks results we ran 2 experiments, the first based on correctly specified models, and the second based on mispecified models. In both experiments,

[8] Direct application of Wright's method of path coefficients (Wright, 1934) to the causal diagram $A \rightarrow X \leftarrow Y$, shows that the marginal correlations among these three variables can be decomposed as $Cor(A, Y) = \gamma_{YA}$, $Cor(A, X) = \gamma_{XA} + \gamma_{YA}\,\gamma_{XY}$, and $Cor(X, Y) = \gamma_{XY} + \gamma_{XA}\,\gamma_{YA}$ in terms of path coefficients.

we simulated correlated error terms, $\boldsymbol{W}_A$ and $\boldsymbol{W}_X$, from bivariate normal distributions,

$$\boldsymbol{W}_A^o \sim \mathrm{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho_A\\ \rho_A & 1\end{pmatrix}\right), \tag{14}$$

$$\boldsymbol{W}_X^o \sim \mathrm{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho_X\\ \rho_X & 1\end{pmatrix}\right). \tag{15}$$

In the first experiment, the confounders, output and input variables were generated according to,

$$A_j^o = \mu_{A_j} + W_{A_j}^o, \tag{16}$$

$$Y^o = \mu_Y + \beta_{YA_1} A_1^o + \beta_{YA_2} A_2^o + W_Y^o, \tag{17}$$

$$X_j^o = \mu_{X_j} + \beta_{X_jA_1} A_1^o + \beta_{X_jA_2} A_2^o + \beta_{X_jY} Y^o + W_{X_j}^o, \tag{18}$$

while in the second (i.e., the mispecified case) they were generated according to,

$$A_j^o = \mu_{A_j} + W_{A_j}^o, \tag{19}$$

$$Y^o = \mu_Y + \beta_{YA_1} A_1^{o2} + \beta_{YA_2} A_2^{o2} + W_Y^o, \tag{20}$$

$$X_j^o = \mu_{X_j} + \beta_{X_jA_1} A_1^{o2} + \beta_{X_jA_2} A_2^{o2} + \beta_{X_jY} Y^{o2} + W_{X_j}^o, \tag{21}$$

where $j = 1, 2$ and $W_Y^o \sim \mathrm{N}(0\,,\,\sigma_Y^2)$.

For each experiment, we performed 1,000 simulations as follows:

1. Randomly sampled the simulation parameters from uniform distributions, with the intercept and slope parameters μ_{A_1}, μ_{A_2}, μ_Y, μ_{X_j}, β_{YA_1}, β_{YA_2}, β_{X_jY}, $\beta_{X_jA_1}$, and $\beta_{X_jA_2}$ drawn from a $\mathrm{U}(-3, 3)$ distribution; the error variance, σ_Y^2, drawn from a $\mathrm{U}(1, 3)$ distribution; and the correlations ρ_A and ρ_X from a $\mathrm{U}(-0.8, 0.8)$ distribution.

2. Simulated the original data $\boldsymbol{A}^o$, Y^o, and $\boldsymbol{X}^o$ using the simulation parameters sampled in step 1, according to the models in equations (16)-(18) in the first experiment, and equations (19)-(21) in the second, and then standardized the data to obtain $\boldsymbol{A}$, Y, and $\boldsymbol{X}$. (Each simulated dataset was composed of 10,000 training and 10,000 test examples.)

3. For each simulated feature, X_j, we generated the respective residualized and causality-aware features as described in Sections 3.3 and 3.4, and computed $Cov(\hat{X}_{r,j}, Y)$ and $Cov(\hat{X}_{c,j}, Y)$.

4. Finally, we trained linear regression models using the residualized and the causality-aware inputs, and computed the respective test set mean squared errors, MSE_r and MSE_c.

Figures 5 reports the results. Panels a and b illustrate the result from Theorem 3, showing that $|Cov(X_{c,j}, Y)| \geq |Cov(X_{r,j}, Y)|$ for both input variables X_1 and X_2, while panels d and e illustrate that the results still hold under model mispecification. Panels c and f show that $\mathrm{MSE}_c \leq \mathrm{MSE}_r$, illustrating the results from Theorem 2.

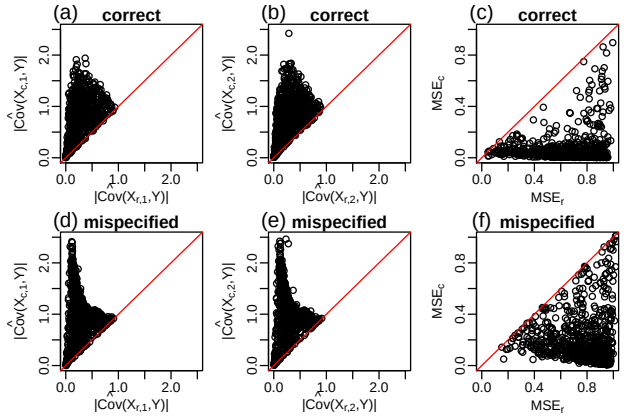

Figure 5. Regression task experiments.

5.2. Classification task experiments

Similarly to in the previous subsection, we ran two experiments. The first based on correctly specified models, and the second based on mispecified models. In both experiments, we simulated correlated input error terms, $\boldsymbol{W}_X$, as described in eq. (15) and simulate correlated binary confounder variables from a bivariate Bernoulli distribution (Dai et al., 2013), with probability density function,

$$p(A_1^o, A_2^o) = p_{11}^{a_1\,a_2}\, p_{10}^{a_1\,(1-a_2)}\, p_{01}^{(1-a_1)\,a_2}\, p_{00}^{(1-a_1)\,(1-a_2)}, \tag{22}$$

where $p_{ij} = P(A_1^o = i, A_2^o = j)$, and the covariance between A_1^o and A_2^o is given by $Cov(A_1^o, A_2^o) = p_{11}\,p_{00} - p_{01}\,p_{10}$.

The binary output data Y^o was generated according to a logistic regression model where, $P(Y^o = 1 \mid A_1^o = a_1, A_2^o = a_2) = 1/(1 + \exp\{-(\mu_Y + \beta_{YA_1}\,a_1 + \beta_{YA_2}\,a_2)\})$. For the correctly specified experiments, the features X_j^0, $j = 1, 2$, where generated according to $X_j^o = \mu_{X_j} + \beta_{X_jA_1}\,A_1^o + \beta_{X_jA_2}\,A_2^o + \beta_{X_jY}\,Y^o + W_{X_j}^o$. For the incorrectly specified experiments, on the other hand, the features were generated as $X_j^o = \mu_{X_j} + \beta_{X_jYA_1}\,Y^o\,A_1^o + \beta_{X_jYA_2}\,Y^o\,A_2^o + W_{X_j}^o$, containing only interaction terms between A_k^o and Y^o.

As before each experiment was based on 1,000 replications with simulation parameters μ_Y, μ_{X_j}, β_{YA_1}, β_{YA_2}, β_{X_jY}, $\beta_{X_jA_1}$, $\beta_{X_jA_2}$, $\beta_{X_jYA_1}$, and $\beta_{X_jYA_2}$ drawn from a $\mathrm{U}(-3, 3)$ distribution; $\rho_X \sim \mathrm{U}(-0.8, 0.8)$; and p_{11}, p_{10}, p_{01}, and p_{00} sampled by randomly splitting the interval $(0, 1)$ into 4 pieces. For each simulated input we: generated the respective residualized and causality-aware inputs;

trained logistic regression classifiers using the processed features; and computed the respective test set classification accuracies, ACC_r and ACC_c.

Figure 6 reports the results. As before, panels a, b, d, and e illustrate the result from Theorem 3. Panels c and f provide empirical evidence for our conjecture that the causality-aware approach tends to outperform the residualization adjustment in classification tasks.

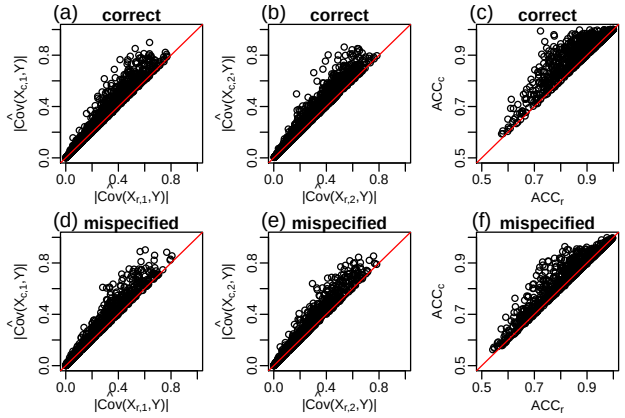

Figure 6. Classification task experiments.

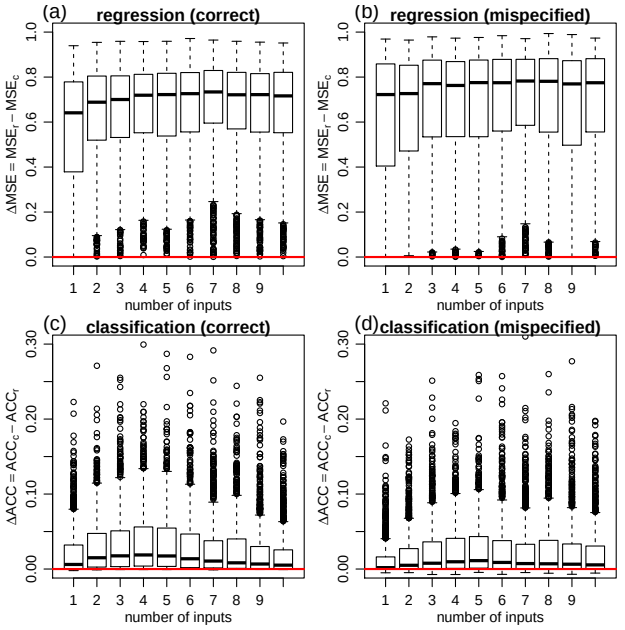

Figure 7. Results from additional experiments.

5.3. Simulation experiments on larger dimensions

Here, we present further simulation experiments based on models trained with larger number of inputs. The synthetic data was generated as described in the previous subsections, except that the input error terms were sampled according to $W_X^o \sim N_p(\mathbf{0}, \mathbf{\Sigma})$, where the ijth entry of $\mathbf{\Sigma}$ is given by 1 for $i = j$, and by $\rho^{|i-j|}$ for $i \neq j$, and $p = 1, \ldots, 10$ represents the number of inputs.

Figure 7 reports the results for both correctly specified and mispecified regression and classification models. In all panels, the x-axis represents the number of inputs, while the y-axis represents $\Delta MSE = MSE_r - MSE_c$ in panels a and b, and $\Delta ACC = ACC_c - ACC_r$ in panels c and d. Note that positive $\Delta METRIC$ values indicates a better performance of the causality-aware approach. The red horizontal line is set at 0.

In accordance with Theorem 2, panels a and b show that the causality-aware approach tended to dominated residualization for regression tasks trained with regression models and evaluated with the MSE metric. Panels c and d show that while the residualization produced slightly higher ACCs than the causality-aware adjustment on a small fraction of the simulated data sets (note how some of that ΔACC values are slightly below the red line, especially in panel d), the causality-aware approach outperformed the residualization adjustment on the majority of the simulations. These experiments again provide empirical evidence for the better performance of the causality-aware adjustment.

Finally, observe that all simulations reported in this section were based in large sample sizes. Supplementary Section 3 reports the results for the same experiments based on smaller sample sizes (training and test sets containing 100 examples each). The causality-aware approach still tended to outperform (although it no longer dominated) the residualization approach in this setting too (see Supplement).

6. Real data illustrations

We also compared these approaches using real data from the Parkinson's Disease (PD) Digital Biomarker Dream Challenge (Sieberts et al, 2021). We build classifiers of disease status (i.e., PD vs non-PD) adjusted for age confounding, using 23 distinct feature datasets submitted by participants of the challenge (these feature sets were extracted from raw accelerometer data using distinct signal processing techniques or deep learning models to automatically learn the features from the raw data). Each available feature set was processed using the causality-aware and residualization adjustments. Figure 8 reports the results from logistic regression classifiers trained on 100 distinct i.i.d. training/test data splits (with 2,772 samples equally split between train/test sets). In all 23 experiments, the causality-aware approach outperformed residualization in terms of AUROC. See Supplementary Section 4 for further details.

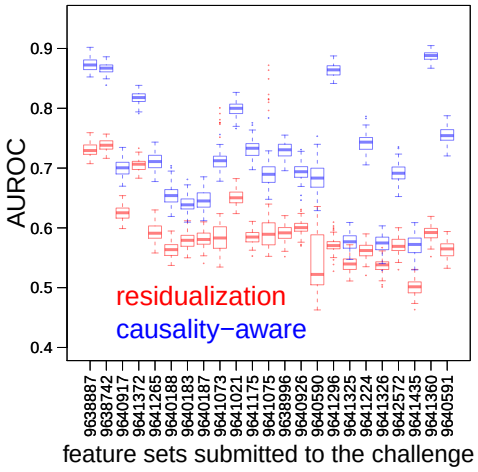

Figure 8. Real data illustrations.

7. Stability results

So far, our investigations have focused on the case where there is no dataset shift between the training and test sets. In this section, we show that the causality-aware approach is more stable than residualization under dataset shifts of the joint distribution of the confounders and outcome variables generated by selection mechanisms. The stability properties of the causality-aware approach were presented in (Chaibub Neto, 2020) where it was shown that the expected MSE of the causality-aware approach does not depend on $Cov(\boldsymbol{A}^{ts}, Y^{ts})$, and, therefore, is stable w.r.t. dataset shifts in this quantity. The next result shows that this is not the case for the residualization approach.

Theorem 4. *For an anticausal prediction task influenced by a set of observed confounders $\boldsymbol{A}$, we have that, contrary to the causality-aware approach, the linear residualization approach is not stable with respect to dataset shifts in $Cov(\boldsymbol{A}^{ts}, Y^{ts})$, when the predictive performance is measured by MSE.*

See Supplementary Section 2 for the proof. Because dataset shifts in $Cov(\boldsymbol{A}^{ts}, Y^{ts})$ caused by selection mechanisms are commonly observed in real word applications, this observation has important practical implications.

Next, we illustrate this result in a couple of synthetic data experiments. For simplicity, we generate data from the toy model in Figure 9, where the double arrow represents an association generated by a selection mechanism, and where $Cov(A, Y) = \sigma_{AY}$, $Var(A) = \sigma_{AA}$, and $Var(Y) = \sigma_{YY}$. As shown in Supplementary Section 2, the expected MSE for the residualization approach is given by,

$$E[MSE_r] = Var(Y^{ts}) + (\hat{\beta}_r^{tr})^2 Var(X_r^{ts}) - \\ - 2\hat{\beta}_r^{tr} Cov(X_r^{ts}, Y^{ts}) \,,$$

where both,

$$Var(X_r^{ts}) = \sigma_X^2 + \beta_{XY}^2\, \sigma_{YY}^{ts} - \frac{\beta_{XY}^2\, (\sigma_{AY}^{ts})^2}{\sigma_{AA}^{ts}} \,, \quad (23)$$

$$Cov(X_r^{ts}, Y^{ts}) = \beta_{XY}\, \sigma_{YY}^{ts} - \frac{\beta_{XY}\, (\sigma_{AY}^{ts})^2}{\sigma_{AA}^{ts}} \,, \quad (24)$$

are still functions of σ_{AY}^{ts}, showing that the expected MSE of the residualization approach will be unstable w.r.t. shifts in $Cov(A^{ts}, Y^{ts})$[9].

In our experiments, we generated dataset shift in $P(A, Y)$ by varying $Cov(A, Y) = \sigma_{AY}$, $Var(A) = \sigma_{AA}$, and $Var(Y) = \sigma_{YY}$ between the training and test sets. We, nonetheless, use the same values of β_{XA}, β_{XY}, and σ_X^2 in the generation of the training and test features, so that only the joint distribution $P(A, Y)$ differs between the training and test sets (while $P(X \mid A, Y)$ is stable).

We performed two stability experiments. In the first we kept $Var(Y^{ts})$ constant across the test sets, while in the second we let $Var(Y^{ts})$ vary across the test sets. Each experiment was based in 1,000 replications. In our first simulation experiment, for each replication we:

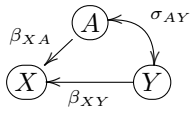

Figure 9.

1. Sampled the causal effects β_{XY} and β_{XA} from a $U(-3, 3)$ distribution, and the training set covariance σ_{AY}^{tr} from a $U(-0.8, 0.8)$ distribution.

2. Generated training data ($n = 10,000$) by first sampling,

$$\begin{pmatrix} A^{tr} \\ Y^{tr} \end{pmatrix} \sim \mathrm{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{AY}^{tr} \\ \sigma_{AY}^{tr} & 1 \end{pmatrix} \right) , \quad (25)$$

and then generating $X^{tr} = \beta_{XA}\, A^{tr} + \beta_{XY}\, Y^{tr} + U_X^{tr}$ with $U_X^{tr} \sim N(0, 1)$.

3. Generated 9 distinct test sets, where each test set dataset ($n = 10,000$) was generated by first sampling,

$$\begin{pmatrix} A^{ts} \\ Y^{ts} \end{pmatrix} \sim \mathrm{N}_2 \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{AA}^{ts} & \sigma_{AY}^{ts} \\ \sigma_{AY}^{ts} & \sigma_{YY}^{ts} \end{pmatrix} \right) , \quad (26)$$

and then generating $X^{ts} = \beta_{XA}\, A^{ts} + \beta_{XY}\, Y^{ts} + U_X^{ts}$ with $U_X^{ts} \sim N(0, 1)$. In order to generate dataset shifts, the covariances between A^{ts} and Y^{ts} and the variances of A^{ts} were set, respectively, to $\sigma_{AY}^{ts} =$

[9] For the causality-aware approach, on the other hand, we have that $Var(X_c^{ts}) = \sigma_X^2 + \beta_{XY}^2\, \sigma_{YY}^{ts}$ and $Cov(X_c^{ts}, Y^{ts}) = \beta_{XY}\, \sigma_{YY}^{ts}$, so that the expected MSE, $E[MSE_c] = Var(Y^{ts}) + (\hat{\beta}_c^{tr})^2 Var(X_c^{ts}) - 2\hat{\beta}_c^{tr} Cov(X_c^{ts}, Y^{ts})$ does not depend on $Cov(A^{ts}, Y^{ts})$.

$\{-0.8, -0.6, -0.2, 0, 0.2, 0.4, 0.6, 0.8\}$ and $\sigma^{ts}_{AA} = \{1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00\}$ across the 9 distinct test sets, while the variance of Y_{ts} was fixed at $\sigma_{YY} = 1$.

4. Processed the training and the test features as described in Sections 3.3 and 3.4 to generate the residualized and causality-aware inputs.

5. Trained regression models using the residualized and causality-aware inputs and evaluated the performance of each of the trained models on each of the 9 test sets.

Figure 10 reports the results and clearly shows that while the predictive performance of the causality-aware approach was stable across the test sets, the residualization approach was fairly unstable. Panel b shows the results of the first 3 simulations in more detail. Each line presents the MSE of the same trained model across the 9 distinct test sets, showing that the residualization results (red lines) vary widely across the test sets, while the causality-aware (blue lines) are fairly stable. Panel c reports the distributions of stability error (i.e., the standard deviation of the MSE scores across the 9 test sets) for both approaches.

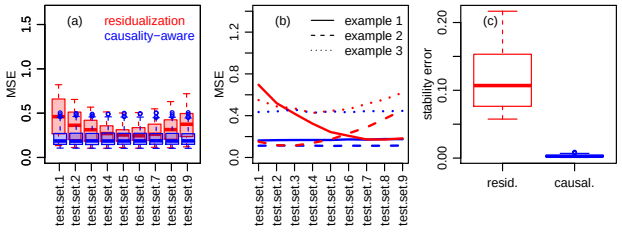

Figure 10. Stability illustrations, with fixed $Var(Y^{ts})$.

Because the expected MSE of any approach will, in general, depend on the variance of Y^{ts} we performed an additional simulation study (Figure 11) where the data was generated as before except that we varied $Var(Y^{ts})$ according to $\{1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, 3.00\}$ across the 9 test sets. The results show that, while MSE_c also changed across the test sets, the causality-aware approach is still much more stable than residualization.

These observations suggest that, in safety-critical applications where large shifts in performance are concerning, and where stability is a desirable property, the causality-aware adjustment is again more appropriate than residualization[10].

[10]Observe that we do not compare the causality-aware approach against alternative stable prediction approaches because the approaches proposed by Kuang et al. (2018; 2020) are tailored to causal prediction tasks, rather than to anticausal ones. Note, as well, that while counterfactual normalization (Subbaswamy et al., 2018) can also be applied in anticausal tasks, we have that its application to the particular model used in our simulations would produce identical results as the causality-aware adjustment (since the counterfactual $X(A = \emptyset)$, used by the counterfactual normalization method as the stable set for predicting Y, corresponds to causality-aware input $X^* = X - \hat{\beta}_{XA}\, A$ in our example).

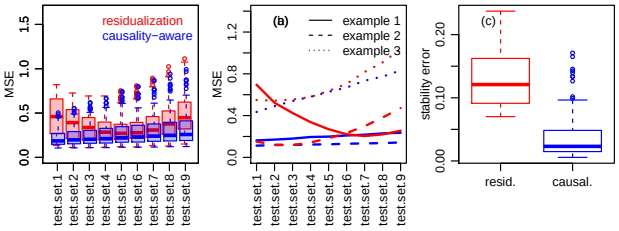

Figure 11. Stability illustrations, with increasing $Var(Y^{ts})$.

8. Final remarks

In this paper, we compare linear residualization against the causality-aware confounding adjustment. In situations where the training and test sets come from the same distribution, our results suggest that the causality-aware approach outperforms residualization. In situations where the training and test sets are shifted due to selection biases, we show that the causality-aware approach generates more stable predictions than the linear residualization adjustment.

Our main goal in this paper was to show how a more principled, causality-inspired approach, can better handle confounding in comparison with a widely used ad hoc statistical adjustment approach. Given that confounding is a causal concept, its is not really surprising that a causality-inspired approach will outperform a purely statistical one. In any case, we believe that this point is not well appreciated by the larger ML community (outside causality experts) and by researchers in applied fields, which still make use of linear residualization. In anticausal ML applications where a linear model provides an adequate fit to the data, there is really no good reason to keep using linear residualization (given that the causality-aware approach does not require knowledge about how the input variables are causally related, and that its implementation is as trivial as the implementation of residualization).

Finally, we point out that even though the causality-aware approach also outperforms the residualization technique when the true data generation process for the inputs cannot be adequately modeled by a linear model, the causality-aware adjustment might still fail to fully deconfound the predictions in applications for which linear models are not really adequate. In Supplementary Section 5 we describe a simple approach (Chaibub Neto et al., 2019) that can be used to evaluate the adjustment's effectiveness, and provides an useful sanity check to determine if more sophisticated adjustments are necessary.

References

[1] Abdulkadir, A., Ronneberger, O., Tabrizi, S.J., Kloppel, S. (2014). Reduction of confounding effects with voxel-wise Gaussian process regression in structural MRI. In: 2014 *International Workshop on Pattern Recognition in Neuroimaging*, 1-4.

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

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

[4] Chaibub Neto, E. (2020) Towards causality-aware predictions in static anticausal machine learning tasks: the linear structural causal model case. In *Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems 2020. arXiv:2001.03998*.

[5] 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.

[6] Dai B, Ding S, and Wahba G. 2013. Multivariate Bernoulli distribution. *Bernoulli* **19**: 1465-1483.

[7] Doan N. T.. et al. (2017) Dissociable diffusion MRI patterns of white matter microstructure and connectivity in Alzheimer's disease spectrum. *Scientific Reports*, **7**:45131, DOI: 10.1038/srep45131

[8] Dubois, J., Galdi, P., Han, Y., Paul, L.K., Adolphs, R. (2017). Predicting Personality Traits from Resting-State fMRI. *bioRxiv*. https://doi.org/10.1101/215129.

[9] Dukart, J., Schroeter, M.L., Mueller, K., et al. (2011) Age correction in dementia-matching to a healthy brain. *PloS One*, **6**, e22193.

[10] Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., and Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human Brain Mapping*, **2**, 189-210.

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

[12] Greenstein D., Malley J. D., Weisinger B., Clasen L., and Gogtay N. (2012) Using multivariate machine learning methods and structural MRI to classify childhood onset schizophrenia and healthy controls. *Frontiers in Psychiatry*, **53**: 1.

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

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

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

[16] Kostro, D., Abdulkadir, A., Durr, A., Roos, R., Leavitt, B.R., Johnson, H., et al. (2014) Correction of inter-scanner and within-subject variance in structural MRI based automated diagnosing. *Neuroimage*, **98**, 405-415.

[17] Maglanoc, L. A., et al. (2020) Multimodal fusion of structural and functional brain imaging in depression using linked independent component analysis. *Human Brain Mapping*, **41**, 241-255.

[18] 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*.

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

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

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

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

[23] 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.

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

[25] 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/.

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

[27] Rao, A., Monteiro, J.M., Mourao-Miranda, J., Alzheimer's Disease Initiative (2017) Predictive modelling using neuroimaging data in the presence of confounds. *Neuroimage*, **150**, 23-49.

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

[29] Sieberts, S. K., Schaff, J., Duda, M., Pataki, B. A., Sun, M., et al. (2021) Crowdsourcing digital health measures to predict Parkinson's disease severity: the Parkinson's Disease Digital Biomarker DREAM Challenge. *npj Digital Medicine*, **4:** 53; https://doi.org/10.1038/s41746-021-00414-7

[30] Snoek L., Miletic S., Steven Scholte H. S. (2019) How to control for confounds in decoding analyses of neuroimaging data. *NeuroImage*, **184**, 741-760.

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

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

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

[34] Subbaswamy, A., Chen, B., Saria, S. (2019b) A universal hierarchy of shift-stable distributions and the tradeoff between stability and performance. arXiv:1905.11374.

[35] Todd, M.T., Nystrom, L.E., Cohen, J.D., (2013) Confounds in multivariate pattern analysis: theory and rule representation case study. *Neuroimage*, **77**, 157-165.

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