Supporting information Appendix for Pichler & Hartig – Can machine learning be used for causal inference?

Table of contents

**Summary:** This document provides supporting information on Pichler & Hartig – Can machine learning be used for causal inference.

## 1 Hyper-parameter tuning

We performed a hyper-parameter search to check if hyper-parameters can influence the unbiasness of the models. For that, we created simulation scenarios 100 and 1,000 observations, and each with 10 and 100 features. In each simulation scenario, two features (X1 and X5) had an effect on Y and all other features had no effect on Y. Features were sampled from a multivariate normal distribution and all features were randomly correlated (Variance-covariance matrix was sampled from a Wishart distribution (degrees of freedom = number of features (10 or 100), scale matrix was the identity matrix).

3,000 combinations of hyper-parameters were randomly drawn (Table S1). For each draw, the data was sampled, and bias for X1, X2, and X5 were calculated (bias = true effect - estimated effect, effects were estimated by MCE). In each draw, holdout data of the same size as the training data were used used to calculate the root mean squared error (RMSE) of the models.

**Table S** 1: Overview over hyper-parameters for Neural Network, Boosted Regression Tree, and Random Forest

| Algorithm | Hyper-parameter | Range |
| --- | --- | --- |
| Neural Network | activation function | [relu, leaky\_relu, tanh, selu, elu, celu, gelu] |
|  | depth | [1, 20] |
|  | width | [2, 171] |
|  | batch size (sgd) | [1, 100] |
|  | dropout rate | [0, 0.3] |
| Boosted Regression Tree | eta | [0.01, 0.4] |
|  | max depth | [2, 25] |
|  | subsample | [0.5, 1] |
|  | max tree | [30, 125] |
|  | lambda | [1, 20] |
| Random Forest | mtry | [2, 99] |
|  | min node size | [2, 70] |
|  | max depth | [2, 50] |
|  | regularization factor | [0, 1] |

### 1.1 Neural Network

|  |
| --- |
| **Figure S** 1: Trade-off between bias (feature X1, true effect = 1) and RMSE of the fittend NNs. Each point is a simulated scenario with a different set of hyper-parameters. Scenarios with 1,00 and 1,000 observations and with 100 or 10 features (P) were simulated. In each scenario, only two features (X1 and X5) had an effect (effect size = 1) on the response variable Y. Hyper-parameters (activation functions, depth and width of the hidden layers, batch\_size (sgd), and dropout) were randomly sampled. |

In general, we found only a weak correlation between the bias of X1 and the RMSE on the holdout data (Fig. S1). This correlation was slightly stronger in the data-poor situations (N = 100, P = 100, and N = 1000, P = 100; Fig. S1).

|  |
| --- |
| **Figure S** 2: Effects of hyper-parameters of neural networks on bias of feature X1. Scenarios with 1,00 and 1,000 observations and with 100 or 10 features (P) were simulated. In each scenario, only two features (X1 and X5) had an effect (effect size = 1) on the response variable Y. Hyper-parameters (activation functions, depth and width of the hidden layers, batch\_size (sgd), and dropout) were randomly sampled. Effects of hyper-parameters were estimated by linear regression models. |

We found significant differences between the activation functions for the bias of X (true effect = 1) and the RMSE (N = 100, P = 100). Here, the ‘selu’ activation function had the smallest biases and RMSE (Fig. S2). With more data, or less features, the differences between the activation functions got smaller and for data-rich situations (N = 1000, P = 10), the confidence intervals of the different activation functions were in the same range (Fig. S2).

Depth had a strong effect on the bias and the RMSE, and this effect was stronger for the data-poor scenario (N = 100, P = 100; Fig. S2). Lower depth had smaller biases and lower RMSE. Width had a weak effect in the data-poor situation, but stronger effects in data-rich scenarios on the bias. But width had strong effects on the RMSE, increasing the width led to lower RMSE (this effect was weaker for the data-rich scenarios, Fig. S2). Dropout had only small effects on the biases and the RMSE.

Hyper-parameters had only small effects on the bias of the zero-effect (X2), thus it seems that the NN can identify, reliable, zero-effects independent of the hyper-parameters (Fig. S2).

### 1.2 Boosted Regression Trees

The correlation between RMSE and bias was stronger for data-rich scenarios than for the data-poor scenarios (Fig. S3). We found a stronger relation between RMSE and bias of X1 for BRT than for NN (Fig. S1, S3).

|  |
| --- |
| **Figure S** 3: Trade-off between bias (feature X1, true effect = 1) and RMSE of the fittend BRTs. Each point is a simulated scenario with a different set of hyper-parameters (3,000 samples). Scenarios with 1,00 and 1,000 observations and with 100 or 10 features (P) were simulated. In each scenario, only two features (X1 and X5) had an effect (effect size = 1) on the response variable Y. Hyper-parameters (max\_depth, subsample, lambda, and max\_tree) were randomly sampled. |

Overall, the effects on bias (X1 ) and RMSE were similar, although small (Fig. S4). Lambda and max tree had the largest effects on the biases and the RMSE (Fig. S3). Effects on the bias of the zero-effect, X2, were small and had large confidence intervals (Fig. S3).

|  |
| --- |
| **Figure S** 4: Effects of hyper-parameters of BRT on bias of feature X1. Scenarios with 1,00 and 1,000 observations and with 100 or 10 features (P) were simulated. In each scenario, only two features (X1 and X5) had an effect (effect size = 1) on the response variable Y. Hyper-parameters (max\_depth, subsample, lambda, and max\_tree) were randomly sampled. Effects of hyper-parameters were estimated by linear regression models. |

#### 1.2.1 Confounder unequal scenario

As we found large biases in the proof-of-concept simulations for BRT, we repeated the hyper-paramter tuning for the worst scenario (unequal confounder, Fig. 2C in the main text).

We used the second best set (eta = 0.31, max depth = 5, subsample = 0.55, lambda = 4.29, max tree = 116) for the rest of analyses. because it had the best trade-off for the biases of three features and the RMSE (Table S2).

**Table** **:** Trade-off between bias (feature X~1~, true effect = 1) and RMSE of the fittend BRT Each point is a simulated scenario with a different set of hyper-parameters. Scenarios with 1,00 and 1,000 observations and with 100 or 10 features (P) were simulated. In each scenario, only two features (X~1~ and X~5~) had an effect (effect size = 1) on the response variable Y. Hyper-parameters (max\_depth, subsample, lambda, and max\_tree) were randomly sampled (3,000 draws).

| **eta** | **max\_depth** | **subsample** | **lambda** | **max\_tree** | **Bias X1** | **Bias X2** | **Bias X3** | **rmse** | **weighted\_bias** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.38 | 2 | 0.53 | 2.77 | 77 | 0.01 | -0.10 | -0.04 | 0.34 | 0.08 |
| 0.31 | 5 | 0.55 | 4.29 | 116 | 0.04 | 0.02 | -0.02 | 0.37 | 0.08 |
| 0.35 | 16 | 0.55 | 6.97 | 111 | 0.03 | 0.05 | -0.03 | 0.39 | 0.09 |
| 0.27 | 2 | 0.57 | 2.88 | 105 | 0.05 | -0.05 | -0.02 | 0.36 | 0.09 |
| 0.20 | 4 | 0.69 | 1.36 | 76 | 0.03 | -0.04 | -0.04 | 0.36 | 0.10 |
| 0.37 | 3 | 0.51 | 6.14 | 120 | -0.01 | 0.04 | -0.04 | 0.36 | 0.10 |

### 1.3 Random Forest

RF showed a stronger correlation between bias of X1 and the RMSE (Fig. S5) than NN and BRT.

|  |
| --- |
| **Figure S** 5: Trade-off between bias (feature X1, true effect = 1) and RMSE of the fittend RF. Each point is a simulated scenario with a different set of hyper-parameters. Scenarios with 1,00 and 1,000 observations and with 100 or 10 features (P) were simulated. In each scenario, only two features (X1 and X5) had an effect (effect size = 1) on the response variable Y. Hyper-parameters (max\_depth, subsample, lambda, and max\_tree) were randomly sampled. |

Bias of X1 was mostly only affect by the mtry parameter (Fig. S6), which was within our expectations because it controls the spill-over effect for collinear features.

|  |
| --- |
| **Figure S** 6: Effects of hyper-parameters of neural networks on bias of feature X1. Scenarios with 1,00 and 1,000 observations and with 100 or 10 features (P) were simulated. In each scenario, only two features (X1 and X5) had an effect (effect size = 1) on the response variable Y. Hyper-parameters (max\_depth, subsample, lambda, and max\_tree) were randomly sampled. Effects of hyper-parameters were estimated by linear regression models. |

#### 1.3.1 Confounder unequal scenario

As we found large biases in the proof-of-concept simulations for RF, we repeated the hyper-paramter tuning for the worst, the last, scenario (unequal confounder, Fig. 2C in the main text).

We tried the second best set of hyper-parameters in the proof-of-concept simulations but it actually led to larger biases in Fig. 2A and 2B and we decided to fall back to the default values.

**Table** **:** Trade-off between bias (feature X~1~, true effect = 1) and RMSE of the fittend RF Each point is a simulated scenario with a different set of hyper-parameters. Scenarios with 1,00 and 1,000 observations and with 100 or 10 features (P) were simulated. In each scenario, only two features (X~1~ and X~5~) had an effect (effect size = 1) on the response variable Y. Hyper-parameters (max\_depth, subsample, lambda, and max\_tree) were randomly sampled.

| **mtry** | **min.node.size** | **max.depth** | **regularization.factor** | **bias\_1** | **Bias X1** | **Bias X2** | **Bias X3** | **weighted\_bias** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 5 | 9 | 38 | 0.30 | 0.21 | -0.05 | -0.23 | 0.37 | 0.68 |
| 4 | 7 | 43 | 0.04 | 0.23 | -0.09 | -0.26 | 0.38 | 0.74 |
| 4 | 3 | 11 | 0.22 | 0.26 | -0.05 | -0.24 | 0.38 | 0.75 |
| 5 | 2 | 48 | 0.25 | 0.27 | -0.02 | -0.24 | 0.37 | 0.76 |
| 3 | 7 | 25 | 0.01 | 0.27 | -0.09 | -0.25 | 0.36 | 0.77 |
| 5 | 2 | 50 | 0.31 | 0.26 | -0.07 | -0.26 | 0.36 | 0.78 |

## 2 Proof of concept - Additional results

All models (Fig. 2) showed small variances (Fig. S7)

|  |
| --- |
| **Figure S** 7: Variances of effect estimates for different ML algorithms in three different simulated causal simulations (a, b, and c). Sample sizes are so large that stochastic effects can be excluded. Effects of the ML models were inferred using marginal conditional effects. Row a) shows results for simulations with uncorrelated features with effect sizes (x1: 1.0, x2: 0.5, x3: 1.0). Row b) shows the results for simulations with x1 and x2 being strongly correlated (Pearson correlation factor = 0.9) but only x1 has an effect on y (mediator) and row c) shows the results for x1 and x2 being strongly correlated (Pearson correlation factor = 0.9 with x1 and x2 having effects on y (confounder scenario) |

### 2.1 NN with Dropout, LASSO, and Ridge

We additionally tested NN with dropout (rate = 0.2), LASSO regression, and Ridge regression for the proof-of-concept simulations.

|  |
| --- |
| **Figure S** 8: Bias on effect estimates for additional ML algorithms (NN with Dropout, LASSO, and Ridge regression) in three different simulated causal simulations (a, b, and c).Sample sizes are so large that stochastic effects can be excluded. Effects of the ML models were inferred using marginal conditional effects. Row a) shows results for simulations with uncorrelated features with effect sizes (x1: 1.0, x2: 0.5, x3: 1.0). Row b) shows the results for simulations with x1 and x2 being strongly correlated (Pearson correlation factor = 0.9) but only x1 has an effect on y (mediator) and row c) shows the results for x1 and x2 being strongly correlated (Pearson correlation factor = 0.9 with x1 and x2 having effects on y (confounder scenario) |

All three methods showed biased estimates in the first scenario for effects without collinearity (Fig. S8). With collinearity, all three showed larger biases but Ridge regression showed the largest biases (Fig. S8).

|  |
| --- |
| **Figure S** 9: Variances of effect estimates for additional ML algorithms (NN with Dropout, LASSO, and Ridge regression) in three different simulated causal simulations (a, b, and c). Sample sizes are so large that stochastic effects can be excluded. Effects of the ML models were inferred using marginal conditional effects. Row a) shows results for simulations with uncorrelated features with effect sizes (x1: 1.0, x2: 0.5, x3: 1.0). Row b) shows the results for simulations with x1 and x2 being strongly correlated (Pearson correlation factor = 0.9) but only x1 has an effect on y (mediator) and row c) shows the results for x1 and x2 being strongly correlated (Pearson correlation factor = 0.9 with x1 and x2 having effects on y (confounder scenario) |

While estimates from NN with dropout had the smallest biases, they had the largest variances (Fig. S9), though these variances were still small (Fig. S9).

### 2.2 RMSE on holdout

In addition to bias and variance, we calculated the predictive error using the RMSE on holdout data which had the same size as the training data (N = 1000).

RF, BRT, and Dropout showed the highest RMSE in all three scenarios (Fig. S10). LASSO and elastic-net showed the smallest RMSE in all three scenarios (Fig. S10).

|  |
| --- |
| **Figure S** 10: Root mean squared error (RMSE) for different ML algorithms in three different simulated causal simulations (a, b, and c). Sample sizes are so large that stochastic effects can be excluded. 1,000 observations were used to train the models and 1,000 observations were used to evaluate the predictive performance of the models. Column ‘effects’ shows results for simulations with uncorrelated features with effect sizes (x1: 1.0, x2: 0.5, x3: 1.0). Column ‘collinearity\_0.90’ shows the results for simulations with x1 and x2 being strongly correlated (Pearson correlation factor = 0.9) but only x1 has an effect on y (mediator) and column ‘confounder\_unequal’ shows the results for x1 and x2 being strongly correlated (Pearson correlation factor = 0.9 with x1 and x2 having effects on y (confounder scenario). |

## 3 Data-poor simulation

### 3.1 Without interactions

#### 3.1.1 RMSE

|  |
| --- |
| **Figure S** 11: Root mean squared error (RMSE) of models for data-poor simulations. N = 100 (low), 600 (med), 2,000 (large) was simulated with 100 features. All features were weakly correlated. Two features had main effects (effect sizes = 1, X1 and X5) on the response and the other 95 features had no effects. Models (LM, RF, BRT, NN, Elastic-net) were fitted on the simulated data and holdout data of the same size as the training data was used to estimate predictive error (RMSE). Simulation and model fitting were repeated 100 times. |

### 3.2 With interactions

#### 3.2.1 With Collinearity

##### 3.2.1.1 Bias

|  |
| --- |
| **Figure S** 12: Bias and variance of estimated effects in a data-poor situation. N = 100, 600, 2,000 was simulated with 29 features. All features were weakly correlated. Two features had main effects (effect sizes = 1) on the response and 4 features had two interaction effects (effect sizes = 1) on the response. Models (LM, RF, BRT, NN, Elastic-net) were fitted on the simulated data and effect estimates were approximated using MCEs. LM and elastic-net received all possible two-way interactions (406) as features. Simulation and model fitting were repeated 100 times. For all effects (Main = 1, Inter = 1) and non-effects (Main = 0, Inter = 0), bias and variance of the effect estimates were calculated. |

##### 3.2.1.2 RMSE

|  |
| --- |
| **Figure S** 13: Root mean squared error (RMSE) of models for data-poor simulations. N = 100 (low), 600 (med), 2,000 (large) was simulated with 29 features. All features were weakly correlated. Two features had main effects (effect sizes = 1, X1 and X5) and there were two interaction effects (X1:X2 and X3:X4) on the response and the other 95 features had no effects. Models (LM, RF, BRT, NN, Elastic-net) were fitted on the simulated data and holdout data of the same size as the training data was used to estimate predictive error (RMSE). LM and elastic-net received all possible two-way interactions (406) as features. Simulation and model fitting were repeated 100 times. |

#### 3.2.2 Without collinearity

##### 3.2.2.1 Bias

|  |
| --- |
| **Figure S** 14: Bias and variance of estimated effects in a data-poor situation. N = 100, 600, 2,000 was simulated with 29 features. All features were uncorrelated. Two features had main effects (effect sizes = 1) on the response and 4 features had two interaction effects (effect sizes = 1) on the response. Models (LM, RF, BRT, NN, Elastic-net) were fitted on the simulated data and effect estimates were approximated using MCEs. LM and elastic-net received all possible two-way interactions (406) as features. Simulation and model fitting were repeated 100 times. For all effects (Main = 1, Inter = 1) and non-effects (Main = 0, Inter = 0), bias and variance of the effect estimates were calculated. |

##### 3.2.2.2 RMSE

|  |
| --- |
| **Figure S** 15: Root mean squared error (RMSE) of models for data-poor simulations. N = 100 (low), 600 (med), 2,000 (large) was simulated with 29 features. All features were uncorrelated. Two features had main effects (effect sizes = 1, X1 and X5) and there were two interaction effects (X1:X2 and X3:X4) on the response and the other 95 features had no effects. Models (LM, RF, BRT, NN, Elastic-net) were fitted on the simulated data and holdout data of the same size as the training data was used to estimate predictive error (RMSE). LM and elastic-net received all possible two-way interactions (406) as features. Simulation and model fitting were repeated 100 times. |

## 4 Weighted MCE

|  |
| --- |
| **Figure S** 16: Comparison of weighted versus unweighted MCE. |