Modified Causal Forest: Estimation, Sensitivity Analysis

Section 1: General information

Welcome to the mcf estimation and optimal policy package.
This report provides you with a summary of specifications and results. More detailed information can be found in the respective output files. Figures and data (in csv-format, partly to recreate the figures on your own) are provided in the output path as well.

Output information for MCF ESTIMATION
Path for all outputs:
Q:\SEW\Projekte\MLechner\Projekte und Angebote\Unicef\Kasachstan\Workshops\Astana\Wednesday_examples070/example/output
Detailed text output:
Q:\SEW\Projekte\MLechner\Projekte und Angebote\Unicef\Kasachstan\Workshops\Astana\Wednesday_examples070/example/output/mcf.py.0.7.0.txt
Summary text output:
Q:\SEW\Projekte\MLechner\Projekte und Angebote\Unicef\Kasachstan\Workshops\Astana\Wednesday_examples070/example/output/mcf.py.0.7.0_Summary.txt

Output information for SENSITIVITY ANALYSIS
Path for all outputs:
Q:\SEW\Projekte\MLechner\Projekte und Angebote\Unicef\Kasachstan\Workshops\Astana\Wednesday_examples070/example/outputsensitivity
Detailed text output:
Q:\SEW\Projekte\MLechner\Projekte und Angebote\Unicef\Kasachstan\Workshops\Astana\Wednesday_examples070/example/outputsensitivity/mcf.py.0.7.0.txt
Summary text output:
Q:\SEW\Projekte\MLechner\Projekte und Angebote\Unicef\Kasachstan\Workshops\Astana\Wednesday_examples070/example/outputsensitivity/mcf.py.0.7.0_Summary.txt

BACKGROUND

ESTIMATION OF EFFECTS
The MCF is a comprehensive causal machine learning estimator for the estimation of treatment effects at various levels of granularity, from the average effect at the population level to very fine grained effects at the (almost) individual level. Since effects at the highler levels are obtained from lower level effects, all effects are internally consistent. Recently, the basic package has been appended for new average effects as well as for an optimal policy module. The basis of the MCF estimator is the the causal forest suggested by Wager and Athey (2018). Their estimator has been changed in several dimensions which are described in Lechner (2018). The main changes relate to the objective function as well as to the aggreation of effects. Lechner and Mareckova (2024) provide the asymptotic guarantees for the MCF and compare the MCF, using a large simulation study, to competing approaches like the Generalized Random Forest (GRF, Athey, Tibshirani, Wager, 2019) and Double Machine Learning (DML, Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey,

Robins, 2018, Knaus, 2022). In this comparison the MCF faired very well, in particular, but not only, for heterogeneity estimation. Some operational issues of the MCF are discussed in Bodory, Busshof, Lechner (2022). There are several empirical studies using the MCF, like Cockx, Lechner, Boolens (2023), for example.

References

- Athey, S., J. Tibshirani, S. Wager (2019): Generalized Random Forests, The Annals of Statistics, 47, 1148-1178.
- Athey, S., S. Wager (2019): Estimating Treatment Effects with Causal Forests: An Application, Observational Studies, 5, 21-35.
- Bodory, H., H. Busshoff, M. Lechner (2023): High Resolution Treatment Effects Estimation: Uncovering Effect Heterogeneities with the Modified Causal Forest, Entropy, 24, 1039.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, J. Robins (2018): Double/debiased machine learning for treatment and structural parameters, Econometrics Journal, 21, C1-C68.
- Cockx, B., M. Lechner, J. Bollens (2023): Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium, Labour Economics, 80, Article 102306.
- Knaus, M. (2022): Double Machine Learning based Program Evaluation under Unconfoundedness, Econometrics Journal.
- Lechner, M. (2018): Modified Causal Forests for Estimating Heterogeneous Causal Effects, arXiv.
- Lechner, M. (2023): Causal Machine Learning and its Use for Public Policy, Swiss Journal of Economics & Statistics, 159:8.
- Lechner, M., J. Mareckova (2024): Comprehensive Causal Machine Learning, arXiv.
- Wager, S., S. Athey (2018): Estimation and Inference of Heterogeneous Treatment Effects using Random Forests, Journal of the American Statistical Association, 113:523, 1228-1242.

SENSITIVITY

Sensitivity analysis is currently experimental and not (yet) documented here.

Modified Causal Forest: Estimation, Sensitivity Analysis

Section 2: MCF estimation

METHOD
Standard MCF method used. Nearest neighbour matching performed using the Prognostic Score.
Feature selection not is used.
Local centering is used.
Common support is enforced.

VARIABLES
Outcome: outcome
Treatment: treat (with values 0 1 2)
Ordered confounders: x_cont0, x_cont1, x_cont2, x_cont3, x_cont4, x_cont5, x_cont6, x_ord0, x_ord1, x_ord2, x_ord3, x_ord4, x_ord5, x_ord6, x_cont0, x_cont1, x_ord5, x_ord6
Unordered (categorical) confounders: x_unord0, x_unord1
Continuous heterogeneity variables: x_cont0, x_cont1
Ordered heterogeneity variables (few values, continuous variables are discretized): x_ord5, x_ord6, x_cont0catv, x_cont1catv
Unordered heterogeneity variables: x_unord0, x_unord1

EFFECTS ESTIMATED
Average Treatment Effect (ATE), Group Average Treatment Effect (GATE), Individualized Average Treatment Effect (IATE)

NOTE on unordered variables:
One-hot-encoding (dummy variables) is not used as it is expected to perform poorly with trees: It may lead to splits of one category versus all other categories. Instead the approach used is analogous to the one discussed in Chapter 9.2.4 of Hastie, Tibshirani, Friedmann (2013), The Elements of Statistical Learning, 2nd edition.

Section 2.1: MCF Training

Training uses 6 CPU cores.

Section 2.1.1: Preparation of training data (mcf training)

METHOD
Variables without variation are removed.
Variables that are perfectly correlated with other variables are removed.
Dummy variables with less than 10 observations in the smaller group are removed.
Rows with any missing values for variables needed for training are removed.

Modified Causal Forest: Estimation, Sensitivity Analysis

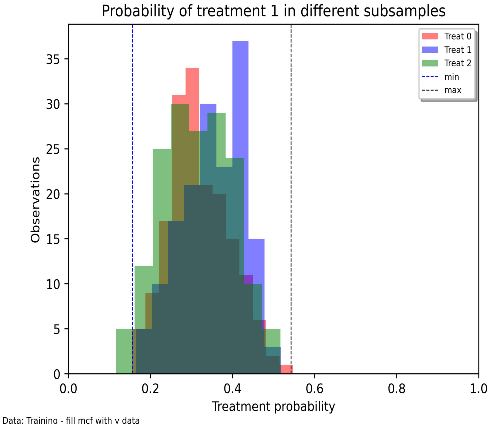
RESULTS
No relevant variables were removed.
Sample size of training data: 1000   (no observations removed).

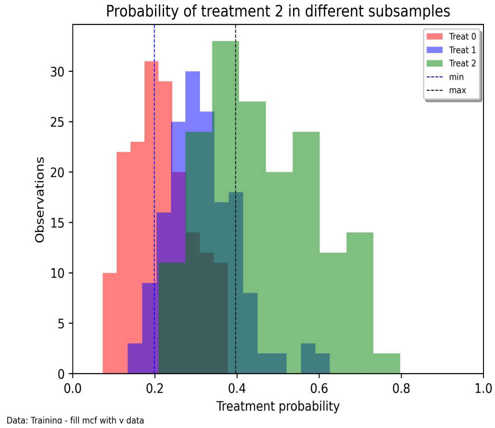
Section 2.1.2: Common support (mcf training)

METHOD
The common support analysis is based on checking the overlap in the out-of-sample predictions of the propensity scores (PS) for the different treatment arms. PSs are estimated by random forest classifiers. Overlap is operationalized by computing cut-offs probabilities of the PSs (ignoring the first treatment arm, because probabilities add to 1 over all treatment arms). These cut-offs are subsequently also applied to the data used for predicting the effects.
Overlap is determined by the min / max rule.
Cut-offs for PS are widened by 0.05.
Out-of-sample predictions are generated by 5-fold cross-validation.

RESULTS
Share of observations deleted: 44.00%
Number of observations remaining: 560
WARNING: Check output files whether the distribution of the features changed due to the deletion of part of the data.

Common support plots





Section 2.1.3: Local centering (mcf training)

METHOD
Local centering is based on training a regression to predict the outcome variable conditional on the

features (without the treatment). The regression method is selected among various versions of Random Forests, Support Vector Machines, Boosting methods, and Neural Networks of scikit-learn. The best method is selected by minimizing their out-of-sample Mean Squared Error using 5-fold cross-validation. The full set of results of the method selection step are contained in Q:\SEW\Projekte\MLechner\Projekte und Angebote\Unicef\Kasachstan\Workshops\Astana\Wednesday_examples070/example/output/mcf.py.0.7.0.txt.
The respective out-of-sample predictions are subtracted from the observed outcome in the training data used to build the forest. These out-of-sample predictions are generated by 5-fold cross-validation.

RESULTS
Out-of-sample fit for Random Forest of Ey|x (R2) for outcome: 12.39%

Section 2.1.4: Forest

METHOD and tuning parameters
Method used for forest building is MSE & MCE Penalty mse_d.
The causal forest consists of 1000 trees.
The minimum leaf size is 5.
The number of variables considered for each split is 8
The share of data used in the subsamples for forest building is 67%.
The share of the data used in the subsamples for forest evaluation (outcomes) is 100%.
Alpha regularity is set to 10%.

outcome_lc is the outcome variable used for splitting (locally centered).
The features used for splitting are x_cont0 x_cont1 x_cont2 x_cont3 x_cont4 x_cont5 x_cont6 x_ord0 x_ord1 x_ord2 x_ord3 x_ord4 x_ord5 x_ord6 x_unord0 x_unord1 x_unord2 x_unord3 x_unord4 x_unord5 x_cont0catv x_cont1catv.

RESULTS
Each tree has on average 27.30 leaves.
Each leaf contains on average 6.9 observations. The median # of observations per leaf is 6.
The smallest leaves have 5 observations.
The largest leaf has 41 observations.
25.33% of the leaves were merged when populating the forest with outcomes from the honesty sample.

Modified Causal Forest: Estimation, Sensitivity Analysis

Section 2: MCF estimation

Section 2.2: MCF Prediction of Effects

Training uses 6 CPU cores.

Section 2.2.1: Common support (mcf prediction)

Share of observations deleted: 42.40%
Number of observations remaining: 576
WARNING: Check output files whether the distribution of the features changed due to the deletion of part of the data.

Section 2.2.2: Results

GENERAL REMARKS
The following results for the different parameters are all based on the same causal forests (CF). The combination of the CF with the potentially new data provided leads to a weight matrix. This matrix may be large requiring some computational optimisations, such as processing it in batches and saving it in sparse matrix format. One advantage of this approach is that aggregated effects (ATE, GATE) can be computed by aggregation of the weights used for the IATE. Thus a high internal consistency is preserved in the sense that IATE will aggregate to GATEs, which in turn will aggregate to ATEs.

ESTIMATION
Weights of individual training observations are truncated at 5.00%.Aggregation of IATE to ATEs and GATEs may not be exact due to weight truncation.

INFERENCE
Inference is based on using the weight matrix. Nonparametric regressions are based on k-nearest neighbours.

NOTE
Treatment effects for specific treatment groups (so-called treatment effects on the treated or non-treated) can only be provided if the data provided for prediction contains a treatment variable (which is not required for the other effects).

Section 2.2.2.1: ATE

Modified Causal Forest: Estimation, Sensitivity Analysis

RESULT

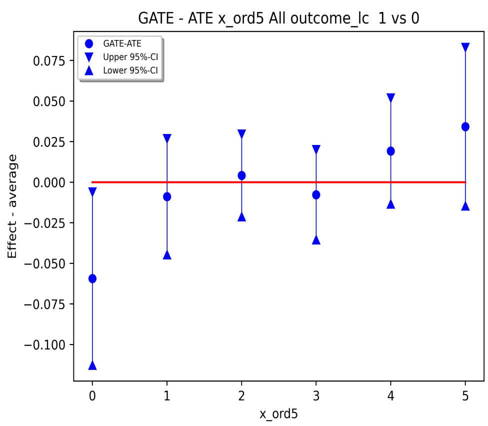
ATE for outcome

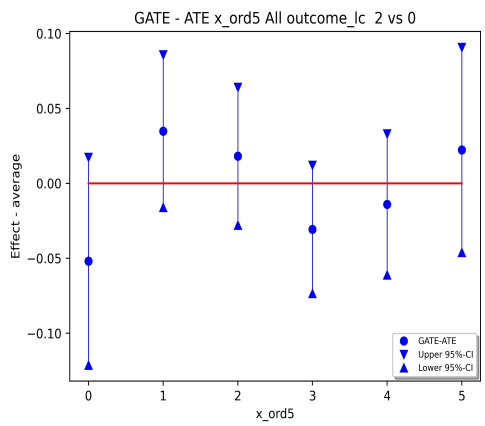
Comparison	*Effect*	*SE*	*t-value*	*p-value (%)*	*Sig.*
1 vs 0	0.807	0.189	4.27	0.0	****
2 vs 0	0.682	0.268	2.54	1.11	**
2 vs 1	-0.124	0.251	0.49	62.41	

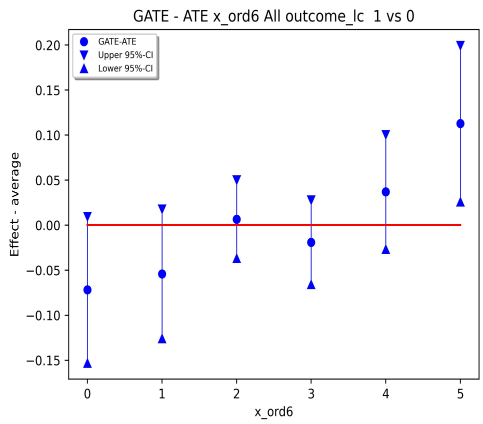
Note: *, **, ***, **** denote significance at the 10%, 5%, 1%, 0.1% level. The results for the potential outcomes can be found in the output files.

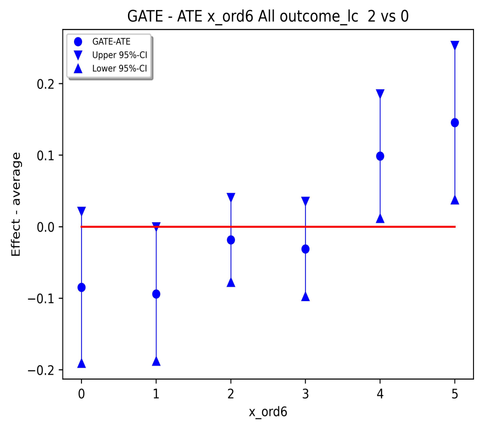
Section 2.2.2.2: GATE

GATE

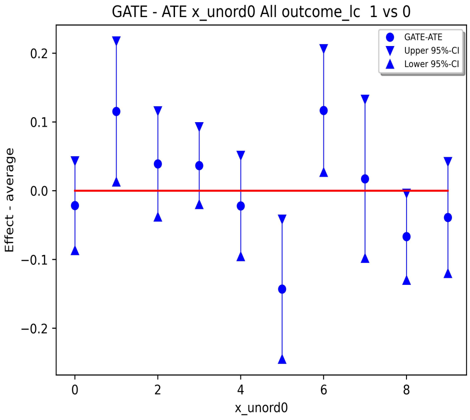


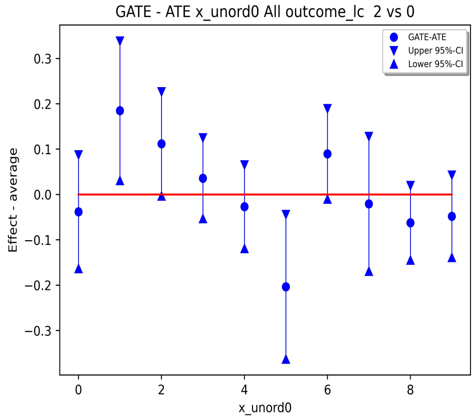


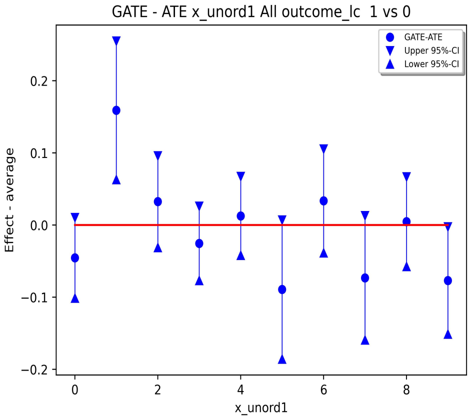


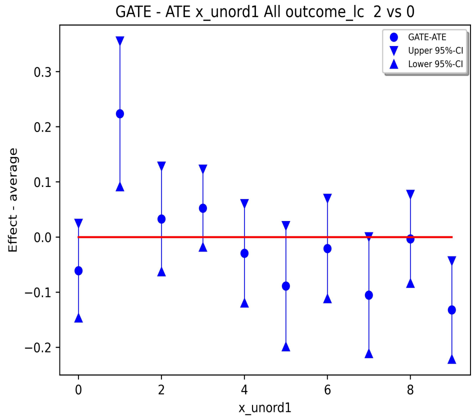


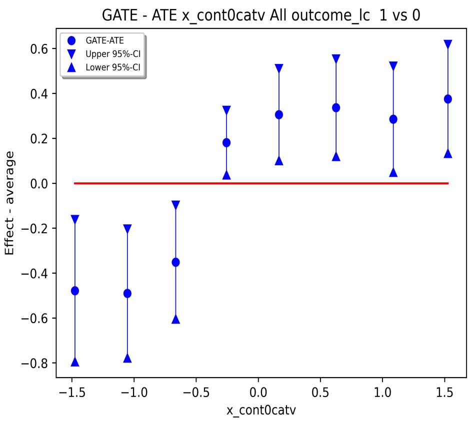
Modified Causal Forest: Estimation, Sensitivity Analysis

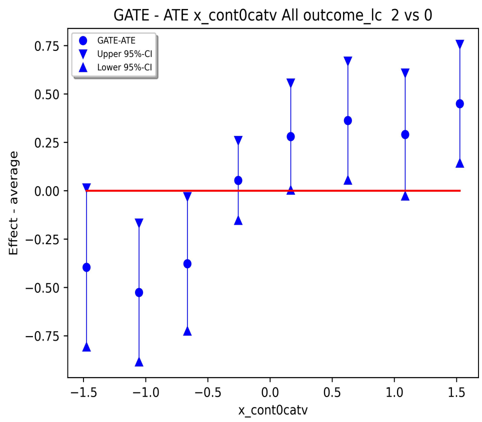


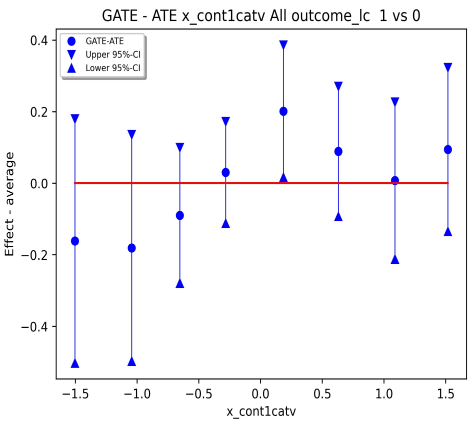


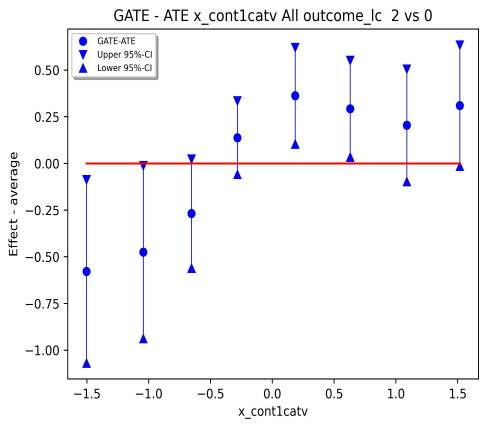


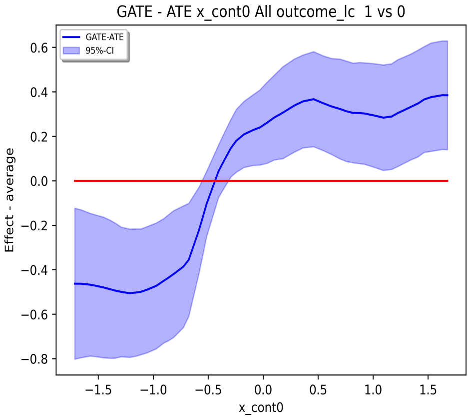


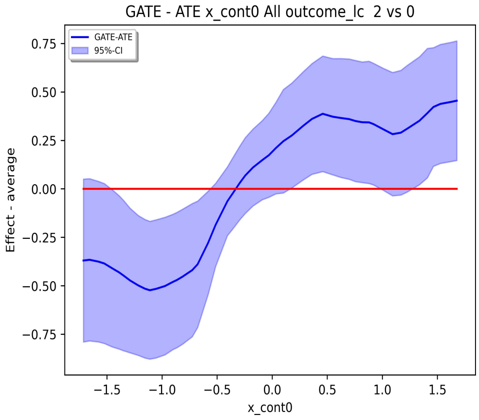


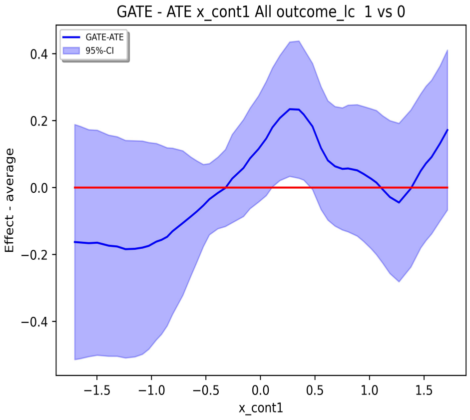


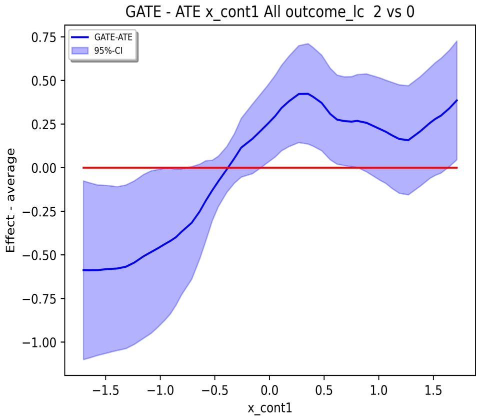












Note: Detailed tables and figures for additional effects are contained in the output files and output directories.

Section 2.2.2.3: IATE

This section contains parts of the descriptive analysis of the IATEs. Use the analyse method to obtain more descriptives of the IATEs, like their distribution, and their relations to the features.

RESULTS

Outcome variable: outcome

Comparison	Mean	Median	Std	Effect > 0
1 vs 0	0.80384	0.80201	0.41800	100.00%
2 vs 0	0.65816	0.45853	0.51262	93.92%
2 vs 1	-0.14569	-0.09673	0.23607	32.29%

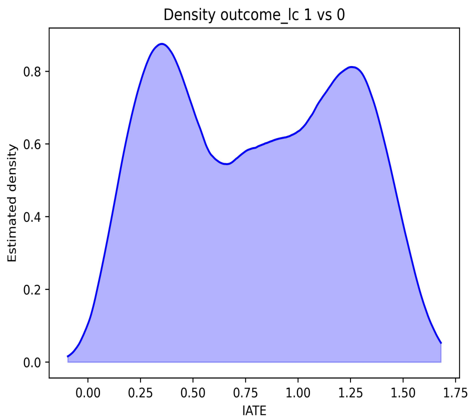
Modified Causal Forest: Estimation, Sensitivity Analysis

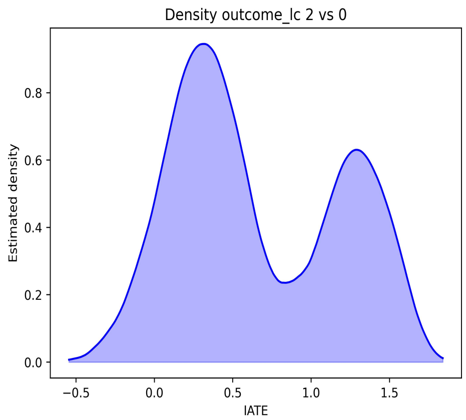
Section 2: MCF estimation

Section 2.3: Analysis of Estimated IATEs

This section contains parts of the descriptive analysis of the IATEs. More detailed tables and figures are contained in the output files and directories. These additional results include variable importance plots from a regression random forest as well as and linear regression results. Both estimators use the estimated IATEs as dependent variables and the confounders and heterogeneity variables as features.

IATEs





K-MEANS CLUSTERING
The sample is divided using k-means clustering based on the estimated IATEs. The number of clusters is determined by maximizing the Average Silhouette Score on a given grid. The following table shows the means of the IATEs, potential outcomes, and the features in these clusters, respectively.

Number of observations in the clusters

Cluster 0: 215
Cluster 1: 152
Cluster 2: 209

Modified Causal Forest: Estimation, Sensitivity Analysis

IATE

Comparison	*0*	*1*	*2*
outcome_lc1vs0_iate	0.34	0.82	1.27
outcome_lc2vs0_iate	0.21	0.43	1.28

Note: Mean of variable in cluster.

Potential Outcomes

Comparison	*0*	*1*	*2*
outcome_lc0_un_lc_pot	-0.17	-0.14	0.32
outcome_lc1_un_lc_pot	0.17	0.68	1.59
outcome_lc2_un_lc_pot	0.04	0.3	1.6

Note: Mean of variable in cluster.

Features

Comparison	*0*	*1*	*2*
x_unord0	4.08	4.47	3.79
x_unord1	4.29	3.78	3.79
x_unord2	3.66	3.64	4.11
x_unord3	2.88	3.11	3.66
x_unord4	4.34	4.16	4.0
x_unord5	3.86	3.31	3.97
x_cont0	-1.13	0.41	0.65
x_cont1	0.02	-0.81	0.62
x_cont2	0.03	-0.03	-0.04
x_cont3	-0.03	0.13	-0.09
x_cont4	0.06	0.15	-0.01
x_cont5	0.14	-0.01	0.05
x_cont6	0.05	-0.03	0.12
x_ord0	0.52	0.43	0.5
x_ord1	0.44	0.59	0.46
x_ord2	0.5	0.51	0.56
x_ord3	0.5	0.46	0.46
x_ord4	2.45	2.6	2.49
x_ord5	2.46	2.64	2.55
x_ord6	2.33	2.66	2.67
x_unord0_0	0.11	0.08	0.1
x_unord0_1	0.05	0.04	0.09
x_unord0_2	0.07	0.06	0.09
x_unord0_3	0.19	0.28	0.24
x_unord0_4	0.27	0.14	0.22
x_unord0_5	0.08	0.09	0.03
x_unord0_6	0.03	0.05	0.06
x_unord0_7	0.02	0.05	0.03
x_unord0_8	0.08	0.07	0.06
x_unord0_9	0.09	0.13	0.08

x_unord1_0	0.11	0.1	0.08
x_unord1_1	0.05	0.13	0.14
x_unord1_2	0.07	0.07	0.09
x_unord1_3	0.22	0.16	0.21
x_unord1_4	0.21	0.28	0.21
x_unord1_5	0.04	0.03	0.03
x_unord1_6	0.05	0.07	0.06
x_unord1_7	0.07	0.04	0.04
x_unord1_8	0.06	0.04	0.06
x_unord1_9	0.13	0.09	0.08
x_unord2_0	0.13	0.12	0.08
x_unord2_1	0.09	0.11	0.11
x_unord2_2	0.09	0.14	0.1
x_unord2_3	0.22	0.19	0.16
x_unord2_4	0.21	0.16	0.22
x_unord2_5	0.05	0.07	0.04
x_unord2_6	0.03	0.05	0.06
x_unord2_7	0.02	0.03	0.07
x_unord2_8	0.06	0.05	0.06
x_unord2_9	0.09	0.09	0.11
x_unord3_0	0.09	0.04	0.04
x_unord3_1	0.15	0.13	0.08
x_unord3_2	0.09	0.14	0.12
x_unord3_3	0.34	0.27	0.25
x_unord3_4	0.23	0.31	0.25
x_unord3_5	0.04	0.04	0.08
x_unord3_6	0.02	0.03	0.12
x_unord3_7	0.04	0.03	0.03
x_unord3_8	0.0	0.01	0.03
x_unord4_0	0.1	0.13	0.07
x_unord4_1	0.08	0.07	0.1
x_unord4_2	0.09	0.08	0.11
x_unord4_3	0.18	0.22	0.23
x_unord4_4	0.18	0.14	0.19
x_unord4_5	0.03	0.05	0.05
x_unord4_6	0.07	0.05	0.06
x_unord4_7	0.06	0.06	0.03
x_unord4_8	0.07	0.03	0.06
x_unord4_9	0.14	0.16	0.11
x_unord5_0	0.08	0.16	0.1
x_unord5_1	0.13	0.09	0.09
x_unord5_2	0.08	0.14	0.1
x_unord5_3	0.23	0.22	0.2
x_unord5_4	0.19	0.2	0.19
x_unord5_5	0.03	0.03	0.09
x_unord5_6	0.06	0.03	0.06
x_unord5_7	0.04	0.03	0.03
x_unord5_8	0.05	0.01	0.06
x_unord5_9	0.1	0.1	0.1

Note: Mean of variable in cluster. Categorical variables are recoded to indicator (dummy) variables.

Modified Causal Forest: Estimation, Sensitivity Analysis

Section 3: MCF Sensitivity Analysis

METHOD
This sensitivity analysis is based on a placebo-like experiment:
(1) Using a random forest classifier, probabilities into the different treatments are estimated. Predictions are based on 5-fold cross-fitting.
(2) All observations that do not belong to the largest treatment are deleted.
(3) The conditional treatment probabilities are used to simulate treatments on the remaining observations. These treatments should feature the same selectivity as the original data but with a zero treatment effect.
(4) Training is performed on this reduced data and the usual effects are estimated. Ideally, they are close to zero.
(5) If the results dictionary of the prediction method is passed to the sensitivity method, and if it includes estimated IATEs, the same data as in the estimation is used. In this case the estimated IATEs are compared to the placebo IATEs. This indicates in which regions of the IATEs violations may take place.

APPLICATION
The following scenarions are investigated: basic
Simulations are repeated 2 times and results are averaged to reduce simulation noise.
Path for all output files:
Q:\SEW\Projekte\MLechner\Projekte und Angebote\Unicef\Kasachstan\Workshops\Astana\Wednesday_examples070/example/outputsensitivity

IATEs are investigated. If estimated IATEs are available, then plots comparing placebo IATEs with estimated IATEs are presented below. Further statistics are contained in the output files.

IMPORTANT REMARK
This sensitivity analysis is experimental, rudimentary, and not yet fully tested (although it appears to work fine in tests so far). In the future, it will be expanded and improved.

RESULTS: ATE

- -
ATEs (only if p-value < 10%)
- -
Scenario: basic ATE (sensitivity) Outcome variable: outcome_lc   Reference population: All
- -
Comparison   Estimate   Standard error t-value   p-value
- -
- -

RESULTS: Plots of estimated and placebo IATEs

Modified Causal Forest: Estimation, Sensitivity Analysis

Placebo and estimated IATEs