Orthogonal Random Forest for Causal Inference

Miruna Oprescu — Microsoft Research, Cambridge, MA

Vasilis Syrgkanis — Microsoft Research, Cambridge, MA

Zhiwei Steven Wu — University of Minnesota, Twin Cities, MN

github.com/Microsoft/EconML/tree/master/prototypes/orthogonal_forests

ALICE — Automated Learning and Intelligence for Causation and Economics

Motivation

Dynamic Pricing — Clinical Trials — Targeted Advertising

Heterogeneous Treatment Effect Applications

Related Work

- Generalized Random Forest (GRF) [1] – a flexible method for estimating θ_0 in the absence of f_0, g_0 and high-dimensional W.
- Orthogonalization[2] – a technique that removes the confounding effect of W via two-stage estimation for $\theta_0(x) = \theta_0 = const.$ only.

Our Contribution

- The Orthogonal Random Forest (ORF), an algorithm that combines generalized random forests and orthogonalization in a non-trivial way to leverage both the flexibility of the random forest framework and the robustness of the double ML technique.
- New consistency results in the partially linear regression model with non standard nuisance functions:

$$E[Y_i|x_i, W_i] = \langle W_i, \theta_0(x_i)\beta_0 + \gamma_0\rangle, \qquad E[T_i|W_i] = \langle W_i, \gamma_0\rangle$$

where β_0 and γ_0 are k-sparse.

Formal Model

- $\theta(x)$ is the solution to the conditional moment equation:

$$E\big[\psi(Z;\theta, h_0(x,W))\big|X = x\big] = 0$$

ψ – a score function, h_0 – unknown nuisance function. We wish to estimate $\theta(x)$ non-parametrically, for potentially high-dimensional W.

- We require that ψ is locally orthogonal w.r.t h_0:

$$E\big[\nabla_h\psi(Z;\theta, h_0(x,W))\big(\hat{h}(x,W) - h_0(x,W)\big)\,\big|\,x\big] = 0$$

We can write down an orthogonal ψ for many applications, including *quantile regression, instrumental variable regression, continuous and discrete treatment effects*.

- For heterogenous treatment effects, take $Z = (T_i, Y_i, W_i, x_i)$ and:

$$Y_i = \underbrace{\theta_0(x_i)}_{\text{treatment effect}}\, T_i + \underbrace{f_0(x_i, W_i)}_{\text{unknown}} + \epsilon_i$$

$$T_i = \underbrace{g_0(x_i, W_i)}_{\text{unknown}} + \eta_i$$

T_i – treatment policy, Y_i – outcome of intervention, x_i – features that capture heterogeneity, W_i – high-dimensional confounders.

ORF Algorithm

- The following orthogonal moment[2] ψ for continuous treatment effects:

$$\psi(Z;\theta,h(x,w)) = \{Y - \langle x_\parallel W, q\rangle - \theta(T - \langle x_\parallel W,\gamma\rangle)\}(T - \langle x_\parallel W,\gamma\rangle)$$

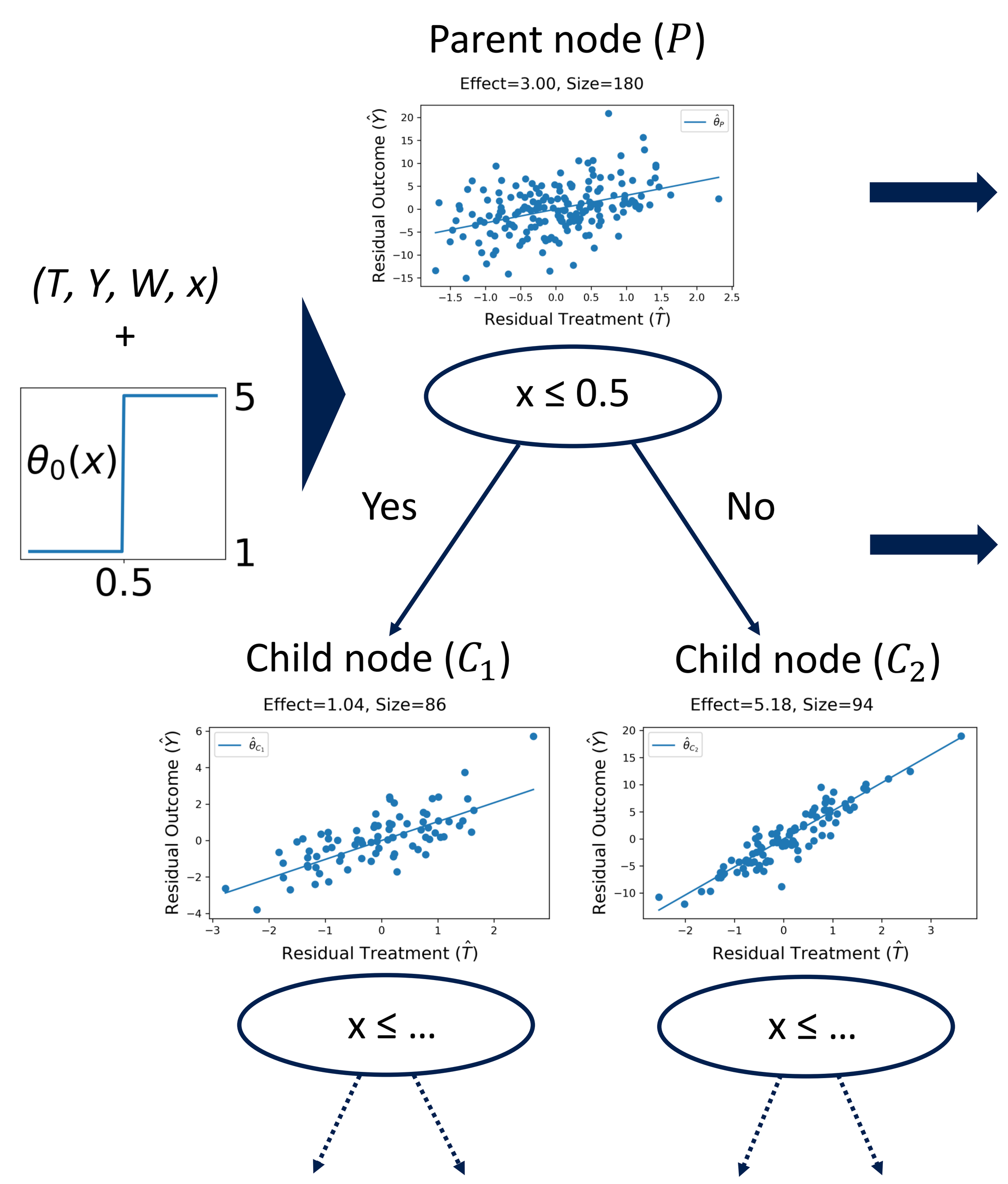
- For discrete treatments, we can employ *a doubly robust*[3] ψ.

The ORF algorithm is the following two-step procedure for estimating $\theta(x)$:

1. Forest Learner

- Partition the dataset D into D_1 and D_2
- Train a forest learner on each partition

We build a forest of "orthogonal" trees as follows:



a) Use Double ML[2] to remove confoundedness:

$$\tilde{T}_i = T_i - \langle x_\parallel W_i, \hat{\gamma}_P\rangle$$
$$\tilde{Y}_i = Y_i - \langle x_\parallel W_i, \hat{q}_P\rangle$$
$$\hat{\theta}_P = \operatorname*{argmin}_\theta \sum_i \frac{1}{2}\big(\langle\tilde{T}_i,\theta\rangle - \tilde{Y}_i\big)^2$$

b) Perform the split that maximizes the heterogeneity score:

$$\Delta(C_1, C_2) = \sum_{j=1}^{2}\frac{1}{|C_j|}\big(\hat{\theta}_{C_j} - \hat{\theta}_P\big)^2$$

where $\hat{\theta}_{C_j}$'s are obtained from applying double ML on the child nodes.

Fig. 1: Example construction of orthogonal tree

2. Kernel Two-Stage Estimation

We estimate the treatment effect $\hat{\theta}(x)$ using a two-stage procedure:

a) Weighted Lassos $T, Y \sim x_\parallel W_i$ with weights :

$$\omega_i = K_{D_1}(x_i, x) = \frac{1}{\#\,trees}\sum_{trees\,\epsilon\,D_1}\frac{1(x \text{ and } x_i \text{ in same leaf})}{\text{leaf size}}$$

b) Weighted regression on residuals from a):

$$\hat{\theta} = \operatorname*{argmin}_\theta \textstyle\sum_{i=1}^n a_i\big(\theta\tilde{T}_i - \tilde{Y}_i\big)^2$$

$$a_i = K_{D_2}(x_i, x), \tilde{T}_i = T_i - \langle x_\parallel W_i, \hat{\gamma}\rangle, \tilde{Y}_i = Y_i - \langle x_\parallel W_i, \hat{q}\rangle$$

where $\hat{\gamma}, \hat{q}$ are the results from the Lassos in a).

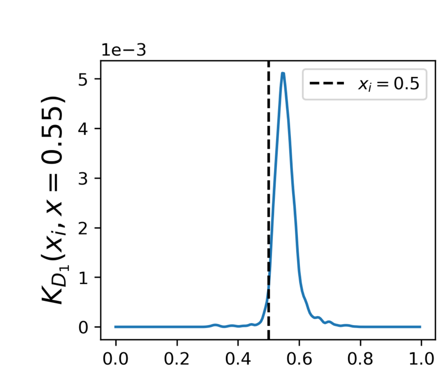


Fig. 2: Kernel for $x = 0.55$ for θ_0 discontinuous at 0.5.

Selected References

1. Athey, S., Tibshirani, J., & Wager, S. *Generalized Random Forests*. ArXiv e-prints. Oct. 2017.
2. Chernozhukov, V., Goldman, et al. *Orthogonal Machine Learning for Demand Estimation: High Dimensional Causal Inference in Dynamic Panels*. arXiv preprint. December 2017.
3. Robins, J.M. and Rotnitzky, A. Semiparametric efficiency in multivariate regression models with missing data. Journal of the American Statistical Association, 1995.

Real-world Application

- We wish to estimate the effect of orange juice price on demand
- The dataset contains several covariates W, but we want to learn the elasticity of demand as a function of income alone (x).
- The ORF results unveil the natural phenomenon that lower income consumers are more price-sensitive.

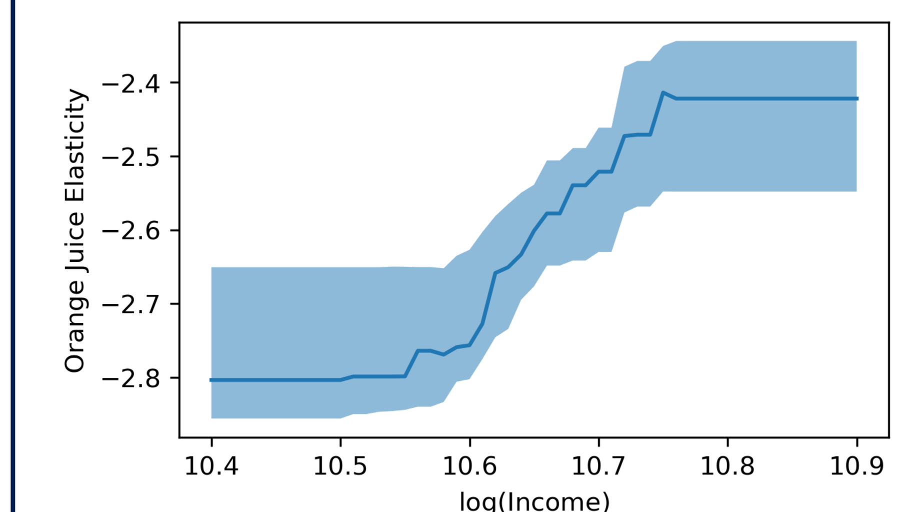


Fig. 3: ORF estimates of orange juice elasticity by income from a high-dimensional dataset. The shaded region depicts the 1%-99% confidence interval obtained via bootstrap.

Monte Carlo Experiments

We compare the performance of the ORF with other methods in the literature:

- GRF on residualized treatments and outcomes
- Variants of double ML – an adaptation of double ML that allows for parametric heterogeneity

Fig. 4 and 5 show the results for continuous treatments and a piecewise linear treatment effect.

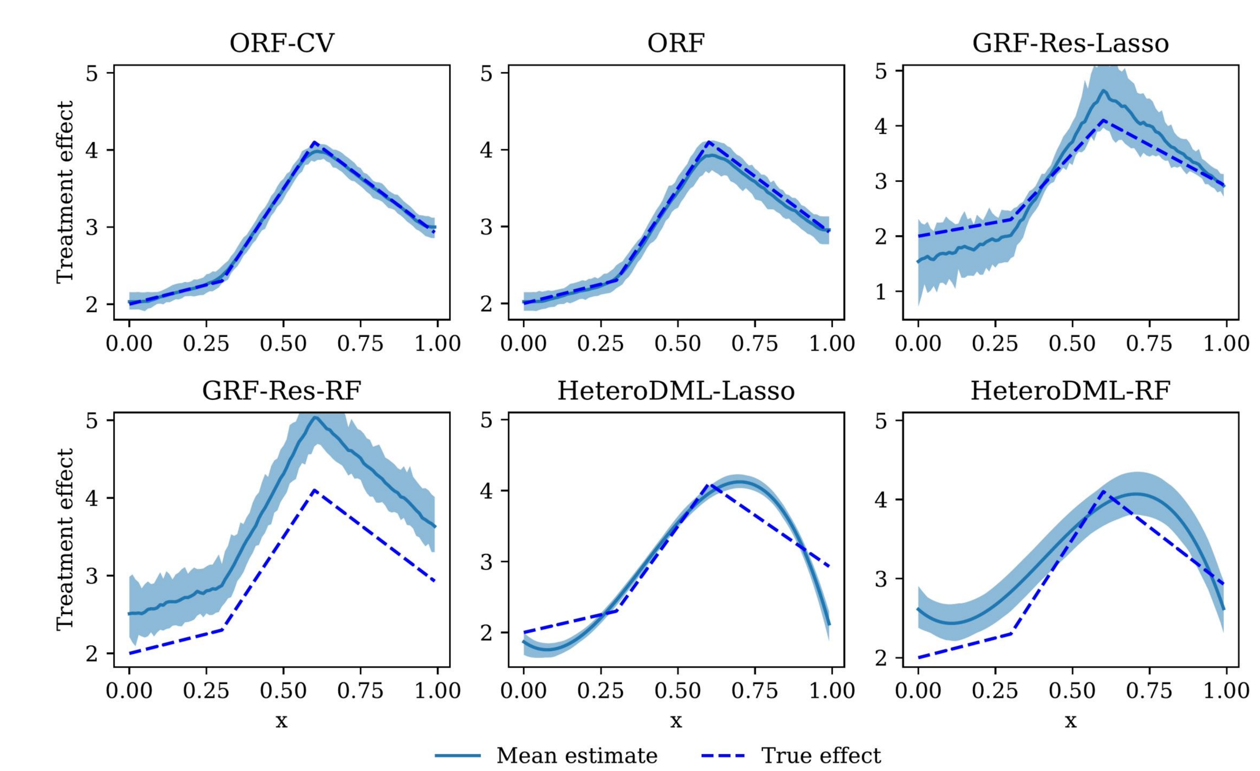


Fig. 4: Monte Carlo treatment effect estimations. The shaded regions depict the mean and the 5%-95% interval for the 100 experiments.

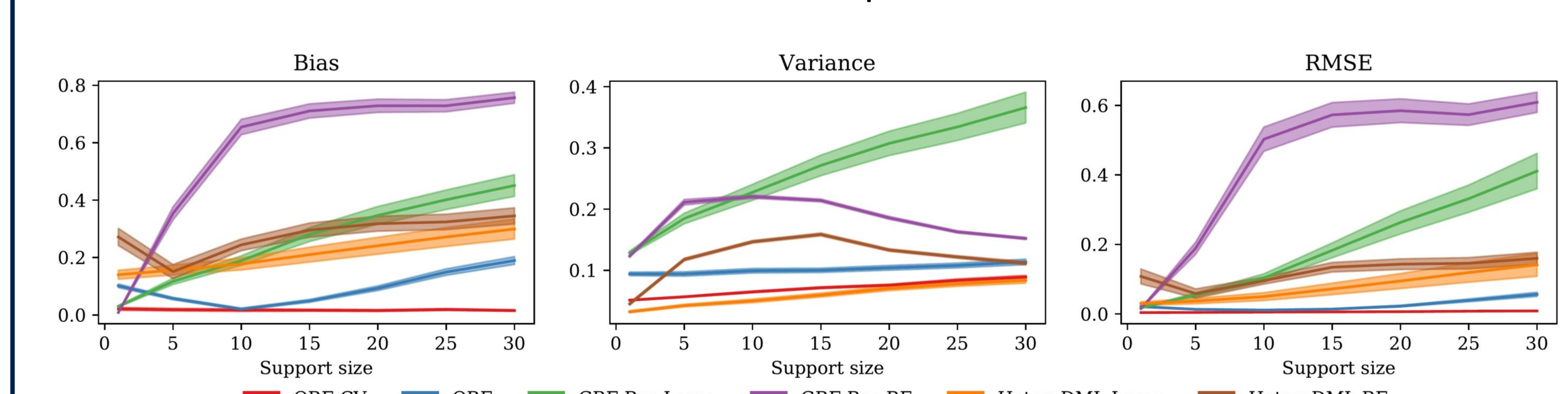


Fig. 5: Mean and standard deviation (scaled by a factor of 3 for clarity) of the bias, variance and RMSE as a function of support size k.