CVtreeMLE: Efficient Estimation of Mixed Exposures using Data Adaptive Decision Trees and Cross-Validated Targeted Maximum Likelihood Estimation in R

David McCoy[1], **Alan Hubbard**[2], **and Mark Van der Laan**[2]

1 Division of Environmental Health Sciences, University of California, Berkeley **2** Department of Biostatistics, University of California, Berkeley

Summary

Statistical causal inference of mixed exposures has been limited by reliance on parametric models and, in most cases, by researchers considering only one exposure at a time, usually estimated as a beta coefficient in a regression model (glm). This independent assessment of exposures poorly estimates the joint impact of a collection of exposures in a realistic exposure setting. Non-parametric methods such as decision trees are a useful tool to evaluate combined exposures by finding partitions in the mixture space that best explain the variance in an outcome. However, current methods using decision trees to assess statistical inference for interactions are biased and are prone to overfitting by using the full data to both identify nodes in the tree and make statistical inference given these nodes. The `CVtreeMLE` R package provides researchers in (bio)statistics, epidemiology, and environmental health sciences with access to state-of-the-art statistical methodology for evaluating the causal effects of a mixed exposure using decision trees. `CVtreeMLE` builds off the general theorem of cross-validated minimum loss-based estimation (CV-TMLE) which allows for the full utilization of loss-based ensemble machine learning to obtain the initial estimators needed for our target parameter without risk of overfitting. Additionally, `CVtreeMLE` uses V-fold cross-validation and partitions the full data in each fold into a parameter-generating sample and an estimation sample. Decision trees are applied to a mixed exposure to obtain rules and estimators for our statistical target parameter using the parameter-generating sample. The rules from decision trees are then applied to the estimation sample where the statistical target parameter is estimated. `CVtreeMLE` makes possible the non-parametric estimation of the causal effects of a mixed exposure producing results that are both interpretable and asymptotically efficient.

Statement of Need

In many disciplines there is a demonstrable need to ascertain the causal effects of a mixed exposure. Advancement in the area of mixed exposures is challenged by real-world joint exposure scenarios where complex agonistic or antagonistic relationships between mixture components can occur. More flexible methods which can fit these interactions are less biased, but results are difficult to interpret which leads many researchers to use more biased methods based on glms. Current software tools for mixtures show minimal performance tests using data that reflect the complexities of real-world exposures. In many instances, new methods are not tested against ground-truth target parameter under various mixture conditions. New areas of statistical research, rooted in non/semi-parametric efficiency theory for statistical functionals, allow for robust estimation of data-adaptive parameters. That is, it is possible to use the data to both define and estimate a target parameter. This is important in mixtures when the most

41 important set of variables and levels in these variables is unknown and therefore, statistical
42 inference given exposure to these variables are unknown. Thus, the development of asymptot-
43 ically linear estimators for data-adaptive parameters are critical for the field of mixed exposure
44 statistics. However, the development of open-source software which translates semipara-
45 metric statistical theory into well-documented functional software is a formidable challenge.
46 Such implementation requires understanding of causal inference, semiparametric statistical
47 theory, machine learning, and the intersection of these disciplines. The `CVtreeMLE` R package
48 provides researchers with an open-source tool for evaluating the causal effects of a mixed
49 exposure by treating decision trees as a data-adaptive target parameter to define exposure.
50 The `CVtreeMLE` package is well documented and includes a vignette detailing semi-parametric
51 theory for data-adaptive parameters, examples of output, results with interpretations under
52 various real-life mixture scenarios, and comparison to existing methods.

Background

54 In most research scenarios, the analyst is interested in causal inference for an **a priori** specified
55 treatment or exposure. However, in the evaluation of a mixed exposure, such as air pollution
56 or pesticides, it is not possible to estimate the expected outcome given every combination of
57 exposures due to non-identification, violations of the assumption of positivity, and inefficiency.
58 Even still, this approach lacks a pre-specified target parameter. In such a setting, it is helpful
59 to map a set of continuous mixture components into a lower dimensional exposure using
60 a pre-determined algorithm to estimate a target parameter that has easier interpretation.
61 Decision trees provide a useful solution by mapping a set of exposures into a rule which can
62 be represented as a binary vector. This binary vector indicates whether an individual has been
63 exposed to a particular rule estimated by the decision tree. Our target parameter is then
64 defined as the mean difference in counterfactual outcomes for those exposed to the mixture
65 rule compared to those unexposed, or the average treatment effect for the mixture.

`CVtreeMLE`'s Scope

67 Building on prior work related to data-adaptive parameters (**Hubbard2016?**) and CV-TMLE
68 (**Zheng2010?**), `CVtreeMLE` is a novel approach for estimating the joint impact of a mixed
69 exposure by using cross-validated targeted minimum loss-based estimation which guarantees
70 consistency, efficiency, and multiple robustness despite using highly flexible learners to estimate
71 a data-adaptive parameter. `CVtreeMLE` summarizes the effect of a joint exposure on the
72 outcome of interest by first doing an iterative backfitting procedure, similar to general additive
73 models, by fitting two discrete Super Learning (SL) algorithms, an unrestricted SL for E(Y|W)
74 and a SL restricted to only decision tree learners for E(Y|A), where A is a vector of exposures.
75 After initialization, for every iteration, each algorithm is fit with an offset using predictions
76 from the complementary algorithm; this is done until convergence, when there is no difference
77 in predictions between each model. In this way, we can data-adaptively find the best fitting
78 decision tree model which has the lowest cross-validated model error while flexibly adjusting
79 for covariates. This procedure is done to find rules for the mixture modeled collectively and
80 for each mixture marginal component individually. There are two types of results, 1. an
81 average treatment effect (ATE) comparing those exposed to a mixed exposure of potentially
82 many mixture variables to those unexposed to the mixture rules and 2. the ATE for each
83 data-adaptively identified quantile of an individual mixture component when compared to the
84 lowest identified quantile. In the marginal space, results are analogous to a dose-response
85 analysis but with data-adaptively identified quantiles. The `CVtreeMLE` software package, for
86 the R language and environment for statistical computing (R Core Team, 2020), implements
87 this methodology for deriving causal inference from ensemble decision trees.

CVtreeMLE is designed to provide analysts with both V-fold specific and pooled results for ATE causal effects of a joint exposure determined by decision trees. `CVtreeMLE` integrates with the sl3package (Coyle et al., 2020) to allow for ensemble machine learning to be leveraged in the estimation of nuisance parameters.

Availability

The `CVtreeMLE` package has been made publicly available both via GitHub and the Comprehensive R Archive Network. Use of the `CVtreeMLE` package has been extensively documented in the package's README, two vignettes, and its pkgdown documentation website.

Acknowledgments

References

Coyle, J. R., Hejazi, N. S., Malenica, I., & Sofrygin, O. (2020). *sl3: Modern pipelines for machine learning and Super Learning*. https://github.com/tlverse/sl3. https://doi.org/10.5281/zenodo.1342293

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/