The Journal of Open Source Software

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

DOI: 10.21105/joss.0XXXX

Software

- Review
- Repository
- Archive

Editor: Editor Name

Submitted: 01 January XXXX
Published: 01 January XXXX

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 generalized linear regression model (glm). This independent assessment of exposures poorly estimates the joint impact of a collection of the same 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 joint-exposure (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. Other methods have used an independent test set to derive inference which does not use the full data. 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 data-adaptively determined 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. Thus, `CVtreeMLE` allows for discovery of important mixtures of exposure *and also* provides robust statistical inference for the impact of these mixtures.

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 may be less biased, but results are typically difficult to interpret, which has lead researchers to favor more biased methods based on glm's. Current software tools for mixtures rarely report performance tests using data that reflect the complexities of real-world exposures. In many instances,

The Journal of Open Source Software

new methods are not tested against a 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 important set of variables and levels in these variables are almost always unknown. Thus, the development of asymptotically linear estimators for data-adaptive parameters are critical for the field of mixed exposure statistics. However, the development of open-source software which translates semiparametric statistical theory into well-documented functional software is a formidable challenge. Such implementation requires understanding of causal inference, semiparametric statistical theory, machine learning, and the intersection of these disciplines. The `CVtreeMLE` `R` package provides researchers with an open-source tool for evaluating the causal effects of a mixed exposure by treating decision trees as a data-adaptive target parameter to define exposure. The `CVtreeMLE` package is well documented and includes a vignette detailing semi-parametric theory for data-adaptive parameters, examples of output, results with interpretations under various real-life mixture scenarios, and comparison to existing methods.

Background

In most research scenarios, the analyst is interested in causal inference for an **a priori** specified treatment or exposure. However, in the evaluation of a mixed exposure, such as air pollution or pesticides, it is not possible to estimate the expected outcome given every combination of exposures due to high-dimensionality and sparsity. Even if this approach were possible, a target parameter that can inform public policy is still ill-defined. In such a setting, it is helpful to map a set of continuous mixture components into a lower dimensional represenation of exposure using a pre-determined algorithm to estimate a target parameter that has a meaningful interpretation. Decision trees provide a useful solution by mapping a set of exposures into a rule which can be represented as a binary vector. This binary vector indicates whether an individual has been exposed to a particular rule estimated by the decision tree. Our target parameter is then defined as the mean difference in counterfactual outcomes for those exposed to the mixture rule compared to those unexposed, or the average treatment effect for the mixture.

`CVtreeMLE`'s Scope

Building on prior work related to data-adaptive parameters (Hubbard et al., 2016) and CV-TMLE (Zheng & Laan, 2010), `CVtreeMLE` is a novel approach for estimating the joint impact of a mixed exposure by using cross-validated targeted minimum loss-based estimation which guarantees consistency, efficiency, and multiple robustness despite using highly flexible learners to estimate a data-adaptive parameter. `CVtreeMLE` summarizes the effect of a joint exposure on the outcome of interest by first doing an iterative backfitting procedure, similar to general additive models, to fit f(A), a Super Learner of decision trees, and h(W), an unrestricted Super Learner, in a semi-parametric model; $E(Y|A,W) = f(A) + h(W)$, where A is a vector of exposures and W is a vector of covariates. In this way, we can data-adaptively find the best fitting decision tree model which has the lowest cross-validated model error while flexibly adjusting for covariates. This procedure is done to find rules for the mixture modeled collectively and for each mixture component individually. There are two types of results, 1. an average treatment effect (ATE) comparing those who fall within a subspace of the joint exposure versus those in the complement of that space and 2. the ATE for each data-adaptively identified threshold of an individual mixture component when compared to the lowest identified exposure level. The `CVtreeMLE` software package, for the `R` language and environment for statistical computing (R

Core Team, 2020), implements 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 `sl3`package (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` and a vignette.

Acknowledgments

David McCoy's contributions to this work were supported in part by a grant from the National Institutes of Health: T32#####.

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

Hubbard, A. E., Kherad-Pajouh, S., & Van Der Laan, M. J. (2016). Statistical Inference for Data Adaptive Target Parameters. *International Journal of Biostatistics*, *12*(1), 3–19. https://doi.org/10.1515/ijb-2015-0013

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

Zheng, W., & Laan, M. van der. (2010). Asymptotic theory for cross-validated targeted maximum likelihood estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*, *273*. http://biostats.bepress.com/ucbbiostat/paper273/