Title
Targeted Learning for Estimating Mediation and Moderation in Toxic Mixtures

Author
McCoy, David Brenton

Publication Date
2023

Peer reviewed|Thesis/dissertation

Targeted Learning for Estimating Mediation and Moderation in Toxic Mixtures

by

David McCoy

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Environmental Health Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Martyn Smith, Chair
Professor Alan Hubbard
Professor Rosemarie de la Rosa
Professor Jay Graham

Summer 2023

Targeted Learning for Estimating Mediation and Moderation in Toxic Mixtures

Abstract

Targeted Learning for Estimating Mediation and Moderation in Toxic Mixtures

by

David McCoy

Doctor of Philosophy in Environmental Health Sciences

University of California, Berkeley

Professor Martyn Smith, Chair

Understanding how exposures from our environment, diet, and lifestyle interact with unique genetic, physiologic, and epigenetic profiles to impact health is a main objective of environmental health sciences. However, causal inference is a formidable challenge in many environmental health contexts such as mixed exposures, multiple mediating pathways, heterogeneity in exposure effects, and interactions in high-dimensional data. Existing causal inference methodologies in these settings make too many simplifying assumptions that do not represent complex real-world patterns. Commonly used statistical methods based on general linear models (GLMs) fail to untangle the exposures truly affecting health due to multi-collinearity, high-dimensional interactions, and complex joint distributions. Today's scientific endeavors in environmental health require adoption of new non-parametric methods using flexible machine learning for causal inference of mixed exposures, mixture-mediation, and heterogeneity of exposure effects. Causal inference in these arenas can answer critical questions such as: What mixture of metal exposures during pregnancy influence maternal and child health? At what levels are these impacts most severe? How do oxidative stress and inflammatory biomarkers mediate action mechanisms of these exposure interactions? How do we both identify parts of a mixture that are important and estimate the expected outcome if this part of the mixture changed? Are there certain subpopulations that are more susceptible to changes in parts of a mixture?

Unlike many setting such as medicine where the treatment/exposure is known *a priori* and propensity score-based methodologies can be deployed with relative ease, the issue with mixtures is that, not only are many measured on the continuous scale (where propensity score methodologies break down) but that there are many of these exposures. We do not have the expected outcome under every combination of multiple continuous exposures. Even if this were possible, still some interpretable representation of this gradient is necessary. As such, subspaces of the exposure or specific variable subsets of the mixture that are impactful on the outcome must be identified and are not known *a priori*. We must use the data to both identify

these mixture regions and derive estimates given exposure to this region. This requires *data adaptive target parameters*, or the mapping of a mixture into a lower dimensional exposure in one part of the data and estimation of a target parameter given this exposure is done in another part of the data. Data adaptive target parameters therefore provide a unifying framework for causal inference mixture problems, each non/semi-parametric method presented leverages data adaptive target parameters to first find areas of the mixture space that are most impactful and then estimate a target parameter given that space. This dissertation is divided into five chapters, each aiming to estimate causal inference of a mixture under the larger theory of data adaptive target parameters which also extends to decomposing effects into mediating pathways, estimates of heterogeneity, and interaction.

Statistical advancement in estimating mixtures is key to furthering the progress of environmental health science to understand the health impacts of environmental exposures. Current statistical methodology lacks the ability to realistically capture the complexity of mixed exposures in an interpretable and informative summary measure. The question then becomes what is the mapping of multiple continuous, multinomial, and/or binary exposures into an interpretable summary measure and what estimation given that summary measure are we interested in? The different statistical aims provided represent different ways of answering this question. Each can be thought of as a statistical machine where the analyst simply inputs the data for exposures, covariates and an outcome and the rest is automatic. From the data the impactful areas of the mixture are identified using the best fitting model chosen from an ensemble and a target parameter is estimated with proper estimates of variance for this mixture subregion. In this way, rather than relying on human choice of modeling which introduces bias, results are data-driven.

Chapter 1 considers the problem of both identifying exposure variables and thresholds of these variables in a mixture and estimating the expected outcome if individuals were all exposed to this exposure combination compared to if they were not. To meet this challenge, the best fitting decision tree from an ensemble is treated as a data-adaptive parameter. Using the subregions of the mixture delineated by the tree which best explains an outcome, we then develop an estimator which compares the expected outcome if all individuals were exposed to this region compared to unexposed while flexibly adjusting for covariates. We apply this novel approach to the NIEHS synthetic mixtures data which allows us to compare interactions identified and estimated in the mixture to ground-truth interactions built into the data-generating system. Furthermore, we apply our method to National Health and Nutrition Examination Survey (NHANES) data to understand what metal mixtures, if any, contribute to shorter leukocyte telomere length. Telomeres are sensitive to various environmental factors, including exposure to metals and metal mixtures. Several studies have explored the relationship between metal exposures and telomere length, particularly in occupational and environmental settings.

With both synthetic and real-world data we compare our findings to other commonly used mixture methods. Our goal is to show that when using other methods to test for interactions,

the combinatorial problem explodes, reducing power. The analyst may be interested in testing the effects of different possible interactions in the mixture, the question becomes what degree of interaction? What variables are included in the interaction? Does the definition of the interaction even make sense? Quickly it becomes clear that comparing results to our approach is difficult because other methods require user choice of model parameters which may induce bias. Our approach automatically identifies the correct interactions built into the data generating process whereas methods like quantile g-computation require users to select interaction, which are not known. Therefore these interactions are missed and estimates are incorrect.

Chapter 2 examines new semi-parametric definitions for interaction and effect modification that exist outside the scope of linear modeling. Consider the analyst is interested in assessing for interactions using a GLM; the question becomes what do the beta coefficients in front of this interaction term mean if the model itself is inherently misspecified? We need a definition of interaction and effect modification that can be estimated from a large class of non/semi-parametric functions which best estimate nonlinearities in a mixture. Even with these definitions, we need a method that both identifies variable sets used in the best fitting estimator, selected from a large class of flexible functions, and then applies these interaction and effect modification target parameters to these variable sets. Here we again rely on the general framework of data-adaptive target parameters. We expand work done in *stochastic interventions* to create definitions for interaction and effect modification and use the same sample splitting techniques used in **Chapter 1** to identify variable sets in one part of the data and apply our target parameters in another. Again, we apply our this method to NHANES data to investigate the interactions in persistent organic pollutants (POPs) on leukocyte telomere length. We focus on POPs because this dataset is publicly available and has been used in mixtures workshops. This allows us to compare our findings to those published on this dataset. Although this example and the NIEHS synthetic data focus on interaction **Chapter 2** also investigates heterogeneity of treatment/exposure effects. For instance, our causal target parameter in **Chapter 1** is the average regional exposure effect or the average difference in outcomes if all individuals were exposed to a subspace of the mixture compared to if no individuals were exposed to this subspace. Likewise, in **Chapter 2**, for the marginal case, we are interested in the expected disease outcome if say exposure to certain metals decreased by 1 nanogram; we then compare this expected outcome to the outcome under observed metal levels (not decreased). In both situations we are averaging across our sample but what if certain subpopulations exist whose impacts are much greater? After a target parameter, which approaches the truth at a certain rate, is determined, how do we find regions in the covariate-exposure space where these impacts vary the most? Here, we are interested in identifying populations that are vulnerable to a mixed exposure. **Chapter 2** also describes a novel approach for finding types of people who are differentially impacted by chemical exposures.

Chapter 3 extends the data-adaptive work for variable set identification and stochastic

interventions developed in **Chapter 2** used for interaction and effect modification discovery and estimation. **Chapter 3** describes how, using the same framework, mediating pathways can be discovered and estimated. Mediation analysis in causal inference has traditionally focused on one binary exposure using deterministic interventions, decomposing the average treatment effect into direct and indirect effects through one mediator. As discussed, in more realistic exposure settings, individuals are exposed to multiple continuous valued exposures that have effects on health outcomes through different mediating pathways. The exposures that impact health outcomes and their possibly mediating pathways are unknown *a priori* in most instances. Even if the analyst wants to test an exposure-mediator pathway based on domain knowledge, this may not be the strongest pathway in the underlying data. To address this, we propose a methodological framework that both identifies exposure-mediation pathways and delivers unbiased estimates for direct (not through a mediator) and indirect (through a mediator) effects given intervention on exposure subsets. Our approach follows the same framework described in **Chapter 2** but estimates direct and indirect effects in the presence of high-dimensional continuous, binary, and categorical exposures and mediators. To uncover the exposure-mediation pathways, we propose a cross-validation procedure where in the path identification portion of the data, sequential semi-parametric regressions, one for mediators given exposures and covariates, and another for the outcome given exposure, mediators, and covariates are applied to find pathways. In the estimation portion of the data, we apply stochastic interventions to exposures with targeted learning to create efficient estimators based on flexible regression techniques. Our efficient estimator is asymptotically linear under a condition requiring $n^{1/4}$-consistency of certain regression functions.

Chapter 4 discusses the importance of maintained open source software which makes new methodologies available and reproducible for analysts. We discuss the two software packages which house the proposed methods. The first, called `CVtreeMLE`, stands for cross-validated decision trees with targeted maximum likelihood estimation, and makes the statistical causal inference parameters in **Chapter 1** available to researchers. The second, `SuperNOVA`, which stands for Analysis of Variance using Super Learner, makes the statistical causal inference parameters in **Chapter 2** and **Chapter 3** available. **Chapter 5** concludes with a discussion on the future of statistical research using data-adaptive target parameters.

To my mother, who, despite our late reunion and the absence of a traditional family structure, instilled in me an enduring spirit of grit, curiosity and kindness. These qualities have become an inner beacon, guiding me through life's darkest hours.

To the invisible forces that paved the way for me to tap into and flow with my highest creativity.

Contents

List of Figures

List of Tables

Acknowledgments

This dissertation, a milestone in my academic journey, owes its existence to the guidance, support, and inspiration provided by a host of exceptional mentors, colleagues, friends, and loved ones.

Foremost, I extend heartfelt gratitude to Alan Hubbard, whose inspirational presence, relentless positivity, and unfaltering curiosity provided the backbone for this work. The completion of this program would not have been possible without his companionship, mentorship, and encouragement. His unwavering belief in my potential paved the way for me to defy my challenging background as a first-generation college student and emerge successful.

I am immensely grateful to Mark van der Laan, whose intellectual brilliance and exemplary patience have been instrumental in the transformation and maturation of my ideas. The privilege of even being in an office with him, talking about non-parametric definitions of interaction and other core concepts in this work, has been nothing short of surreal. His acceptance and nurturing of my unorthodox ideas boosted my confidence in my creative thinking and its place within semi-parametric statistics. Teaching alongside him further solidified my understanding of his effective, structured problem-solving approach that extends well beyond the realm of statistics.

My deepest thanks go to Alejandro Schuler who entered my professional life at a critical juncture. His exceptional skills as an educator, presenter, and writer have greatly influenced my own teaching style and presentation skills. Alejandro's unique ability to elucidate complex ideas with clarity and empathy, while keeping the audience's comprehension at the forefront, is truly commendable. His emphasis on fostering a supportive community within the often solitary world of research has been a game-changer and a philosophy I aim to carry forward as I transition into professorship.

I owe an enormous debt of gratitude to Martyn Smith, who played a pivotal role in ushering me into this program and fostering my ambition of pioneering a new statistical field for exposomics. His provision of uninterrupted financial support and the freedom to delve into biostatistics without the common constraints of ancillary graduate student projects has been invaluable. This independence facilitated my capacity to envision and initiate a paradigm shift in statistics and epidemiological research, a vision that this work represents the inception of.

I look forward to further collaborations and continued growth under the tutelage of Mark, Alan, Martyn and Alejandro during my postdoctoral training and beyond.

I also wish to acknowledge my outstanding colleagues and friends: Rachael Phillips, Yi Li, Jessica Briggs, James Duncan, Philippe Boileau, and Sophie Fuller. Their invaluable inputs, willingness to entertain new ideas, and passion for exploration enriched our joint endeavors and injected a dose of fun into our rigorous academic pursuits. I also want to acknowledge Sean O'Connell for his unwavering decade long friendship, for always staying connected while we were both out on our respective voyages, exploring new worlds.

Lastly, I would like to express my gratitude to Leah, who provided love, support, and stability as I embarked on the daunting journey of obtaining a PhD. The journey of achieving

a PhD, a demanding and often isolated pursuit, is a privilege that necessitates the backing of family and friends. Without a conventional support system in place, the love and encouragement I received from the aforementioned individuals proved indispensable to my success. Their unwavering belief in me fortified my resolve to complete this journey, and for that, I am forever thankful.

Chapter 1

CVtreeMLE

Exposure to a mixture of chemicals, including drugs, pollutants, and nutrients, is often found in real-world exposure or treatment situations. Within the exposure space, there are arbitrary regions that maximize the mean difference in covariate adjusted disease outcomes. An ideal statistical estimator would identify regions maximizing this difference while delivering unbiased estimations of the relevant effect, benefiting public health efforts aiming to understand the combination of pollutant or drug doses that have the strongest effects on disease outcomes. The estimator should take as input a vector of exposures A, baseline covariates W, and outcome Y. It should output a region in the exposure space that optimizes the maximum mean difference and an unbiased estimate of this average regional-exposure effect (ARE). Rectangular regions, which can be expressed as a series of thresholds, are preferred as they facilitate policy implications by helping policymakers decide on appropriate combinations of exposure thresholds. Non-parametric methods like decision trees are valuable for evaluating combined exposures by identifying partitions in the joint-exposure space. Our proposed methodology leverages decision trees, K-fold cross-validation, and targeted learning to estimate the causal effects of a data-adaptively determined mixture region. The approach uses a parameter-generating sample in each fold to obtain the region and estimators for the statistical target parameter, then applies the region indicator and estimators to the estimation sample, where the ARE is estimated. Targeted learning is utilized to update initial estimates of the ARE in the estimation sample, optimizing bias and variance towards the target parameter. This results in a plug-in estimator with an asymptotically normal distribution and minimum variance, allowing the derivation of confidence intervals. Our approach uses the full data without loss of power due to sample splitting. The open-source R package CVtreeMLE implements this methodology, enabling non-parametric estimation of the causal effects of mixed exposures. The approach produces interpretable and asymptotically efficient results, assisting researchers in discovering significant mixtures of exposure and providing robust statistical inference for their impact.

1.1 Introduction

In most environmental epidemiology studies, researchers are interested in how a joint exposure affects an outcome. This is because, in most real world exposure settings, an individual is exposed to a multitude of chemicals concurrently or, a mixed exposure. Individuals are exposed to a range of multi-pollutant chemical exposures from the environment including air pollution, endocrine disrupting chemicals, pesticides, and heavy metals. Because many of these chemicals may affect the same underlying biological pathway which lead to a disease state, the toxicity of these chemicals can be modified by simultaneous or sequential exposure to multiple agents. In these mixed exposure settings, the joint impact of the mixture on an outcome may not be equal to the additive effects of each individual agent. Mixed exposures may have impacts that are greater than expected given the sum of individual exposures or effects may be less than additive expectations if certain exposures antagonize the affects of others. [1, 49, 45, 69] Likewise, the effects of a mixed exposure may be different for subpopulations of individuals based on environmental stressors, genetic, and psychosocial factors that may modify the impact of a mixed exposure. [91, 53]

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 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. Given that most researchers simply add individual effects to estimate the joint impact of an exposure, it is almost certain that the current evidence of the total impact environmental toxicants have on chronic disease is incorrectly estimated. The impact of using linear modeling is not limited to just potential bias: in the case where linearity does not hold, it's not even clear *what* is being estimated.

The limitation in effective estimation of the joint effects of mixed exposure is (in-part) due to the lack of robust statistical methods. There has been some method development for estimation of joint effects of mixed exposures, such as Weighted Quantile Sum Regression [50], Bayesian Mixture Modeling [18], and Bayesian Kernel Machine Regression [8]. However, these mixture methods have strong assumptions built into them, including directional homogeneity (e.g. all mixtures having a positive effect), linear/additive assumptions and/or require information priors. Many methods suffer from human bias due to choice of priors or poor model fit. More flexible models remain more or less a black-box and describe the mixture through a series of plots rather than with an interpretable summary statistic [8]. Given that the National Institute for Environmental Health Sciences (NIEHS) has included the study of mixtures as a key goal in its 2018-2023 strategic plan [70], it is imperative to develop new statistical methods for mixtures that are less biased, rely less on human input, use machine learning (ML) to model complex interactions, and are designed to return an interpretable parameter of interest.

Decision trees are a useful tool for outcome prediction based on exposures because they are fast, nonparametric (i.e. can discover and model interaction effects), and interpretable [9]. However, it is not immediately clear how to adapt outcome *prediction* methods to *inference*

about the effect of some kind of hypothetical intervention on the mixture of exposures-especially because in these settings we don't have a particular intervention in mind.

Rather than leveraging decision trees for a simple prediction model, we introduce a target parameter on top of the prediction model, which is the average outcome within a fixed region of the exposure space. When an ensemble of decision trees is applied to an exposure mixture, this coincides with a leaf in the best fitting decision tree. By cross-estimating the average outcome given exposure to this region which maximizes the outcome difference we are able to build an estimator that is asymptotically unbiased with the smallest variance for our causal parameter of interest. Previous work, in the most naive approach, confidence intervals (CI) and hypothesis testing of decision trees is done by constructing a $(1-\alpha) \times 100\%$ confidence interval for a node mean $\bar{y}_t$ as $\bar{y}_t \pm z_{1-\alpha/2}(\frac{s_t}{\sqrt{n_t}})$ where $\bar{y}_t$ is the node mean and s_t is the standard deviation estimates in the node. Of course, these CI intervals tend to be overly optimistic because 1. decision trees are adaptive and greedy algorithms, meaning that they have a tendency to overfit and 2. the target parameter, in this case the node average, is estimated on the same data by which the node was created. Because of this the estimated CIs are too narrow. The best approach is to use an independent test set to derive inference for the expected outcome in each leaf. However, this approach is costly if additional data is gathered or power is greatly reduced if sample-splitting is done. Sampling splitting is done in previous work for causal inference of decision trees using so-called "honest estimation" for estimation of heterogeneous causal effects of a binary treatment. This approach [2] uses one part of the data for constructing the partition nodes and and and another for estimating effects within leaves of the partition. Our proposed approach follows a similar sample-splitting technique where one part of the data is used to determine the partition nodes and the other is used to estimate the parameter of interest; however, we extend this technique to K-fold cross-validation where we rotate through the full data. Additionally, rather than estimating heterogenous treatment effects, we are interested in mapping a set of exposures that are of a variety of data types (continuous, binary, multinomial) into a set of partitioning rules using the best fitting decision tree from which we can estimate the average regional-exposure effect, or the expected outcome difference if all individuals were exposed to an exposure region compared to if no individuals were exposed to this region.

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 it is not known what mixture components, levels of these components and combinations of these component levels contribute to the most to a change in the outcome. In the ideal scenario, the analyst has knowledge of the full, multidimensional dose-response curve $E[Y(A_1, A_2, ...A_k)]$ where A are the exposures and Y is the outcome. However, even in this case, it is difficult to estimate and/or interpret this curve. Estimation is hard because 1. we need unrealistic assumptions to get identifiability for the full curve and 2. the curve isn't pathwise differentiable which means there aren't any robust methods to build confidence intervals. Therefore, a sensible approach is to instead categorize the joint exposure and compare averages between categories as one would for a binary exposure. This approach is helpful because we can

define interpretable categories like $(A_1 > a_1)\&(A_2 < a_2)$ where a_i are specific values in A (vs complement of this space) which are of clear interest to policymakers. Identification assumptions are also more transparent in this setting. However, we don't know *a priori* what the right categorization of the exposure space are given some objective function. We have to use the data to tell us what regions are determined given a predefined objective function. In our case, we want a categorization that shows a maximal mean difference in outcomes. Regression trees are a nice way to do this while respecting the fact that we want interpretable rules like the above. The idea is to fit a kind of decision tree to figure out what thresholds in the exposure space produce a maximal exposure effect. As discussed, the result can be biased if we use the same data to define the thresholds and to estimate the effects in each leaf. We solve that problem by splitting the data, doing threshold estimation in one part and regional-exposure effect estimation (given the fixed thresholds) in the other. We can even redo the splits in a round-robin fashion (K-fold cross-validation) to efficiently use all of the data. Lastly, once we have thresholds, we want to get the best possible inference for the effect. We could always do a difference in outcome means between the samples in each category/region, but that estimate would be 1. biased by confounding and 2. have a large confidence interval because we haven't used covariates to soak up residual variance. Our approach is thus to use a doubly-robust efficient estimator (TMLE) that simultaneously addresses both these problems.

Building on prior work related to data-adaptive parameters [44] and cross-validated targeted minimum loss-based estimation (CV-TMLE) [121], our method, called CVtreeMLE, is a novel approach for estimating the joint impact of a mixed exposure by using CV-TMLE which guarantees consistency, efficiency, and multiple robustness despite using highly flexible learners (ensemble machine learning) 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 [35], to fit $f(A)$, a Super Learner [59] 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 many public health settings, the analyst is first interested in a parsimonious set of thresholds focusing on the exposure space that best explains some outcome across the whole population rather partitions that also include baseline covariates. This additive model approach allows us to identify partitioning nodes in the exposure space while flexibly adjusting for 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 partitions in the mixture space which allows for an interpretable mixture contrast parameter, "What is the expected difference in outcomes if all individuals were exposed to this region of the mixed exposure vs. if no individuals were exposed?". This approach easily extends to marginal case (partitions on individual exposures) as well. Our approach for integrating decision trees as a data-adaptive parameter with cross-validated targeted minimum loss-based estimation (CV-TMLE) allows for flexible machine learning estimators to be used to estimate nuisance parameter functionals while preserving desirable asymptotic properties of our target parameter. We provide implementations of this methodology in our free and

open source software CVtreeMLE package [63], for the R language and environment for statistical computing [R Core Team, 2022]. The CVtreeMLE software package has undergone a rigorous peer review process by the Journal of Open Source Software, which validates the robust implementation of our innovative methodology for mixed exposure analysis.

This manuscript is organized as follows, in Section 2.1 we give a background of semi-parametric methodology, in section 2.2 we discuss the ARE target parameter for a fixed exposure region and in 2.3 the assumptions necessary for our statistical estimate to have a causal interpretation. In section 3 we discuss estimation and inference of the ARE for a fixed region. In section 4 we discuss data-adaptively determining the region which maximizes the W-controlled mean outcome difference. In section 5 we show how this requires cross-estimation which builds from 2.2 for a fixed region ARE. In section 5 we expand this to cross-estimation to k-fold CV and discuss methods for pooling estimates across the folds. Lastly, in section 5.3, because we may have different data-adaptively identified regions across the CV folds, we discuss the union rule which pairs with the pooled estimates. In section 6 we discuss simulations with two and three exposures and show our estimator is asymptotically unbiased with a normally sampling distribution. In section 7 we apply CVtreeMLE to the NIEHS mixtures workshop data and identify interactions built into the synthetic data. In section 7.1 we compare CVtreeMLE to the popular quantile sum g-computation method. In section 7.2 we apply CVtreeMLE to NHANES data to determine if there is association between mixed metals and leukocyte telomere length. We selected telomere length as our focus due to its predictive value for cellular aging, longevity, and age-related disease risk. Environmental exposures, including metals, are known to significantly contribute to the variation in telomere length among individuals throughout their lifespan. Given the guanine-rich structure of telomeres, they are particularly susceptible to the detrimental effects of oxidative stress. Therefore, our study aims to investigate the association between a mixture of metal exposure levels, capable of inducing oxidative stress, and telomere length. Section 8 describes our CVtreeMLE software. We end with a brief discussion of the CVtreeMLE method in Section 9.

1.2 The Estimation Problem

Setup and Notation

Our setting is an observational study with baseline covariates ($W \in \mathbb{R}^p$), multiple exposures ($A \in \mathbb{R}^m$), and a single-timepoint outcome (Y). Let $O = (W, A, Y)$ denote the observable data. We presume that there exists a potential outcome function $Y(a)$ (i.e. $Y(a)$ is a random variable for each value of a) that generates the outcome that would have obtained for each observation had exposure been forced to the value $A = a$. These potential outcomes are unobserved but the observed outcome Y corresponds to the potential outcome for the observed value A of the exposure, i.e. $Y = Y(A)$. Let $E[Y(a)|W = w] = \mu(a, w)$ denote the causal dose-response curve for observations with covariates w so that $E[\mu(a, W)]$ represents

the average outcome we would observe if we forced treatment to a for all observations.

We use P_0 to denote the data-generating distribution. That is, each sample from P_0 results in a different realization of the data and if sampled many times we would eventually learn the true P_0 distribution. We assume our $O_1, O_2, ..., O_n$ are iid draws of $O = (W, A, Y) \sim P_0$. We decompose the joint density as $p_{Y,A,W}(y, a, w) = p_{Y|A,W}(y, a, w)p_{A|W}(a, w)p_W(w)$ and make no assumptions about the forms of these densities.

Compare this to to many methods which assume a parametric model which is one where each probability distribution $P \in \mathcal{M}$ can be uniquely described with a finite-dimensional set of parameters. Many methods assume O to be identically distributed normal random variables, which means that the model can be described by the mean and standard deviation. Models like GLMs assume normal-linear relationships and assume $Y = X\beta + \mathcal{N}(\mu, \sigma^2)$. Thus, methods for mixtures that use this approach have three parameters: the slope β, mean μ and standard deviation σ of the normally-distributed random noise. This model assumes that the true relationship between X and conditional mean of Y is additive and linear and that the conditional distribution of Y given X is normal with a standard deviation that is fixed and doesn't depend on X. Of course, these are very strict assumptions especially in the case of mixtures where exposures from a common source may be highly correlated, may interact on the outcome in a non-additive way, and may have non-normal distributions. As such, simply adding coefficients attached to variables in a mixture to estimate the overall joint effect may be biased.

Our statistical target parameter, $\Psi(P_0)$, is defined as a mapping from the statistical model, $\mathcal{M}$, to the parameter space (i.e., a real number) $\mathbb{R}$. That is, Ψ: $\mathcal{M} \to \mathbb{R}$. We can think of this as, if Ψ were given the true distribution P_0 it would provide us with our true estimand of interest.

We can think of our observed data $(O_1 \dots O_n)$ as a (random) probability distribution P_n that places probability mass $1/n$ at each observation O_i. Our goal is to obtain a good approximation of the estimand Ψ, thus we need an estimator, which is an a-priori specified algorithm that is defined as a mapping from the set of possible empirical distributions, P_n to the parameter space. More concretely, the estimator is a function that takes as input the observed data, a realization of P_n, and gives as output a value in the parameter space, which is the estimate, $\hat{\Psi}(P_n)$. Since the estimator $\hat{\Psi}$ is a function of the empirical distribution P_n, the estimator itself is a random variable with a sampling distribution. So, if we repeat the experiment of drawing n observations we would every time end up with a different realization of our estimate. We would like an estimator that is provably unbiased relative to the true (unknown) target parameter and which has the smallest possible sampling variance so that our estimation error is as small as it can be on average.

Defining the Differential Effect Given Regional Exposure

In problems with binary treatment $A \in 0, 1$, the standard counterfactual model defines potential outcomes $Y(0)$ and $Y(1)$, describing what would happen to each individual had they been forced onto either treatment. The estimand of interest is most often the average

treatment effect $E[Y(1)] - E[Y(0)]$. In our setting, $A \in \mathbb{R}^m$ is continuous, and we must thus define a potential outcome $Y(\mathcal{A} = a)$, where $\mathcal{A}$ is a region determined by applying a function to the observed A. The assignment of observations to the (in our case, RuleFit)-defined regions ($\mathcal{A} = 1$ or $\mathcal{A}^c = 1$) is not deterministic, as it depends on the learned rules, which are subject to randomness. Thus, the intervention we are considering is not a deterministic assignment of individuals to regions but rather an assignment based on a set of rules learned from the data. As such, our intervention is considered a stochastic intervention.

Under the causal assumption that A is conditionally randomized, we say that the parameter is identified by:

$$E[E[Y|A \in \mathcal{A}, W] - E[Y|A \in \mathcal{A}^c, W]],$$

which is the mean outcome under a stochastic intervention on A that keeps A given W, beyond that it enforces A to fall in the set $\mathcal{A}$. Because we know that $E[Y_{g*_{\mathcal{A}}}]$ (the expectation of Y where A is under stochastic intervention) is identified by:

$$E_W \int_a E[Y|A = a, W] g *_{\mathcal{A}} (a|W) da,$$

and this equals

$$E_W \int_a E[Y|A = a, W] P(A = a|W, A \in \mathcal{A}) da = E_W[E[Y|A \in \mathcal{A}, W]]$$

by iterative conditional expectation, this proves our parameter can be estimated by the observed data under certain assumptions (discussed next). We use $\mathcal{A} = 0$ and $\mathcal{A}^c$ interchangeably moving forward.

Identification and Causal Assumptions

Our target parameter is defined on the causal data-generating process, so it remains to show that we can define it only in reference to observable quantities under certain assumptions. Standard conditioning arguments show that

$$\psi = E[E[Y|\mathcal{A}, W] - E[Y|\mathcal{A}^c, W]]$$

identifies the causal effect as long as the following assumptions hold:

1. Conditional Randomization: $A \perp Y(a) \mid W$ for all a
2. Positivity: $P(\mathcal{A} = 1|W) > 0$ for all w

Our identification result shows that we can get at the *causal ARE* by estimating an *observable "ATE"* under certain conditions. Our goal is now to show how to efficiently estimate the observable ATE without imposing any additional assumptions (e.g. linearity, normality,

etc.). While our identification assumptions may not always hold in all applications, we can at least eliminate all model misspecification bias and minimize random variation. Once we've established how to estimate the ARE for a fixed region, we'll turn our attention to the problem of finding a good region $\mathcal{A}$ and lastly how to do that without incurring selection bias in estimating the ARE for that region.

1.3 Estimating ARE with TMLE

In the previous sections we established that the causal ARE is equivalent to the observable ATE $E[E[Y|\mathcal{A}=1,W]-E[Y|\mathcal{A}=0,W]]$ under standard identifying assumptions. Therefore to estimate it all we need to do is 1) create a new binary random variable $\mathcal{A}_i = 1_{\mathcal{A}}(A_i)$ and 2) proceed as if we were estimating the observable ATE from the observational data structure $(Y, \mathcal{A}, W)$.

There is an extensive literature on estimating the ATEs from observational data [72, 110, 90]. Using split-sample machine learning we can construct estimators that are provably unbiased (modulo bias from any violations of identifying assumptions), have the minimum possible sampling variance, and which are "doubly robust" [121, 125]. Augmented inverse propensity-weighting (AIPW) and targeted maximum likelihood (TMLE) are two established estimation approaches that accomplish these goals. Although they are usually very similar in practice, TMLE is often better for smaller samples [61, 93, 57, 56] and should generally be preferred. In what follows we use the TMLE estimator of the ATE, which we briefly describe here.

The TMLE estimator is inspired by the fact that if we knew the true conditional mean $Q(\mathcal{A}, W) = E[Y|\mathcal{A}, W]$ we could estimate the ATE with the empirical average $\frac{1}{n}\sum_i Q(1, W_i) - Q(0, W_i)$. Of course, we do not know Q, but we can estimate it by regressing the outcome Y onto the exposure $\mathcal{A}$ and covariates W. However a detailed mathematical analysis shows that we incur bias if we use our estimate $\hat{Q}$ instead of the truth. This bias might decrease as sample size increases, but it dominates relative to random variability, making it impossible to establish p-values or confidence intervals. TMLE solves this problem by computing a correction to the regression model $\hat{Q}$ that removes the bias. In other words, it "targets" the estimate $\hat{Q}$ to the parameter of interest (here the ATE). The process is as follows:

1. Use cross-validated ensembles of machine learning algorithms (a "super learner") to generate estimates of the conditional means of treatment: $\hat{g}(\mathcal{A}=a, W) \approx P(\mathcal{A}=a|W)$ (i.e. propensity score) and outcome: $\hat{Q}(\mathcal{A}, W) \approx E[Y|\mathcal{A}, W]$

2. Regress Y (scaled to $[0,1]$) onto the "clever covariate" $H_i = \frac{1_1(\mathcal{A}_i)}{\hat{g}(1,W_i)} - \frac{1_0(\mathcal{A}_i)}{\hat{g}(0,W_i)}$ using a logistic regression with a fixed offset term $\text{logit}(\hat{Q}(\mathcal{A}, X))$. The (rescaled) output of this is our *targeted* regression model $\hat{Q}^*$

3. Compute the plug-in estimate using the targeted model: $\hat{\psi} = \frac{1}{n}\sum_i \hat{Q}^*(1, W_i) - \hat{Q}^*(0, W_i)$

An estimated standard error for $\hat{\psi}$ is given by

$$\hat{\sigma}^2 = \frac{1}{n^2} \sum_i \left[\left(\frac{1_1(\mathcal{A}_i)}{\hat{g}(1, W_i)} - \frac{1_0(\mathcal{A}_i)}{\hat{g}(0, W_i)} \right) \left(Y_i - \hat{Q}^*(\mathcal{A}_i, W_i) \right) + \left(\hat{Q}^*(1, W_i) - \hat{Q}^*(0, W_i) \right) - \hat{\psi} \right]^2$$

with corresponding 95% confidence interval $\hat{\psi} \pm 1.96\hat{\sigma}$.

Explaining why the targeting step takes the form of a logistic regression and how the estimated standard error is derived are beyond the scope of this work. [58, 57, 56] offer explanations targeted to audiences with varying levels of mathematical sophistication.

To obtain these estimates we need only to specify the ensemble of machine learning algorithms used to estimate the propensity and initial outcome regressions $\hat{g}$ and $\hat{Q}$. The theoretical guarantees hold as long as a sufficiently rich library is chosen.

For estimating the ARE, we must also specify the region $\mathcal{A}$ so that we can compute our binary "exposure" variable. The issue of course is that we have been treating $\mathcal{A}$ as a known region, whereas in many applications the important question is figuring out what guidelines to impose in the first place. This is the focus of the next section.

1.4 Defining the Target Region

Thus far we have not focused on how we define the target region $\mathcal{A}$. First, let's think of $\mathcal{A}$ as nonparametrically defined as the maximizer of some criterion, independent of an estimator. We can think of this as any region on the exposure gradient that maximizes the outcome (can take any shape). However, such a region isn't interpretable. Therefore, it is easier to constrain the optimization so $\mathcal{A}$ is a rectangle in the exposure space. This is because these sections can be easily described using $\geq$ and $\leq$ rules. This is also important from a public health standpoint where these rules effectively are thresholds of exposures found to have the most severe (or least severe) affects. For this purpose, regression trees are an ideal estimator to get at such a region. Each decision tree algorithm uses some objective function to split a node into two or more sub-nodes. Of course, it is generally impossible to know *a priori* which learner will perform best for a given prediction problem and data set. Decision trees have many hyper-parameters such as the maximum depth, minimum samples in a leaf, and criteria for splitting amongst others. As such, we need to find the decision tree estimator that best fits the data given a set of nodes. We do this by creating a library of decision tree estimators to be applied to the exposure data and use cross-validation to select nodes based on the best fitting decision tree. This CV selection of the best fitting decision tree algorithm defines our exposure Super Learner $f(A)$ in our additive semi-parametric model $E(Y|A, W) = f(A) + h(W)$. This additive model is needed because we are interested in finding regions that maximize an outcome within only an exposure space not including the baseline covariates.

Discovering Regions in Multiple Exposures using Ensemble Trees

To discover regions in multiple exposures and therefore discover interactions in the exposure space, we use predictive learning via rule ensembles [28]. Thus, as part of the data-adaptive procedure the $f(A)$ is a regression model constructed as a linear combinations of simple rules derived from the exposure data. Each rule consists of a conjunction of a small number of simple statements concerning the values of individual input exposures. Machine learning using rule ensembles have not only been shown to have predictive accuracy comparable to the best methods but also result in a linear combination of interpretable rules. Prediction rules used in the ensemble are logical if [conditions] then [prediction] statements, which in our case the conditions are regions in the exposure space that are predictive of the outcome. Learning ensembles have the structure:

$$F(x) = a_0 + \sum_{m=1}^{M} a_m f_m(x)$$

where M is the size of the ensemble (total number of trees) and each ensemble member $f_m(x)$ is a different function of the input exposures A derived from the training data in the cross-estimation procedure (discussed later). Ensemble predictions from $F(x)$ are derived from a linear combination of the predictions of each ensemble member with ${a_m}_0^M$ being the parameters specifying the linear combination. Given a set of base learners, trees constructed using the exposures, ${f_m(x)}_1^M$ the parameters for the linear combination are obtained by a regularized linear regression using the training data. Ideally, each tree in the ensemble is limited to including only 2-3 exposures at a time which enhances interpretability. For instance, given the noise and small sample size in most public health studies, it is unlikely that signal is strong enough to detect interactions with 4 or more variables. Not only that but trees with partitions across many variables become less interpretable. Therefore, in our case we are interested in using an ensemble algorithm that creates a linear combination of smaller trees but also shows optimal prediction performance. To accomplish this goal we use the PRE package [26] which is similar to the original RuleFit algorithm [28] with some enhancements including 1. unbiased recursive partitioning algorithms, 2. complete implementation in R, 3. capacity to handle many outcome types and 4. includes a random forest approach to generating prediction rules in addition to bagging and boosting methods. Here, the package PRE is fit to the exposure data in the training sample. Mechanically the procedure is 1. generate an ensemble of trees using exposure data, 2. fit a lasso regression using these trees to predict the outcome, 3. extract the tree basis with nonzero coefficients, 4. store these basis as rules which when evaluated on the exposure space demarcate an exposure region $\mathcal{A}$. There may then be many exposure regions such as $\mathcal{A}_{X_1,X_2}$ which is a region including exposures X_1, X_2 or another region in the exposure space which uses variables X_4, X_5, $\mathcal{A}_{X_4,X_5}$ etc. The ARE is then calculated for each of these regions which are based off of trees found to be predictive in the ensemble. In the case that multiple trees are included in the ensemble which are composed of the same set of exposures, we select the tree with the largest coefficient.

This procedure is done in each fold of the cross-validation procedure.

Discovering Regions in Single Exposures using Decision Trees

In addition to finding interactions in the exposure space, the analyst may also be interested in identifying what exposures have a marginal impact and at what levels the outcome changes the most in these exposures. To answer this question we include a marginal tree fitting procedure which is very similar to the method described in 4.1. Here, $f(A)$ is a Super Learner of decision trees fit onto one exposure at a time. We then extract the rules determined from the best fitting tree. Each terminal leaf demarcates a region in the exposure and thus similarly we may have several $\mathcal{A}$ for 1 exposure which are the regions found when creating the partitions which best explains the outcome. Here, rather than calculating an ARE for each region we calculate an ARE comparing each region to the reference region. The reference region is defined as the region that captures the lowest values of A. For example, consider our resulting decision tree when fit to variable A_1 resulted in terminal leaves $A_1 < 0.6$, $A_1 > 0.6$ & $A_1 < 0.9$, $A_1 > 0.9$, in this case the reference region would be $A_1 < 0.6$. We would then have two ARE estimates for the two regions above the reference region. Mechanically, in the training sample, we find the best fitting decision tree which finds partitions in one exposure that best explains an outcome, these rules are evaluated as if statements on the exposure to create $\mathcal{A}_i = 1_{\mathcal{A}}(A_i)$ for the respective exposure. We then subset the reference level out and row bind it with each region above the reference region and pass that data to our estimators of the ARE. This approach was chosen to give users a dose-response type estimate for data-adaptively determined thresholds in the univariate exposure space.

Iterative Backfitting

We need an algorithm that will allow us to fit $f(A)$ while controlling for W but not including W in the partitions (trees). As such, we iteratively backfit two Super Learners $f(A)$ a Super Learner of decision trees and $h(W)$ an unrestricted Super Learner applied to the covariates. However, both algorithms need to use the same convergence criteria (here maximum likelihood estimation). Thus, $f(A)$ uses an ensemble of regression trees and $h(W)$ uses an ensemble of flexible MLE based algorithms (MARS, elastic net, Highly Adaptive Lasso amongst others). The algorithm first initializes by getting predictions from $f(A)$ and $h(W)$, that is, simply fitting a Super Learner to the exposures and covariates separately and then getting predictions. Then we begin fitting each algorithm offset by the predictions of the other. So at iteration 1 we fit $f(A, \text{offset} = h(W)_{\text{iter } 0})$ and likewise $h(W, \text{offset} = f(A)_{\text{iter } 0})$; where the offsets are predictions of the models fit individually (without offset at iteration 0). The predictions of these models without an offset then gives us $f(A)_{\text{iter } 1}$ and $h(W)_{\text{iter } 1}$. These predictions are then used as offsets at iteration 2. This process continues until convergence where convergence is is defined as the absolute mean difference between the two models being less than some very small number δ where δ by default is 0.001. In this way, for both where A in $f(A)$ is a vector of exposures (resulting rules include combinations of different exposure levels)

and when A is a single exposure, we are able to identify cut-points in the exposure space while controlling for W in the additive model that converges in maximum likelihood. We evaluate the best fitting decision tree onto the exposure space which results in an indicator of the exposure region and calculate our k-fold specific and pooled target parameters give this region.

Asymptotic Properties of Rule Fitting on Mixtures

As discussed, an implementation of the RuleFit algorithm, originally proposed by Jerome H. Friedman [28], is used to determine regions in the mixture space. RuleFit is an ensemble learning method that combines decision trees with Lasso (Least Absolute Shrinkage and Selection Operator) regression to generate a set of rules (conditions) and learn their importance in predicting the outcome variable. The algorithm aims to create a more interpretable model that offers improved generalization performance compared to traditional decision trees. We use this model because it allows us to extract trees used for each unique variable set and calculate the respective ARE.

In terms of asymptotic properties, the RuleFit algorithm does not have any established convergence guarantees to the true regions in the underlying data generating process. It is an empirical method and, like many other machine learning algorithms, its performance depends on the quality of the training data, the complexity of the problem, and the tuning of its hyperparameters. However, it is important to note that the Lasso regression component of the RuleFit algorithm has some desirable asymptotic properties, such as consistency and variable selection consistency, under certain conditions such as in sparse additive data generating processes. In practice, this means that as the sample size grows, Lasso regression can recover the true sparse model and provide accurate predictions. Nevertheless, these properties are not directly transferable to the RuleFit algorithm as a whole, since convergence also depends on how partitions are generated in the tree ensembles. That being said, while the RuleFit algorithm demonstrates good performance and interpretability in various applications, it does not have any established asymptotic properties regarding convergence to the true regions in the underlying data generating process and therefore our parameter is not going for an oracle region. Its performance depends on several factors and, like other machine learning algorithms, it is not guaranteed to find the optimal solution in all cases.

Overall, this means that we cannot theoretically guarantee that regions identified approximate a true max ARE in the population. Interpretation is rather a region for a variable set that has a nonzero coefficient in a penalized model controlling for other identified regions for other variable sets in the mixture. In certain situations this may approximate a true region that maximizes the ARE such as the case in simulations we show where there is a region in a mixture, based on certain thresholds, that has a much larger w-controlled outcome compared to the complimentary space.

This being said, although going after an oracle target parameter such as max ARE in a mixture would provide more interpretable results for the data-adaptive target parameter it is not undermine the proposed approach using RuleFit. In our case, regions are interpreted

as thresholds found for a set of exposures in a mixture while controlling for regions of other exposures and covariates in a semi-parametric additive model (not allowing interactions between exposures and covariates), for interpretability. Users can interpret such regions as, "thresholds that best explain the outcome in a penalized regression model controlling for other exposures and covariates".

1.5 K-fold Cross-Estimation

Of course, the mixture region used to estimate the ARE is not defined *a priori*. If we were to use the same data to both identify the region and make the ARE our estimates will be biased. Thus, for desirable asymptotic properties to hold without additional assumptions, we need our conditional means to be cross-estimated from the observed data. We split the data into $P_{n_{-k}}$ (parameter-generating) and P_{n_k} (estimation) samples. These splits or folds are part of a k-fold cross-validation framework. K-fold cross-validation involves: (i) $1, ..., n$, observations, is divided into K equal size subgroups, (ii) for each k, an estimation-sample, notationally P_k, is defined by the k-th subgroup of size n/K, while the parameter-generating sample, $P_{n_{-k}}$, is its complement. In this round robin manner we rotate through our data and thus, in the case of $K = 10$ get 10 difference target parameter mappings $\mathcal{A}_n$, outcome estimators Q_n and propensity estimators g_n. We want one summary measure of the target parameter found across the folds, such as the average.

With $P_{n_{-k}}$ we find thresholds in our exposure space (using the results of a decision tree) which designates exposure region. Then given this exposure region using the same $P_{n_{-k}}$ we train our g_n and Q_n estimators which are needed for our TMLE update step to debias our initial estimates of the ARE and give us an asymptotically unbiased estimator. We then plug-in our P_{n_k} to this unbiased estimator to get our ARE estimate in this estimation sample.

Let $\bar{Q}_n$ denote a substitution estimator that plugs in the empirical distribution with weight 1/n for each observation which approximates the true conditional mean $\bar{Q}_0$ in P_0, this estimator, in our case is a Super Learner, or ensemble machine learning algorithm, our substitution estimator looks like:

$$\Psi(Q_{P_{n_k}}) = \frac{1}{V}\sum_{v=1}^{V} \bar{Q}_{n_{-k}}(\mathcal{A}_{n_{-k}} = 1, W_v) - \bar{Q}_{n_{-k}}(\mathcal{A}_{n_{-k}} = 0, W_v)$$

Let's focus first on the k subscripts, we split data into $k \in 1...K$ non-overlapping folds and fit K different models. Thus, $\bar{Q}_{n_{-k}}$ denotes our outcome regression function fit when excluding the data for fold k. P_{n_k} denotes our estimation-data and $P_{n_{-k}}$ is the parameter-generating sample, that is, our parameter-generating sample is used to train our estimators and then we pass our estimation-data in to get estimates. $\bar{Q}_{n_{-k}}$ then, in our case, is a Super Learner fit using the parameter-generating data. Likewise, $\mathcal{A}_{n_{-k}}$ is a decision tree fit using the parameter-generating data. $\Psi(Q_{P_{n_k}})$ then indicates that we pass the estimation-sample data into our estimators trained with the parameter-generating data; so here we first fit a decision

tree to the exposure space of the parameter-generating data, then apply the rules found to the estimation-sample data to create an exposure region indicator. Then using this exposure and the estimation-sample covariates, we feed this into the outcome regression model trained on the parameter-sample data. We then get predicted outcomes under different counterfactuals for a data-adaptively determined exposure using our estimation-sample data. Our cross-estimated TMLE estimator for this data-adaptively defined exposure produces an unbiased, efficient substitution estimator of target parameters of a data-generating distribution we are interested in. This estimator looks like:

$$\Psi(Q^{\star}_{P_{n_k}}) = \frac{1}{V}\sum_{v=1}^{V}\{\bar{Q}^{\star}_{n_{-k}}(\mathcal{A}_{n_{-k}} = 1, W_v) - \bar{Q}^{\star}_{n_{-k}}(\mathcal{A}_{n_{-k}} = 0, W_v)\}$$

Here we can see the only change to our above equation is $\bar{Q}^{\star}$ which is the TMLE augmented estimate. This new function, $f(\bar{Q}^{\star}_{n_{-k}}(A, W)) = f(\bar{Q}_{n_{-k}}(A, W)) + \epsilon_{n_{-k}} \cdot h_{n_{-k}}(A, W)$, where $f(\cdot)$ is the appropriate link function (e.g., logit), ϵ_n is an estimated coefficient and $h_n(A, W)$ is a "clever covariate" which is now cross-estimated. Here what we mean is that, the initial estimates for the estimation-sample using models trained using the parameter-generating data are updated through this so-called, least-favorable submodel. The cross-estimated clever covariate looks like:

$$h_{n_{-k}}(\mathcal{A}, W) = \frac{\mathbb{I}(\mathcal{A}_{n_{-k}} = 1)}{g_{n_{-k}}(\mathcal{A}_{n_{-k}} = 1|W)} - \frac{\mathbb{I}(\mathcal{A}_{n_{-k}} = 0)}{g_{n_{-k}}(\mathcal{A}_{n_{-k}} = 0|W)}$$

Here, $g_{n_{-k}}(W) = \mathbb{P}(\mathcal{A}_{n_{-k}} = 1 \mid W)$, the propensity score of the data-adaptively determined exposure region, is being estimated using a Super Learner with the parameter-generating data. That is, in our parameter generating sample we get the exposure region, and an estimator g_n we apply this exposure region to the estimation sample and then get predictions for the probability of that exposure region indicator using the estimation sample, we then plug these estimates into the above cross-estimated clever covariate used in the TMLE update.

We can see that by using v-fold cross-validation, we can do better than traditional sample splitting as v-fold allows us to make use of the full data which results in tighter confidence intervals because our variance is estimated over the full data. Similarly, our estimate is an average of the v-fold specific estimates:

$$\Psi_n(P) = Ave\{\Psi_{P_{n_{-k}}}(P)\} \equiv \frac{1}{V}\sum_{v=1}^{V}\Psi_{P_{n,-k}}(P)$$

We do this in a pooled TMLE update manner where we stack the estimation-sample estimates for each nuisance parameter and then do a pooled TMLE update across all the initial estimates using clever covariates across all the folds to get our estimate ϵ we then update our counterfactuals across all the folds and take the average. More concretely, in each fold we have our initial estimates from that fold from $Q_{n_{-k}}(Y|\mathcal{A}, W)$ and the fold specific clever covariate $h_{n_{-k}}(\mathcal{A}|W)$ of length k for a fold specific exposure found using $\mathcal{A}_{n_{-k}}$. We

stack all the $Q_{n_{-k}}$'s and $h_{n_{-k}}(\mathcal{A}|W)$'s together along with the outcomes in each validation fold and do our fluctuation step:

$$f(\bar{Q}_n^\star(\mathcal{A}, W)) = f(\bar{Q}_n(\mathcal{A}, W)) + \epsilon_n \cdot h_n(\mathcal{A}, W)$$

Notice here the k subscripts are removed, this is because we are using our cross-estimates for all of n. Using the ϵ from this model, we then update the counterfactuals across all the folds and take the difference for our final ARE. In a similar fashion, we use the updated conditional means, counterfactuals, and clever covariates to solve the IC across the whole sample. By pooling the cross-estimates across the folds and then calculating the SE for this pooled IC we are able to derive more narrow confidence intervals compared to if we were to average the IC estimated in each of the folds (because the IC is scaled by n and not n/K). This pooled estimate still provides us with proper intervals because all estimates in its construction were cross-estimated.

An alternative to this pooled approach is to simply report the k-fold specific estimates of the ARE and fold specific variance estimates for this ARE using the fold specific IC. We do this as well. We do this because, if the exposure region $\mathcal{A}$ identified in each region is highly variable, that is, if the region that that maximizes the difference for sets of exposure variables are very different across the folds, then interpreting the pooled ARE is difficult. By calculating and providing both k-fold specific and pooled results users can investigate how variable a pooled result is across the folds.

Inverse-Variance Method for Combining K-fold Results

In addition to the pooled TMLE approach to aggregate k-fold specific data-adaptive target estimates, we also calculate the inverse-variance method (IVM) commonly used in meta-analyses. We call this method the k-fold harmonic mean. Here each fold is given a weight defined as:

$$w_{k_i} = \frac{1}{SE(\hat{\theta}_{k_i})^2}$$

Which is simply an inverse of the standard error such that estimates with smaller SE are given a higher weight. The inverse-variance pooled ARE across the folds is given as:

$$\hat{\theta}_{IVM} = \frac{\sum w_{k_i} \hat{\theta}_{k_i}}{\sum w_{k_i}}$$

And lastly, the pooled SE is calculated as:

$$SE(\hat{\theta}_{IVM}) \frac{1}{\sqrt{\sum w_{k_i}}}$$

For which confidence intervals and p-values are derived for the pooled IVM estimate.

This pooled estimate is given because, in the event of high inconsistency of the k-fold estimates in lower sample size, the confidence intervals from pooled influence curve may not cover the true ARE if the pooled ARE was applied to P_0. This is because the union rule attached to the pooled ARE is a conservative rule which covers all observations across the folds (discussed later). The IVM derived CIs are wider and provide better coverage in the event of high inconsistency (which we show in simulations). We explain rule stability metrics and establishing a union rule across the folds in the next section.

Defining the Union Region

The pooled TMLE ARE is matched with a pooled region that encompasses all the observation indicated by each fold specific regions. We group the trees across the folds according to what variable sets the trees are composed of. That is, a linear combination of tree ensembles is fit to each training sample specific to the fold. There may be variability in where the partition is set for trees with the same variable sets across the folds, or certain ensembles don't use certain variable sets at all in some folds but used in others. We need a method of creating a pooled region and give stability metrics for how consistently trees with a respective variable set are found in the cross-validation procedure. For this we create a union region. There are other possible ways of pooling the regions, such as averaging the partitions per exposure variable across the folds. Here we choose a conservative approach. This is the union region of the k-fold regions in the sense that, we create a new region that is the OR combination of each k-fold specific tree. For three folds and therefore three partitions say, $X_1 < 2$ & $X_2 > 5$, $X_1 < 2.3$ & $X_2 > 5.2$ and $X_1 < 1.9$ & $X_2 > 5.3$, the union rule is $X_1 < 2$ & $X_2 > 5$ OR $X_1 < 2.3$ & $X_2 > 5.2$ OR $X_1 < 1.9$ & $X_2 > 5.3$ forms the rule: $X_1 < 2.3$ & $X_2 > 5$ because this region covers all the observations indicated in the fold specific regions. For variables where the logic is $>$ we take the minimum value across the folds and likewise for $<$ we take the maximum. This union region is conservative and sensitive to outlier partition points found across the folds and therefore higher K folds will lead to more stable partitions if there is signal in the data. Additionally, the analyst should investigate the fold specific regions to determine the interpretability of the pooled region. If there is high variability or outliers, there may be bias in the TMLE pooled estimate when compared to the expected difference in outcomes if the respective pooled region was applied to the true population P_0.

Stability Metrics

Given a pooled region, we simply give the proportion of folds trees with a respective variable set are found across the folds. For example, consider a study of mixed metals that uses CVtreeMLE and the results across three folds are: 1. lead $>$ 2.2 & arsenic $>$ 1.3, 2. lead $>$ 2.1 & arsenic $>$ 1.2, 2. lead $>$ 2.0 & arsenic $>$ 1.1. Our pooled region is lead $>$ 2.0 & arsenic $>$ 1.1 because this region contains all the fold specific regions. The stability metric here is 100% because a tree with lead and arsenic was found in all three folds. If however, this tree was only found in 2 of three folds, the stability metric is 67%.

1.6 Simulations

In this section, we demonstrate using simulations that our approach identifies the correct exposure region which maximizes the difference in conditional means and estimates the correct difference built into a DGP for this region.

Data-Generating Processes

Because a two dimensional exposure space is easier to visualize and describe compared to higher dimensional spaces, we start by investigating a squared dose-response relationship between two exposure variables where an interaction occurs between the exposures when each meets a particular threshold value. We extend simulations to the three dimensional case. In both 2-D and 3-D exposure simulations there are specific outcome values generated for each subspace of the mixture based on split points $\mathcal{D}_d$ but there exists one region with the maximum outcome (the truth that we want). In both scenarios, the goal is to determine if our data adaptive target parameter is targeting the region that maximizes the conditional mean outcome for the given sample and evaluate how CVtreeMLE approaches this desired oracle parameter as sample size increases. To meet this goal, we construct a data-generating process (DGP) where Y is generated from a tree-structured covariate-adjusted relationship of a mixture consisting of components, A_1, A_2, A_3, A_n. That is, generally in each simulation we generate exposure regions, where the density of the region is driven by covariates and there is one region that has the maximum difference compared to outside the respective region. More details for each simulation are given below.

Two-Dimensional Exposure Simulations

This DGP has the following characteristics, $O = (W, A, Y)$. W are three baseline covariates

$$W_1 \sim \mathcal{N}(\mu = 37, \sigma = 3), W_2 \sim \mathcal{N}(\mu = 20, \sigma = 1), W_3 \sim \mathcal{B}(\mu = 0.5)$$

Where $\mathcal{B}$ is a Bernoulli distribution and $\mathcal{N}$ is normal. These distributions and values were chosen to represent a study with covariates for age, BMI and sex. Our generated exposures were likewise created to represent a chemical exposure quantized into 5 discrete levels. The values and range of the outcome were chosen to represent common environmental health outcomes such as telomere length or epigenetic expression.

We are interested in sampling observations into a 2-dimensional exposure grid. Here a 5×5 grid is based on combinations of two discrete exposure levels with values 1-5. We want the number of observations in each of these cells to be affected by covariates. To do this we define a conditional categorical distribution $P\{(A_1, A_2) = (a_1, a_2)|W = w\}$ and sample from it.

$$P\{(A_1, A_2) = (a_k, a_l)|W\} = \frac{e^{W^\top \beta_{k,l}}}{1 + \sum_{k,l} e^{W^\top \beta_{k,l}}}$$

Here the β's attached to each covariate were drawn from a normal distribution with means 0.3, 0.4, 0.5 and 0.5 respectively all with a standard deviation of 2. This then gives us 25 unique exposure regions with densities dependent on the covariates. We then want to assign an outcome in each of these regions based on main effects and interactions between the exposures. We use the relationship

$$Y = 0.2A_1^2 + 0.5A_1A_2 + 0.5A_2^2 + 0.2 * \text{age} + 0.4 * \text{sex} + \epsilon(0, 0.1)$$

Which indicates there is a slightly weaker squared effect for A_2 relative to A_1 and a strong interaction between the exposures and confounding due to age and sex. The resulting data distribution and generating process is shown in **Figure 1.1**.

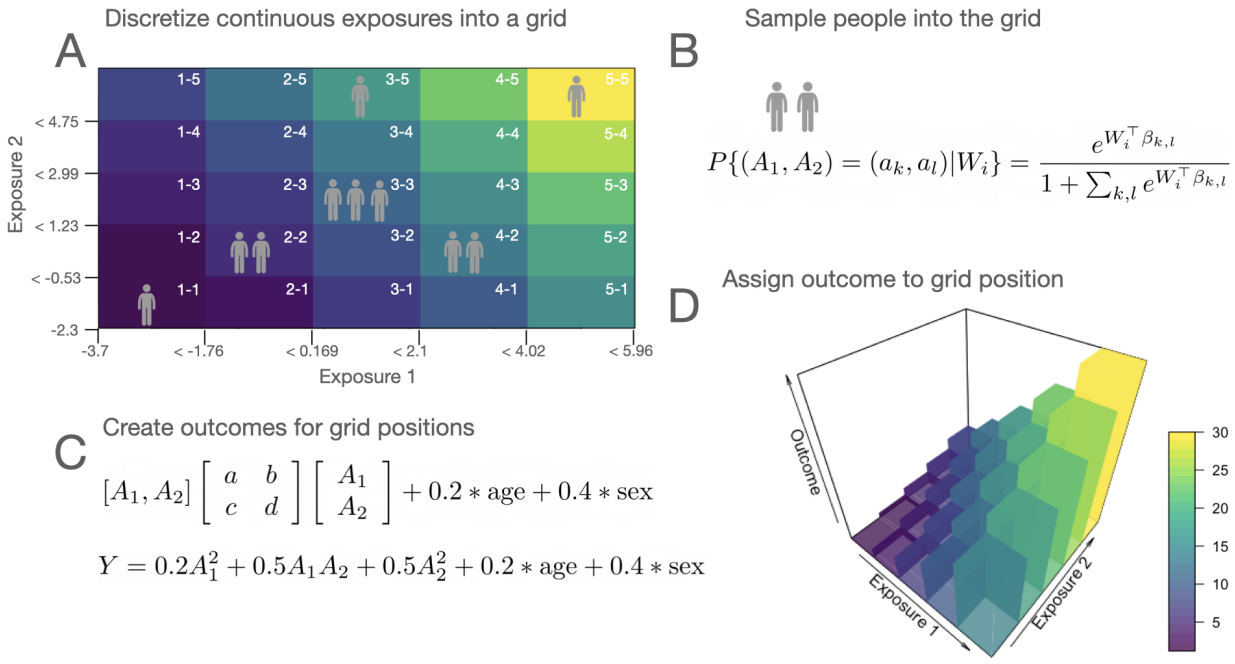


Figure 1.1: 2D Exposure Simulation

Of course, it is also possible to explore other dose-response relationships (such as logarithmic) by changing the coefficient matrix.

Computing Ground Truth The fact that our exposures are discrete in this simulation lets us easily compute the ground-truth ARE for any region $\mathcal{A}$ because we can explicitly compute the conditional mean function m

$$m(\mathcal{A}=1,w) = \int_{a\in\mathcal{A}} \mu(a,w)\frac{p_{A|W}(a,w)}{\pi_{\mathcal{A}}(w)}\, da$$
$$= \frac{\sum_{a\in\mathcal{A}} \mu(a,w)p_{A|W}(a,w)}{\sum_{a\in\mathcal{A}} p_{A|W}(a)}$$

Therefore to approximate the ARE to arbitrary precision we can

1. Sample a large number of times (e.g. $b = 100,000$) from the covariate distribution to obtain $W_{\{1,\ldots i,\ldots b\}}$.
2. Compute the values $m(\mathcal{A}=1,W_i)$ using the above formula. This is possible because the functions μ and $pA|W$ are known for the data-generating process [1]. In a similar fashion compute $m(\mathcal{A}=0,w)$.
3. Compute $\text{ARE}(\mathcal{A}) = \frac{1}{b}\sum_i^b m(1,W_i) - m(0,W_i)$.

Three-Dimensional Exposure Simulations

This DGS has the same general structure, $O = W, A, Y$. W and baseline covariates

$$W_1 = \mathcal{N}(\mu=37,\sigma=3), W_2 = \mathcal{N}(\mu=20,\sigma=1), W_3 = \mathcal{B}(\mu=0.5)$$

In this 3D simulation we are interested in keeping the exposures continuous as this is more realistic compared to the 2D simulation.

Here A are three continuous mixtures from a multivariate normal distribution:

$$\begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1.0 & 0.5 & 0.8 \\ 0.5 & 1.0 & 0.7 \\ 0.8 & 0.7 & 1.0 \end{pmatrix}\right]$$

We assign one partition point value to each exposure which creates 8 possible regions in the $2x2x2$ 3D grid for which we want to assign outcomes. Just as the first simulation we want the number of observations in each of the cells in the "mixture cube" to be affected by covariates. To do this we define a conditional categorical distribution $P\{(A_1,A_2,A_3) = (a_1,a_2,a_3)|W=w\}$ and sample from it.

$$P\{(A_1,A_2,A_3) = (a_k,a_l,a_j)|W_i\} = \frac{e^{W_i^\top \beta_{k,l,j}}}{1+\sum_{k,l,j} e^{W_i^\top \beta_{k,l,j}}}$$

[1] If the exposure space were not discrete, this step would require numerical approximation of an integral for each different value of w which would be generally impractical.

For each of these categories which defines a region in the exposure space we need to assign exposure values while also preserving the local correlation structure within that region. To do this, we convert the cumulative distribution function of the exposures to a uniform distribution then back transform this uniform distribution to the original exposure distribution with bounds for each exposure region. So for instance, in the region where each exposure is less than each threshold value, we back transform the uniform distribution with the minimum value set as the minimum for each exposure and max as the partition value for each exposure. These values then are attached to the categorical variables generated which represent the mixture region. This then generates continuous exposure values with a correlation structure in each region.

The outcome Y is then generated via a linear regression of the form:

$$Y = \beta_0 + \beta_1 \mathbb{1}(A = 1) + \beta_2 \mathbb{1}(A = 2) + ... + \beta_7 \mathbb{1}(A = 7) + \beta_{W_1} W_1 + \beta_{W_2} W_2 + \epsilon, \epsilon \ N(0, \sigma)$$

Where the β_j are chosen so some mixture groups have a high mean, some have a low mean and $\mathbb{1}$ represent indicators of each of the possible 8 regions. Thus, the outcome in each region of the mixture cube is determined by the β assigned to that region. Given this formulation of a DGP it is possible to then generate Y by shifting the drivers or "hot spots" around the mixture space, thereby simulating possible agonist and antagonistic relationships. We could assign something like $\beta_2 = 2$ with all other regions having a $\beta_{\neq 2} = 0$. This then would mean the ARE in the true DGP is 2. Likewise we could assign β's in each region in which case the truth by our definition is the region with the max ARE. The process for this DGP is shown in **Figure 1.2**.

Overall, our 3D example is very similar to the 2D exposure simulation but we aim to test CVtreeMLE in identifying thresholds used to generate an outcome in a space of three continuous exposures. Also, because we keep the space of possible outcomes relatively simple here, we simply generate individual outcomes for each mixture subspace. This allows us to create situations where only one region drives the outcome while the complementary space is 0 or there is an outcome in each region and we are interested in identifying the region with the maximum outcome. In each simulation we are interested in the bias/variance of our estimates compared to the truth, the bias of our rule compared to the true rule and the bias of our data-adaptive rule compared to the expected ARE if that rule was applied to the true population. We discuss this next.

Computing Ground Truth Previously, in the discrete exposure case, we could directly estimate ground truth by inverse weighting given the summed probability in the exposure region multiplied by the outcome. This is not possible in the continuous case. To make things simpler, we z-score standardize the covariates so the mean of each covariate is 0. Therefore we can directly compute the mean in the region indicated by the ground-truth rule and the mean outcome in the complementary space and take the difference. This is the same as the max coefficient minus the mean of the other coefficients in the linear model, this is the true ARE.

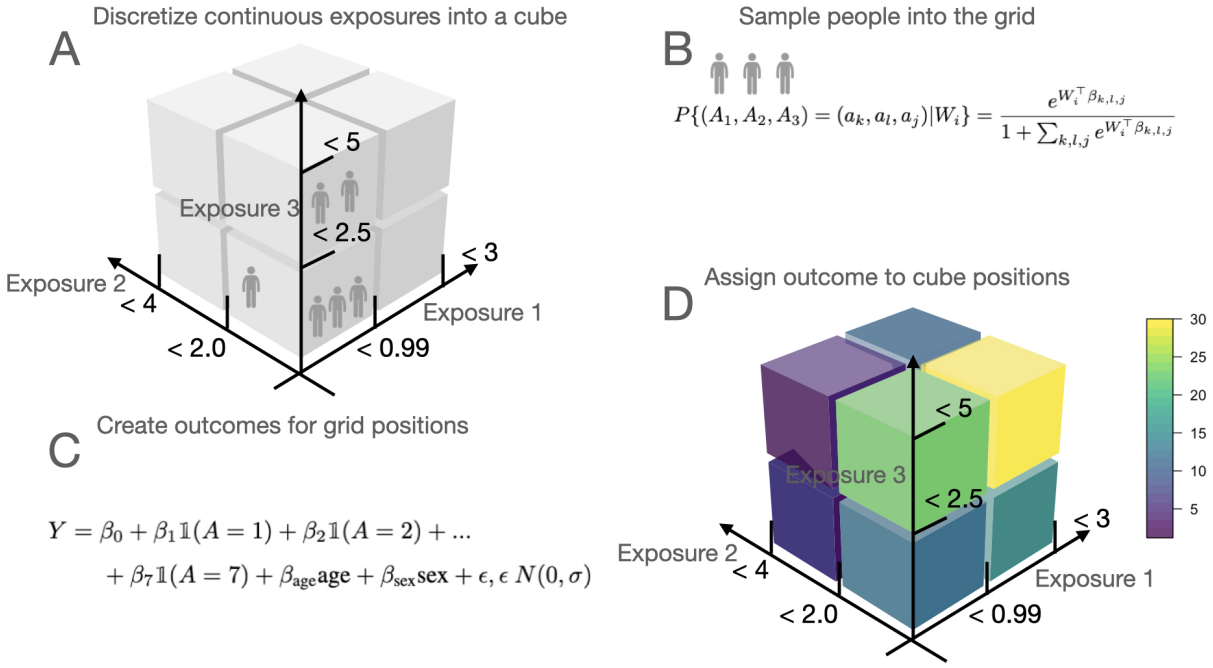


Figure 1.2: 3D Exposure Simulation

Evaluating Performance

The following steps breakdown how each simulation was tested to determine 1. asymptotic convergence to the true mixture region used in the DGP, 2. convergence to the true ARE based on this true region and 3. convergence to the true data-adaptive ARE, that is CVtreeMLE's ability to correctly estimate the ARE if the data-determined rule was applied to the population. We do this by:

1. To approximate P_0, we draw a very large sample (500,000) from the above described DGP.

2. We then generate a random sample from this DGP of size n which is broken into K equal size estimation samples of size $n_k = n/K$ with corresponding parameter generating samples of size $n - n/K$.

3. At each iteration the parameter generating fold defines the region and is used to create the necessary estimators. The estimation fold is used to get our TMLE updated causal parameter estimate, we then do this for all folds.

4. For an iteration, we output the ARE estimates given pooled TMLE, k-fold specific TMLE and the harmonic mean. The region identified in the fold is applied to the large

sample P_0 to estimate the data-adaptive bias. Likewise, each estimate is compared to the ground-truth ARE and region.

For each iteration we calculate metrics for bias, variance, MSE, CI coverage, and confusion table metrics for the true maximal region compared to the estimated region. For each type of estimate (pooled TMLE, k-fold specific TMLE estimates, and harmonic mean) we have bias when comparing our estimate to 1. the ARE based on the true region in the DGP that maximizes the mean difference and 2. the ARE when the data-adaptively determined region is applied to the population. Therefore, when comparing to the true "oracle" region ARE we have:

1. $\psi^0_{\text{pooled tmle bias}}$: This is the bias of the pooled TMLE ARE compared to the ground-truth ARE for the true region built into the DGP which maximizes the mean difference in adjusted outcomes.

2. $\psi^0_{\text{mean v-fold tmle bias}}$: This is the bias of the mean k-fold specific AREs compared to the ground-truth ARE for the true region built into the DGP.

3. $\psi^0_{\text{harmonic mean v-fold tmle bias}}$ This is the bias of the harmonic mean of k-fold specific AREs compared to the ground-truth ARE for the true region built into the DGP.

The above bias metrics are each compared to the true ARE for the oracle region in the DGP. We are also interested in the ARE if the data-adaptively determined region, the region estimated to maximizes the difference in outcomes in the sample data, were applied to P_0 the true population. Therefore, there are also bias estimates for:

1. $\psi^{DA}_{\text{pooled tmle bias}}$: This is the bias of the pooled TMLE ARE compared to the ARE of the union region across the folds applied to P_0.

2. $\psi^{DA}_{\text{mean v-fold tmle bias}}$: This is the bias of the mean k-fold specific AREs compared to the mean ARE when all the k-fold specific rules are applied to P_0.

3. $\psi^{DA}_{\text{harmonic mean k-fold tmle bias}}$ This is the bias of the harmonic mean of k-fold specific AREs compared to the ARE of the union region across the folds applied to P_0.

We multiple each bias estimate by $\sqrt{n}$ to ensure the rate of convergence is at or faster than $\sqrt{n}$. For each ARE estimate we calculate the variance and subsequently the mean-square error as: MSE = bias^2 + variance. MSE estimates were also multiplied by n. For each ARE estimate we calculate the confidence interval coverage of the true ARE parameter givent the oracle region and the ARE given the data-adaptively determined region applied to P_0. For the TMLE pooled estimates these are lower and upper confidence intervals based on the pooled influence curve. For the k-fold specific coverage, we take the mean lower and upper bounds. For the harmonic pooled coverage, we calculate confidence intervals from the pooled standard error. In each case, we check to see if the ground-truth rule ATE and data-adaptive

rule ATE are within the interval. Lastly, we compare the data-adaptively identified region to the ground-truth region using the confusion table metrics for true positive, true negative, false positive and false negative to determine whether, as sample size increases, we converge to the true region.

These performance metrics were calculated at each iteration, where 50 iterations were done for each sample size n = (200, 350, 500, 750, 1000, 1500, 2000, 3000, 5000). It was ensured that, for each data sample, at least one observation existed in the ground-truth region to ensure confusion table estimates could be calculated. CVtreeMLE was run with 5 fold CV (to speed up calculations in the simulations) with default learner stacks for each nuisance parameter and data-adaptive parameter. Our data-adaptive parameter for interactions was the tree with the max ARE (positive coefficient) for each variable set in the ensemble.

Default Estimators

CVtreeMLE needs estimators for $\bar{Q} = E(Y|A, W)$ and $g_n = P(A|W)$. CVtreeMLE has built in default algorithms to be used in a Super Learner [59] that are fast and flexible. These include random forest, general linear models, elastic net, and xgboost. These are used to create Super Learners for both $\bar{Q}$ and g_n. CVtreeMLE also comes with a default tree ensemble which is fit to the exposures during the iterative backfitting procedure. These trees are built from the partykit package [43] in R. By default we include 7 trees in the tree Super Learner that have various levels for the hyper-parameters alpha (p-value to partition on), max-depth (maximum depth of the tree), bonferroni correction (whether to adjust alpha by bonferroni) and min-size (minimum number of observations in terminal leaves). These trees are used during the iterative backfitting in estimating partitions for each individual exposure. For the rule ensemble, the predictive rule ensemble package (pre) [25] is used with default settings and 10-fold cross-validation. Users can pass in their own libraries for these nuisance and data-adaptive parameters. For these simulations, we use these default estimators in each Super Learner.

Results

CVtreeMLE Algorithm Identifies the True Region with Maximum ARE

First we describe results for identifying the true region built into the DGP. It is obviously necessary for this to converge to the truth as sample size increases in order for the ψ^0 estimates to be asymptotically unbiased. Overall we find the tree algorithm identifies the true region in the DGP and therefore provides results which have high-value for treatment policies. **Figure 1.3** shows metrics comparing observations covered by the estimated pooled region to those indicated by the true region in the DGP for two discrete exposures. From this figure it can be seen that, at around 1500 observations, the pooled region is the true region. **Figure 1.4** shows the confusion table metrics comparing the data-adaptive pooled region to the oracle region in the three continuous exposure scenario. As sample size increases, the

false positives approach 0 which is what we would desire in this continuous case. From this, we see that in both instances of discrete and continuous exposures, CVtreeMLE is able to identify the correct region in the exposure data which has the maximum ARE. There is some small disparity in the discovered region compared to the truth in the continuous case, this is because for false-positives to perfectly match the true region, the tree search algorithm must identify the exact set of continuous digits that delineate the region which is very difficult. In our case, this region is is $M_3 <= 2.5$ & $M_1 >= 0.99$ & $M_2 >= 2.0$. In this three exposure case there is antagonism of M_3. Given the exposures are continuous, it is likely that the tree search algorithm gets very close but not absolutely exact to these boundaries. In the two exposure case where the exposures were discretized finding the boundaries is easier. As such, our future evaluation is focused more on the data-adaptive estimates (comparing estimates to the ARE given applying the data-adaptive rule to P_0). Ultimately, the data-adaptive target parameter theory only holds for the data-adaptive parameter and not the parameter given an oracle rule; however, we include both again to investigate how CVtreeMLE approaches the oracle rule as sample size increases.

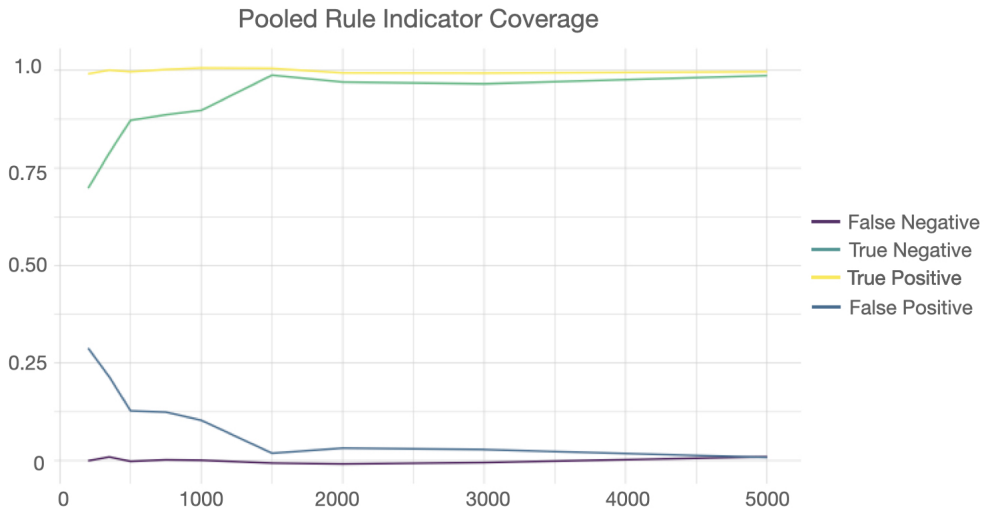


Figure 1.3: 2D Exposure Confusion Table Metrics of Rule Coverage

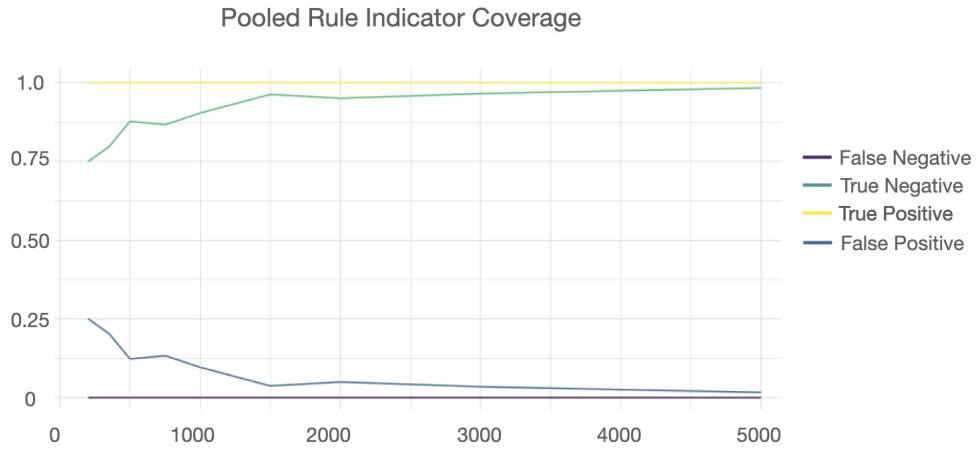


Figure 1.4: Three Exposure Confusion Table Metrics of Rule Coverage

CVtreeMLE Unbiasedly Estimates the Data-Adaptive Parameter

Looking at the bias for the ARE estimate given two discrete exposures compared to the data-adaptively discovered region applied to P_0 TMLE unbiasedly estimates the data-adaptive parameter at root n rates with good coverage. Below, **Figure 1.5 A** shows the data-adaptive rule ARE bias ($\psi^{DA}_{\text{pooled tmle bias}}$, $\psi^{DA}_{\text{mean k-fold tmle bias}}$, $\psi^{DA}_{\text{harmonic mean k-fold tmle bias}}$) and MSE (B).

In **Figure 1.5 A** the data-adaptive rule ARE bias is larger for the pooled estimates (pooled TMLE ARE and harmonic mean ARE compared to ARE if the pooled rule was applied to P_0) compared to the average folds bias (mean k-fold ARE compared to the mean of each k-fold rule applied to P_0). This is because inconsistent rule estimates in lower sample sizes can bias the pooled ARE compared to the pooled region. Consider a 3-fold situation where for variables X_1 and X_2 the region was designated by $X_1 > 4$ & $X_2 > 4$, $X_1 > 4$ & $X_2 > 4$, and $X_1 > 2$ & $X_2 > 4$; because $X_1 > 2$ & $X_2 > 4$ is found in one of the folds, this is the pooled region (as it covers observations for $X_1 > 4$) and thus (if the true ARE for $X_1 > 4$ & $X_2 > 4$ applied to P_0) is higher, our pooled results would be biased to this higher ARE because two of three of our folds have an ARE for this region. This bias converges to average k-fold bias at a sample size of 1500. Effectively, once the trees across the folds stabilizes there is less bias in the pooled estimate compared to the pooled region ATE. This similar pattern is reflected in the pooled estimates MSE (given higher bias in smaller samples). For the user, this indicates that, in smaller sample sizes ($n < 1000$) the analyst should look at fold specific results to ensure the trees are close in the cut-off values in order to interpret the pooled result. If not, k-fold specific results should be reported as these show very low bias/MSE even in smaller sample sizes. The bias and MSE for all estimates compared to the ground-truth rule ATE show an $1/\sqrt{n}$ reduction as sample size increases. In sum, as sample size increases the bias for all estimates converge to 0 which which is necessary

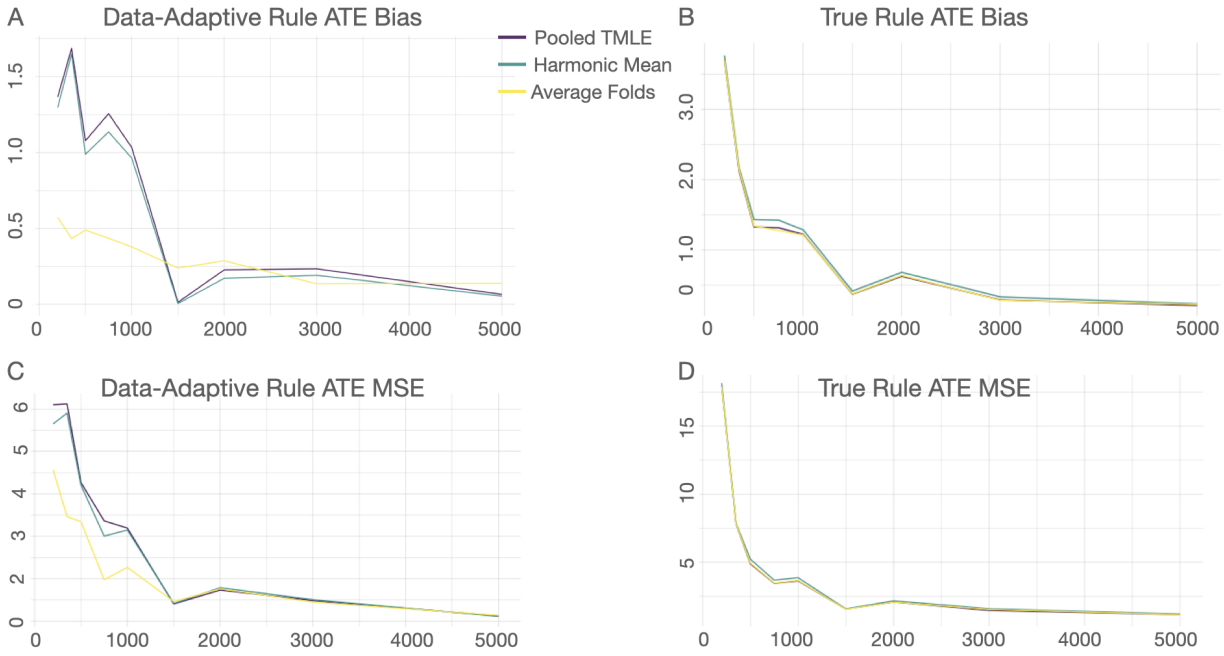


Figure 1.5: 2D Exposure Bias and MSE

for our estimator to have valid confidence intervals.

Figure 1.6 A and C likewise show the asymptotic bias in the three continuous exposure case. All estimates show bias decreasing when evaluated against the ARE when the data-adaptively determined region is applied to P_0; however, these estimates do not go to 0 exactly (at max sample size equal to 5000) as the data-adaptive rule is still not exactly the true rule, which is expected. **Figure 1.7 A** shows the confidence interval coverage for each estimate compared to the data-adaptive region applied to P_0.

For coverage of the ARE of the pooled rule applied to P_0, the CIs calculated from the pooled k-fold standard errors showed coverage between 95% - 100%. The pooled TMLE CIs showed poorer coverage at lower sample sizes, this is likely due to the bias of the pooled ARE estimate compared to the pooled region applied to P_0 paired with the more narrow confidence intervals calculated across the full sample. The harmonic pooled k-fold CIs were wider and thus covered the truth in this pooled setting. Coverage for the k-fold specific CIs were almost always at or above 90% and converge to 95% at higher sample sizes.

Figure 1.8 A shows the CI coverage of the data-adaptive rule in the three exposures simulation. As expected, the average k-fold CI converges to 95%. The pooled estimates are lower given the conservative pooled rule.

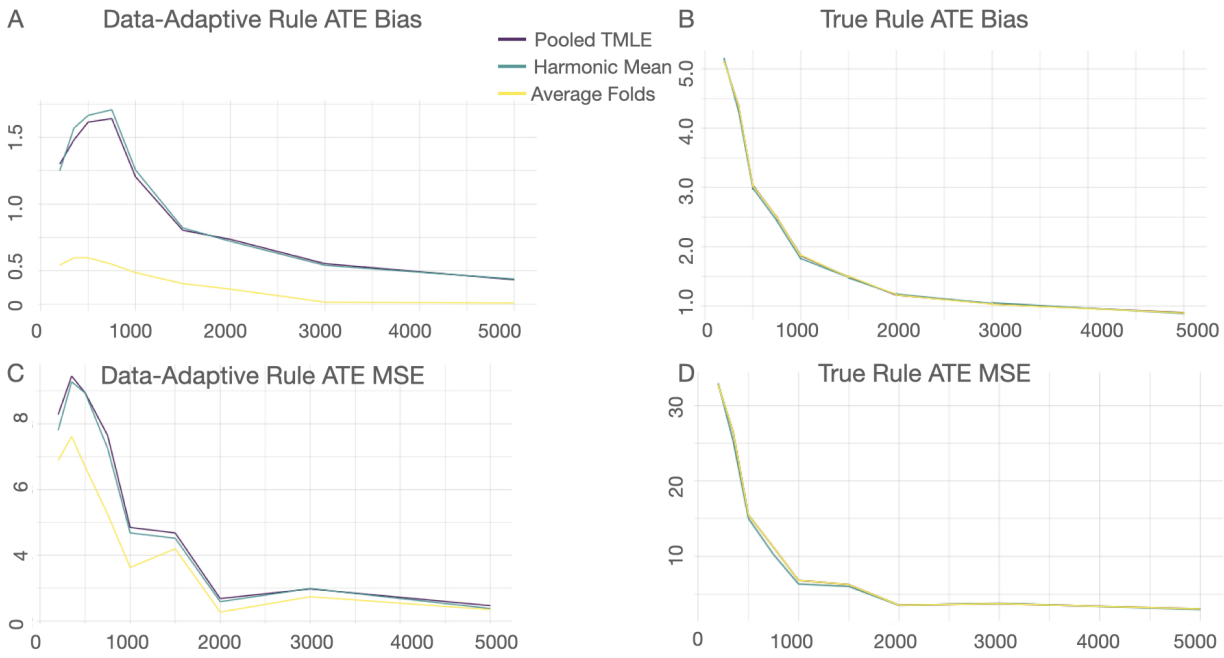


Figure 1.6: 3D Exposure Bias and MSE

CVtreeMLE Unbiasedly Estimates the Oracle Target Parameter

Now we look at comparing estimates to the true ARE given the oracle region in the DGP. **Figure 1.5 B and D** show the bias and MSE for this comparison in the two discrete exposures. As can be seen, both decrease at root n rate for all estimates. **Figure 1.6 B and D** likewise show this same rate of convergence for the three continuous exposure case. Based on these simulations, CVtreeMLE unbiasedly estimates the oracle target parameter at root n rates. We next look at the coverage. **Figure 1.7 B** shows coverage of the true ARE given the true region. The CIs calculated from the harmonic pooled k-fold standard errors had consistent 95% coverage, the k-fold specific CIs converged to 95% when sample sizes reached 1500 and the pooled TMLE CIs converged to 75% coverage. The same is shown for the three continuous exposures in **Figure 1.8 B** the inverse variance CI converges to 95% for coverage of the true ATE with the mean k-fold slightly lower around 82%. **Table 1.2** gives the bias, SD, MSE, and coverage for sample sizes 200, 1000, and 5000, comparing estimates to the data-adaptive truth.

CVtreeMLE has a Normal Sampling Distribution for Valid Inference

For our estimator to have valid inference, we must ensure that the estimator has a normal sampling distribution centered at 0 that gets more narrow as sample size increases. To

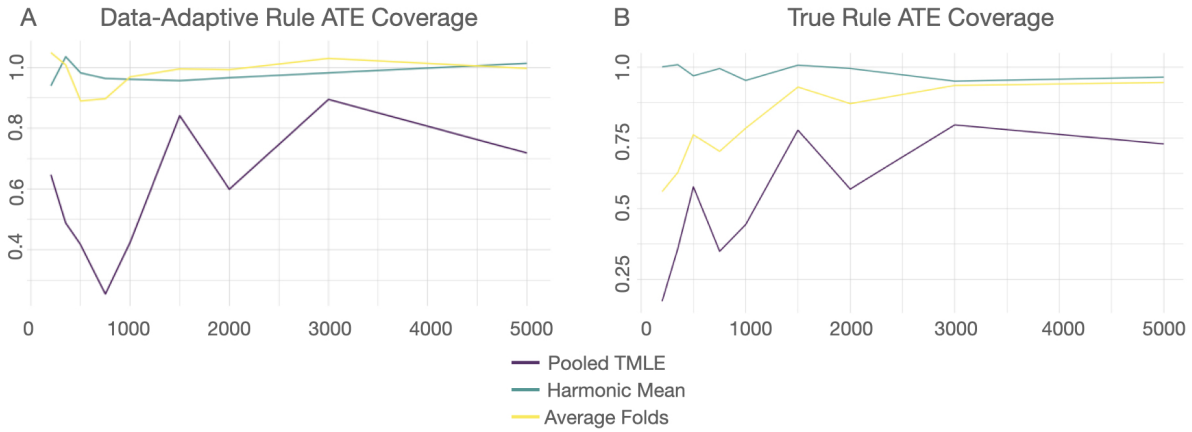


Figure 1.7: 2D Exposure Confidence Interval Coverage

confirm this, we next examine the empirical distribution of the standardized differences, (ψ_n - ψ^0)/SE(ψ_n), this is the ARE estimate bias compared to the true ARE given the true region divided by the standard error of the estimates over the iterations and ψ_n - ψ^{DA})/SE(ψ_n) which is the same standardized difference but compared to the resulting ARE when the data-adaptive region is applied to P_0. **Figure 1.9** shows the sampling distribution for each sample size with 50 iterations per sample size to estimate the probability density distribution of the standardized bias compared to the data-adaptive ARE. We see convergence to a mean 0 normal sampling distribution as sample size increases for all estimates. **Figure 1.9 A** shows the sampling distribution of the standardized bias of the mean k-fold AREs compared to the ground-truth ARE. We can see that this sampling distribution is quite tight around 0. **Figures 1.9 B and C** show the sampling distribution for the harmonic mean and the pooled TMLE estimates which are mirror reflections of each other. For both estimates, lower sample sizes (such as in purple $n = 200$) there is a wider spread of bias (estimates vary more widely) with z-scores out to 2 or 4 but this distribution gets tighter as sample size increases.

Likewise, **Figure 1.10 A-C** show the standardized bias of each estimate compared to the ground-truth region ARE. All estimates generally follow the same distribution and converge to a 0 mean normal distribution as sample size increases.

Table 1.1 shows the results of the simulations based on comparing the mean fold estimated ARE to the mean ARE of data-adaptive rules applied to P_0. It can be seen that the estimation

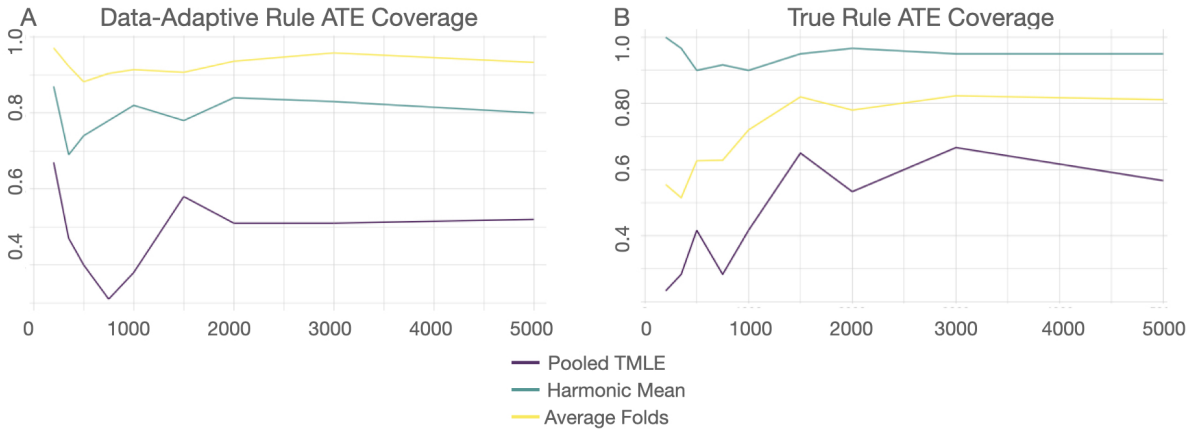


Figure 1.8: 3D Exposure CI Coverage

N	Absolute Bias	SD	MSE	Coverage
200	0.574	2.058	4.565	1
1000	0.379	1.458	2.268	0.97
5000	0.140	1.058	1.138	0.97

Table 1.1: Simulation results for Estimating the Data-Adaptive ARE using the Average k-fold Estimates

is unbiased, and the coverage of confidence intervals based IC-based estimates of the standard errors is slightly high. **Figure 1.11** shows the sampling distribution for each sample size for each type of estimate in the three continuous exposures. We see each estimate converge to a mean 0 normal sampling distribution as sample size increases with the average k-fold estimate having a tighter distribution.

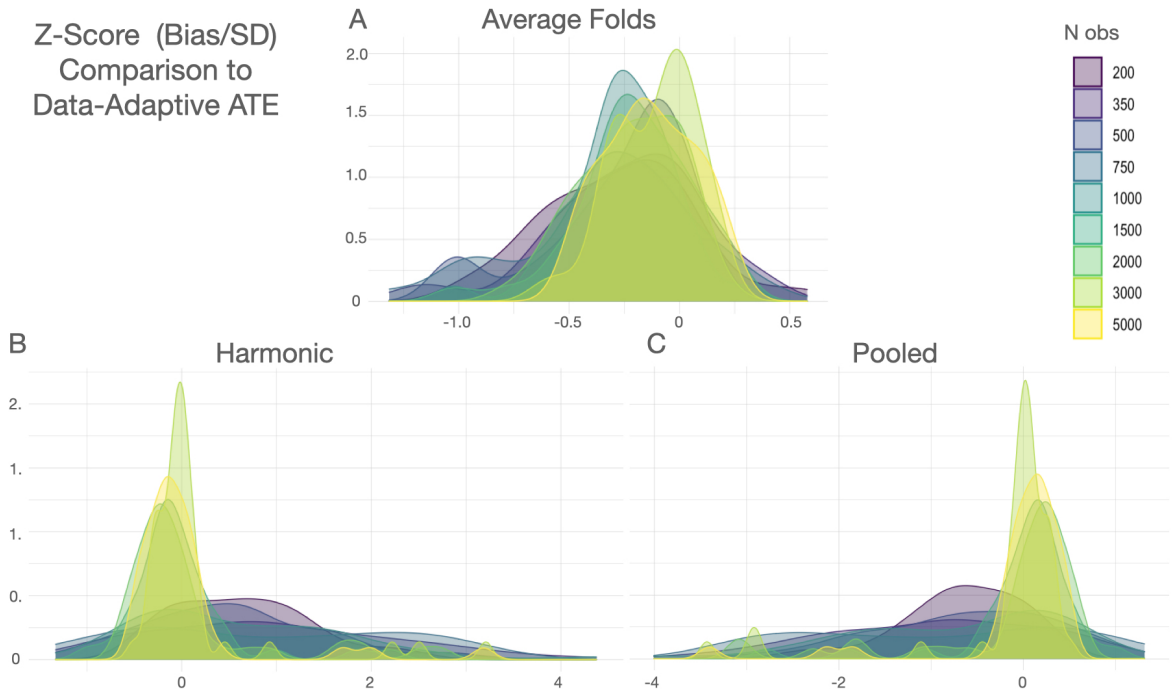


Figure 1.9: Bias Standardized by Standard Error Compared to ATE of Data-Adaptive Rule

N	Absolute Bias	SD	MSE	Coverage
200	0.608	2.10	4.797	0.95
1000	0.382	1.437	2.210	0.95
5000	0.178	0.894	0.831	0.96

Table 1.2: Simulation results for Estimating the Data-Adaptive ARE using the Average k-fold Estimates in Three Exposure Simulations

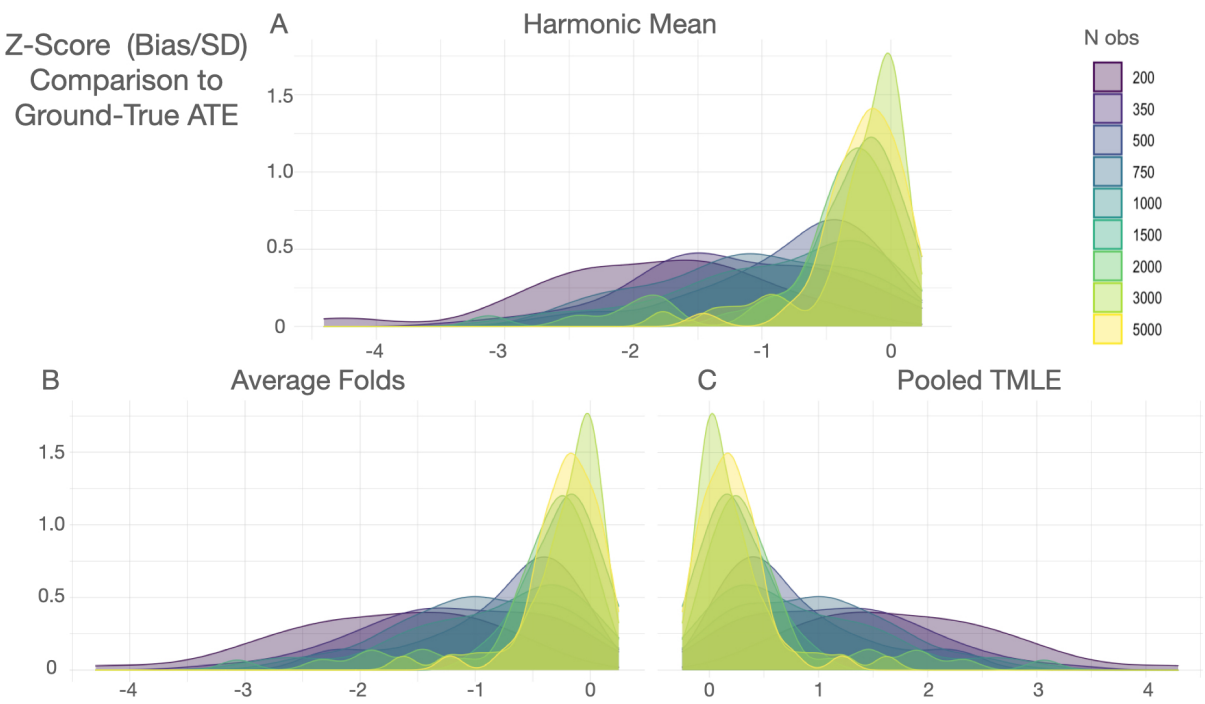


Figure 1.10: Bias Standardized by Standard Error Compared to ATE of True Rule

1.7 Applications

NIEHS Synthetic Mixtures

The NIEHS synthetic mixtures data (found here on github) is a commonly used data set to evaluated the performance of statistical methods for mixtures. This synthetic data can be considered the results of a prospective cohort study. The outcome cannot cause the exposures (as might occur in a cross-sectional study). Correlations between exposure variables can be thought of as caused by common sources or modes of exposure. The nuisance variable Z can be assumed to be a potential confounder and not a collider. There are 7 exposures ($X_1 - X_7$) which have a complicated dependency structure with a biologically-based dose response function based on endocrine disruption. For details the github page synthetic data key for data set 1 (used here) gives a description as to how the data was generated. Largely, there are two exposure clusters (X_1, X_2, X_3 and X_5, X_6). And therefore, correlations within these clusters are high. X_1, X_2, X_7 contribute positively to the outcome; X_4, X_5 contribute negatively; X_3 and X_6 do not have an impact on the outcome which makes rejecting these variables difficult given their correlations with cluster group members. This correlation and effects structure is biologically plausible as different congeners of a group of compounds (e.g., PCBs) may be highly correlated, but have different biological effects. There

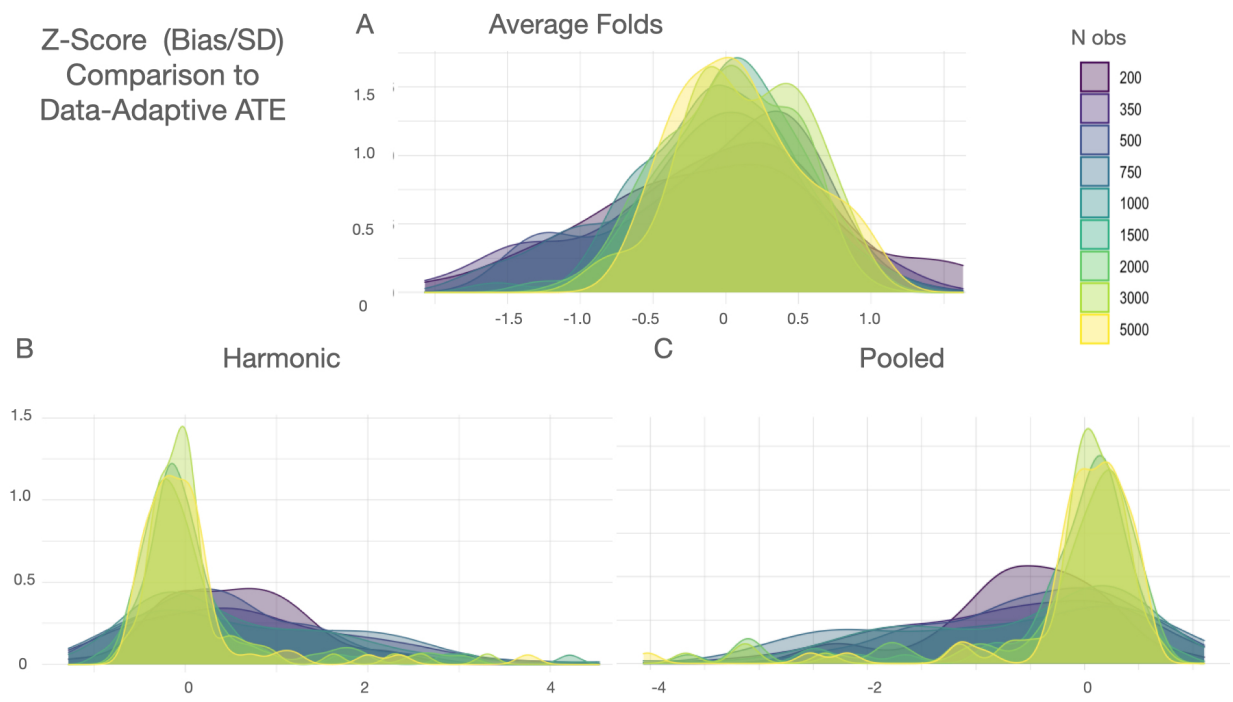


Figure 1.11: Bias Standardized by Standard Error Compared to ARE of Data-Adaptive Rule for Three Exposures

are various agonistic and antagonistic interactions that exist in the exposures. **Table 1.3** gives a breakdown of the variable sets and their relationships.

Variables	Interaction Type
X1 and X2	Toxic equivalency factor, a special case of concentration addition (both increase Y)
X1 and X4	Competitive antagonism (similarly for X2 and X4)
X1 and X5	Competitive antagonism (similarly for X2 and X4)
X1 and X7	Supra-additive ("synergy") (similarly for X2 and X7)
X4 and X5	Toxic equivalency factor, a type of concentration addition (both decrease y)
X4 and X7	Antagonism (unusual kind) (similarly for X5 and X7)

Table 1.3: NIEHS Synthetic Data Interactions

Given these toxicological interactions we can expect certain statistical interactions determined as cut-points for sets of variables from CVtreeMLE. For example, we might expect a positive ARE attached to a rule for $X_1 >= x_1$ & $X_2 >= x_2$ where x_1, x_2 are certain values

for the respective exposures because these two exposures both have a positive impact on Y. Likewise, in the case for antagonistic relationships such as in the case of X_2, X_4, we would expect a positive ARE attached to a rule $X_2 >= x_2$ & $X_4 <= x_4$. This is because we might expect the outcome to be highest in a region where X_2 is high and X_4 is low given the antagonistic interaction.

The NIEHS data set has 500 observations and 9 variables. Z is a binary confounder. Of course, in this data there is no ground-truth, like in the above simulations, but we can gauge CVtreeMLE's performance by determining if the correct variable sets are used in the interactions and if the correct variables are rejected. Because many machine learning algorithms will fail when fit with one predictor (in our case this happens for g(Z)), we simulate additional covariates that have no effects on the exposures or outcome but prevent these algorithms from breaking.

We apply CVtreeMLE to this NIEHS synthetic data using 10-fold CV and the default stacks of estimators used in the Super Learner for all parameters. We select for trees with positive coefficients in the ensemble during the data-adaptive estimation and therefore report results as positive AREs. We parallelize over the cross-validation to test computational run-time on a newer personal machine an analyst might be using.

Mixture ARE	Standard Error	Lower CI	Upper CI	P-value	P-value Adj	Region
8.24	0.56	7.14	9.34	0.00	0.00	X1 >= 0.267 & X5 <= 3.189
8.16	0.56	7.07	9.26	0.00	0.00	X1 >= 0.326 & X7 >= 0.22
6.68	0.62	5.46	7.89	0.00	0.00	X2 >= 0.602 & X5 <= 3.189
6.82	0.58	5.68	7.95	0.00	0.00	X2 >= 0.619 & X7 >= 1.171
7.29	0.51	6.29	8.29	0.00	0.00	X5 <= 3.269 & X7 >= 0.138

Table 1.4: NIEHS Synthetic Data Consistent Interaction Results

Table 1.4 shows the results from CVtreeMLE when applied to this NIEHS synthetic data set using the aforementioned settings. We filter results to only interactions that were found in all 10-folds, that is, trees with variable sets found across all the folds and therefore have consistent "signal" in the data. Let's focus on the second row with variables X_1 and X_7. **Table 1.3** shows that these two variables have a supra-additive or synergistic non-additive relationship. The union rule for trees including these two variables was $X1 >= 0.326$ & $X7 >= 0.22$ meaning this rule covers all observations indicated by the fold-specific rules. The mixture ARE is then interpreted as, if all individuals were exposed to X_1 at levels at or greater than 0.326 and exposed to levels of X_7 at or greater than 0.22 the outcome would be **8.16** units greater compared to if all individuals were exposed to levels less than these respective levels. The subsequent standard errors derived from the pooled influence curve (column 2) are used to derive the confidence intervals and p-values for hypothesis testing. Overall, comparing these statistical interactions to the toxicological interactions listed CVtreeMLE identifies 5 of 9 interactions. The other interactions in the above table are interpreted in the same way as the X_1 and X_7 interaction.

We next can investigate how consistent the results are across the folds by looking at the k-fold specific results, this gives us a sense of how reliable our ARE estimates are for the pooled rule. Let's dig deeper into this X_1 and X_7 interaction. **Table 1.5** shows the k-fold specific results for the interactions found for the variables X_1 and X_7. Each row is the results for each fold and the final row is the inverse variance weighted pooled result, pooling estimates across the folds. Estimates show stability across the folds with only one fold, fold 8, deviating from the trend. Cut-points at X_1 were either at 0.991 or 0.998 with fold 8 having a lower cut-point of 0.319. Likewise, X_7 was partitioned at 0.48 in most folds. Each fold-specific result has valid inference however it is also necessary to evaluate how consistent results were across the folds and thus determine if partitions are stable. Here we see the $X1 >= 0.99$ & $X7 >= 0.48$ partition for these two variables is stable and found in 8 of the folds. The ARE estimate for these rules ranges from 8-9 all with a significant effect. CVtreeMLE also provides plots of k-fold estimates to more easily assess for trends, **Figure 1.12** gives an example of this plot for the interaction X_1 and X_7.

Mixture ARE	SE	Lower CI	Upper CI	P-Value	P-Value Adj	Mix Rule	Fold
8.09	1.03	6.08	10.10	0.00	0.00	X1 > 0.998 & X7 > 0.48	1
7.12	1.63	3.92	10.32	0.00	0.00	X1 > 0.998 & X7 > 2.108	2
8.38	1.51	5.41	11.34	0.00	0.00	X1 > 0.991 & X7 > 0.48	3
9.42	2.23	5.05	13.80	0.00	0.00	X1 > 0.991 & X7 > 0.441	4
8.81	1.76	5.36	12.26	0.00	0.00	X1 > 0.998 & X7 > 0.439	5
8.43	1.53	5.42	11.43	0.00	0.00	X1 > 1.049 & X7 > 0.482	6
8.84	1.84	5.23	12.44	0.00	0.00	X1 > 0.991 & X7 > 0.413	7
5.20	2.69	-0.07	10.48	0.05	0.53	X7 > 0.217 & X1 > 0.319	8
8.41	1.04	6.36	10.46	0.00	0.00	X1 > 0.991 & X7 > 0.482	9
8.94	1.39	6.22	11.67	0.00	0.00	X1 > 0.991 & X7 > 0.482	10
8.30	5.49	-2.45	19.05	0.13	0.13	X1 >= 0.326 & X1 <= 4.687 & X7 >= 0.22 & X7 <= 4.886	Pooled

Table 1.5: X_1 and X_7 k-fold Interaction Results

Overall, CVtreeMLE is able to determine subspaces in the respective variables that have the most impact on the endocrine disrupting outcome. Of note is the fact that no interactions include the variables X_3 and X_6 both of which have no impact on the outcome.

Comparison to Existing Methods

Currently, quantile g-computation is a popular method for mixture analysis in environmental epidemiology. The method yields estimates of the effect of increasing all exposures by one quantile, simultaneously under linear model assumptions. Quantile g-computation looks like:

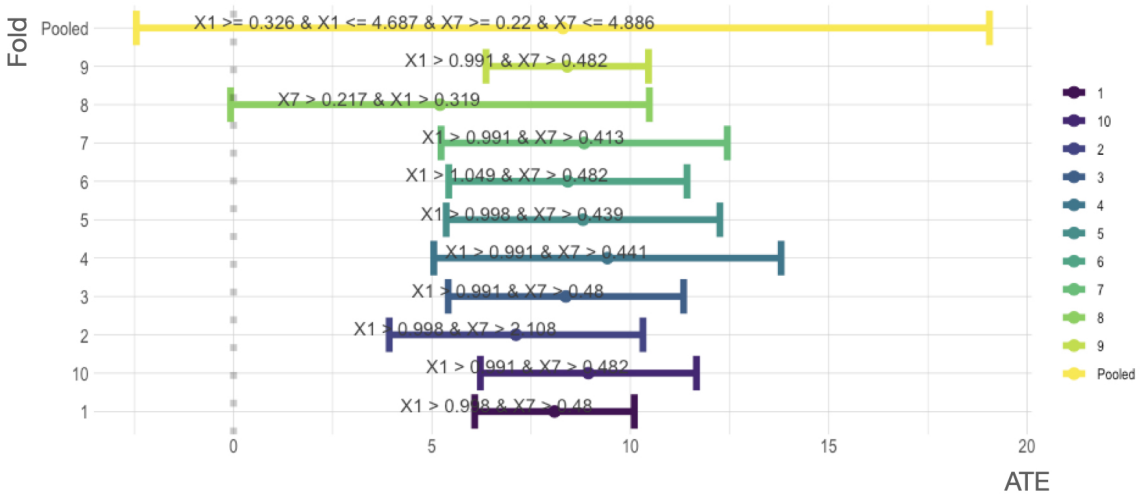


Figure 1.12: K-fold specific results for the interaction X_1 and X_7

$$Y_i = \beta_0 + \sum_{j=1}^{d} \beta_j X_{ji}^q + \beta Z_i + \epsilon_i$$

Where X^q are the quantized mixture components and Z are the covariates. Which works by first transforming mixture components into quantiles. Then the negative and positive coefficients from a linear model for the mixture components are summed to give a mixture (Ψ) summary measure which characterizes the joint impact. There are many assumptions that should be poignant after our discussion of mixtures. Firstly, quantiles may not characterize the exposure-response relationship (could be non-monotonic) which occurs in endocrine disrupting compounds. For interpretable weights and mixture estimate Ψ, assumes additive relationship of quantiles (Ψ is just sum of β's in front of mixture components). After our discussion, in mixtures our main goal is model possible interactions in the data because we expect exposures to have non-additive, possible non-monotonic, antagonistic and agonistic relationships. Therefore, we should expect interactions in our mixture data. In quantile g-computation, with the inclusion of interactions, the proportional contribution of an exposure to the overall effect then varies according to levels of other variables and therefore weights cannot be estimated. Because we can never assume no interactions, quantile g-computation then boils down to getting conditional expectations when setting mixtures

to quantiles through a linear model with interaction terms specified by the analyst. After our discussion of mixtures this should feel incorrect. As we argue, the important variables, relationships, and thresholds in a mixture are all unknown to the analyst which makes this a data-adaptive target parameter problem. Even testing quantile g-computation on the NIEHS data is difficult because we don't know what interactions to include *a priori*. The best we can do is run it out of the box and with two-way interactions and compare results to the ground-truth measures. Lastly, quantile g-computation does not flexibly control for covariates.

We run quantile g-computation on the NIEHS data using 4 quantiles with no interactions to investigate results using this model. The scaled effect size (positive direction, sum of positive coefficients) was 6.28 and included X_1, X_2, X_3, X_7 and the scaled effect size (negative direction, sum of negative coefficients) was -3.68 and included X_4, X_5, X_6. Compared to the NIEHS ground-truth, X_3, X_6 are incorrectly included in these estimates. However the positive and negative associations for the other variables are correct.

Next, because we expect interactions to exist in the mixture data, we would like to assess for them but the question is which interaction terms to include? Our best guess is to include interaction terms for all the exposures. We do this and show results in **Table 1.6**.

	Estimate	Std. Error	Lower CI	Upper CI	Pr(>|t|)
(Intercept)	21.29	1.58	18.19	24.39	0.00
psi1	0.02	1.62	-3.16	3.20	0.99
psi2	0.59	0.67	-0.71	1.90	0.37

Table 1.6: Quantile G-Computation Interaction Results from NIEHS Synthetic Data

In **Table 1.6** Ψ_1 is the summary measure for main effects and Ψ_2 for interactions. As can be seen, when including all interactions neither of the estimates are significant. Of course this is to be expected given the number of parameters in the model and sample size $n = 500$. However, moving forward with interaction assessment is difficult, if we were to assess for all 2-way interaction of 7 exposures the number of sets is 21 and with 3-way interactions is 35. We'd have to run this many models and then correct for multiple testing. Hopefully this example shows why mixtures are inherently a data-adaptive problem and why popular methods such as this, although succinct and interpretable, fall short even in a simple synthetic data set.

Bayesian kernel machine regression (BKMR) is a flexible method for mixed exposure analysis [8], implemented in the R package bkmr. We are able to directly compare our results to a study that applied BKMR to the same dataset, using their provided workbook of results. While BKMR, like our approach, identifies X_3 and X_6 as non-predictors, it does so by showing nonvarying cross-sectional plots for these exposures. Similarly, the user must choose the interactions of interest, which are then analyzed as bivariate plots. However, BKMR lacks

statistically rigorous summary measures for these marginal or interaction effects, unlike what we report in our CVtreeMLE method.

In BKMR, joint results are given for a quantile increase in all exposures. However, no information is provided about which exposures (and in which directions) these joint impacts occur. In contrast, our CVtreeMLE method explicitly considers and accounts for antagonistic relationships that nullify one another, providing a more nuanced understanding of the data.

Moreover, while BKMR offers flexibility, its results primarily consist of a series of plots and comparisons for chosen exposure quantile changes, with no succinct thresholds like the ones provided by CVtreeMLE. BKMR's interpretability relies heavily on the user's query, which can lead to selective reporting and potentially biased interpretations.

Although BKMR is a versatile method for mixed exposure analysis, it has limitations compared to our CVtreeMLE approach. BKMR lacks statistically rigorous summary measures based on proven asymptotic theory, does not provide comprehensive information about joint impacts of exposures, and relies on user queries, potentially introducing subjectivity and selective reporting. These limitations highlight the advantages of our CVtreeMLE method, which offers more robust and interpretable results for analyzing mixed exposures.

NHANES Data

Environmental chemical and metal exposure can affect telomere length, a biological marker that has been recognized as a significant mediator in the pathogenesis of adverse health outcomes, including several chronic diseases and cancers. Telomeres, the protective end caps of chromosomes, are crucial for maintaining genome stability. Their length, particularly in leukocytes (LTL), has been regarded as a barometer of biological aging, with implications for human health.

Shortened LTL has been linked to increased all-cause mortality and a range of age-related diseases such as cardiovascular disease and some types of cancer. Paradoxically, some studies also suggest longer LTL might be associated with an increased risk of certain cancers, pointing to a complex relationship that may vary across disease types. This counterintuitive link between longer LTL and certain cancers is thought to result from increased cellular proliferation and potential for malignant transformation.

Studies investigating the association of metals with LTL have predominantly focused on single-metal effects [114, 126, 15]. Some studies have examined the overall joint associations of metal mixtures with LTL using parametric models, such as multiple linear regression or quantile-sum g-computation [50, 60]. These methods, however, may not fully capture the complexity of exposure effects across the entire exposure space, especially when considering threshold effects.

Metal mixtures can cause oxidative stress which disrupts telomere length homeostasis, impacting cellular aging and disease. Given the significant environmental health burden of metal exposure, and the known involvement of LTL in disease pathogenesis, it's critical to extend our understanding of these joint associations. Our study aims to investigate potential thresholds in mixed metal exposure which might influence LTL. Such thresholds

could represent critical points of biological interaction, providing novel insights into exposure-related health risks and informing effective interventions.

Moreover, our objectives are two-fold: 1) to demonstrate the application of CVtreeMLE results on real-world data, and 2) to provide the data and data processing code through the open-source CVtreeMLE package. As part of this initiative, we develop a pipeline to download and clean National Health and Nutrition Examination Survey (NHANES) dataset, and provide this as a resource in the CVtreeMLE package.

We download and format the relevant NHANES 1999–2002 dataset containing demographic data, disease history, nine urine metals, and LTL. The demographic data used as possible confounders (W) include age, gender, race, education level, marital status, alcohol, smoking (cotinine) , body mass index (BMI), family poverty ratio (PIR), fasting glucose, systolic and diastolic blood pressure, exercise and birth country. Urine metal contained barium (Ba), cadmium (Cd), cobalt (Co), cesium (Cs), molybdenum (Mo), lead (Pb), antimony (Sb), thallium (Tl) and tungsen (W). These metals sampled in urine (as opposed to blood samples) were available in the NHANES data with accompanied LTL. The outcome is LTL. The number of observations in this test data is 2510. The coding pipeline and data are available in the CVtreeMLE package.

We apply CVtreeMLE using the default learners in each stack. We use 10-fold CV and set the max number of iterations in the iterative backfitting to 10 as well. Because previous research has shown the exposure to metals shortens LTL, we set the ATE direction to negative to select trees in the data-adaptive procedure which have the minimum (negative) impact and thus return negative ATEs for each fold.

Mixture ARE	Standard Error	Lower CI	Upper CI	P-value	P-value Adj	Union Region	% Fold
0.06	0.04	-0.02	0.13	0.17	1.00	cadmium >= 0 & cadmium <= 0.715 & molybdenum >= 19.8 & molybdenum <= 436.8	0.80
-0.03	0.01	-0.06	-0.01	0.02	0.37	cadmium >= 0.027 & cadmium <= 36.777 & thallium >= 0.01 & thallium <= 0.38	1.00

Table 1.7: Consistent Pooled TMLE Results NHANES Metal Mixture-LTL

Table 1.7 shows the pooled TMLE ARE results for rules found in more than 75% of the folds. Here we see rules including cadmium and thallium were found in all the folds and rules including cadmium and molybdenum were found in 80% of the folds. The cadmium-thallium interaction had a significant ARE of -0.03 and the cadmium-molybdenum was borderline significant with an ARE of 0.06. These results show that, exposure to high levels of cadmium >= 0.027 and low levels of thallium <= 0.38 is associated with a reduced telomere length of 0.03 compared to exposure levels of cadmium levels lower than 0.027 and thallium levels

greater than 0.38. This result implies an antagonistic relationship between cadmium and thallium. Likewise, telomere length was longer (0.06) for those exposed to low levels of cadmium <= 0.715 and high levels of molybdenum >= 19.8 compared to those exposed to the inverse exposure region for these two metals.

Like the NIEHS synthetic data results, we can investigate the k-fold specific results for these pooled results. Let's look at the cadmium and thallium interaction in each fold to see how stable the partition points were for each metal.

ARE	SE	Lower CI	Upper CI	P-Value	P-Value Adj	Region
-0.03	0.06	-0.15	0.09	0.67	1.00	thallium <= 0.21 & cadmium > 0.243
-0.04	0.03	-0.10	0.03	0.24	1.00	cadmium > 0.295 & thallium <= 0.31
-0.03	0.03	-0.09	0.03	0.32	1.00	thallium <= 0.21 & cadmium > 0.101
-0.05	0.04	-0.13	0.02	0.14	1.00	cadmium > 0.097 & thallium <= 0.38
-0.03	0.03	-0.09	0.03	0.37	1.00	cadmium > 0.295 & thallium <= 0.21
-0.03	0.05	-0.12	0.06	0.54	1.00	thallium <= 0.21 & cadmium > 0.143
-0.04	0.08	-0.20	0.12	0.62	1.00	cadmium > 0.29 & thallium <= 0.21
-0.04	0.05	-0.14	0.06	0.44	1.00	thallium <= 0.36 & cadmium > 0.092
-0.02	0.06	-0.13	0.10	0.76	1.00	cadmium > 0.027 & thallium <= 0.14
-0.03	0.04	-0.11	0.05	0.52	1.00	thallium <= 0.22 & cadmium > 0.254
-0.03	0.16	-0.34	0.27	0.83	0.83	cadmium >= 0.027 & thallium <= 0.38

Table 1.8: K-fold specific results for cadmium-thallium interactions associated with LTL

Table 1.8 shows the k-fold specific results for cadmium and thallium interaction. This interaction was found in all the folds with an ARE ranging from -0.02 to -0.05. None of the fold specific results were significant due to the variance estimates being calculated on the 251 observations in each validation fold, making standard errors high. However, we see consistent partitioning of thallium between 0.14 and 0.38 and partitioning of cadmium between 0.027 and 0.29. Overall, we see consistent cut-points across the folds which indicates this interaction is stable. The last row in this table is the inverse weighted pooled results. Here we can see that we gain much power by using the pooled influence curve in the pooled

TMLE procedure which is able to borrow variance information across the folds because all estimates are cross-estimated. Here, we can see the pooled estimated has much higher variance and wider confidence intervals.

ARE	SE	Lower CI	Upper CI	P-Value	P-Value Adj	Mix Rule	fold
0.05	0.04	-0.04	0.14	0.30	1.00	molybdenum > 55.2 & cadmium <= 0.384	2
0.04	0.10	-0.15	0.23	0.68	1.00	molybdenum > 52.9 & cadmium <= 0.368	3
0.01	0.04	-0.07	0.09	0.77	1.00	cadmium <= 0.715 & molybdenum > 19.7	4
0.11	0.14	-0.16	0.37	0.42	1.00	cadmium <= 0.35 & molybdenum > 102.5	5
0.05	0.05	-0.04	0.14	0.27	1.00	cadmium <= 0.292 & molybdenum > 57.2	6
0.02	0.22	-0.41	0.46	0.92	1.00	cadmium <= 0.124 & molybdenum > 21.7	7
0.02	0.06	-0.09	0.14	0.71	1.00	cadmium <= 0.429 & molybdenum > 44.5	8
0.14	0.28	-0.42	0.69	0.63	1.00	cadmium <= 0.131 & molybdenum > 55.6	9
0.04	0.41	-0.77	0.84	0.93	0.93	cadmium >= 0 & cadmium <= 0.715 & molybdenum >= 19.8 & molybdenum <= 436.8	Pooled

Table 1.9: K-fold specific results for cadmium-molybdenum interactions associated with LTL

Lastly, we look at the cadmium-molybdenum interactions in **Table 1.9**. As we can see here, interactions are not found in every fold and the partition points have a larger range although they all point in the same direction (low cadmium and high molybdenum) and all fold specific results are positive. This makes sense given that molybdenum processes proteins and genetic material like DNA and helps break down drugs and toxic substances that enter the body. Therefore, we would expect low cadmium and high molybdenum to be associated with longer telomere length. An association between molybdenum and longer LTL was found in [114].

Overall, in this NHANES example, we show that in real world data, CVtreeMLE can answer questions regarding expected outcomes under different exposure levels of a mixture which are otherwise occult given the limitation of existing methods.

1.8 Software

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 semi-parametric statistical theory into well-documented functional software is a formidable challenge. Such implementation requires understanding of causal inference, semi-parametric statistical theory, machine learning, and the intersection of these disciplines. The CVtreeMLE R package [63] provides researchers with an open-source tool for evaluating the causal effects of a mixed exposure using the methodology described here. 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. The NIEHS synthetic data and the NHANES mixed metal exposure data are provided. The NIEHS synethetic data application is used in the vignette of the package which makes these results reproducible to any researcher and likewise the NHANES data and code are provided for reproducibility. CVtreeMLE can run sequentially or parallelized across folds using the furrr package [104]. New statistical software using machine learning often presume the availability of significant computational resources in order to run in a timely manner. Here, our applications of NIEHS and NHANES were all run on a personal macbook machine in under 30 minutes by utilizing parallelization and using flexible yet efficient estimators. Of course, for the simulations high performance computing was used to parallelize iteration over clusters. To-date in scientific publication, the release of reproducible software is the exception rather than the rule. In an effort to make robust statistical software adopted in the future, rather than reliance of simple parameteric models, we make CVtreeMLE available with clear, easily accessible, highly detailed documentation of the coding methods. We also make all functions user-accessible, and develop numerous tests and examples. Coding notebooks show simulations of mixed exposure data and CVtreeMLE output with detailed summaries of interpretation. Lastly, the CVtreeMLE package is well maintained to ensure accessibility with ongoing improvements tested at each iteration. The CVtreeMLE package has been made publicly available via GitHub. A schematic that describes the CVtreeMLE method is shown in **Figure 1.13**.

1.9 Discussion

In this paper we introduce a new method for estimating the effects of a mixed exposure. Our approach treats ensemble decision trees as a data-adaptive target parameter for which we estimate the average effects of exposure for regions identified in the best fitting decision trees. This is done within a cross-validated framework paired with targeted learning of our target parameter which provides estimates that are asymptotically unbiased and have the lowest variance for studies which satisfy the unconfoundedness and positivity assumptions. Our proposed method provides valid confidence intervals without restrictions on the number of exposure, covariates, or the complexity of the data-generating process. Our method first partitions the exposure space into subspaces or regions that best explains the outcome. The output of our method is the exposure effect and respective confidence intervals if all individual were exposed to the exposure region compared to if all individual were not exposed to this region. Our approach has potentially many important applications including identifying

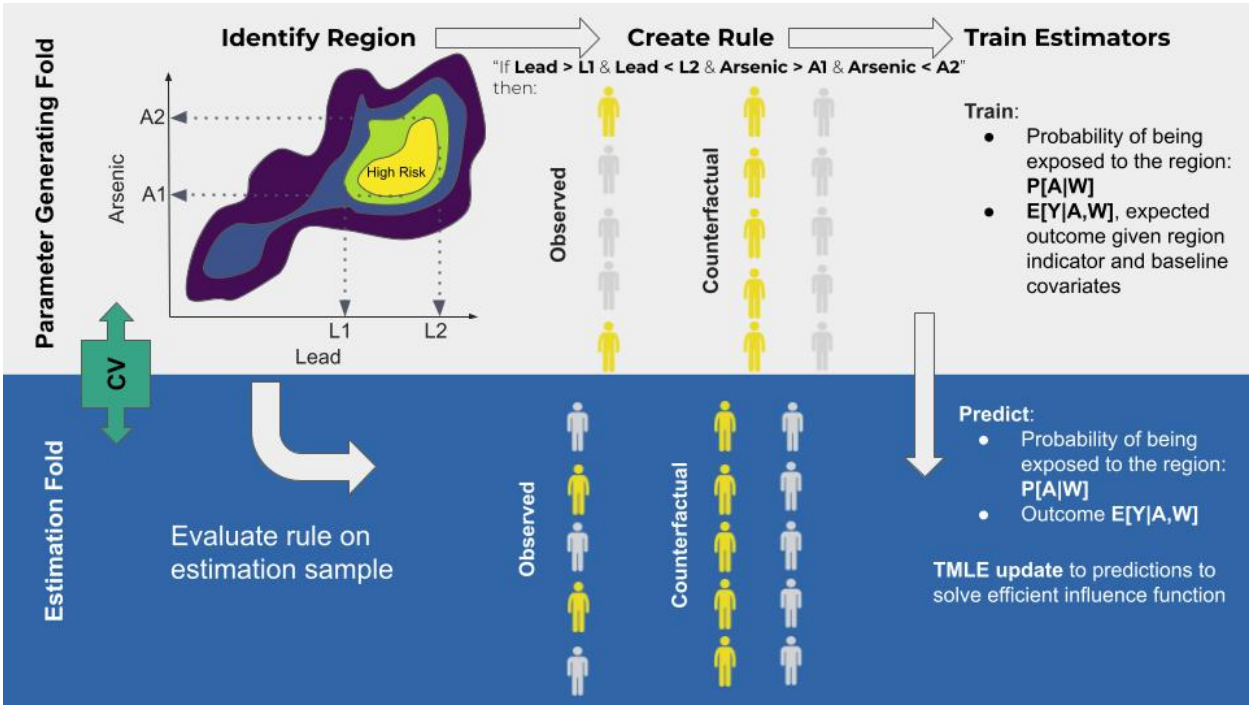


Figure 1.13: CVtreeMLE Schematic

what combinations of drugs lead to the most beneficial patient outcomes as well as finding what combinations of pollution chemicals have the most deleterious outcomes on public health. Our approach allows for "dredging with dignity" wherein exposure regions can be discovered in the data which are not known *a priori* and still provide unbiased estimates for the target parameter with valid confidence intervals. This approach of course comes with some cost as construction of a pooled region across the folds is rather ad hoc. This is the main limitation in the proposed method and other alternatives may exist such as using the average partitioning values of each exposure variable rather than our union approach which is conservative. Our simulations with ground-truth, NIEHS synethetic data and real-world data application show the robustness and interpretability of our approach. In an effort to make adoption of semi-parametric methods such as this more seamless we provide the CVtreeMLE R package on github which is well documented for analysts to apply to their respective data.

Chapter 2

SuperNOVA

In many fields, including environmental epidemiology and medicine, researchers strive to understand the joint impact of a mixture of exposures or treatments. This involves analyzing a vector of exposures or treatments rather than a single exposure, with the most significant individual exposures and exposure sets being unknown. Examining every possible interaction or effect modification in a high-dimensional vector of candidates can be challenging or even impossible. To address this challenge, we propose a method for the automatic identification and estimation of individual exposures in a mixture with explanatory power, baseline covariates that modify the impact of a mixture exposure or set of mixture variables, and sets of mixture exposures that have synergistic non-additive relationships. We define these parameters in a realistic nonparametric statistical model and use data-adaptive machine learning methods to identify these variables and variable sets while estimating the nuisance parameters for our target parameters of interest to avoid model misspecification. We establish a prespecified target parameter applied to variable sets when identified and use cross-validation procedures to train efficient estimators employing targeted maximum likelihood estimation for our target parameter given these sets. Our approach applies a shift intervention targeting individual variable importance, interaction, and effect modification based on the data-adaptively determined sets of variables. Our methodology is implemented in the open-source SuperNOVA package in R, enabling researchers to discover interaction and effect modification in a mixed exposure, providing robust statistical inference for these estimands without relying on arbitrary parametric assumptions. Our approach has broad applications across various fields and holds the potential to significantly advance our understanding of complex mixtures of exposures. We demonstrate the utility of our method through simulations, showing that our estimator is efficient and asymptotically linear under conditions requiring $n^{1/4}$-consistency of certain regression functions. We apply our method to the National Institute of Environmental Health Science mixtures workshop data, revealing correct identification of antagonistic and agonistic interactions built into the data. Additionally, we investigate the association between exposure to persistent organic pollutants and longer leukocyte telomere length and compare findings to previous results using other mixture methods.

2.1 Introduction

Individual health outcomes are influenced by the complex interplay of our environment and biology. We are exposed to multiple pollutants simultaneously, such as air pollution and contaminants in water. As a result, we should expect complex relationships to exist between these exposures, where certain exposures synergize to create super-additive effects on outcomes, and others antagonize or nullify the effects of exposure. Despite this, most epidemiological studies have focused on analyzing one pollutant at a time [66, 106, 115], partly due to the reliance on traditional generalized linear models (GLMs) that estimate each mixture component after controlling for others [4, 29, 120]. Such models are inadequate for accurately modeling complex multi-pollutant mixtures and understanding their relationships. The inadequacy of the use of parametric models to efficiently explore the space of possible interactions is a significant limitation in environmental exposure data analysis. Traditional methods, such as parametric models, struggle to capture complex relationships due to their rigid assumptions and constrained structure. GLMs attempt to model complex effects; however, their constrained nature restricts their ability to accurately represent the true underlying relationships. Furthermore, alternative approaches like tree-based methods do not provide explicit estimators of interactions, limiting their utility in deciphering intricate relationships between various factors. Consequently, there is a growing need for more flexible and adaptive methods that can better capture the intricacies of environmental exposure data, allowing for a more comprehensive understanding of the potential causal relationships and interactions at play.

In the context of analyzing mixture data using GLMs, it is critical to consider the limitations of this modeling approach. Mixture data is often characterized by high-dimensionality, multicollinearity, nonlinearity, and complex interactions among exposures. These features are well-established in the literature [10, 106, 117, 7]. When applying GLMs to such data, the analyst is essentially imposing a linear function on a non-linear relationship, which raises important questions about the interpretation of the resulting model coefficients. Specifically, what is the meaning of the beta coefficients obtained from this linear projection, and is there any utility in interpreting the coefficients in the presence of non-linearity and interactions? Furthermore, when estimating the joint impact of co-exposures, does it make sense to simply add up the coefficients obtained from the GLM, such as what is done in weighted quantile sum regression? These are important considerations that we address in this study as we introduce a new statistical method for analyzing mixture data.

In order to derive meaningful interpretations from statistical quantities, it is important to begin by framing the research question as a causal one. For example, when considering the impact of mixed exposures to persistent organic pollutants, the relevant questions may involve the effect of a specific dioxin level on leukocyte telomere length, the difference in leukocyte telomere length between individuals exposed to dioxin levels above and below EPA regulations, or the combined effect of increases in dioxin and polychlorinated biphenyl (PCB) exposure on leukocyte telomere length. To answer these questions, it is necessary to use estimators that make minimal assumptions and can capture the underlying functional forms,

including interactions and nonlinearity, of the mixture data. This requires the use of flexible algorithms to avoid unrealistic linear assumptions.

Analysts faced with high-dimensional and multicollinear exposure data often turn to methods like principal component analysis (PCA) [123, 82] or penalized regression [71, 116, 118, 65] to address these challenges. PCA produces a set of linearly uncorrelated variables that represent combinations of the original exposures, and these are then used as predictors in a linear model. Penalized regression models, such as least absolute shrinkage and selection operator (LASSO) or Ridge regression, can actively select a smaller set of exposures from the full exposure profile, based on the strength of their associations with the outcome. While these methods can help with issues of multicollinearity and high-dimensionality, the resulting quantities are often not easily interpretable in the context of causal questions that is informative for chemical regulation and public policy. For example, with PCA, it can be difficult to link the resulting principal components to specific exposures, and it may not be clear how the outcome changes with an increase in a particular exposure. Similarly, with penalized regression models, the selected subset of exposures may not represent a causal set, and interpreting the resulting coefficients may not be straightforward given the penalization imposed. Another popular approach in mixed exposure analysis involves: 1) reducing dimensionality through a technique such as PCA, 2) using a regression model on the reduced set of variables, 3) summing up the coefficients to estimate the joint effect of the mixture, and 4) possibly using Bayesian kernal machine regression (BKMR) to assess for nonlinearity/interaction. However, given evidence of nonlinearity and interaction found in many studies using the more flexible BKMR model, the question remains on how to interpret individual or summed coefficients in a misspecified linear model, and how these results address the proposed questions about mixed exposure.

In general, researchers need methods based on modified treatment policies. That is, methods that allow understanding of how various health related outcomes change given a change in exposure where the interventions go through a model that most accurately captures the underlying functional forms of the data. Therefore, the field needs non-parametric definitions for marginal effects, interaction and effect modification based on modified treatment policies. These estimates would provide answers for questions such as 1. what is the joint impact of multiple exposures on an outcome, 2. how does the joint impact of co-exposure compare to individual exposure (interaction)? and 3. is there heterogeneity in the impact of joint exposure/interaction across different subpopulations of individuals?. In each of these cases we need the target parameter to be defined outside considerations of the statistical model being used. For example, traditionally it has been taught that to assess for interaction to simply include an interaction term in a GLM. Of course, this is assessing for a multiplicative interaction between (normally) two exposures in a projected linear model with only main terms for the remaining exposures and covariates. Give there is likely nonlinearity in the exposures and covariates with possible interactions between both, once again we ask what does multiplying two exposures and estimating the coefficient of this term mean in a misspecified linear model? Furthermore, effect modification and interaction are measured in the same way with GLMs. Within the context of linear models, there is no statistical distinction

between interaction and effect modification. The difference arises from interpreting the interaction term used in the regression. However, within the non-parametric counterfactual framework these target parameters are different. Interaction is defined as the effects given an intervention on two or more exposures whereas effect modification is defined as the effect of one intervention varying across strata of a second unperturbed variable. [100]. If we consider two binary exposures, with observations defined as $O = (W, A, Y)$ where W are covariates, A is a binary exposure and Y is the outcome, we might define interaction in a mixture as the causal target parameter $E[E[Y|A_1 = 1, A_2 = 1|W] - [E[Y|A_1 = 1, A_2 = 0|W] - E[Y|A_1 = 0, A_2 = 1|W]] + E[Y|A_1 = 0, A_2 = 0|W]]$ which in the first term states we are taking the whole population and forcing individuals to be exposed to both exposures, the second forcing individuals to be exposed to A_1 and not A_2, the third term forcing individuals to be exposed to A_2 but not A_1 and the fourth term forcing individuals to not be exposed to A_1 or A_2. By subtracting the individual exposures from the joint we can estimate the greater impact a joint exposure has on the outcome. For such a target parameter we would then show that we can construct an unbiased and maximally efficient estimator under certain causal assumptions (i.e. no unmeasured confounding), but without imposing additional *modeling* assumptions. Targeted maximum likelihood and efficient estimating equations are two "recipes" for constructing such estimators that we could apply in this setting [57, 56] [31].

Of course, in most studies the mixture data is not composed of two binary exposures. Our target parameter should be able to handle a variety of data types (binary, multinomial, continuous) and many exposures. For a continuous exposure, there is no pathwise differentiable dose-response parameter in a nonparametric model. However, we can use stochastic exposure regimes, or stochastic interventions, which replaces f_A, a function that gives rise to the exposure A, and $g(A \mid W)$, the natural conditional density of A given covariates W, with a candidate density $g_{A_\delta}(A \mid W)$. More simply, we could replace the observed probability density distribution of, say, persistant organic pollutant (POP) exposure with that of a distribution if everyone were exposed an additional 1 pg/g unit. Because we are interested in the change in probability density, this approach works for all data-types and easily extends to multiple exposures, for example, observing how the joint density changes given an increase in a dioxin and furan, which were exposures of interest in mixture analysis workshops to study impacts on longer telomere length [68]. Using $d_x(A_x)$ notation to denote a shift in an exposures conditional density distribution, we can extend the original definition of interaction under joint binary intervention to: $E[Y(d_1(A1), d_2(A2)] - [E[Y(d_1(A1), A2)] + E[Y(A1, d_2(A2))]$ where now we are forcing the whole population to experience a shift in exposure(s) A_x for some shift δ. This is simply comparing the outcome under joint exposure shifts to the sum of individual shifts. This provides a non-parametric definition of interaction that is not dependent on a particular estimator.

The same limitations to binary treatment also exists in methods for estimating the heterogeneous treatment/exposure effects. One approach to estimating heterogenous treatment effects is through looking at the average treatment effect (ATE) within a specific subgroup of the population defined by a set of covariates, this is known as the conditional average treatment effect (CATE). In the context of environmental health and mixed exposures,

estimation of heterogenous exposure effects can help identify vulnerable populations that are differentially impacted by certain exposures or groups of exposures. These effects are typically associated with exposure effect modifiers (EEMs), which are covariates that modify the association of exposure on the outcome. Traditional parametric methods, such as GLMs, can detect EEMs under stringent conditions about the data-generating process, but when the posited functional relationship does not correspond to reality, inference is invalid. Detecting interactions in data involves a data-adaptive approach, which when relying on regression techniques, involves exploring various combinations of main effects and their tensor products. This process can involve trying different polynomial terms, interaction terms, and other transformations of the predictor variables to capture the complexity of the relationships between them. However, the resulting parameters reported from such an approach are heavily dependent on the model selection process.

More flexible semi-parametric methods have been developed and are described here [55]. However, the metaalgorithms discussed in this paper, such as the T-learner and S-learner, are limited to only binary treatment. The T-learner is a machine learning approach for estimating the CATE, which involves training separate prediction models on treated and control groups and then calculating the difference in their predictions for individual observations. The S-learner is a machine learning approach for estimating the CATE, which involves training a single prediction model on the entire dataset with treatment as an additional covariate and then calculating the difference in predictions with and without treatment for individual observations. While some researchers estimate heterogeneous treatment effects by analyzing ATEs for meaningful subgroups or using causal forests [105], [55] propose a new metaalgorithm, the X-learner, which builds on the T-learner and uses observed outcomes to estimate unobserved individual treatment effects. This involves training separate prediction models for treated and control groups, estimating individual treatment effects using both models, and then using a third model to learn the relationship between covariates and these estimated treatment effects to make individualized predictions. However, it is important to note that the T-, S-, and X-learners may not be suitable for estimating the CATE in the context of environmental exposures, where most chemicals are continuous, and alternative methods are needed.

Our approach is to investigate how the conditional average treatment effect for various stochastic shift interventions is different for certain vulnerable populations as an estimation strategy for effect modification. To do this, we can use an *a priori* specified estimator to find regions in the covariate space that best explain variance in the conditional treatment effects. That is, we simply regress the vector of expected individual counterfactual differences under a shift intervention onto the the covariates space using the optimal decision tree to identify interpretable regions that explain variability in the conditional average treatment effect. This approach then provides a semi-parametric definition of effect modification.

The problem only gets more complicated in mixtures. Non-parametric estimation of main effects, interactions, effect modification, and mediation for a continuous mixture of exposures (with more than two components), high-dimensional baseline covariates, and intermediates is impossible without large sample sizes. As discussed, existing methods reduce the dimension of

the problem artificially by imposing highly constrained models, resulting in biased estimations. No method currently exists to simultaneously estimate all these estimands motivated by causal parameters. Given these limitations, it is impossible to reduce the problem's dimension non-data-adaptively. Therefore, it is necessary to use the data both to define the specific parameters of interest and to estimate the data-adaptively defined parameters, which is an unavoidable part of the estimation problem. Addressing this issue with simplified models is not a viable solution.

Recent developments in both data-adaptive parameters and new estimators for the impact of relevant interventions on continuous exposures have paved the way for methodological research advancements. The goal of this paper is to develop an estimation machine that, under highly flexible (optimal) estimates of the data-generating distribution, can both identify which potential causal parameters have support in the data and use cross-validated targeted minimum loss-based estimation (CV-TMLE) to estimate and provide inference about them without overfitting. This approach will allow for a more accurate and unbiased analysis of the complex relationships in environmental exposure data. Furthermore, these estimates must be interpretable using modified treatment policies.

Identifying these relationships and estimating the target parameter is challenging, as both cannot be achieved using the same data without bias. To address this, sample splitting is necessary where one part of the data is used to identify the variable relationships, while the other is used to estimate the target parameters. We need a data-driven search of exposure space to find the sub-spaces that, the change of which, result in the largest change in the outcome, in a very large statistical model capable of capturing very complex relationships. To do this, we propose using an ensemble of basis spline models and selecting the best one. With this best fitting model, which is a linear combination of basis functions, we propose a nonparametric way to do ANOVA. This is analogies to a simple type III ANOVA and generalizes to big statistical models with joint exposures. Bases with interaction and effect modification variables that meet a specific F-statistic threshold are used to estimate the target parameters in the estimation data.

Our proposed method called SuperNOVA uses data-adaptive parameters [44] and cross-validated targeted minimum loss-based estimation (CV-TMLE) [121], resulting in a powerful approach for estimating the joint impact, interaction, and effect modification of a mixed exposure. By using CV-TMLE our estimates are guaranteed to have consistency, efficiency, and multiple robustness despite using highly flexible learners (ensemble machine learning) estimate a data-adaptive parameters. Data-adaptive parameters refer to parameters of interest that are identified and estimated through a data-driven process, in our case, the exposures are identified and a stochastic shift in these variables are estimated using the observed data. SuperNOVA is a statistical method that estimates the adjusted joint total impacts of multivariate exposures by searching for the sub-spaces that are "most important" for the outcome. This method generalizes associations away from discretized exposures to interpretable parameters that can handle continuous joint exposures. SuperNOVA has a built-in data-adaptive way for both estimation and parameter-generation to optimize for "complexity," and is capable of modeling complex nonlinear functions while still being able

to fit simpler models if supported by the data. The method estimates interactions using a combination of shift interventions among the subspaces of joint exposures found in the parameter-generating part and can also estimate effect modification for particular covariates data-adaptively identified. Additionally, SuperNOVA provides marginal impacts of shift exposures for the relevant subset of them determined in the parameter-generating part and provides robust inference despite using the data to both define the parameter and estimate it using CV-TMLE.

This manuscript is organized as follows, in Section 2.1 we give a background of semi-parametric methodology for stochastic interventions, in section 2.2 we discuss the target parameter for interaction given a fixed set of variables and in 2.3 we describe the effect modification parameter under a fixed set of variables, 2.4 describes assumptions necessary for our statistical estimates to have a causal interpretation. In section 3.1 we discuss estimation and inference of the interaction for a fixed variable set. In 3.2 we discuss estimation of the effect modification for a fixed variable set. In section 4 we discuss data-adaptively determining the variable set. In section 5 we show how this requires cross-estimation which builds from 2.2 for a fixed variable set. In section 5 we expand this to cross-estimation to k-fold CV and discuss methods for pooling estimates across the folds. Lastly, in section 5.3, because we may data-adaptively choose different deltas (when the choice of shift is also data-adaptively determined) across the folds and different variable sets chosen across the CV folds, we discuss the pooled estimates across the folds. In section 6 we discuss simulations of interactions and effect modification and show our estimates are asymptotically unbiased with normal sampling distributions. In section 7 we apply SuperNOVA to the NIEHS mixtures workshop data and identify interactions built into the synthetic data. In section 7.1 we compare SuperNOVA to exising mixture methods. In section 7.2 we apply SuperNOVA to NHANES data to determine if there is association between mixed POPs and leukocyte telomere length. Section 8 describes our SuperNOVA software. We end with a brief discussion of the SuperNOVA method in Section 9.

2.2 The Estimation Problem

Setup and Notation

Our setting is an observational study with baseline covariates ($W \in \mathbb{R}^p$), multiple exposures ($A \in \mathbb{R}^m$), and a single-timepoint outcome (Y). Let $O = (W, A, Y)$ denote the observable data. We use P_0 to denote the data-generating distribution. That is, each sample from P_0 results in a different realization of the data and if sampled many times we would eventually learn the true P_0 distribution. We assume our $O \sim P_0$ are n independent identically distributed (i.i.d.) observations of the random variable. We decompose the joint density as $p_{Y,A,W}(y, a, w) = p_{Y|A,W}(y, a, w)p_{A|W}(a, w)p_W(w)$ and make no assumptions about the forms of these densities. Our structural causal model (SCM) implied by the time ordering:

$$W = f_W(U_W), A = f_A(W, U_A), Y = f_Y(A, W, U_Y)$$

where $\{f_W, f_A, f_Y\}$ specify deterministic functions generating each variable $\{W, A, Y\}$ based on those preceding it and exogenous (unobserved) variables $\{U_W, U_A, U_Y\}$. We denote $g_0(A|W) \equiv p_0(A|W)$, the conditional probability density of exposure, $\bar{Q}(A, W) \equiv E_0(Y|A, W)$, the conditional outcome given exposure and covariates, and $q_{W,0}(W) \equiv P_0(W)$ the probability of covariates. Our statistical target parameter, $\Psi(P_0)$, is defined as a mapping from the statistical model, $\mathcal{M}$, to the parameter space (i.e., a real number) $\mathbb{R}$. That is, Ψ: $\mathcal{M} \to \mathbb{R}$. We can think of this as, if Ψ were given the true distribution P_0 it would provide us with our true estimand of interest.

We can think of our observed data $(O_1 \dots O_n)$ as generating a (random) probability distribution P_n that places probability mass $1/n$ at each observation O_i. Our goal is to obtain a good approximation of the estimand Ψ, thus we need an estimator, which is an a-priori specified algorithm that is defined as a mapping from the set of possible empirical distributions, P_n to the parameter space. More concretely, the estimator is a function that takes as input the observed data, a realization of P_n, and gives as output a value in the parameter space, which is the estimate, $\hat{\Psi}(P_n)$. Since the estimator $\hat{\Psi}$ is a function of the empirical distribution P_n, the estimator itself is a random variable with a sampling distribution. So, if we repeat the experiment of drawing n observations we would every time end up with a different realization of our estimate. We would like an estimator that is provably unbiased relative to the true (unknown) target parameter and which has the smallest possible sampling variance so that our estimation error is as small as it can be on average.

Defining the Differential Effect of a Modified Exposure Policy

In problems with a single exposure or treatment we can apply a stochastic shift on the exposure and get the average outcome under this post-intervention distribution under the existing stochastic shift framework. We denote counterfactual outcomes under a stochastic interventions as Y_{A_δ}. Stochastic interventions modify components of the SCM by replacing the equation that defines A and its natural conditional density $g(A \mid W)$ with a candidate density $g_{A_\delta}(A \mid W)$. In the absence of the intervention, A would be determined by a random draw from the distribution $g(A \mid W)$. With the intervention, A is stochastically modified by being drawn from the distribution defined by the candidate density $g_{A_\delta}(A \mid W)$. This can include static interventions, which place all mass on a single value. Static intervention parameter include the average treatment effect where all observations are set to receive or not receive treatment and we estimate the post-intervention outcome difference.

In terms of stochastic interventions, there are few restrictions on the choice of the candidate post-treatment density $g_{A_\delta}(A \mid W)$, but it is usually chosen based on the pre-intervention density $g(A \mid W)$. A popular approach is choose a function $d(A, W; \delta)$ which maps a pair $\{A, W\}$ to the post-intervention quantity A_δ. In these cases, the stochastic intervention is referred to as a modified treatment policy (MTP). For example, [73] examined how county-

level mobility affected COVID-19 cases. They estimated how much COVID-19 rates would change if the average mobility in each county increased by some amount (e.g. 5%) using stochastic interventions. In this setup, the policy does not dictate the mobility in each county: each county's post-intervention mobility depends on what level that county would otherwise have "self-selected" to without the policy (i.e. the level that was actually observed). The interpretations of the causal effects under MTPs and general stochastic interventions are discussed in detail in [48, 40, 38]. In our case, we are only additively changing the distribution such as shifting all POP exposure down by some δ and so we will use $g_0(a - \delta \mid w)$ to denote the shifted conditional probability density of $A|W$. We focus on a reduction since, in the context of most environmental exposures, a reduced expossure policy is normally of interest.

A stochastic intervention gives rise to a counterfactual random variable $Y_{A_\delta} := f_Y(A_\delta, W, U_Y)$, where the counterfactual outcome $Y_{A_\delta} \sim P_0^{A_\delta}$ arises from replacing the natural value of a treatment A with a shifted value A_δ. This shift is defined by a degree $\delta \in \mathbb{R}$, which describes how much the exposure A should be shifted in the context of the stratum W (the individual characteristics of a person). For example, if A is a continuous-valued dosage of a chemical exposure such as POPs, then the degree of shift δ can be interpreted as the reduction in pg/g of POPs an individual is exposed to. This reduction is from the natural exposure of POPs they would normally be exposed to, based on their baseline characteristics W.

We can evaluate the causal effect of our intervention by considering the counterfactual mean of the outcome under our stochastically modified intervention distribution. This target causal estimand is $\psi_{0,\delta} \coloneqq E_{P_0}^{A_\delta}\{Y_{A_\delta}\}$, the mean of the counterfactual outcome variable Y_{A_δ}. [48] describe the identification of this parameter for one exposure. In our case, we want to expand this identification to at least two variables. As such, for notational convenience we use $\mathcal{A}$ to denote a subset of exposures from A. That is, if A are air pollution exposures: carbon monoxide, lead, nitrogen oxides, ozone, particle matter 2.5, 10, and sulfur oxides, $\mathcal{A}$ may be subsets of these such as particle matter 2.5 and lead or carbon monoxide and ozone etc. Each of our target parameters involves a shift in $\mathcal{A}$, which may be one or two variables. We limit discussion to two exposures but estimating a shift in more than two exposures naturally follows.

We describe it briefly here. We must assume that the data is generated by independent and identically distributed units, and that there is no unmeasured confounding, consistency, or interference (discussed in more detail in subsequent sections). Under these assumptions, $\psi_{0,\delta}$ can be identified by a functional of the distribution of O:

$$\psi_{0,\delta} = \int_{\mathcal{W}} \int_{\mathcal{A}} E_{P_0}\{Y(\mathcal{A} = d(a, w)) \mid W = w\} g_{0,}(A = a \mid W = w) q_{0,W}(w)\ da\ dw$$

or

$$\psi_{0,\delta} = \int_{\mathcal{W}} \int_{\mathcal{A}} E_{P_0}\{Y(a) \mid W = w\} g_\delta(\mathcal{A} = a \mid W = w) q_{0,W}(w)\ da\ dw$$

which is the observed outcome under observed exposure integrated over the exposure density shifted by δ. In more compact notation:

$$\psi_{0,\delta} = \int_{\mathcal{W}} \int_{\mathcal{A}} \bar{Q}(A, W) g_\delta(A \mid W) q_W(w) \; da \; dw$$

Mechanically this is the outcome predictions from our Q model integrated over density predictions from our g model under δ shift integrated over our covariate density.

Our target parameters for interaction and effect modification build off this univariate stochastic shift intervention. Specifically our conditional average treatment effect is defined as the average of counterfactual differences in a covariate region. This is determined by regressing the individual expected outcome differences onto the covariate space: $Y_{A_\delta} - Y_a | W$. Thus identification for this parameter is the same but we are integrating in a subset of W. Our interaction target parameter requires estimates of the counterfactual outcome under a joint shift and individual shifts. The above identification framework holds for a joint intervention as well.

Target Parameter Causal Assumptions for Identification

Our above target parameter is defined on the causal data-generating process, so it remains to show that we can define it only in reference to observable quantities under certain assumptions. Each of our target parameters compares a shift on some subset of exposures to the observed outcome under observed exposures to understand how the outcome changes under a modified treatment policy. Here we give a brief description of each target parameter to show that each requires either an individual shift or joint shift in exposure. With this identified each target parameter can be estimated using the functional delta method.

For $\mathcal{A} = A_1, A_2$ our interaction parameter looks like:

$$E[E[Y_{\mathcal{A}_{\delta_1,\delta_2}}, W] - [E[Y_{A1_{\delta_1},A2}, W] + E[Y_{A1,A2_{\delta_2}}, W]]]$$

Which is the expected outcome under the joint shift of $\mathcal{A}$ compared to the expected additive outcome under each individual shift in $\mathcal{A}$.

For our effect modification parameter:

$$E[E[Y|\mathcal{A}_{\delta_1,\delta_2}, W)] - E[Y|\mathcal{A}, W)]]|W \in \mathcal{V}$$

Which is the expected mean difference in outcomes under a shift in $\mathcal{A}$ compared to the observed outcome in a region of the covariates W.

Each of these parameters requires a shift in one or two exposures or:

$$\psi = E[E[Y|\mathcal{A}_{\delta_1,\delta_2}, W]]$$

Where, in the univariate exposure shift case, one of the δ's is simply set to 0.

The above causal effects for this parameter are identified as long as the following assumptions hold:

1. Unconfoundedness: $Y(a)A|W$ for all $a \in \mathcal{A}$, where W is a set of pre-treatment variables that satisfies the backdoor criterion for $\mathcal{A}$ and Y.
2. Positivity: $\Pr(A = a|W = w) > 0$ for all $a \in \mathcal{A}$ and $w \in \mathcal{W}$, where $\mathcal{W}$ is the support of W.
3. Consistency: $Y = Y(a)$ for all $a \in \mathcal{A}$.
4. Conditional Exchangeability: $Y(a)Y(a')|A = a''$ for all $a, a', a'' \in \mathcal{A}$.

Our identification result shows that we can get at the *causal* counterfactual outcomes under stochastic interventions by estimates in *observable* data under certain conditions. Our goal is now to show how to efficiently estimate the observable interaction without imposing any additional assumptions (e.g. linearity, normality, etc.). While our identification assumptions may not always hold in all applications, we can at least eliminate all model misspecification bias and minimize random variation. Now that we've established how to estimate the interaction for a fixed set of exposures, we'll turn our attention to the problem of data-adaptively determining the variable set $\mathcal{A}$ and how to do so without incurring selection bias in estimating the interaction for those respective variables.

Estimating the Modified Exposure Policy for One Exposure or Treatment

In 2012, [48] derived the efficient influence function (EIF) for stochastic interventions in a nonparametric model $\mathcal{M}$, which is the essential quantity needed for constructing pathwise differetiable parameters and deriving variance estimates. Given this EIF derivation they also developed substitution, inverse probability weighted, one-step, and targeted maximum likelihood (TML) estimators. These estimators allow for semiparametric-efficient estimation and inference on the target quantity of interest $\psi_{0,\delta}$. The EIF is expressed as:

$$D(P_0)(x) = H(a, w)(y - \overline{Q}(a, w)) + \overline{Q}(d(a, w), w) - \Psi(P_0),$$

with,

$$H(a, w) = \mathbb{I}(a + \delta < u(w))\frac{g_0(a - \delta \mid w)}{g_0(a \mid w)} + \mathbb{I}(a + \delta \geq u(w))$$

Here, the auxiliary covariate $H(a, w)$ is a ratio of conditional densities $p(a|W)$. $g_0(a \mid w)$ is the conditional density of A under the observed values and $g_0(a - \delta \mid w)$ is the density under a decreased of δ to a. Thus, the ratio indicates how much the conditional density "moves" under a δ shift with larger values indicating large discrepancies between the shifted and unshifted conditional densities.

Because TMLE is a plug-in estimator and often performs better for smaller samples relative to alternatives [61, 93, 57, 56] we use the TMLE estimator to solve this EIF for our stochastic intervention target parameters.

Construction of the TMLE estimator has the following steps:

1. Use data-adaptive regression techniques to create initial estimates of $g_0(A, W)$ and $\overline{Q}_0(A, W)$.

2. For each observation, calculate an estimate of the auxiliary covariate $H(a_i, w_i)$ as described.

3. Using the estimates of the auxiliary covariate, create a one-dimensional logistic regression model, and estimate the parameter ϵ in the model by:

   a) Regressing Y onto H w/ offset for initial predictions $\bar{Q}(A, W)$.

   b) Update the counterfactual estimates under shift using $\text{logit}\overline{Q}_{\epsilon,n}(a - \delta, w) = \text{logit}\overline{Q}_n(a - \delta, w) + \epsilon H_n(a - \delta, w)$ once we know ϵ

   The outcome of this regression model yields $\overline{Q}_n^\star$.

4. Calculate the TML estimator Ψ_n of the target parameter by taking the average of the estimates of $\overline{Q}_n^\star$ for each observation.

The constructed TMLE estimator for a stochastic intervention is asymptotically linear and doubly robust. The central limit theorem then states that the distribution of the estimator ψ_n is centered at ψ_0 and is Gaussian. An estimate of the variance σ_n^2 can be computed, allowing for Wald-style confidence intervals to be computed at a coverage level of $(1 - \alpha)$ as $\psi_n \pm z_{(1-\alpha/2)} \cdot \sigma_n/\sqrt{n}$. Additionally, resampling based on the bootstrap may also be used to calculate the variance σ_n^2 in certain conditions.

Extension of the TMLE Estimator for Multiple Exposure Shifts

In a mixed exposure setting with multiple continuous exposures $A \in \mathbb{R}^m$, we define a potential outcome $Y(a)$ for each of the infinite possible values of each exposure. Rather than focusing on a stochastic shift in one exposure, we examine the differential policy effect of shifting different exposure sets in the exposure space by some amount δ_i. For instance, we might be interested in the effect of a regulation that concurrently reduces dioxins by δ_1 and furans by δ_2. Both dioxins and furans are classes of toxic environmental pollutants that often coexist due to their similar sources, such as industrial processes, waste incineration, and natural disasters like forest fires. Importantly, prior research suggests that these chemicals may contribute to the shortening of leukocyte telomere length (LTL), a biomarker associated with cellular aging and increased disease risk [101]. Therefore, a joint reduction in these pollutants due to policy intervention could be associated with observable changes in LTL, providing a biologically plausible and measurable outcome of interest for our analysis.

From a policy standpoint, we may want to know how a joint reduction in both dioxins and furans affects an outcome compared to the summed effect if dioxin and furan were changed individually. This information could inform policy to concurrently enforce regulations on

both toxic POPs if a simultaneous decrease in exposure results in better health outcomes compared to a sequential reduction. In the bivariate case we can express this parameter as:

$$\underbrace{E[Y_{\mathcal{A}_{\delta_1,\delta_2}}]}_{\text{1. Joint Shift of A1 and A2}} - \underbrace{E[Y_{\mathcal{A}_{\delta_1,0}}]}_{\text{2. Individual Shift of A1}} - \underbrace{E[Y_{\mathcal{A}_{0,\delta_2}}]}_{\text{3. Individual Shift of A2}} + \underbrace{E[Y_{\mathcal{A}}]}_{\text{4. No Shift of A1 and A2}}$$

$$\begin{aligned}
&\int_{\mathcal{W}}\int_{\mathcal{A}} Q(a,w)g_{\delta_1,\delta_2}(a,w)p(w)\ da\ dw - \\
&\Bigg[\int_{\mathcal{W}}\int_{\mathcal{A}} Q(a,w)g_{\delta_1,0}(a,w)p(w)\ da\ dw \\
&+ \int_{\mathcal{W}}\int_{\mathcal{A}} Q(a,w)g_{0,\delta_2}(a,w)p(w)\ da\ dw\Bigg]
\end{aligned}$$

In this expression, the first term represents the expected outcome under a joint shift of exposures to dioxins A_1 by δ_1 and furans A_2 by δ_2. The terms inside the brackets represent the expected outcomes under shifts for each exposure. By estimating this parameter, we can evaluate the combined impact of interventions on both exposures simultaneously and compare it to the separate effects of each exposure individually.

For instance, both dioxins and furans can originate from waste incineration processes. If an intervention aimed at reducing both dioxin and furan emissions from waste incinerators leads to substantially better health outcomes than interventions targeting dioxin and furan emissions individually, policy measures might focus on implementing stricter controls on waste incineration. On the other hand, if dioxins and furans are emitted from different sources, but their combined effect on health is considerable, policies might need to target multiple sources simultaneously to achieve significant health improvements. Thus, our approach can inform policies that target multiple exposures at once and pinpoint the specific sources of these exposures, ultimately contributing to more effective strategies for public health protection.

With this goal in mind, the parameter above compares the expected outcome under joint shift to the sum of the individual shifts. The 2. and 3. terms in our above parameter are estimated given the methodology described in section 2.3. The only addition we need to make to our estimator is for the joint shift. However, our TMLE estimator is easily extended to the joint shift case. As discussed, in section 2.3, we need a density estimator $g_0(a \mid w)$ to obtain density values of A given W under both observed and shifted conditions - this is used in construction of the "clever covariate". In the joint shift case this clever covariate then becomes a ratio of joint densities. Rather than go after the joint density directly, for which there are very few if any estimators, we estimate joint density by taking the product of conditional and marginal densities. In the bivariate case, the joint density of $A1, A2$ is estimated by:

$$g_0(A1, A2|W) = g_0(A1|W) \cdot g_0(A2|A1, W) \tag{2.1}$$

Here we simply construct two density estimators using ensemble machine learning 1. which estimates the density of $A1$ given W and the other which estimates the density of $A2$ given $A1$ and W, the product of these two densities gives us the joint density.

This forms the g portion of the likelihood we need for our TMLE estimator; we also need the Q. In the joint shift scenario, $\overline{Q}(d(a,w),w)$ is the expected Y under a joint shift. Here, we simply fit a Super Learner and get predictions under a joint shift. These are then used as our initial estimates in the least favorable submodel along with the clever covariate constructed as the ratio of joint densities. As we can see, estimating $E[Y_{A1_{\delta_1},A2_{\delta_2}}]$ follows the existing TMLE framework for stochastic shift interventions with only slight modifications to Q and g. We can then plug-in our individual estimates for each term in our interaction target parameter to get the interaction estimate and use the delta method to combine EIFs to get estimates of variance.

2.3 Estimating Effect Modification

Conditional Average Exposure Effects after Stochastic Intervention

As discussed, we aim to investigate the impact of a stochastic shift intervention in a subregion of covariates, denoted as $\mathcal{V}$ as a measure of conditional average treatment effects (CATE). To measure the impact of the intervention, we compute the ATE in the subpopulation within the region $\mathcal{V}$. We describe our approach below:

1. Define $Q(a,w) = E[Y(a)|W=w]$ as the expected outcome under treatment level a given covariates $W=w$.

2. Define $g_\delta(a,w) = p_\delta(A=a|W=w) = p(A=a-\delta|W=w) = g(a-\delta,w)$ as the probability density function of the shifted treatment assignment, where δ represents the amount of shift.

3. Let $p(w)$ be the marginal density of W at $W=w$.

4. Define the covariate space as $\mathcal{W}$, where the random variable W takes values. Similarly, let $A \in \mathcal{A}$.

5. The shift ATE is computed as the difference between the expected outcomes under the shifted treatment assignment and the original treatment assignment:

$$\int_{\mathcal{W}} \int_{\mathcal{A}} Q(a,w) g_\delta(a,w) p(w)\ da\ dw$$
$$- \int_{\mathcal{W}} \int_{\mathcal{A}} Q(a,w) g(a,w) p(w)\ da\ dw$$

6. To compute the shift ATE for a subpopulation defined by the region $\mathcal{V} \subseteq \mathcal{W}$, we restrict the integration to the region $\mathcal{V}$ and normalize by the probability of $W \in \mathcal{V}$:

$$\int_{\mathcal{V}} \int_{\mathcal{A}} Q(a,w) g_\delta(a,w) p(w)\ da\ dw - \int_{\mathcal{V}} \int_{\mathcal{A}} Q(a,w) g(a,w) p(w)\ da\ dw$$

More concisely, let $P(W,A,Y) = P(Y|A,W)P(A|W)P(W)$ represent the original distribution, and let $P_{\delta,\mathcal{V}}(W,A,Y) = P(Y|A,W)P(A-\delta|W)P(W|W \in \mathcal{V})$ represent the distribution under the shift δ and restricted to the subpopulation with $W \in \mathcal{V}$. Then, the regional shift ATE is given by:

$$\psi_{\delta,\mathcal{V}} = E_{P_{\delta,\mathcal{V}}}[Y]$$

The difference $\psi_{\delta,\mathcal{V}} - \psi_{0,\mathcal{V}}$ represents the regional shift ATE.

Importantly, $\mathcal{V}$ is not known *a priori* and must be discovered in the data. It is most impactful to public policy or pharmaceutical development if this $\mathcal{V}$ was a region in W where the effects of an intervention are most different compared to the complimentary covariate space $\mathcal{V}^c$. With this goal in mind, we can find regions in the covariate space that best explain the variability of the individual exposure effects across observations. In these respective covariate regions $(v_1, v_2, v_3..., v_p)$ we want to take calculate the average stochastic intervention effects:

$$\hat{\tau}_{v_j} = \frac{1}{n_j} \sum_{i \in v_j} Y_i(\mathcal{A}_\delta) - Y_i(\mathcal{A}), \qquad j = 1, 2, ..., p$$

where $\hat{\tau} v_j$ is the estimated stochastic exposure effect for region j, n_j is the number of individuals in region j, $Y_i(\mathcal{A}_\delta)$ and $Y_i(\mathcal{A})$ are the potential outcomes for individual i under shift and no shift, respectively, and v_j is the set of covariate values that define region j. The estimator $\hat{\tau} v_j$ takes the average treatment effect of all individuals in region j, which is the difference between the expected outcome under changed exposure and the expected outcome under unchanged exposure. The regions $v_1, v_2, ..., v_p$ are defined by the covariate values that explain the most variance in the treatment effects.

Using a shallow decision tree to find $\mathcal{V}$ is a convenient approach in this case to uncover heterogeneity in the stochastic shift parameter for several reasons 1. Interpretability: shallow decision trees are often more interpretable than deeper trees or other complex models. This is because they involve fewer splits, making it easier for researchers to understand the relationships between the covariates and the resulting subgroups. This interpretability can be valuable for communicating results to stakeholders and for guiding policy decisions; 2. Reduced risk of overfitting: a shallow decision tree is less likely to overfit the data compared to deeper trees, as it captures only the most significant splits in the covariates. By limiting the complexity of the tree, we avoid capturing noise in the data and better generalize to new, unseen data; 3. Computational efficiency: shallow decision trees are computationally efficient to train and evaluate, as they involve fewer nodes and splits than deeper trees. This efficiency can be especially beneficial when working with large datasets or when computational resources are limited; 4. Detection of interacting effects: shallow decision trees prioritize

the most important interactions between covariates, which can help to identify the primary sources of heterogeneity in the stochastic shift parameter. This focus on combinations of covariates (such as men older than a particular age who smoke) which can help researchers to understand which combination pf factors have the most significant influence on the treatment effect; 5. Robustness to multicollinearity: decision trees are less sensitive to multicollinearity compared to other methods such as linear regression. This means that a shallow decision tree can still perform well even when covariates are highly correlated, making it more robust in finding the region $\mathcal{V}$ that exhibits heterogeneity in the stochastic shift parameter.

We rely on the fact that shallow decision trees don't "overfit", that is, trees fall in a Donsker class. [98] to avoid fitting trees in each fold of the cross-validation procedure (discussed next). Instead we identify these regions on the full data and estimate effects in each $\mathcal{V}$, we descrive the cross-validation procedure first and then return to our estimation approach for the CATE.

2.4 Discovering Variable Relationships

Basis Function Estimators for Variable Relationship Discovery

As discussed, we do not know *a priori* what variable sets to apply our various target parameters. That is, we need a non-parametric method of finding variable sets such as A_1, A_2 (two exposures), A_4 (one exposure), A_2, W_5 (one exposure and one baseline covariate) that are predictive of an outcome. To accomplish this, we use a discrete Super Learner which only considers models for the conditional mean of Y^* given A, W that are a function of basis spline terms and their tensor products. These estimators construct non-linear models using linear combinations of basis functions. In the most flexible setting we would create indicator variables can be used to indicate if a variable X is less than or equal to a specific value x_s. The same can be done for X_1, X_2 to determine if both variable are less than or equal to a specific value. Thus, a function of our outcome is written as:

$$\psi_\beta = \beta_0 + \sum_{s \subset \{1,2,\ldots,p\}} \sum_{i=1}^{n} \beta_{s,i} \phi_{s,i}, \text{ where } \phi_{s,i} = I(X_{i,s} \leq x_s), A \in \mathbb{R}^p$$

and s denotes indices of subsets of the Z (e.g., both functions of single variables and two variables).

Estimators used in SuperNOVA that return tensor products of arbitrary order include the earth [67], polySpline [81] and hal9001 [16] package. Each of these methods uses a linear combination of basis functions to estimate the conditional outcome and we can therefore extract variable sets used in these basis functions as our data-adaptively identified variable sets.

Non-Parametric Analysis of Variance

Given the best fitting basis spline estimator has been selected, we need a method for ranking the "important" variable sets used by the algorithm. We propose a non-parametric approach to the analysis of variance (ANOVA). This approach is analogies to a simple ANOVA (and augmentations that adjust for covariates/interaction in a linear model such as Type III ANOVA). However, we generalize the ANOVA to big statistical models with joint exposures and variety of parameters that are relevant. Here, we use ANOVA decomposition as a general tool for partitioning the variance of a response variable into various components based on different basis factors. For example, in the case of MARS being selected as the best estimator, the ANOVA is used to decompose the variance of the response variable into the contributions of the individual linear basis functions, allowing for an assessment of their relative importance in explaining the response. In the case of HAL being selected, the variance of the response variable can be decomposed into the contributions of the individual zero order basis functions (indicators of exposure-covariate regions).

In this case, the F-statistic is calculated using the same formula as in a traditional ANOVA, but with a few modifications to account for the fact that the model is not linear in the original covariates.

The basic idea is to partition the variance in the response variable into two components: the variance explained by the linear combination of basis functions, and the residual variance that is not explained by the model. The F-statistic is then calculated as the ratio of the explained variance to the residual variance, adjusted for the degrees of freedom.

More specifically, suppose we have a linear combination of basis functions of the form:

$$\hat{f}(\boldsymbol{x}) = \sum_{p=1}^{P} \beta_p B_p(\boldsymbol{x})$$

where $B_p(x)$ are the basis functions, and β_p are the corresponding coefficients to be estimated. We can then fit this model using least squares regression, and calculate the

residual sum of squares (RSS) [1] which measures the proportion of the total variance in the response variable that is explained by the model. The degrees of freedom for the model and the residual are given by p and $n - p - 1$, respectively, where n is the sample size and P is the number of basis functions used in the linear model. The F-statistic is then calculated as:

$$\text{F-statistic} = \frac{\frac{ESS}{P-1}}{\frac{RSS}{n-P}}$$

which follows an F-distribution with p and $n - p - 1$ degrees of freedom under the null hypothesis that the model has no predictive power. Here, the basis functions need not be linear in the original covariates, but the relationship between the response variable and the basis functions is still assumed to be linear.

An F-statistic value is calculated for each basis function. We then sum the F-statistics for basis functions that include the same variable sets. This aggregated statistic is used to rank the variable sets in order of importance for model fit. The variable sets can be subseted based on the F-statistic quantile to select for a more concise list of variables. The output from this process is a list of variable sets that meet the F-statistic threshold. For example, $(A_2\&A_3, A_1\&W_2, A_4)$ would indicate the basis functions for two esposures $(A_2\&A_3)$, an exposure and effect modifier $(A_1\&W_2)$ and one exposure (A_4) met the specified F-statistic threshold and will be used to estimate the interaction parameter (for the two exposures), effect modification parameter (for the exposure and effect modifier) and marginal impact (for the one exposure). This variable set procedure is conducted using the parameter-generating sample in a V-fold cross-validation framework with data. This framework is discussed next.

2.5 Cross-Estimation

To achieve a consistent estimator of our various stochastic shift target parameters, it is necessary to assume certain complexity conditions of the nuisance functions. Specifically, they

[1] Below we give a quick refresher for how RSS, TSS, and ESS are calculated:

$$RSS = \sum_{i=1}^{n} \left(y_i - \hat{f}(\boldsymbol{x}_i)\right)^2$$

where y_i are the observed values of the response variable. We can also calculate the total sum of squares (TSS) as:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

where $\bar{y}$ is the mean of the response variable. The explained sum of squares (ESS) is then given by:

$$ESS = TSS - RSS$$

must be smooth (i.e., differentiable) and their entropy must be sufficiently small to satisfy the Donsker conditions (e.g., if we assume Lipschitz parametric functions or VC classes). However, in high-dimensional settings (p > n) or when using ML methods that are complex or adaptive, the Donsker conditions may not hold ([87, 14]). Verifying the entropy condition is currently only possible for certain machine learning methods, such as lasso. For methods that involve cross-validation or for hybrid methods (like the Super Learner), it is difficult to verify such conditions. Here, we employ a convenient solution by sample splitting, where two independent sets are used for estimating the nuisance functions and constructing the stochastic target parameter. This approach has been used since Bickel [6] and further developed by Schick [92]. We extend the method of cross-fitting to k-fold cross validation, which averages our stochastic shift estimates obtained from different partitions of the data not used for nuisance parameter estimation, in this way we can use ML methods with semi-parametric estimation problems while preserving efficiency and making use of the full-data.

Not only do we need to employ cross-validation to ensure convergence properties of our estimators but we also need to embed data-adaptive discovery of our variable sets and relationships into the cross-validation procedure. As discussed, in a mixture the mixture marginal exposures, interactions and exposures with effect modifiers are unknown. If we were to use the same data to both identify these variable sets and estimate our stochastic shift parameters our estimates will be biased. Thus, we need to discover these variable sets in our cross-validation framework for desirable convergence properties to hold. We split the data into $P_{n_{-k}}$ (parameter-generating) and P_{n_k} (estimation) samples. With $P_{n_{-k}}$ we find the variable sets in our exposure space (using the basis functions from the best fitting b-spline estimator). These variable sets determine which target parameter is applied. For individual exposures found to be predictive of the outcome A, we estimate the individual shift parameter, for A_1, A_2 we estimate the interaction parameter (which includes the individual and joint shift inherently), for A, W we estimate the effect modification parameter which includes the individual shift regressed onto the covariate W using the best fitting decision tree. In any case, we have g_n and Q_n nuisance estimators which only differ in the joint estimation. Given the identified variable sets we use the same $P_{n_{-k}}$ to train our g_n and Q_n estimators which are needed for our TMLE update step to debias our initial stochastic shift estimates and give us an asymptotically unbiased estimator. We then plug-in our P_{n_k} to this unbiased estimator to get our stochastic shift estimate in this estimation sample.

Let $\bar{Q}_n$ denote a substitution estimator that plugs in the empirical distribution with weight 1/n for each observation which approximates the true conditional mean $\bar{Q}_0$ in P_0, this estimator, in our case is a Super Learner, or ensemble machine learning algorithm, our substitution estimator looks like:

$$\Psi(Q_{P_{n_k}}) = \frac{1}{V}\sum_{v=1}^{V} \bar{Q}_{n_{-k}}(\mathcal{A}_{n_{-k}}^{\delta_{n_{-k}}}, W_v) - \bar{Q}_{n_{-k}}(A, W_v)$$

We split the data into K non-overlapping folds and fit K distinct models, denoted $\bar{Q}_{n_{-k}}$. Let P_{n_k} and $P_{n_{-k}}$ be the estimation- and parameter-generating samples, respectively. $\bar{Q}_{n_{-k}}$ is

a Super Learner model fit with the parameter-generating data, and $\mathcal{A}_{n_{-k}}$ is the variable set found with the same. Likewise, in the case where δ (the shift amount) is also data-adaptively determined, this quantity is also found in the parameter-generating sample. Then, using this exposure as well as the estimation-sample covariates, the predicted outcomes under different counterfactual shifts can be obtained through the outcome regression model fit on the parameter-sample data. The resulting cross-estimated TMLE estimator is an unbiased, efficient substitution estimator of the target parameters of the data-generating distribution of interest. This estimator looks like:

$$\Psi(Q^{\star}_{P_{n_k}}) = \frac{1}{V}\sum_{v=1}^{V}\{\bar{Q}^{\star}_{n_{-k}}(\mathcal{A}^{\delta_{n_{-k}}}_{n_{-k}}, W_i)$$

The only alteration to the equation is $\bar{Q}^{\star}$, the TMLE augmented estimate. This is expressed as $f(\bar{Q}^{\star}_{n_{-k}}(A,W)) = f(\bar{Q}_{n_{-k}}(A,W)) + \epsilon_{n_{-k}} \cdot h_{n_{-k}}(A,W)$, with $f(\cdot)$ being the logit function, ϵ_n an estimated coefficient, and $h_n(A,W)$ being a "clever covariate" which is then cross-estimated. The initial estimates for the estimation-sample, which were created using parameter-generating data and models, are altered via the least-favorable submodel. The cross-estimated clever covariate looks like:

$$H(a,w) = \mathbb{I}(a_{n_{-k}} + \delta < u(w))\frac{g_{n_{-k}}(a_{n_{-k}} - \delta \mid w)}{g_{n_{-k}}(a_{n_{-k}} \mid w)} + \mathbb{I}(a_{n_{-k}} + \delta \geq u(w))$$

In this step, we are using a Super Learner to estimate the density of the data-adaptively determined exposure given a stochastic shift, denoted as $g_{n_{-k}}(a_{n_{-k}} - \delta \mid w) = p(a_{n_{-k}} - \delta \mid w)$. Specifically, we are using a parameter-generating sample to obtain the exposure, δ, and an estimator g_n. We then apply the δ to the exposure in the estimation sample and obtain predictions for the density of that exposure. These predictions are plugged into the above cross-estimated clever covariate used in the targeted maximum likelihood estimation (TMLE) update.

K-fold Cross-Validation

Up to this point, we have discussed using a simple sample splitting technique for cross-estimation of our target parameters. However, we can improve our approach by using k-fold cross-validation, which allows us to make use of the full data. This involves dividing our observations $1, ..., n$ into K equal size subgroups and defining an estimation sample P_k for each k, which is the k-th subgroup of size n/K, while the parameter-generating sample $P_{n_{-k}}$ is its complement. We rotate through the data in this round-robin manner, and for $K = 10$, we obtain 10 different target parameter mappings $\mathcal{A}_n$ (exposures found in the fold), outcome estimators Q_n, and density estimators g_n.

To obtain a summary measure of the target parameter found across the folds, such as the average, we use a pooled targeted maximum likelihood estimation (TMLE) update. We stack the estimation-sample estimates for each nuisance parameter and then perform a pooled

TMLE update across all the initial estimates using clever covariates across all the folds to get our estimate ϵ. We then update our counterfactuals across all the folds and take the average.

In each fold, we have initial estimates $Q_{n_{-k}}(Y|\mathcal{A}, W)$ and a fold-specific clever covariate $h_{n_{-k}}(p(\mathcal{A}|W))$ of length k for a fold-specific exposure found using $\mathcal{A}n{-}k$. We stack all the $Q_{n_{-k}}$'s and $h_{n_{-k}}(\mathcal{A}|W)$'s together, along with the outcomes in each validation fold, and perform our fluctuation step.

$$f(\bar{Q}n^{\star}(\mathcal{A}, W)) = f(\bar{Q}n(\mathcal{A}, W)) + \epsilon_n \cdot h_n(\mathcal{A}, W)$$

Note that we have removed the k subscripts as we are now using cross-estimates for all of n. We obtain the ϵ values from this model and then update the counterfactuals across all the folds, taking the difference for our final counterfactual outcome under a shift. Similarly, we use the updated conditional means, counterfactuals, and clever covariates to solve the influence curve (IC) across the whole sample. By pooling the cross-estimates across the folds and calculating the standard error (SE) for this pooled IC, we can derive narrower confidence intervals than if we were to average the IC estimated in each of the folds, since the IC is scaled by n and not n/K. This pooled estimate still provides us with proper intervals because all estimates in its construction were cross-estimated.

An alternative approach is to report the k-fold specific estimates of the stochastic shift parameters and fold-specific variance estimates for this target parameter using the fold-specific IC. We do this as well. This is because, if the exposure $\mathcal{A}$ identified in each fold is highly variable, the pooled estimates can be difficult to interpret. By providing both k-fold specific and pooled results, users can investigate how variable a pooled result is across the folds. For example, if exposure A_2 is found in only one of ten folds, the estimates for this exposure are likely inconsistent and should not be reported. On the other hand, if the exposure is found in all ten folds, the estimates given a shift are likely robust and reliable in predicting the outcome.

Pooled Estimates under Data-Adaptive Delta

Stochastic interventions, especially joint stochastic shift interventions (because we are taking the product of two conditional density distributions), can be sensitive to positivity violations. This can lead to bias and increased variance in exposure effect estimation given some shift. This occurs if the shift is too large, where some subgroups have no chance of receiving a specific exposure. This bias and variance occurs even when using an efficient estimator for a target parameter such as TMLE. One way to address this issue is to use a reduced version of the user provided δ, which constrains the magnitude of the shift to avoid positivity violations. Reduced δ can help to reduce the bias and variance in exposure effect estimation by reducing positivity violations. To do this, we treat δ as a data-adaptive parameter as well. That is, we must identify a δ in the parameter-generating sample that meets some criteria of positivity estimation and apply this δ to the estimation sample data.

Let $H(a_\delta, w)_i$ be the ratio of probability densities for observation i given a shift in exposure by δ. To ensure that all observations have a ratio below a specific threshold λ, we can repeatedly reduce δ by a small step ϵ until the condition $H(a_\delta, w)_i < \theta$ is satisfied for all observations i. This is expressed as:

$$\forall i,\ H(a_\delta, w) = \frac{g_{n_{-k}}(a_{n_{-k}} - \delta \mid w)}{g_{n_{-k}}(a_{n_{-k}} \mid w)} \leq \lambda$$

Where λ is a specified threshold value. The δ is reduced until all values for the clever covariate of density ratios is less than or equal to λ. This process is done with the parameter-generating data using estimators trained on the same data. At each iteration we reduce the δ by $\epsilon = 10\%$ and the default λ in the package is 50. Meaning if any of the predicted conditional probabilities are 50 times greater than the probability under observed exposure the δ is reduced. This parameter is optional in the Super Learner package.

In the event that δ is held constant across the folds then pooling is simply the average estimate for each target parameter across the folds. If δ is data-adaptive, then we provide the average estimates under the average δ, that is we simply average the delta and pair it with the average estimates and the pooled variance calculations described previously.

Estimating Regional Covariate Effect Modification

As discussed, our effect modification parameter is:

$$\psi_{\delta,\mathcal{V}} = E_{P_{\delta,\mathcal{V}}}[Y]$$

Where $\mathcal{V}$, a subset of W need to be discovered from the data. If we were to treat $\mathcal{V}$ as a data-adaptive parameter then this could result in K slightly different regions where K is the number of cross-validation folds explained above. Because we are using a very shallow decision tree on a small subset of the covariates (1 or 2) it is unlikely our tree will overfit and thus given these conditions we instead regress TMLE updated individual exposure effects across the full data onto the covariate space rather than in each fold. That is, the stochastic exposure effects are not regressed onto the full covariate space, instead the covariates that modify the impact of an exposure are discovered in the parameter-generating sample and we regress on these $V \in W$ only. The V for which to find partitions in is treated as a data-adaptive parameter but the partitions within each v_i are determined using the full-data to ensure interpretability.

The steps to this process are:

1. Stack the TMLE updated vector of counterfactual differences under stochastic shift intervention δ across the folds for the parameter-generating samples.

2. Regress these exposure effects onto the covariates V in the parameter-generating sample to identify regions that explain the most variability in the exposure effects.

3. Create rule(s) for these regions.

4. Using the estimation sample exposure effects, similarly stacked across the folds, evaluate the rules for regions V and take the average of the estimation sample exposure effects in each region, similarly estimate variances based on the efficient influence function in each region.

In order to estimate the variance using the EIF we need to weigh the EIF by the inverse proportion of observations in a given region. This looks like:

$$Var(\hat{\theta}r) \approx \frac{1}{n} \sum_{i=1}^{n} I(W_i \in r) \cdot [D(P_0)(\hat{\theta}(x_i))]^2$$

In this equation, $\text{Var}(\hat{\theta}_r)$ represents the estimated variance of the target parameter θ within the region r, $\hat{\theta}_r$ represents the estimate of θ within the region r, n represents the sample size, W_i represents the covariates of the i-th observation, $I(W_i \in r)$ is an indicator function that equals 1 if W_i is in the region r and 0 otherwise, $\hat{\theta}(x_i)$ represents the estimate of θ at the data point x_i, and $D(P_0)(\hat{\theta}(x_i))$ represents the efficient influence function (EIF) of the estimate at the data point x_i.

The weighting by the inverse proportion of observations indicated by the indicator for the region is achieved by multiplying the EIF squared by the indicator function $I(W_i \in r)$ and dividing by the sample size n. This ensures that the variance estimate is properly weighted by the number of observations in the region. Note that this is an approximation and assumes that the observations are independent and identically distributed (i.i.d.).

We find the regions r_p using a Super Learner of shallow decision trees. Because the complexity of these trees is very low, in high folds the single or dual partitions points are likely the same as if applied to the full data. We rely on this complexity assumption to hold in order to do one regression on the full data rather than estimating fold specific trees.

2.6 Simulations

In this section, we demonstrate using simulations that our approach identifies the correct exposures, interactions and effect modifications used to generate the outcome and correctly estimates the counterfactual mean difference in outcome under shift interventions.

Data-Generating Processes

We construct a data-generating process (DGP) where Y is generated from a linear combination of different effect including marginal, interaction and effect modification. This DGP was constructed to represent the complexity of a mixed exposure where mixture variables are correlated but only some have an impact on the outcome and where certain effects are modified by baseline covariates.

Mixed Exposure Simulation

This DGP has the following characteristics, $O = W, A, Y$. Let $W = (W_1, W_2, W_3)$ denote a random vector of covariates, where W_1 and W_2 are bivariate normal with mean vector $\boldsymbol{\mu} = (6, 7)$ and covariance matrix

$$\boldsymbol{\Sigma}_W = \begin{pmatrix} 1 & 0.4 \\ 0.4 & 1 \end{pmatrix}$$

and W_3 is a binary variable with Bernoulli distribution with probability 0.5.

The mean values for the components of the mixture distribution A are computed as $\mu_1 = \mathrm{E}[e^{W_1/2}]$, $\mu_2 = \mathrm{E}[W_2/2]$, and $\mu_3 = 5$. That is exposure 3 is not associated with the covariates. A exposure vector $A = (A_1, A_2, A_3)$ is then generated from a trivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ and covariance matrix:

$$\boldsymbol{\Sigma}_A = \begin{pmatrix} 1 & 0.5 & 0.8 \\ 0.5 & 1 & 0.7 \\ 0.8 & 0.7 & 1 \end{pmatrix}$$

A fourth mixture variable was created via $A_4 \sim \mathcal{N}(4, 2)$. The outcome Y is then generated by:

$$Y = 1.3A_4 + \left(\begin{cases} A_3^2 & \text{if } W_3 = 1 \\ A_3 & \text{if } W_3 = 0 \end{cases} \right) + 0.4A_4A_1 + 0.1W_1 + 0.3W_2$$

Where the third component of A depends on W_3, A_3 is squared if W_3 is equal to 1, and is left unchanged if it is equal to 0. In generating the outcome, this variable represents effect modification or the differential impact of A_3 depending on W_3. Overall, built into this outcome generating process there is a marginal impact of A_4, a multiplicative interaction between A_1 and A_4 and effect modification of A_3 based on W_3. There is no effect of A_2 on the outcome.

To calculate the ground-truth counterfactual mean under a shift intervention we simply generate a very large data set from this DGP and apply various shifts to the exposures. We apply a shift of 1 unit increase to each exposure. In the first simulation we deterministically set the variable sets to evaluate estimates of the counterfactual outcomes under these shifts. These is done with the var_sets parameter in SuperNOVA which, if is not null, bypasses the data-adaptive determination of the variable sets in the basis functions. In the second simulation we include the data-adaptive discovery of the variable sets to provide estimates for evaluating the variable relationship identification through the basis function procedure.

Evaluating Performance

We assessed the asymptotic convergence to the true exposure relationships used in the DGP, as well as the convergence to the true counterfactual differences for these exposures, in each simulation. To do so, we followed the following steps:

1. From this sample, we generated a random sample of size n from the DGP.
2. At each iteration, we used the parameter generating sample to define the exposure(s) and create the necessary estimators for the target parameter dependent on the variable sets. We then used the estimation sample to obtain the updated causal parameter estimate using TMLE. We repeated this process for all folds.
3. At each iteration, we output the stochastic shift estimates given the pooled TMLE.

To evaluate the performance of our approach, we calculated several metrics for each iteration, including bias, variance, MSE, confidence interval (CI) coverage, and the proportion of instances in which the true variable relationships were identified. To ensure the rate of convergence was at least as fast as $\sqrt{n}$, we multiplied each bias estimate by $\sqrt{n}$. We then calculated the variance for each estimate and used it to compute the mean squared error (MSE) as MSE = bias^2 + variance. To account for different sample sizes, we multiplied the MSE estimates by n. For each estimate, we also calculated the CI coverage of the true stochastic shift parameter given the data-adaptively determined exposure. We calculated these performance metrics at each iteration, performing 50 iterations for each sample size n = (250, 500, 1000, 1500, 2000, 2500, 3000, 5000). We used SuperNOVA with 2-fold cross-validation (to speed up calculations in the simulations) and default learner stacks for each nuisance parameter and data-adaptive parameter. Additionally, the quantile threshold to filter the basis functions based on F-statistic was set to 0 to include all basis functions used in the final best fitting model.

Default Estimators

To use SuperNOVA, we need estimators for $\bar{Q} = E(Y|A, W)$ and $g_n = p(A|W)$ (the conditional density). SuperNOVA provides default algorithms to be used in a Super Learner [59] that are both fast and flexible. For our data-adaptive procedure, we include learners from the packages earth [67], polspline [81], and hal9001 [16]. The results from each of these packages can be formed into a model matrix, on which we can fit an ANOVA to obtain the resulting linear model of basis functions.

To estimate $\bar{Q}$, we include estimators for the Super Learner from glm [62], elastic net [27], random forest [111], and xgboost [13]. For the semiparametric density estimator g_n, we create estimators based on homoscedastic errors (HOSE) and heteroscedastic errors (HESE) from the same estimators used in $\bar{Q}$.

Results

Target Parameters Estimated by SuperNOVA Converge to Truth at $1/\sqrt{n}$

We can determine the accuracy of our estimator based on its convergence rate in simulations, which describes how fast the estimator approaches the true value of the parameter as the

sample size increases in a DGP where there are marginal, effect modifying, interacting and confounding effects. For our estimator to have desirable asymptotic properties this convergence needs to be at $1/\sqrt{n}$ rate. [2]

Figure 2.1 shows the absolute bias (**A**), MSE (**B**), CI coverage (**C**) and estimate standard deviation (**D**) as sample size increases to 5000. For bias and MSE there is a converge to zero at sample size 5000 apart from the effect modification parameter where there is still residual bias at this sample size. For coverage, the average coverage for each target parameter were: individual shift: 98%; effect modification: 95%; joint shift: 98%; interaction: 98%.

Although these plots show generally a reduction in bias, MSE, and SD as sample size grows we need to ensure the rate of reduction is at $1/\sqrt{n}$. To show this we multiply the bias estimates by $\sqrt{n}$ and we scale the MSE by n. [3]

Figure 2.2 shows the scaled bias and MSE. The estimates for each target parameter look relatively flat across sample size apart from the effect modification parameter which shows some small variability. Generally these results show that the precision and error magnitude remain consistent when accounting for sample size.

SuperNOVA's Valid Inference Assessed Through Simulation

To ensure proper inference, we need to demonstrate that SuperNOVA's estimator has a normal sampling distribution centered at 0 and narrows as the sample size increases. We assess this by examining the empirical distribution of standardized differences. Figure 3.7 shows the probability density distribution of the standardized bias compared to the true estimates using 50 iterations per sample size. All estimates converge to a mean 0 normal distribution with increasing sample size.

---

[2] Briefly, in semi-parametric statistics, an estimator is said to be asymptotically linear if its bias approaches zero as the sample size n approaches infinity, and its variance is proportional to $1/n$. That is, the estimator is linear with respect to the sample size n, and the error between the estimator and the true value decreases at a rate proportional to $1/\sqrt{n}$. The $1/\sqrt{n}$ convergence rate is necessary for asymptotically linear estimators because it ensures that the estimator is consistent and efficient. Consistency means that the estimator converges to the true value of the parameter as the sample size increases, and efficiency means that the estimator achieves the smallest possible variance among all unbiased estimators. The $1/\sqrt{n}$ convergence rate is also important because it determines the trade-off between bias and variance in the estimator. As the sample size increases, the variance of the estimator decreases, but the bias may increase if the estimator is not well-designed (such as if we were to apply a simple GLM to our DGP). The $1/\sqrt{n}$ convergence rate ensures that the estimator is well-balanced, with small bias and variance, and is able to converge to the true value of the parameter at a reasonable rate.

[3] This quantity is called the Mean Squared Error of the Estimator (MSEE), the MSEE measures the expected squared difference between the estimator and the true value of the parameter, normalized by the sample size. The reason we scale the MSE by n instead of $\sqrt{n}$ is that the MSE measures the average squared difference between the estimator and the true value of the parameter, which is a measure of the error magnitude. The sample size n represents the amount of information available to the estimator, which affects the precision of the estimator's estimate. When the estimator is more precise, the error magnitude is smaller, and the MSEE will be smaller as well. Scaling the MSE by n reflects this relationship between precision and error magnitude.

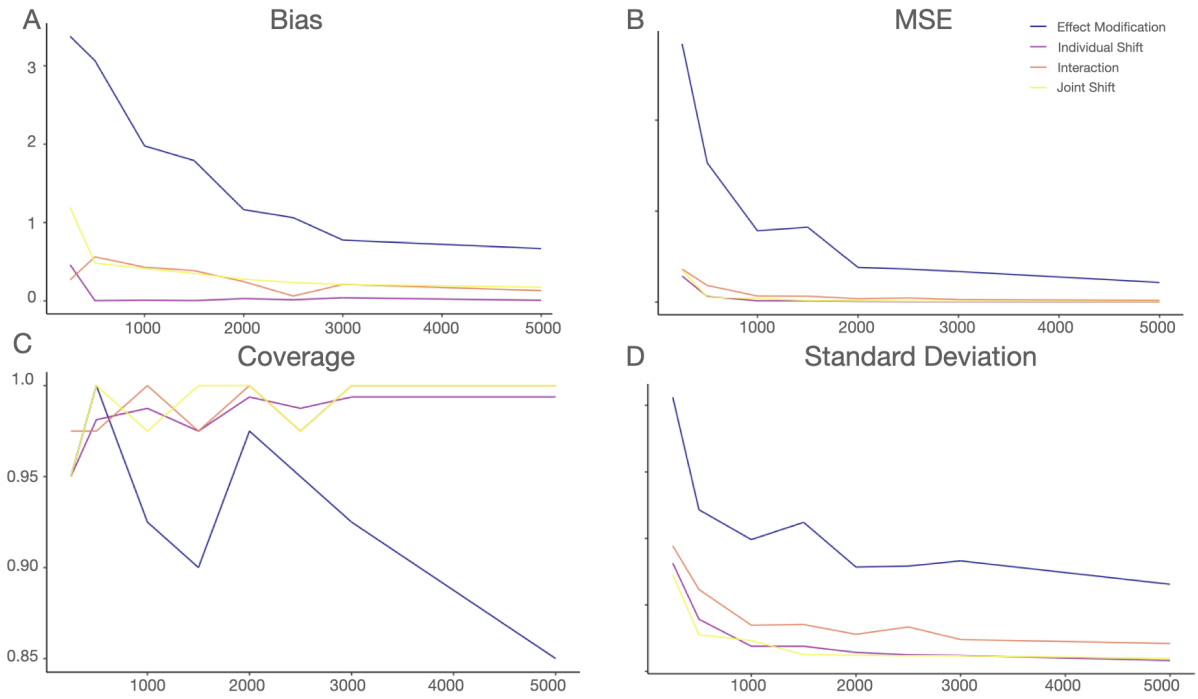


Figure 2.1: Bias, MSE, CI Coverage and Standard Deviation for Each Parameter Across Sample Sizes

For example, **Figure 3.7 A** shows the sampling distribution of the standardized bias for the marginal parameter or an individual shift. We can see that this sampling distribution converges to a normal mean 0 distribution with standard deviation 1. Similarly, **Figure 3.7 B** shows the sampling distribution for estimates of a dual shift of two variables to measure the joint impact. **Figure 3.7 C** shows estimates for effect modification, which breaks down to estimates of individual exposure shifts within data-adaptively identified regions of a covariate, and **Figure 3.7 D** shows the z-score distribution of the interaction parameter estimates. In each case, as sample size increases, the estimates converge to a normal 0, 1 distribution, indicating proper inference for SuperNOVA. Because the effect modification parameter is an average of the counterfactual differences within a covariate region, we still see slight bias at sample size 5000, this is due to averaging over fewer number of observations.

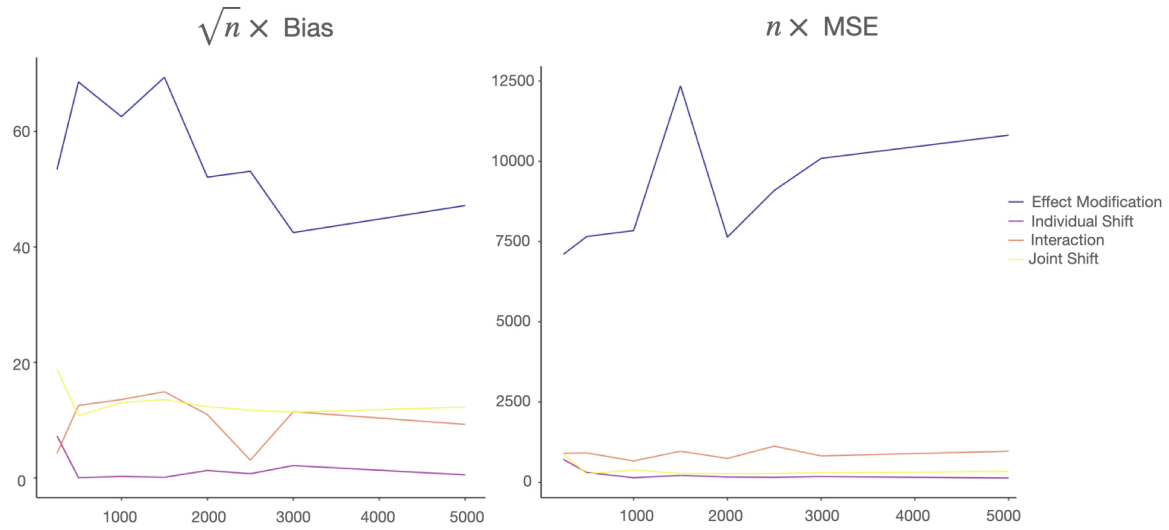


Figure 2.2: Scaled Bias and MSE

2.7 Applications

NIEHS Synthetic Mixtures

The NIEHS synthetic mixtures data is a commonly used dataset to evaluate the performance of statistical methods for mixtures. This synthetic data can be considered the results of a prospective cohort study, where the outcome cannot cause the exposures, and correlations between exposure variables can be thought of as caused by common sources or modes of exposure. The nuisance variable Z can be assumed to be a potential confounder and not a collider. The dataset has 7 exposures ($X_1 - X_7$) with a complex dependency structure based on endocrine disruption. Two exposure clusters (X_1, X_2, X_3 and X_5, X_6) lead to high correlations within each cluster. X_1, X_2, X_7 positively contribute to the outcome, X_4, X_5 have negative contributions, while X_3 and X_6 have no impact on the outcome. Rejecting X_3 and X_6 is difficult because of their correlations with cluster group members. This correlation and effects structure is biologically plausible, as different congeners of a group of compounds may be highly correlated but have different biological effects. The exposures have various agonistic and antagonistic interactions, and Table 2.1 provides a breakdown of the variable sets and their relationships. The synthetic data and key for dataset 1 are available on GitHub. Figure 2.4 shows the marginal dose-response relationships.

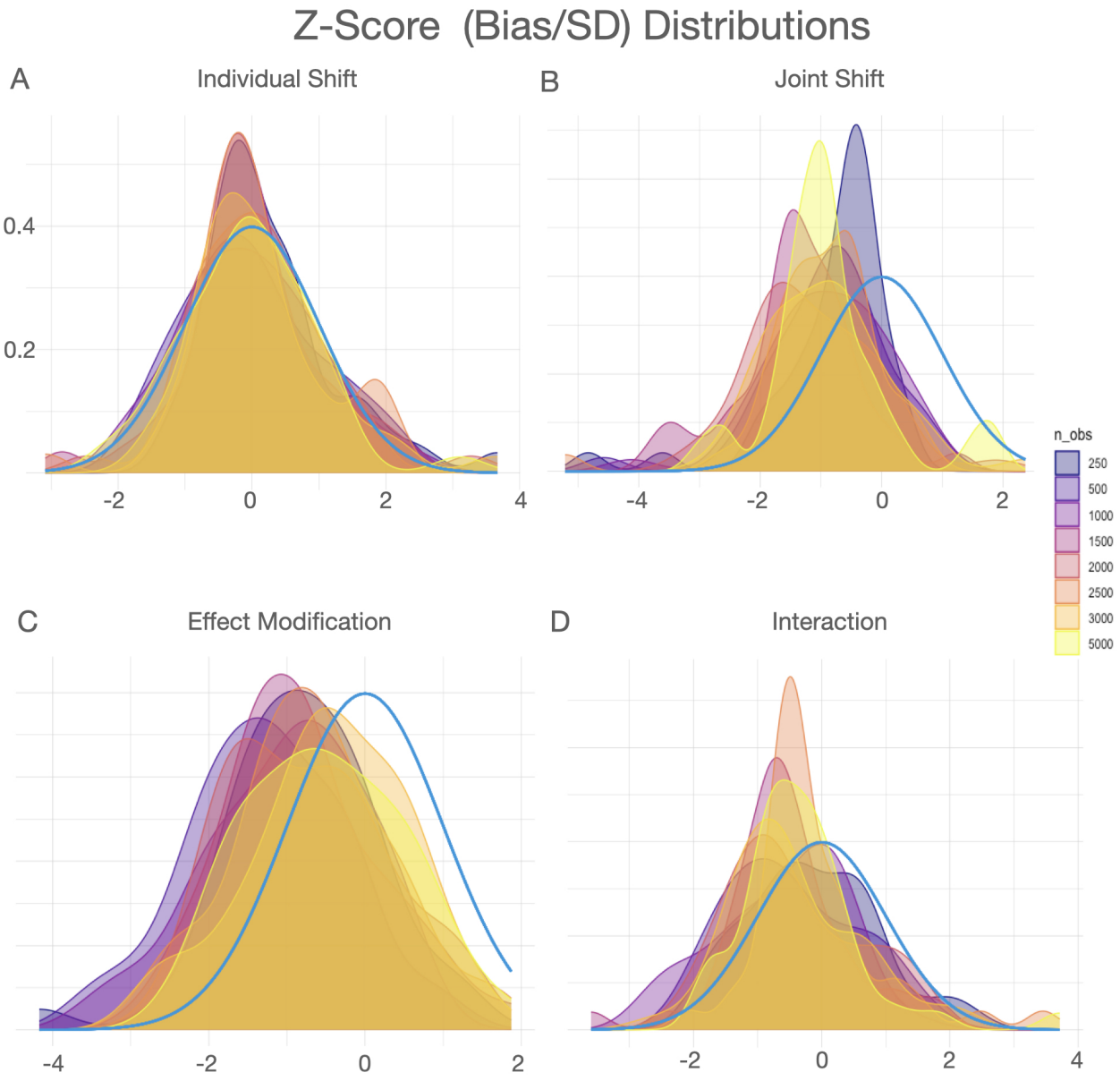


Figure 2.3: Bias Standardized by Standard Error Compared to Ground-Truth Outcome Under Shift Interventions

Given these toxicological interactions we expect these variable sets to be determined in SuperNOVA. For example, we might expect a positive counterfactual result for X_1, X_2, X_7 and negative results for X_4, X_5. Likewise, in the case for antagonistic relationships such as X_2, X_4, we would expect a joint shift to get closer to the null given X_4 antagonizes the

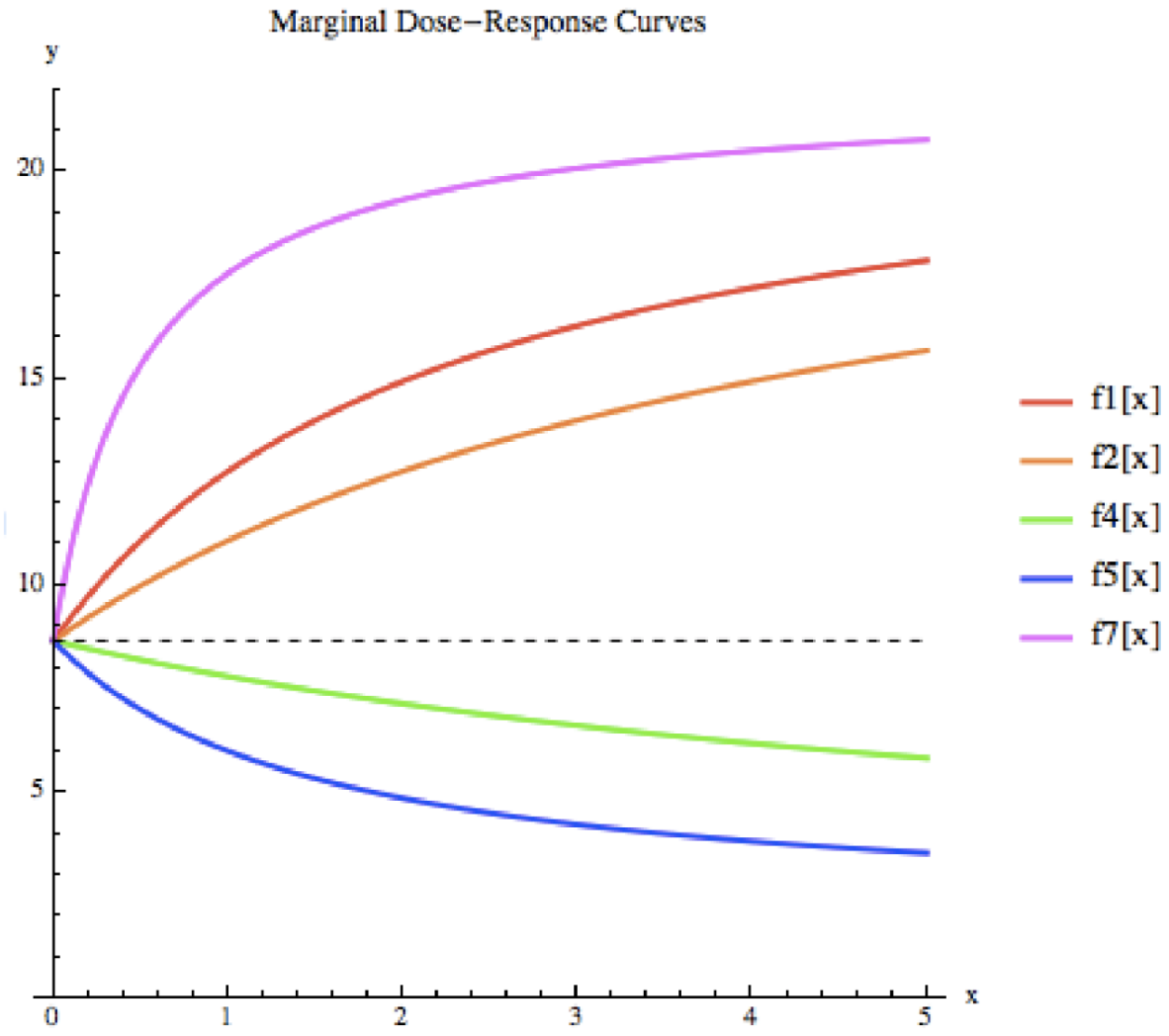


Figure 2.4: Marginal Dose-Response Relationships

positive effects of X_2. For X_1, X_2 we would expect the joint shift to be close to the sum of individual shifts (not much interaction) but for X_1, X_7 there to be a more than additive effect (some interaction).

The NIEHS data set has 500 observations and 9 variables. Z is a binary confounder. Of course, in this data there is no ground-truth, like in the above simulations, but we can gauge SuperNOVA's performance by determining if the correct variable sets are used in the interactions and if the correct variables are rejected. Because many machine learning algorithms will fail when fit with one predictor (in our case this happens for g(Z)), we simulate

Variables	Interaction Type
X1 and X2	Toxic equivalency factor, a special case of concentration addition (both increase Y)
X1 and X4	Competitive antagonism (similarly for X2 and X4)
X1 and X5	Competitive antagonism (similarly for X2 and X4)
X1 and X7	Supra-additive ("synergy") (similarly for X2 and X7)
X4 and X5	Toxic equivalency factor, a type of concentration addition (both decrease y)
X4 and X7	Antagonism (unusual kind) (similarly for X5 and X7)

Table 2.1: NIEHS Synthetic Data Interactions

additional covariates that have no effects on the exposures or outcome but prevent these algorithms from breaking.

We apply SuperNOVA to this NIEHS synthetic data using 4-fold CV and the default stacks of estimators used in the Super Learner for all parameters. We parallelize over the cross-validation to test computational run-time on a newer personal machine an analyst might be using. For this set-up, we intentionally use fewer folds (4) which may lead to less consistent findings because less data (75%) is used for the parameter-generating sample to identify the variable relationships compared to say 10 fold where 90% of the data is used.

Table 2.2 shows the marginal results from SuperNOVA when applied to this NIEHS synthetic data set using the aforementioned settings. The Condition column shows the variables identified across the folds. Psi shows the counterfactual mean difference under a shift value listed under Delta compared to the observed average outcome. Fold indicates the fold the specific result was found in and N is the number of observations in the fold.

Overall, SuperNOVA accurately rejects exposures X_3, X_6, finds marginal effects for X_1, X_5, X_7 consistently in all the folds and effects for X_4 in three folds and X_2 in two folds. Directions of the estimates are as expected based on ground-truth where X_1, X_2, X_7 are positive and X_4, X_5 are negative. As seen in the dose-response curve, X_2, X_4 have the weakest relationships which resulted in fewer fold detections and non-significant findings. The largest effects are for X_7 which is also true based on the dose-response relationships.

Comparison to Existing Methods

Currently, quantile g-computation is a popular method for mixture analysis in environmental epidemiology. The method yields estimates of the effect of increasing all exposures by one quantile, simultaneously under linear model assumptions. Quantile g-computation looks like:

$$Y_i = \beta_0 + \sum_{j=1}^{d} \beta_j X_{ji}^q + \beta Z_i + \epsilon_i$$

Where X^q are the quantized mixture components and Z are the covariates. Which works by first transforming mixture components into quantiles. Then the negative and

Condition	Psi	Variance	SE	Lower CI	Upper CI	P-value	Fold	N	Delta
X1	2.00	0.58	0.76	0.51	3.49	0.01	1	125	0.50
X1	5.87	1.08	1.04	3.83	7.91	0.00	2	125	0.50
X1	0.30	1.39	1.18	-2.01	2.61	0.80	3	125	0.50
X1	-0.05	2.56	1.60	-3.18	3.09	0.98	4	125	0.50
X1	1.23	0.28	0.53	0.19	2.27	0.02	Pooled TMLE	500	0.50
X5	-2.12	0.48	0.69	-3.48	-0.76	0.00	1	125	0.50
X5	-1.66	0.75	0.87	-3.36	0.04	0.06	2	125	0.50
X5	-1.20	1.23	1.11	-3.37	0.98	0.28	3	125	0.50
X5	-1.36	0.59	0.77	-2.86	0.14	0.08	4	125	0.50
X5	-1.99	0.24	0.49	-2.94	-1.04	0.00	Pooled TMLE	500	0.50
X7	2.08	0.52	0.72	0.66	3.49	0.00	1	125	0.50
X7	1.98	0.79	0.89	0.24	3.72	0.03	2	125	0.50
X7	2.38	0.97	0.98	0.45	4.31	0.02	3	125	0.50
X7	2.13	0.54	0.73	0.69	3.57	0.00	4	125	0.50
X7	2.11	0.23	0.47	1.18	3.04	0.00	Pooled TMLE	500	0.50
X4	-0.30	0.56	0.75	-1.77	1.18	0.69	1	125	0.50
X4	-0.20	0.79	0.89	-1.94	1.55	0.83	2	125	0.50
X4	0.25	1.15	1.07	-1.86	2.35	0.82	3	125	0.50
X4	-0.29	0.34	0.58	-1.43	0.85	0.62	Pooled TMLE	375	0.50
X2	1.63	0.68	0.83	0.01	3.25	0.05	1	125	0.50
X2	0.00	7.15	2.67	-5.24	5.25	1.00	2	125	0.50
X2	0.15	2.24	1.50	-2.78	3.09	0.92	Pooled TMLE	250	0.50

Table 2.2: Marginal Results from NIEHS Mixtures Data

positive coefficients from a linear model for the mixture components are summed to give a mixture (Ψ) summary measure which characterizes the joint impact. There are many assumptions that should be poignant after our discussion of mixtures. Firstly, quantiles may not characterize the exposure-response relationship (could be non-monotonic) which occurs in endocrine disrupting compounds. For interpretable weights and mixture estimate Ψ, assumes additive relationship of quantiles (Ψ is just sum of β's in front of mixture components). After our discussion, in mixtures our main goal is model possible interactions in the data because we expect exposures to have non-additive, possible non-monotonic, antagonistic and agonistic relationships. Therefore, we should expect interactions in our mixture data. In quantile g-computation, with the inclusion of interactions, the proportional contribution of an exposure to the overall effect then varies according to levels of other variables and therefore weights cannot be estimated. Because we can never assume no interactions, quantile g-computation then boils down to getting conditional expectations when setting mixtures to quantiles through a linear model with interaction terms specified by the analyst. After our discussion of mixtures this should feel incorrect. As we argue, the important variables,

relationships, and thresholds in a mixture are all unknown to the analyst which makes this a data-adaptive target parameter problem. Even testing quantile g-computation on the NIEHS data is difficult because we don't know what interactions to include *a priori*. The best we can do is run it out of the box and with two-way interactions and compare results to the ground-truth measures. Lastly, quantile g-computation does not flexibly control for covariates.

We run quantile g-computation on the NIEHS data using 4 quantiles with no interactions to investigate results using this model. The scaled effect size (positive direction, sum of positive coefficients) was 6.28 and included X_1, X_2, X_3, X_7 and the scaled effect size (negative direction, sum of negative coefficients) was -3.68 and included X_4, X_5, X_6. Compared to the NIEHS ground-truth, X_3, X_6 are incorrectly included in these estimates. However the positive and negative associations for the other variables are correct.

Next, because we expect interactions to exist in the mixture data, we would like to assess for them but the question is which interaction terms to include? Our best guess is to include interaction terms for all the exposures. We do this and show results in **Table 2.4**.

In **Table 2.4** Ψ_1 is the summary measure for main effects and Ψ_2 for interactions. As can be seen, when including all interactions neither of the estimates are significant. Of course this is to be expected given the number of parameters in the model and sample size $n = 500$. However, moving forward with interaction assessment is difficult, if we were to assess for all 2-way interaction of 7 exposures the number of sets is 21 and with 3-way interactions is 35. We'd have to run this many models and then correct for multiple testing. Hopefully this example shows why mixtures are inherently a data-adaptive problem and why popular methods such as this, although succinct and interpretable, fall short even in a simple synthetic data set.

NHANES Data

Data Description

The Columbia University Mailman School of Public Health Department of Environmental Health Sciences held a two-day Mixtures Workshop in August 2018 to teach environmental health researchers various statistical methods for studying mixtures. The workshop covered unsupervised methods like clustering, principal component analysis, and exploratory factor analysis, as well as supervised methods like variable selection, weighted quantile sum regression, and Bayesian kernel machine regression. A publicly available real dataset was used to illustrate the methods, taken from a study investigating the association between exposure to persistent organic pollutants (POPs) and leukocyte telomere length [68]. We chose POPs as the mixed exposure so that we can directly compare our results with those found in [68]. The study's results provided insight into a potential mechanism underlying carcinogenesis mediated by activation of AhR and subsequent telomerase expression, but certain mixtures methods were not utilized resulting in loss of information. The paper by Gibson et al. [30] discusses the

results obtained using each method on this data during this workshop and how they compare to one another, focusing on understanding how exposure to this mixture may impact LTL.

For our analysis, we used the same population as in the original paper by Mitro et al. [68] which investigated the association between exposure to (POPs) with high affinity to the aryl hydrocarbon receptor (AhR) and longer leukocyte telomere length (LTL). The exclusion criteria used in the study are described in Gibson et al. [30]. This data is from the 2001-2002 National Health and Nutrition Examination Survey (NHANES) cycle, which interviewed 11,039 people, of whom 4,260 provided blood samples and consented to DNA analysis. Sufficient stored samples to estimate telomere length were available for 1,003 participants after excluding individuals without environmental chemical analysis data (n=2,850), those who were missing data on covariates such as body mass index (BMI) (n=70), education (n=2), and serum cotinine (n=8), and those with missing values for individual PCBs, dioxins, or furans (n=327). After filtering for only complete exposures our final study population is nearly identical to the smallest sub-sample included in the original analyses by Mitro et al, our study has 1007 observations compared to 1003 in the original paper.

Exposure assessment was performed as previously described [68]. Congeners were adjusted for serum lipids which were calculated using an enzymatic summation method. 18 cogeners were used as exposures in our analysis. These breakdown to 8 non-dioxin like PBCs (PCB 74, PCB 99, PCB 138, PCB 153, PCB 170, PCB 180, PCB 187, PCB 194); 2 non-ortho PCBs (PCB 126, PCB 169); 1 mono-ortho PCB (PCB 118); 4 Dioxins (1,2,3,6,7,8-hxcdd, 1,2,3,4,6,7,8-hpcdd, 1,2,3,4,6,7,8,9-ocdd) and 4 Furans (2,3,4,7,8-pncdf, 1,2,3,4,7,8-hxcdf, 1,2,3,6,7,8-hxcdf, 1,2,3,4,6,7,8-hxcd).

Telomere length measurement was performed as previously described [68].The quantitative polymerase chain reaction (qPCR) method was used to measure telomere length relative to standard reference DNA (T/S ratio). Samples were assayed three times in duplicate wells, producing six data points which were averaged to calculate mean T/S ratios. Analytical runs were blinded, and the CDC conducted a quality control review. We adjusted for the same covariates as in previous modeling using supervised methods: age, sex, race/ethnicity, educational attainment, BMI, serum cotinine, and blood cell count and distribution. Race/ethnicity was categorized as non-Hispanic white, non-Hispanic black, Mexican American, or other. Educational attainment was categorized as less than high school, high school graduate, some college, or college or more. BMI was categorized as $< 25, 25 - 29.9, \geq 30$. Blood cell count and distribution were included as individual covariates: white blood cell count, percent lymphocytes, percent monocytes, percent neutrophils, percent eosinophils, and percent basophils. We include this data as example data in the SuperNOVA package.

We apply SuperNOVA using the default learners in each stack. We use 20-fold CV with no quantile limit for the F-statistic of basis functions. Our default algorithms used in the data-adaptive Super Learner to identify variable relationships include both basis functions and linear terms with up to 2-way interactions. We allow for interactions/effect modification to occur between all 18 POPs and 13 covariates. Given the range of the data, we set a δ to 2 for all exposures, that is, we investigate the counterfactual differences in telomere length for an increase in 2 ng/g or pg/g for all exposures data-adaptively identified to predict telomere

length.

Furan Findings on Telomere Length

Of the 18 POPs we find only one chemical had consistent results across all the folds. No other chemical was used in any of the folds. The furan 2,3,4,7,8-pncdf was identified in all 20 folds. No other marginal effects or interactions were discovered. **Table 2.5** gives a breakdown of the fold specific results and the final pooled TMLE result for an increase in furan 2,3,4,7,8-pncdf by 2 pg/g. The final pooled TMLE estimate shows that a 2 pg/g increase in furan 2,3,4,7,8-pncdf leads to a 0.02 decrease in telomere length.

The [68] summarize results from the workshop. They state that clustering methods identified high, medium, and low POP exposure groups with longer log-LTL observed in the high exposure group. Principal component analysis (PCA) and exploratory factor analysis (EFA) revealed positive associations between overall POP exposure and specific POPs with log-LTL. Penalized regression methods identified three congeners (PCB 126, PCB 118, and furan 2,3,4,7,8-pncdf) as potentially toxic agents. WQS identified six POPs (furans 1,2,3,4,6,7,8-hxcdf, 2,3,4,7,8-pncdf, and 1,2,3,6,7,8-hxcdf, and PCBs 99, 126, 169) as potentially toxic agents with a positive overall effect of the POP mixture. BKMR found a positive linear association with furan 2,3,4,7,8-pncdf, suggestive evidence of linear associations with PCBs 126 and 169, a positive overall effect of the mixture, but no interactions among congeners. These results (in the supervised methods) controlled for the same covariates. Interestingly, although we corroborate the finding of furan 2,3,4,7,8-pncdf, specifically in the BKMR method (the most flexible method used) our results show a negative association. Additionally, we do not find associations of any other POP used in any fold meaning that, in the best fitting spline model, no basis functions for any of the other POPs were used. The inverse association we find compared to these other methods could be a result of our Q model being a Super Learner of many estimators which can model relationships perhaps missed in these other methods.

In this NHANES analysis, we demonstrate that SuperNOVA can effectively identify variable relationships in real-world mixture data and provide estimates for these variables. The resulting summary statistics offer a meaningful estimate of how the outcome would change under an intervention on the identified exposures, providing an interpretable understanding of the results.

2.8 Software

The development of open-source software for semi-parametric statistical theory is critical for consistent and reproducible results in mixed exposure research. SuperNOVA is an R package that provides an open-source tool for evaluating the causal effects of a mixed exposure using asymptotically linear estimators. This package includes a detailed vignette, documentation of semi-parametric theory, examples of output, and comparison to existing

methods. The NIEHS synthetic data and NHANES mixed POP exposure data are also provided for reproducibility. SuperNOVA can run sequentially or in parallel [104], and its efficient estimators make it suitable for use on personal machines. The SuperNOVA package is well-maintained, easily accessible, and highly detailed, with coding notebooks that show simulations of mixed exposure data and SuperNOVA output with detailed summaries of interpretation. The package has been made publicly available via GitHub. By making robust statistical software widely accessible, we aim to move beyond simple parametric models and ensure more consistent and accurate results in mixed exposure research.

2.9 Discussion

In this paper we introduce a new method for estimating the effects of a mixed exposure. Our approach fits a very large statistical model to the exposure-covariate space and treats the basis functions as a data-adaptive target parameter for which we estimate the average change in outcome under stochastic shift interventions. This is done within a cross-validated framework paired with targeted learning of our target parameter which provides estimates that are asymptotically unbiased and have the lowest variance for studies which satisfy the unconfoundedness and positivity assumptions. Our proposed method provides valid confidence intervals without restrictions on the number of exposure, covariates, or the complexity of the data-generating process. Our method first identifies relevant exposure-covariate subspaces that best explains the outcome. The output of our method is the counterfactual mean difference in outcomes if all individual were exposed to a change exposure compared to the observed outcome under observed exposure. Our approach has potentially many important applications including identifying what interacting drugs lead to the most beneficial patient outcomes as well as finding what pollution chemicals interact which leads to deleterious outcomes on public health. Our approach allows for "dredging with dignity" wherein exposure regions can be discovered in the data which are not known *a priori* and still provide unbiased estimates for the target parameter with valid confidence intervals. The major limitation of our proposed method is the computational burden of density estimation. There are alternative approaches which can be explored. For example, [77] first propose alternative ways for estimating the probability density ratio needed in our proposed methodology. Instead of relying on estimates of the density estimator gt, the problem is recast as a classification problem in an augmented dataset that contains 2n observations. In this dataset, each observation is duplicated, and one is assigned the observed exposure, At, and the other is assigned the intervened exposure, A_t^d. An indicator variable, Λ, is introduced to identify the treatment under intervention. The density ratio can then be expressed as a function of of the probability of Λ over 1 - probabilty of Λ. This approach can be carried out by estimating it via any classification method available in the machine and statistical learning literature, such as Super Learning. This would improve computational time in the case where δ is not data-adaptive. In the case that it is, this would need to be refit at each iteration and our approach of directly estimating the density may be more efficient as given a density estimator

fit new density distributions can be generated via predictions for any δ_i. Another limitation to the proposed method is regardind inconsistent findings. In the case that a variable set is identified in every fold interpretation is straightforward; however, in the event that findings are inconsistent interpretation is relative to the analyst. In reporting findings it will become important to also report the number of folds the estimates occur in to give a measure of reliability and consistency in the data. Overall, our simulations with ground-truth, NIEHS synethetic data and real-world data application show the robustness and interpretability of our approach. In an effort to make adoption of semi-parametric methods such as this more seamless we provide the SuperNOVA R package on GitHub which is well documented for analysts to apply to their respective data.

Condition	Psi	Variance	SE	Lower CI	Upper CI	P-value	Fold	N
X5	-1.63	0.76	0.87	-3.34	0.08	0.08	2	125
X7	2.06	0.79	0.89	0.32	3.80	0.03	2	125
X5&X7	0.26	0.91	0.95	-1.61	2.13	0.79	2	125
Interaction	-0.17	0.93	0.96	-2.06	1.72	0.86	2	125
X5	-1.39	0.52	0.72	-2.79	0.02	0.10	4	125
X7	2.16	0.53	0.73	0.73	3.59	0.01	4	125
X5&X7	1.08	0.63	0.80	-0.48	2.64	0.23	4	125
Interaction	0.31	0.49	0.70	-1.07	1.69	0.71	4	125
X5	-1.56	0.35	0.59	-2.72	-0.40	0.04	Pooled TMLE	250
X7	2.20	0.36	0.60	1.02	3.38	0.00	Pooled TMLE	250
X5&X7	0.62	0.42	0.65	-0.65	1.89	0.44	Pooled TMLE	250
Interaction	-0.02	0.38	0.61	-1.23	1.18	0.98	Pooled TMLE	250
X2	-0.21	8.41	2.90	-5.89	5.47	0.90	2	125
X7	2.19	0.79	0.89	0.45	3.93	0.02	2	125
X2&X7	3.44	0.74	0.86	1.76	5.12	0.00	2	125
Interaction	1.46	8.64	2.94	-4.30	7.22	0.39	2	125
X2	1.42	1.00	1.00	-0.54	3.38	0.15	4	125
X7	2.21	0.54	0.73	0.77	3.65	0.01	4	125
X2&X7	2.73	0.63	0.80	1.17	4.29	0.00	4	125
Interaction	-0.90	1.06	1.03	-2.92	1.12	0.37	4	125
X2	0.44	3.12	1.77	-3.02	3.91	0.74	Pooled TMLE	250
X7	2.27	0.36	0.60	1.09	3.46	0.00	Pooled TMLE	250
X2&X7	3.11	0.37	0.61	1.91	4.31	0.00	Pooled TMLE	250
Interaction	0.39	3.21	1.79	-3.12	3.90	0.77	Pooled TMLE	250
X1	5.30	1.22	1.11	3.13	7.47	0.00	2	125
X7	2.01	0.82	0.90	0.24	3.78	0.03	2	125
X1&X7	4.18	0.79	0.89	2.44	5.92	0.00	2	125
Interaction	-3.13	1.29	1.13	-5.35	-0.90	0.00	2	125
X1	5.30	1.22	1.11	3.13	7.47	0.00	Pooled TMLE	125
X7	2.01	0.82	0.90	0.24	3.78	0.03	Pooled TMLE	125
X1&X7	4.18	0.79	0.89	2.44	5.92	0.00	Pooled TMLE	125
Interaction	-3.13	1.29	1.13	-5.35	-0.90	0.00	Pooled TMLE	125
X1	3.56	0.67	0.82	1.95	5.16	0.00	4	125
X5	-1.35	0.51	0.72	-2.76	0.06	0.11	4	125
X1&X5	-0.20	0.52	0.72	-1.61	1.21	0.81	4	125
Interaction	-2.41	0.68	0.82	-4.02	-0.80	0.01	4	125
X1	3.56	0.67	0.82	1.95	5.16	0.00	Pooled TMLE	125
X5	-1.35	0.51	0.72	-2.76	0.06	0.11	Pooled TMLE	125
X1&X5	-0.20	0.52	0.72	-1.61	1.21	0.81	Pooled TMLE	125
Interaction	-2.41	0.68	0.82	-4.02	-0.80	0.01	Pooled TMLE	125

Table 2.3: Interaction Results from NIEHS Mixtures Data

	Estimate	Std. Error	Lower CI	Upper CI	Pr(>|t|)
(Intercept)	21.29	1.58	18.19	24.39	0.00
psi1	0.02	1.62	-3.16	3.20	0.99
psi2	0.59	0.67	-0.71	1.90	0.37

Table 2.4: Quantile G-Computation Interaction Results from NIEHS Synthetic Data

Condition	Psi	Variance	SE	Lower CI	Upper CI	P-value	Fold	N	Delta
LBXF03LA	0.08	0.00	0.03	0.01	0.14	0.02	1	51	2.00
LBXF03LA	-0.02	0.00	0.03	-0.07	0.04	0.58	2	51	2.00
LBXF03LA	0.02	0.00	0.04	-0.06	0.10	0.66	3	51	2.00
LBXF03LA	0.01	0.00	0.03	-0.04	0.06	0.71	4	51	2.00
LBXF03LA	-0.04	0.00	0.03	-0.09	0.02	0.20	5	51	2.00
LBXF03LA	0.00	0.00	0.03	-0.06	0.07	0.96	6	51	2.00
LBXF03LA	0.03	0.00	0.03	-0.02	0.09	0.24	7	51	2.00
LBXF03LA	0.04	0.00	0.04	-0.03	0.12	0.28	8	50	2.00
LBXF03LA	-0.02	0.00	0.03	-0.08	0.04	0.53	9	50	2.00
LBXF03LA	0.02	0.00	0.02	-0.02	0.06	0.32	10	50	2.00
LBXF03LA	-0.01	0.00	0.03	-0.06	0.04	0.69	11	50	2.00
LBXF03LA	0.07	0.00	0.03	0.01	0.13	0.03	12	50	2.00
LBXF03LA	0.02	0.00	0.02	-0.03	0.07	0.41	13	50	2.00
LBXF03LA	0.01	0.00	0.05	-0.09	0.11	0.81	14	50	2.00
LBXF03LA	-0.01	0.00	0.02	-0.06	0.04	0.66	15	50	2.00
LBXF03LA	0.05	0.00	0.03	-0.00	0.10	0.06	16	50	2.00
LBXF03LA	0.07	0.00	0.03	0.01	0.12	0.01	17	50	2.00
LBXF03LA	0.07	0.00	0.02	0.03	0.11	0.00	18	50	2.00
LBXF03LA	0.02	0.00	0.02	-0.03	0.06	0.49	19	50	2.00
LBXF03LA	0.05	0.00	0.04	-0.03	0.12	0.23	20	50	2.00
LBXF03LA	-0.02	0.00	0.01	-0.03	-0.00	0.04	Pooled TMLE	1007	2.00

Table 2.5: Furan 2,3,4,7,8-pncdf Lipid Adj (pg/g) Fold Specific and Pooled Findings

Chapter 3

NOVAPathways

Mediation analysis in causal inference typically concentrates on one binary exposure, using deterministic interventions to split the average treatment effect into direct and indirect effects through a single mediator. Yet, real-world exposure scenarios often involve multiple continuous exposures impacting health outcomes through varied mediation pathways, which remain unknown *a priori*. Addressing this complexity, we introduce NOVAPathways, a methodological framework that identifies exposure-mediation pathways and yields unbiased estimates of direct and indirect effects when intervening on these pathways. By pairing data-adaptive target parameters with stochastic interventions, we offer a semi-parametric approach for estimating causal effects in the context of high-dimensional, continuous, binary, and categorical exposures and mediators. In our proposed cross-validation procedure, we apply sequential semi-parametric regressions to a parameter-generating fold of the data, discovering exposure-mediation pathways. We then use stochastic interventions on these pathways in an estimation fold of the data to construct efficient estimators of natural direct and indirect effects using flexible machine learning techniques. Our estimator proves to be asymptotically linear under conditions necessitating $n^{-1/4}$-consistency of nuisance function estimation. Simulation studies demonstrate the $\sqrt{n}$ consistency of our estimator when the exposure is quantized, whereas for truly continuous data, approximations in numerical integration prevent $\sqrt{n}$ consistency. Our NOVAPathways framework, part of the open-source SuperNOVA package in R, makes our proposed methodology for high-dimensional mediation analysis available to researchers, paving the way for the application of modified exposure policies which can delivery more informative statistical results for public policy.

3.1 Introduction

Causal mediation analysis allows for the decomposition of an exposure's total effect on an outcome into direct and indirect pathways operating through an intermediate mediator or set of mediators. Identifying the pathways through which environmental mixtures impact health outcomes is crucial for corroborating causal inference of total effects and for developing

effective public health policies. This information can help to strengthen causal inference by providing evidence for a plausible biological mechanism underlying the observed association between the exposure mixture and the outcome. Additionally, if several chemicals with similar structures are found to operate through the same mediating pathway, it may suggest that other chemicals with similar structures may have the same mediating effects. This type of inference is consistent with coherence used in the Bradford Hill criteria [24]. Such evidence can be used to strengthen regulations of unstudied chemicals which are structurally similar to chemicals which have been found to have both total effects and effects through certain biological pathways leading to disease.

Mediation analysis is a powerful tool in the context of environmental health, aiding in the development of targeted interventions by elucidating the specific pathways through which environmental exposures influence health outcomes. Through the identification of mediator variable(s) that bridge the gap between exposure and outcome, mediation analysis opens up new potential avenues for interventions aiming to alleviate the harmful effects of exposure.

Taking inflammation as an example, if mediation analysis pinpoints this as a key mediator between exposure to air pollution and cardiovascular disease, it not only highlights an area for intervention but also serves as a biomarker of early effect. By observing the levels of inflammation, individuals at higher risk of developing cardiovascular disease due to their exposure to air pollution can be identified. This stratification of risk allows resources to be more efficiently allocated towards those who might benefit most from intervention efforts, such as the application of anti-inflammatory medications or dietary modifications.

In scenarios where immediate reduction of air pollution is not feasible, interventions targeted at this mediator can work to dampen the harmful effects of air pollution on cardiovascular health. Moreover, understanding the role of inflammation in this process provides invaluable insights into the biological mechanism underlying the exposure-outcome relationship, which could lead to the identification of additional intervention targets.

Thus, by delineating specific pathways, mediation analysis not only suggests potentially effective targets for intervention but also enhances the understanding of the biological process. This not only increases the likelihood of the interventions being effective but also optimizes the efficiency of resource usage by identifying those most in need. Through this process, mediation analysis could guide the development of comprehensive, multi-targeted approaches to reduce the harmful effects of environmental exposures.

Decomposing the total effects of a mixed exposures in environmental epidemiology presents unique challenges. Unlike single exposures, we do not know *a priori* which specific exposures or sets of exposures act through which mediators to cause the outcome. There can be multiple such pathways, and using the same data to identify these pathways and estimate a target parameter given these pathways can lead to biased results due to overfitting to the sample data. That is, both the discovered pathway and effects for this pathway are overfit to the sample data and may not generalize to the population level. Additionally, it is possible that multiple exposures use the same pathways, and thus may interact through this pathway, such as multiple heavy metals interacting through epigenetic mediators, which can have more than additive effects through this pathway. This highlights the importance of developing

methods that can identify and estimate the effects of mixed continuous-valued exposures on health outcomes through multiple mediators simultaneously while addressing issues of double-dipping and interactions between exposures. Currently, no such statistical methods exist to capture such complex exposure-outcome systems although, almost in all cases this is the system by which exposure leads to disease.

Building upon the seminal work of Sewall Wright, who introduced path analysis in 1934 [112], researchers gained a foundation for exploring causal relationships among observed variables using path diagrams and standardized path coefficients. This approach enabled the decomposition of the total effect of one variable on another into direct and indirect effects via intermediary variables. In 1972, Arthur Goldberger [32] further advanced the field by developing structural equation models (SEMs) for mediation analysis. By integrating path analysis with factor analysis, SEMs facilitated the modeling of intricate relationships between observed and unobserved (latent) variables. Goldberger's contribution linked path analysis to a more comprehensive statistical framework, providing enhanced precision in estimating causal effects while accounting for measurement error. Consequently, the scope and applicability of mediation analysis were significantly expanded. The initial development of path analysis and SEMs largely focused on parametric models, where assumptions about the distributional properties of the data and the functional form of relationships between variables were made. However, over time, researchers extended SEMs to include nonparametric and semiparametric approaches, allowing for more flexible modeling of relationships without strong distributional assumptions [75].

In recent years, the field of causal inference has witnessed substantial advancements with the introduction of non-parametric structural equation models and directed acyclic graphs. These developments have facilitated the non-parametric estimation of causal effects and the evaluation of conditions that permit causal effect identification from data [74, 83, 84, 89, 86]. While these novel approaches have addressed some limitations of traditional parametric structural equation models for mediation analysis, they also brought forth new challenges. Early non-parametric SEMs struggled with issues such as increased computational complexity; model identification, meaning that, in the absence of parametric assumptions, determining whether a non-parametric estimates are identifiable from the observed data is more challenging; sensitivity to choice of estimator; difficulty in assessing model fit and interpretability and limited available software.

Non-parametric partitioning of the causal influence of a binary treatment into natural indirect and direct impacts began by employing the potential outcomes framework proposed by Robins and Greenland [88]. The indirect impact measures the effect on the outcome variable via the mediator, while the direct impact measures the effect through all other pathways. Pearl [76] derived a similar effect partitioning utilizing non-parametric structural equation modeling. The identification of these natural (in)direct impacts depends on cross-world counterfactual independencies. Essentially, this means that we assume the outcomes of different imaginary scenarios, where intervention on the exposure and mediator, do not influence each other. The cross-world counterfactual independence assumption is not directly falsifiable from experimental data. This is because the assumption involves counterfactual

variables that correspond to different hypothetical interventions, and we can only observe one intervention outcome in a single experiment. Therefore, the natural (in)direct impact is not identifiable in a randomized experiment, which means that even in in randomized experiments we cannot know if these estimated mediation effects actual exist at the population level for a deterministic intervention.

These limitations arise because, in most causal inference research on mediation, deterministic interventions are studied, which assign fixed exposure values. Historically, binary exposures have been investigated for several reasons 1. interpretability: causal effects are easier to understand for binary exposures as they involve comparisons between two distinct groups or switching from one group to another; 2. estimation complexity: binary exposures often lead to simpler functional forms and estimation procedures, even in non-parametric settings; 3. identification: verifying assumptions for causal effects identification can be more straightforward for binary exposures; 4. potential outcomes framework: this framework is more intuitive for binary exposures, as there are only two potential outcomes for each individual.

To avoid limitations of binary exposures while retaining interpretability and relaxed identification assumptions, stochastic interventions can be implemented. Stochastic interventions allow exposures to be a random variable after conditioning on baseline covariates. For example, in the context of air pollution exposure and cardiovascular outcomes, we can consider a stochastic shift intervention where air pollution exposure is reduced by an amount δ for each individual in the population. Therefore, this post-intervention distribution still depends on the originally observed air pollution levels. We then would estimate the impact under this post-intervention distribution and compare the average to the observed outcomes under observed air pollution exposures. Stochastic interventions offer analytical benefits over deterministic approaches by enabling the straightforward definition of causal effects for continuous exposures, providing an interpretation that is easily understood by those familiar with linear regression adjustment. Estimation of total effects for stochastic interventions has been explored in various studies, including methods for modified treatment policies and propensity score interventions for binary exposure distributions [51, 19, 95, 85]. Nevertheless, these studies do not focus on decomposing the effects of stochastic interventions into direct and indirect effects, which was first investigated in [21].

In [21], the authors introduce a decomposition of a stochastic intervention's effect into direct and indirect components. This approach identifies (in)direct effects without necessitating cross-world counterfactual independencies, producing experimentally testable scientific hypotheses that can be empirically tested by intervening on the mediator and exposure. The authors develop a one-step non-parametric estimator based on the efficient influence function, incorporating machine learning regression techniques, and provide $\sqrt{n}$-rate convergence and asymptotic linearity results. Importantly, the proposed method provides definition and estimation of non-parametric mediated effects for continuous exposures. However, in the software implementation of the proposed method, the authors employ a reparameterization of specific integrals as regressions and the authors treat the exposure as binary to reduce computation complexity by avoiding direct estimation of the probability density function

(PDF) and estimating the probability mass function (PMF) instead. Likewise, restricting the software to a binary exposure also avoids numeric integration necessary for the estimator. While this approach enables the inclusion of multiple mediators, it necessitates a binary exposure to function effectively. This limitation motivates the work presented here.

In many environmental epidemiology cases, it is crucial to understand the specific mediators through which particular exposures impact an outcome. Instead of reparameterizing an estimand to avoid high-dimensional density estimation or integrals, identifying individual mediators and estimating stochastic effects solely through these mediating pathways leads to deeper interpretation when dealing with multiple mediators. The random variables driving the outcome can be treated as parameters, where these mediators are identified using one portion of the data, and direct/indirect effects are estimated for this mediator using another part of the data. Estimation becomes considerably more complex with multiple exposures, as the connections between exposures and mediators remain unknown. Thus, these paths must be discovered in the data, and mediation analysis employing stochastic interventions can then be estimated for these paths.

This study presents a methodological approach for estimating mediation effects in the presence of high-dimensional exposures and mediators. We employ a cross-validated framework, where in path-finding folds, a cross-validation process is used to identify the mediating paths through a series of semi-parametric regressions. With these paths established, we estimate the direct and indirect effects of a stochastic intervention on the exposure through the mediator, both identified in the path, in an estimation fold. Drawing on the efficient influence function from Diaz et al. [21], we directly compute the integrals required for each component of the efficient influence function, rather than reparameterizing the estimates when the exposure is continuous. We also build in estimation for the case where the exposure is quantized, for example, into bins which represent quartiles. This approach enables the mediation of continuous/discrete exposures which are unknown *a priori* through mediators which are also unknown *a priori* and can take on multiple variable types.

The use of stochastic interventions in a semi-parametric framework provides a promising approach for estimating direct and indirect effects of exposure mixtures on health outcomes through mediation pathways. To our knowledge, no such methods exist in the causal inference literature which both makes available mediation for a continuous/discrete exposure and data-adaptive discovery of mediating paths. Our method proposed here is available for use in the SuperNOVA package in R which also estimates interaction and effect modification of a mixed exposure using stochastic interventions and data-adaptive target parameters.

3.2 The Estimation Problem

Our mediation parameter of interest for a continuous exposure was first described in [21] and therefore, what follows in the our mediation framework for data-adaptively discovered mediation pathways is based on this previous work. That is, the notation, target parameter, identification and efficient influence function are all the same as in [21]; however we extend

this method to work in the continuous case and data-adaptively identify exposure-mediator pathways. To make this current work more self-contained we review and explain with brevity parts of their estimator in order to describe our approach making estimates work in the fully continuous case of data-adaptively identified pathways.

We consider the causal inference problem involving a multivariate continuous, categorical, or binary exposure (A), a continuous, categorical, or binary outcome (Y), a multivariate continuous, categorical, or binary mediator (Z), and a vector of observed covariates (W) which are also a variety of data types. Let $O = (W, A, Z, Y)$ be a random variable with distribution $\mathbb{P}$. We denote the empirical distribution of a sample of n independent and identically distributed observations $O_1, ..., O_n$ as $\mathbb{P}_n$. For any given function $f(o)$, we denote $\mathbb{P}f = \int f(o)d\mathbb{P}(o)$ and use $\mathbb{E}$ to represent expectations with respect to $\mathbb{P}$ averaging over all randomness. We assume $\mathbb{P}$ belongs to $\mathcal{M}$, the nonparametric statistical model comprising all continuous densities on O with respect to a dominating measure v, with p denoting the corresponding probability density function. We go through the framework first ignoring the data-adaptive selection of subsets of the $\{A, Z\}$. We then introduce the data-adaptive component which follows naturally.

Our approach diverges from previous methods, focusing on data-adaptively identifying which sets of exposures ($\hat{A}$) impact which sets of mediators ($\hat{Z}$). This approach bypasses the need for estimating the high-dimensional joint impact of exposures through the mediators, an effort that often encounters the 'curse of dimensionality,' a phenomenon that complicates accurate modeling and prediction due to exponential increase in volume associated with adding extra dimensions in the exposure space. The discovered $\hat{A}$ and $\hat{Z}$ represent the "estimated" or selected subsets from the full set of A and Z variables.

Probability density functions and regression functions are represented as follows:

- $g(a|w)$: Represents the conditional probability density or mass function of A given $W = w$.
- $Q(a, z, w)$: Represents the expected outcome given the variables A, Z, and W.
- $e(a|z, w)$: Represents the conditional density or probability mass function of A given (Z, W).
- $q(z|a, w)$ and $r(z|w)$: Denote the conditional densities of Z.

To define our counterfactual variables, we use the following nonparametric structural equation model (NPSEM):

$$W = f_W(U_W); A = f_A(W, U_A); Z = f_Z(W, A, U_Z); Y = f_Y(W, A, Z, U_Y).$$

This set of equations signifies a mechanistic model, grounded in nonparametric statistical methods, that is assumed to generate the observed data O. It incorporates several fundamental assumptions. First, an implicit temporal ordering is assumed, with Y occurring after Z, A, and W; Z taking place after A and W; and A happening after W. Second, each variable (i.e.,

W, A, Z, Y) is assumed to be generated from the corresponding deterministic, yet unknown, function (i.e., f_W, f_A, f_Z, f_Y) of the observed variables that precede it temporally, as well as an exogenous variable, denoted by U. Each exogenous variable is assumed to encompass all unobserved causes of the corresponding observed variable.

In the context of nonparametric statistics, the independence assumptions on the exogenous variables $U = (U_W, U_A, U_Z, U_Y)$ necessary for identification will be addressed in the assumptions section. This approach allows the model to accommodate the complexities and nuances of the relationships between the variables without relying on specific functional forms or parametric assumptions.

Our causal effects of interest are characterized by hypothetical interventions on the NPSEM. In our situation, we focus on an intervention where the equation associated with A is changed, and the exposure is drawn from a user-defined distribution $g_\delta(a|w)$. This distribution relies on g (the conditional density under observed exposures) and is indexed by a user-specified parameter δ. We assume that when $\delta = 0$, $g_\delta = g$. Let A_δ represent a draw from $g_\delta(a|w)$.

In our scenario, the distribution g_δ is given by $g(a - \delta|W)$, which indicates a shift of δ in the conditional density of A. This shift corresponds to a modified treatment policy aimed at reducing exposure by δ. Essentially, the intervention involves removing the equation associated with A and establishing the exposure as a hypothetical regime, $d(A, W)$. The regime d depends on the natural exposure level A (i.e., without any intervention) and covariates W. For instance, if A denotes continuous exposures such as various air pollution factors (Carbon Monoxide, Lead, Nitrogen Oxides, Ozone, Particulate Matter, etc.) related to asthma incidence Y, we may be interested in investigating the expected asthma incidence if all individuals experienced a δ-unit reduction in Lead exposure, while keeping other exposures and covariates unchanged.

Assume that the distribution of A given $W = w$ is supported within the interval $(l(w), u(w))$. In other words, the minimum pollution level for an individual with covariates $W = w$ is $l(w)$. We can then define a hypothetical post-intervention exposure, $A_\delta = d(A, W)$, as follows:

$$d(a, w) = \begin{cases} a - \delta & \text{if } a > l(w) + \delta \\ a & \text{if } a \leq l(w) + \delta \end{cases}$$

Here, $0 < \delta < u(w)$ is an arbitrary value provided by the user. This regime can be further refined by allowing δ to be a function of w, thereby enabling the researcher to specify a different change in pollution levels as a function of factors such as demographic characteristics or geographical location. This intervention was initially proposed by [48] and [20] and [34].

We are interested in the population intervention effect (PIE) of A on Y using stochastic interventions. That is, given values for an exposure and mediator (a, z), we examine the counterfactual outcome $Y(a, z) = f_Y(W, a, z, U_Y)$, the expected outcome if all individuals were exposed to these values for the exposure and mediator. We also examine the counterfactual mediator $Z(a) = f_Z(W, a, U_Z)$ or the expected value the mediator takes on given exposure

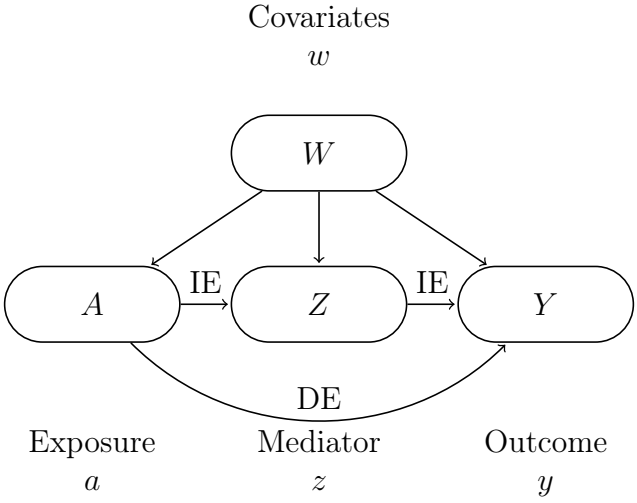
$A = a$. The counterfactual $Y(a, z)$ represents the outcome in a hypothetical scenario where $(A, Z) = (a, z)$ is fixed for all individuals. We are interested in the contrast between the expected outcome given an intervention A_δ which say, reduces exposure to pollution and the expected outcome under no intervention, the observed outcome under observed exposures. This looks like:

$$\psi(\delta) = \mathbb{E}\{Y(A_\delta) - Y\}.$$

Drawing from causal inference literature on mediation, we know that since A is a cause of Z, any intervention altering exposure to A_δ also affects the counterfactual mediator $Z(A_\delta)$. Owing to the consistency ensured by the NPSEM, we obtain $Y(A, Z) = Y$ and $Z(A) = Z$. In addition, from Pearl's [76] law of composition we can express $Y(A_\delta, Z(A_\delta)) = Y(A_\delta)$. In words, this means that the expectation of Y under dual shift is implied by a shift in A ignoring Z. Consequently, the PIE can be decomposed into a population intervention direct effect (PIDE) and a population intervention indirect effect (PIIE). The interepretation of these effects are the same as natural direct and indirect effects but are for a stochastic intervention rather than a deterministic intervention on A.

$$\psi(\delta) = \underbrace{\mathbb{E}\{Y(A_\delta, Z(A_\delta)) - Y(A_\delta, Z)\}}_{\text{PIIE}} + \underbrace{\mathbb{E}\{Y(A_\delta, Z) - Y(A, Z)\}}_{\text{PIDE}}.$$

Essentially, the direct effect demonstrates the impact of an intervention that modifies the exposure distribution while maintaining the mediator distribution at the level it would have been without any intervention. On the other hand, the indirect effect quantifies the influence of an indirect intervention on the mediators, initiated by changing the exposure, while keeping the exposure intervention constant.



Above is a simple directed acyclic graph (DAG) which illustrates the IE through the mediator Z and DE which is the causal effect not through Z. For example, in a study investigating the effects of environmental exposure, such as air pollution, on respiratory

health, the direct effect measures how changing pollution levels impact health outcomes, assuming the mediators (e.g., time spent outdoors) remain unchanged. The indirect effect, conversely, evaluates how health outcomes are influenced by changes in the mediators (e.g., reduced time spent outdoors) that result from modifying pollution levels, while the pollution intervention remains constant.

Above, $Y(A, Z) = \mathbb{E}(Y)$, is simply estimated by the empirical mean in the sample. Moving forward, the optimality theory described in [21] which we review and estimators we present for the truly continuous exposure case focus on $\theta(\delta) = \mathbb{E}\{Y(A_\delta, Z)\}$. These two terms are then used in calculation of the direct effect. Because $\mathbb{E}\{Y(A_\delta, Z(A_\delta))\} = \mathbb{E}\{Y(A_\delta)\}$, which in words is simply the total effect in Y after shifting A ignoring Z. That is, if we were to construct an efficient estimator for a shift in A ignoring Z these estimates encapsulate the indirect effect A has through Z as in the total effect. Ivan Diaz and Mark van der Laan [48] first proposed estimators of the total effect of a stochastic shift intervention including inverse probability weighted, outcome regression, and doubly robust estimators based on the framework of targeted minimum loss-based estimation (TMLE) where in each case data adaptive machine learning can be used to estimate the relevant nuisance parameters. Call the total effect $\theta(\delta)_t$, which is the expected Y given a shift in A and includes the implied shift in $Z(A_\delta)$. Call $\mathbb{E}\{Y(A_\delta, Z) - Y(A, Z)\}$, $\theta(\delta)_d$, the direct effect or the effects of shift in A keeping Z fixed. Lastly, $\mathbb{E}\{Y(A_\delta, Z(A_\delta)) - Y(A_\delta, Z)\}$, the effects of a shift in Z due to a shift in A keeping A fixed we call $\theta(\delta)_i$. Then:

$$\theta(\delta)_t = \theta(\delta)_d + \theta(\delta)_i$$

Which means we can then estimate the indirect effect as:

$$\theta(\delta)_i = \theta(\delta)_t - \theta(\delta)_d$$

Which is simply estimating the indirect effect by subtracting the total effect from the direct effect, this provides us with the point estimate. We can do inference on this difference by utilizing work from [21] which provides an efficient estimator for $\theta(\delta)$ for the construction of the direct effect $\theta(\delta)_d$. We use TMLE or one-step estimators proposed from [48] to estimate $\theta(\delta)_t$ and we use the scalar delta method to estimate $\theta(\delta)_i$. Moving forward we describe $\theta(\delta)$, or $\mathbb{E}\{Y(A_\delta, Z)\}$, the average outcome under a shift in A keeping Z at natural values.

Identification of the Causal Parameter

We can evaluate the causal effect of our intervention by considering the counterfactual mean of the outcome under our stochastically modified intervention distribution. This target causal estimand is $Y(a, z)$, which is the counterfactual outcome we would observe when $\mathbb{P}((A, Z) = (a, z)) = 1$.

Our causal quantitiy is:

$$\theta(\delta) = \int y p_{Y(A_\delta, Z)}(y)\ dy$$

[21] describe identification for this parameter and we briefly review here. We must assume that the data is generated by independent and identically distributed units, and that there is no unmeasured confounding, consistency, or interference (discussed in more detail in subsequent sections). Under these assumptions, $\theta(\delta)$ can be identified by a functional of the distribution of O:

$$\theta(\delta) = \int_{\mathcal{W}} \int_{\mathcal{A}} \int_{\mathcal{Z}} Q(a, z, w) g_\delta(a \mid w) r(z|w) q(w) \; dw \; da \; dz$$

Mechanically this is the outcome predictions from our Q model integrated over density predictions from our g model under δ shift integrated over our the conditional mediator and covariate density.

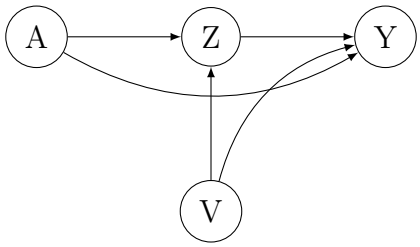
Interpreting the statistical effects in our analysis as causal rests upon two assumptions: common support and conditional exchangeability (or ignorability). These are standard assumptions in causal inference that require consideration in mediation.

Common support, also known as positivity or overlap, is a fundamental assumption in causal inference that ensures that the distribution of the exposure of interest is well defined and supported by the data. For each individual in the population, there should be a non-zero probability of observing the shifted exposure value given their observed covariates. This assumption ensures that the exposure effect is identifiable and that causal inference can be validly conducted. In our case, positivity refers to the probability density of exposure being bounded away from zero or one after an exposure shift. We propose a method that data-adaptively finds a shift which does not lead to positivity violations (described later).

Conditional exchangeability, or ignorability, is related to the assumption made in [102]. In our context, it means that given the observed covariates, the distribution of the potential outcomes, $Y(a, z)$, is independent of the actual exposure, A, and mediator, Z, assignments. This assumption is akin to stating that we have adequately controlled for confounding.

Here, it's essential to note that we need conditional exchangeability both for the exposure-outcome and mediator-outcome relations. This implies that all confounders between the exposure A and outcome Y, and between the mediator Z and outcome Y, should be measured and properly adjusted for. If this assumption is violated—if there are unmeasured confounders—it can lead to biased effect estimates.

Consider the directed acyclic graph (DAG) below:



This DAG illustrates the relations between the exposure A, mediator Z, confounder V, and outcome Y. Here, V can be seen as a confounder that affects both Z (the mediator) and

Y (the outcome). Conditioning on a collider Z when there are unmeasured confounders (V), would open a pathway from A to $Y(a, z)$, introducing bias into our estimates.

Additionally, the methods presented here cannot account for situations where the mediator-outcome confounder V is affected by the exposure A. As this too opens up a backdoor path that would lead to bias [21].

Efficient Estimation of the Direct Effect

In this section we focus on the efficiency theory for estimating $\theta(\delta)$ within the nonparametric model $\mathcal{M}$, with a focus on the efficient influence function (EIF) which was originally derived in [21]. Diaz and Hejazi offer a rigorous breakdown of the EIF for this part of the direct effect and we give a brief overview here to explain our approach for estimating each part of the EIF. The EIF is a fundamental concept in semi-parametric estimation theory. It plays a vital role in determining the asymptotic behavior of all regular and efficient estimators. In simpler terms, the EIF contains the information to predict how these estimators perform when the sample size approaches infinity. Calculating the EIF is crucial for constructing locally efficient estimators for $\theta(\delta)$. Locally efficient estimators are estimators that achieve the best possible asymptotic variance within a specified class of estimators under certain regularity conditions within a stated statisticla model. They are optimal in the sense that, asymptotically, they have the lowest variance among all unbiased estimators in their class. [21] derived the EIF for this problem: we briefly describe each part of the EIF here and describe how we estimate its components for the case where A is a continuous/discrete exposure. The efficient influence function for $\theta(\delta)$ in the nonparametric model $\mathcal{M}$ for a modified treatment policy is $D^Y(o) + D^A(o) + D^{Z,W}(o) - \theta(\delta)$, where:

$$
\begin{aligned}
D^Y(o) &= \frac{g_\delta(a|w)}{e(a|z,w)}\{y - Q(a,z,w)\}, \\
D^{Z,W}(o) &= \int Q(a,z,w) g_\delta(a|w)\ da \\
D^A(o) &= \phi(a,w) - \int \phi(a,w) g(a|w)\ da
\end{aligned}
$$

Where:

$$
\begin{aligned}
\phi(a,w) &= \int Q(d(a,w),z,w) r(z|w)\ dz \\
&= \mathbb{E}\left[\frac{g(A|W)}{e(A|Z,W)} Q(d(A,W),Z,W) \Big| A = a, W = w\right],
\end{aligned}
$$

Constructing an efficient estimator always involves estimating the EIF and so here we describe at a high level how we estimate each component in the rest of the article.

D^Y describes the "weighting factor" in the EIF which adjusts the residuals of the outcome model (Q) based on the differences in exposure distributions between the intervention g_δ and the natural course of exposure e. We calculate this by directly constructing estimators for the conditional densities of g and e. Likewise, Q is simply an outcome regression model which is estimated using flexible machine learning. Therefore, in the case where A is continuous, we use conditional density estimators to estimate the conditional density functions used in this nuisance parameter. When A is discrete, we can also use an ensemble of multinomial regression estimators which provide the probability of exposure falling in each "bin". This probability mass function then replaces the probability density function used when A is continuous.

$D^{Z,W}$ is expected outcome (Q) multiplied by the estimated probability density of the exposure under a shift by δ, and integrating over all possible values of the exposure a. This takes into account the potential shift in the distribution of w (which affects the exposure), to provide a more accurate prediction of the outcome Y. For this estimation we directly integrate the two functions over the exposure using Monte Carlo integration of the exposure variable over the exposure range. That is, exposure values a are shifted until they meet the upper or lower bound in which case they simply take on the min or max value depending on the direction of δ. In the case where A is discrete, g_δ is simply the probability for the bin that corresponds to $a \pm \delta$ depending on the direction. For example, if A is discretized into quartiles and δ is 1, then if a is quartile 1, g_δ is the probability of quartile 2. In this case the integral is simply a weighted sum:

$$D^{Z,W}(o) = \sum_{a_k \in A} Q(a_k, z, w) g_\delta(a_k|w)$$

For D^A the first expression $\phi(a, w)$ can be calculated using either integration or regression. The first line of the expression uses integration to calculate this expected outcome by averaging the outcome model Q over all possible values of z, weighted by the conditional density of z given w, denoted as $r(z|w)$. The second line of the expression uses an alternative formulation to calculate the same expected outcome. It uses the conditional expectation formula to take the conditional expected value of Q given $A = a$ and $W = w$, where the expectation is taken with respect to the conditional density of A given W, denoted as $g(A|W)$, divided by the inverse of the conditional density of A given Z and W, denoted as $e(A|Z, W)$, which effectively regresses out the effect of Z from A. Therefore, it possible to estimate $\phi(a, w)$ by either integrating or using pseudo-regression. We take both approaches to compare finite sample performance in both estimation approaches. For the integration approach, we directly estimate the conditional density of the mediator given covariates and use this function in the integration with Q over z using a Monte Carlo approach. Again if A is discrete this looks like:

$$D^A(o) = \phi(a, w) - \sum_{a_k \in A} \phi(a_k, w) g(a_k|w)$$

Because $\phi(a, w)$ is the integration of Q and r over z and does not include A as an outcome,

it is still necessary to estimate the conditional density of Z given W even when the exposure is discrete. In this discrete exposure case, we use a double integration approach and pseudo regression approach.

Monte Carlo Integration

Monte Carlo (MC) integration is a numerical integration technique that uses random sampling to approximate the integral of a function over a given domain. In our case the range of exposures and/or mediators are the domains to integrate over. MC integration works by first generating random points within the domain. Then, the function values are computed at these points. The average of these values is then multiply by the volume of the domain. As the number of samples increases, the approximation converges to the true integral value.

In our case we are integrating the product of two density/regression estimators, for example in the case of, $D^{Z,W}(o) = \int Q(a, z, w) g_\delta(a|w)\, da$, MC integration can be more advantageous than quadrature methods for several reasons:

1. Handling high-dimensional and non-linear functions: The product of fits using, for example, two Super Learners for Q and g, may result in complex, non-linear, and high-dimensional functions. MC integration is well-suited for handling such functions, as it does not rely on any specific parametric assumptions or require the function to be smooth or continuous.

2. Adaptability to irregular functions: MC integration is adaptive to irregularities in the function being integrated, making it a reasonable method for integrating the product of two flexible Super Learners fits, which can have irregular shapes across covariates. Quadrature methods, on the other hand, often rely on the function being smooth or continuous and may struggle with irregular functions.

3. Scalability: MC integration is easily scalable to high dimensions, making it suitable for problems with a large number of covariates. Quadrature methods, in contrast, can suffer from the curse of dimensionality, where the number of required evaluation points grows exponentially with the dimensionality, leading to an intractable computational burden.

4. Convergence properties: MC integration has desirable convergence properties, meaning that as the number of random samples increases, the accuracy of the approximation improves. This allows for obtaining more accurate estimates, even for complex and irregular functions.

5. Ease of implementation: MC integration is relatively simple to implement and can be easily parallelized for efficient computation on modern hardware. Quadrature methods, on the other hand, can be more complex and challenging to implement, especially for high-dimensional and irregular functions.

For these reasons, we use MC for estimating the necessary integrals of each nuisance function. MC integration is much faster than adaptive quadrature, especially in our case where we need to integrate these functions at every vector of covariates (for each observation). To ensure that the number of iterations is scaled by sample size the number of iterations used in the MC integration is set to four times sample size in this paper.

Estimation

Direct Effect

[21] derive the efficient influence function $D_{\eta,\delta}$ to construct a robust and efficient estimator, which is defined as the solution to the estimating equation $P_n D_{\hat{\eta},\delta} \hat{=} 0$ in θ, given a preliminary estimator $\hat{\eta}$ of η. They advise utilizing cross-fitting in the estimation process to avoid entropy conditions of the initial estimators which we employ in our approach. To do this, the index set $1, \ldots, n$ is randomly partitioned into K equally sized estimation samples, V_k. For each k, the corresponding training sample T_k is obtained by excluding V_k from the index set. The estimator $\hat{\eta}_{T_k}$ is derived by training the prediction algorithm using only the data in T_k. The index of the validation set containing observation i is denoted by $V_k(i)$. The estimator is thus defined as:

$$\hat{\theta}(\delta) = \frac{1}{n}\sum_{i=1}^{n} D_{\hat{\eta}_{k(i)},\delta}(O_i) = \frac{1}{n}\sum_{i=1}^{n}\left[D^{Y}_{\hat{\eta}_{k(i)},\delta}(O_i) + D^{A}_{\hat{\eta}_{k(i)},\delta}(O_i) + D^{Z,W}_{\hat{\eta}_{k(i)},\delta}(O_i)\right]$$

Effectively, the efficient estimator is the average of the cross-estimated sum of each nuisance parameter. Subtracting the mean from this sum of nuisance parameters then gives us the EIF for this shift parameter since the EIF is defined as $D^{Y}_{\eta,\delta}(o) + D^{A}_{\eta,\delta}(o) + D^{Z}_{\eta,\delta}(o) - \theta(\delta)$.

When estimating $\theta(\delta)$ compared to the observed outcome, we employ the scalar delta method by subtracting the two efficient influence functions, resulting in an EIF for $\theta(\delta)_d$ that can be used for constructing confidence intervals and performing hypothesis testing. By subtracting the two EIFs and calculating the variance of the resulting EIF scaled by n observations, we obtain the variance of $\theta(\delta)_d$, which is asymptotically Gaussian and centered around the true difference. Finally, we construct confidence intervals and conduct hypothesis testing using the standard error. This gives us our final point and variance estimates for the $\theta(\delta)_d$.

Indirect Effect

One Mediator We employ one-step estimation or targeted maximum likelihood estimation (TMLE) to estimate the expected outcome of a shift in exposure A without considering the mediator Z. TMLE solves the efficient influence function (EIF) and the delta method is used to estimate the total effect [48] by subtracting this EIF from the observed Y EIF (Y - $Q(a, w)$). By solving the EIF for the total effect parameter $\theta(\delta)_t$ using TMLE/one-step estimation and applying the delta method, we obtain the EIF for the indirect effect parameter

$\theta(\delta)_i$ by subtracting $\theta(\delta)_d$ from $\theta(\delta)_t$, the same is done for the point estimates. Although we use different approaches for estimating $\theta(\delta)_t$ (TMLE) and $\theta(\delta)$ (estimating equations), both result in efficient estimators. According to the central limit theorem, the distribution of each estimator is Gaussian and centered at the true value. We can compute the estimate of the variance σ_n^2, allowing for Wald-style confidence intervals to be computed at a coverage level of $(1-\alpha)$ as $\psi_n \pm z(1-\alpha/2) \cdot \sigma_n/\sqrt{n}$.

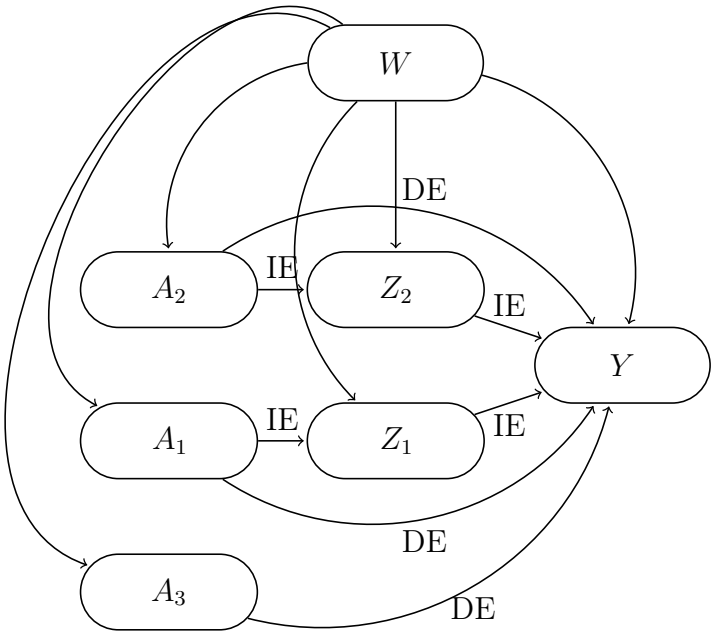
Many Mediators In research situations where multiple mediators are measured, we need to adjust the above described methodology in order to isolate the indirect effect for a given pathway. A simple subtraction of the direct effect from the total effect to derive the indirect effect when multiple potential mediators are present would yield an oversimplified estimation. This approach would instead estimate the collective indirect effect through all potential mediators. This would not provide the specific indirect effect attributable to the pathways of interest. To delineate the specific indirect effect through the mediator of interest, we adopt a slightly different approach. When we estimate the total effect, we adjust for all the other mediators but not the mediator of interest in the model, symbolized as $\mathbb{E}[Y|A, Z_{\neg i}, W]$ where $Z_{\neg i}$ represents all mediators other than the mediator of interest. This enables us to isolate the total effect of A on Y with respect to the particular $A - Z$ pathway under investigation. The rest of the estimation procedure is the same where we subtract this total effect point estimate and EIF from the direct effect to get the indirect effect estimates.

3.3 Finding Mediating Pathways

Mixed Exposures and Mediators

Up to this point, we have focused on fixed exposure A and mediator Z, showing the efficient influence functions (EIFs) necessary for estimating the natural (in)direct effects. However, in scenarios involving mixed exposure and multiple potential mediating paths, the most important exposure-mediator $(A - Z)$ paths among a potentially high-dimensional set are typically unknown.

Consider a hypothetical situation where five exposures represent different pesticides (A_{1-5}), and five measured variables potentially mediate the effects of these pesticides (Z_{1-5}), representing mediation pathways through neurotoxicity, endocrine disruption, oxidative stress, immune system, and DNA damage. The outcome Y is a hypothetical cancer. In this scenario, let us imagine that the effects of A_1 and A_2 are mediated through Z_1 and Z_2 respectively, while A_3 shows no measured indirect effects, and A_4 and A_5 have no impact on the outcome. Additionally, $Z_3 - Z_5$ do not mediate any measured exposures. A directed acyclic graph (DAG) illustrating this situation is presented below.



Scenarios featuring mixed exposures and multiple mediating pathways are not uncommon in real-world contexts. For instance, agricultural workers may encounter multiple pesticides, chemicals, and environmental factors, each acting through different mediating pathways to exert health effects. Similarly, industrial employees can be exposed to various chemicals, urban residents to diverse air pollutants, and individuals practicing unhealthy lifestyle habits to the risk of chronic diseases. Even the spread of infectious diseases and climate change can involve a complex interplay of multiple exposures and mediating pathways.

In all these instances, understanding the complex interplay between various exposures and mediating pathways is crucial. However, since the $A - Z$ pathways are not known *a priori*, and continuously testing different exposure-mediator pathways could lead to type 1 error, a data-driven approach is essential.

Basis Function Estimators for Pathway Discovery

Uncovering mediating pathways in our data requires a non-parametric method, as not only are pathways not known *a priori* but also the functional forms underlying their relationships are not known as well. We leverage a series of discrete Super Learners—best fitting flexible estimators selected from a library of candidate estimators—for this task. These constituent learners used in the Super Learner construct non-linear models through linear combinations of basis spline terms and their tensor products, rendering them ideal for the task at hand.

In the most flexible setting, we form indicator variables for each predictor. These variables denote if a predictor X is less than or equal to a specific value x_s, this approach can be

extended for combinations of predictors, like X_1, X_2. Consequently, a function of our outcome distribution can be represented as:

$$\psi_\beta = \beta_0 + \sum_{s \subset \{1,2,\ldots,p\}} \sum_{i=1}^{n} \beta_{s,i} \phi_{s,i}, \text{ where } \phi_{s,i} = I(X_{i,s} \leq x_s), A \in \mathbb{R}^p$$

Here, s denotes indices of subsets of the X.

This estimator is known as the highly adaptive lasso (HAL) estimator [5]. Its unique attribute is its theoretically proven $n^{-1/4}$ convergence, a necessary condition for the $\sqrt{n}$ rate conditions to hold for our estimator and for convergence to selection of basis functions of true pathways for the underlying yet unknown DGP. However, the HAL estimator is not scalable in high dimensions. Therefore, the estimators employed in NOVAPathways that return tensor products of arbitrary order and approximate this more exhaustive approach include the earth [67], polySpline [81], and hal9001 (under restrinctions) [16] packages. Each method utilizes a linear combination of basis functions to estimate the conditional outcome, allowing us to extract variable sets used in these basis functions as our data-adaptively identified variable sets.

Our process to construct pathways includes:

1. Fitting $\mathbb{E}(Z|A, W)$ as $\beta_0 + \sum[\beta_s \cdot h(A, W)_s]$. Here, β_0 is the intercept, β_s are the coefficients, $h(A, W)_s$ are the basis functions involving A and W, and the sum is over all basis functions in the model.

2. Extracting basis functions for A with non-zero coefficients.

3. Fitting $\mathbb{E}(Y|A, Z, W) = \beta_0 + \sum[\beta_s \cdot g(A, Z, W)_s]$.

4. Matching A to Z pathways: we align the basis functions involving A from the first model ($\mathbb{E}(Z|A, W)$) with the basis functions involving Z from the second model ($\mathbb{E}(Y|A, Z, W)$), if used, to identify the $A - Z$ pathways. Pathways are also AZ basis functions used directly in the second model.

This stepwise approach is necessary. In cases where the effects of A go entirely through Z, or when effects that don't pass through Z are negligible for the model fit, the second model will only contain basis functions for Z. As such, the first model is required to illuminate the underlying A driver, thereby establishing the pathway connection. In summary, this approach non-parametrically identifies mediating pathways in a mixed exposure scenario.

Non-Parametric Analysis of Variance for Identifying "Important" Pathways

Upon identification of the optimal basis spline estimator for each sequential regression segment, we use an ANOVA-like decomposition of basis functions to rank the "important" variable

sets employed by each algorithm; thereby filtering to the most important pathways. This selection process becomes critical in high-dimensional A and Z scenarios, where the possible paths multiply, and the goal is to discern the most influential pathways on Y adaptively. We apply a variant of ANOVA, generalized for large-scale, non-parametric models.

In this context, we partition the response variable's variance based on the contributions from distinct basis factors. For multivariate adaptive regression models, the variance is decomposed into the individual contributions of linear basis functions. For highly adaptive lasso models, zero-order basis functions (exposure-covariate indicators) make these contributions. In both cases, the F-statistic is calculated using the traditional ANOVA formula, albeit with modifications to accommodate the non-linear model concerning the original covariates.

The response variable's variance is split into two: variance explained by the linear combination of basis functions and the residual variance left unexplained by the model. The F-statistic represents the ratio of explained to residual variance, adjusted for degrees of freedom. The F-statistic is computed for each basis using the standard formula, presuming a linear relationship between the response variable and basis functions, though the basis functions themselves need not be linear in the original covariates.

Once we've computed F-statistics for each basis function, we have a measure of each basis function's contribution to the explained variance in the response variable. However, these basis functions represent transformations (which may or may not be linear) of the original variables. Hence, we're interested in getting a measure of the importance of each variable, not just the individual basis functions.

To aggregate these F-statistics to the variable level, we need to map each basis function back to the original variables it was derived from. We do this using the naming conventions of the basis functions, which contain the names of the variables they were derived from. This allows us to identify which F-statistics belong to which variables.

Once this mapping is complete, we have a collection of F-statistics for each variable, with each statistic representing a different basis function of that variable. To aggregate these statistics, we take their sum. The sum provides a measure of the total contribution of all basis functions of a variable to the explained variance in the response variable. In other words, it gives us a measure of the overall importance of that variable.

Finally, we rank the variables according to these sums of F-statistics, which we can then use to filter variables in subsequent analyses. It's important to note that this approach assumes that the F-statistics of basis functions of a variable can be meaningfully added together. This assumption holds true if the basis functions are orthogonal (i.e., uncorrelated), as is the case with splines. However, it may not hold if the basis functions are correlated, which might be the case with other types of basis functions.

We then rank variable sets based on the computed F-statistics, and subsets can be decided based on the F-statistic quantile to yield a concise variable list. The resultant list contains variable sets that meet the F-statistic threshold. This procedure is applied to both $\mathbb{E}(Z|A, W)$ and $\mathbb{E}(Y|A, Z, W)$ models, to filter A based on the F-statistics driving each mediator, and to filter Z, A, and $A - Z$ basis functions, respectively. This variable set process is implemented

within a V-fold cross-validation framework using data, which we discuss in the subsequent section.

It's worth noting that our proposed methodology operates on the principle of heuristics, aiming for an approach that's both computationally practical and effective. It's not designed to achieve theoretical optimality but rather to robustly identify potential pathways, that are part of the data-adaptive target parameter. Theoretical rigor is still maintained during the estimation step. While other methods could be employed, our approach offers a blend of simplicity, speed, and suitability for the task at hand by using basis-function estimators in the two step process that are both flexible but also interpretable, which allows us to construct the pathways. For example, methods could be employed such as using exposure or exposure-mediator sets used in the branches of a best fitting decision tree [63] to identify potential pathways.

3.4 Cross-Estimation

Ensuring the estimators of our mediation target parameters meet the requisite complexity conditions, such as smoothness (differentiability) and entropy small enough to satisfy the Donsker conditions, can be challenging in high-dimensional settings ($p > n$) that necessitate complex/adaptive ML methods. Although verifying entropy conditions is feasible for certain machine learning techniques like lasso, it becomes notably difficult with methods involving cross-validation or hybrid models, such as Super Learner.

To address this, we employ a strategy of sample splitting. This approach separates the data into two independent sets: one for estimating the nuisance functions and the other for constructing the mediation parameters. Originally proposed by Bickel and later refined by Schick, this strategy has been extended to k-fold cross-validation, allowing for the average mediation estimates from different data partitions to be employed.

Sample splitting allows us to handle the more complex task of identifying mediating pathways within high-dimensional data. Typically, we lack prior knowledge of these pathways amidst a diverse mixture of exposures and mediators, necessitating data-adaptive identification methods. The separate data partitions help ensure that pathway discovery and estimation of direct and indirect effects are not overfit to the sample data, thus avoiding the statistical pitfall of double-dipping, which is akin to multiple testing issues.

This process of identification has been termed "dredging with dignity" in the literature [44], recognizing the necessity of exploring the vast array of potential paths in a principled manner. Just as the analyst might be tempted to cherry-pick interesting results from multiple testing, so too can the analyst fall into the trap of selecting intriguing pathways from the same dataset. This separate sample approach steers clear of that, offering a way to explore high-dimensional pathways responsibly.

K-fold Cross-Validation

K-fold cross-validation is a technique that divides our observations, indexed from 1 to n, into K equally sized subgroups. For each k, an estimation sample P_k is defined as the k-th subgroup of size $\frac{n}{K}$, while the complement of P_k, denoted as $P_{n_{-k}}$, serves as the parameter-generating sample. Using $P_{n_{-k}}$, we identify mediation pathways in the exposure-mediator space by employing basis functions from the best-fitting b-spline estimators. In each fold, we have nuisance estimators for every component of the EIF. With these mediation pathways fixed we then train nuisance parameter estimators on the same $P_{n_{-k}}$ samples, which are essential for solving the EIF and providing asymptotically unbiased estimators.

The process is carried out in a round-robin manner. For $K = 10$, we obtain 10 (possibly different) pathways, outcome estimators Q_k, and density estimators g_k, e_k, and r_k, which are used to construct nuisance parameters that comprise $D^Y(o)$, $D^A(o)$, and $D^{Z,W}(o)$. To estimate a pooled $\theta(\delta)$ using the full data, we stack the estimation-sample estimates for each nuisance parameter across the folds. We then calculate the sum and average across the folds to obtain our point estimate and subtract this average from the summed nuisance parameters to obtain the EIF for the full data, yielding our pooled $\theta(\delta)$ estimate. The variance is then calculated by pooling the pooled EIFs. The NDE parameter is obtained by subtracting the pooled $\theta(\delta)$ from the full data mean outcome, and the delta method is applied to the pooled EIFs to obtain the EIF for the pooled NDE ($\theta(\delta)_d$), which is used to derive confidence intervals (CIs).

A similar procedure is employed for the total effect. For the total effect we are using TMLE or one-step estimation. For TMLE, we stack initial estimates and clever covariates across all folds and perform a fluctuation step across the full initial estimates and clever covariate estimates to obtain our estimate ϵ. We then update the counterfactuals across all folds using the ϵ values. The updated conditional means, counterfactuals, and clever covariates are employed to solve the EIF across the entire sample for the shift in A, ignoring Z. The delta method is applied to subtract the EIF for a shift in A, ignoring Z, from the EIF of the observed Y to obtain the EIF for the total effect ($\theta(\delta)_t$), and the same process is used to derive the point estimate for the total effect. The delta method is again used to estimate the pooled NIE ($\theta(\delta)_i$).

In addition to the pooled estimates, we report k-fold specific estimates of the in(direct) effects and fold-specific variance estimates for these target parameters using the fold-specific IC. This is important because if the mediation pathway identified in each fold varies significantly, the pooled estimates can be challenging to interpret (if the same pathway is not found across all folds). By providing both k-fold specific and pooled results, users can assess the robustness of the pooled result across the folds. To visualize the algorithm and what is happening in the parameter generating and estimation folds, we provide a schematic in **Figure 3.1**

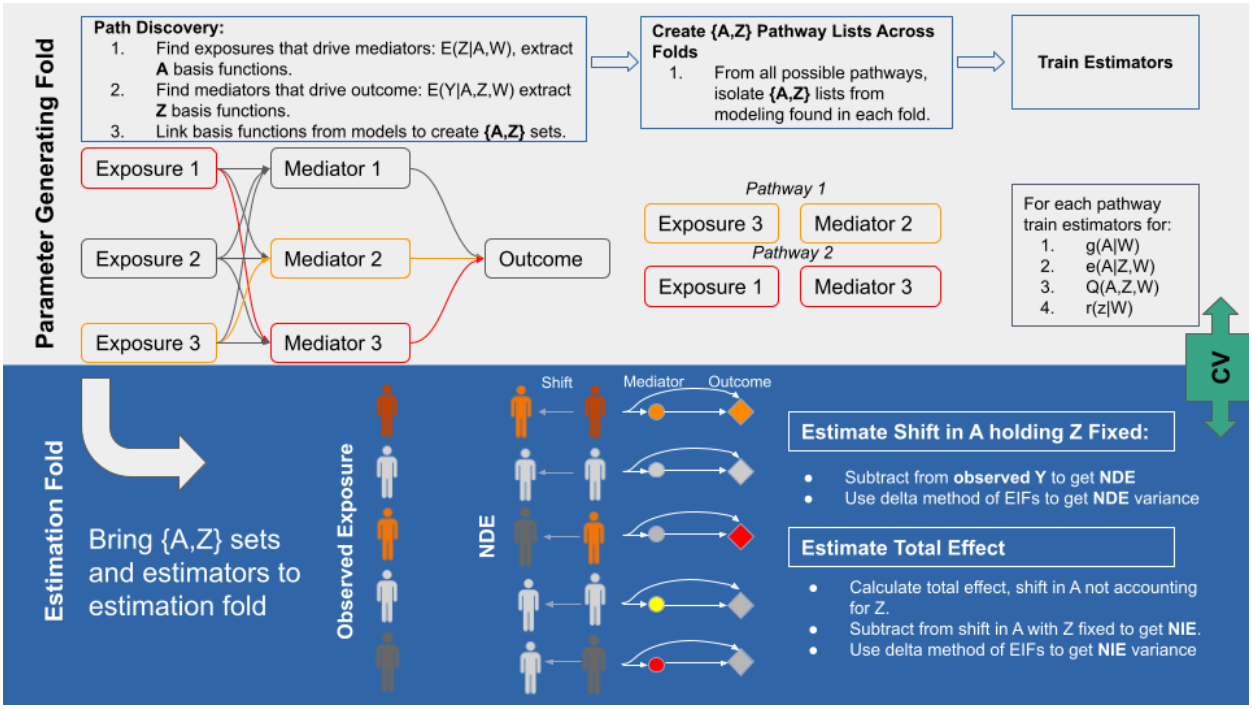


Figure 3.1: Schematic of Operations in the Parameter Generating and Estimation Folds in the NOVAPathways Procedure

Pooled Estimates under Data-Adaptive Delta

Stochastic interventions, particularly those involving significant shifts in exposure, can be susceptible to positivity violations, leading to biased and increased variance in exposure effect estimation. This happens if the exposure shift is so substantial that some subgroups have zero probability of receiving a specific exposure level. This challenge persists even when utilizing an efficient estimator like TMLE.

To mitigate this, we can employ a data-adaptive approach to adjust the exposure shift magnitude, δ. When the exposure is continuous, we modify δ within the parameter-generating sample to meet specific positivity criteria, which helps limit positivity violations. However, when exposure is quantized, a delta of one, signifying an increase in quantiles, is the minimum and most interpretable δ.

Consider $H(a_\delta, w)_i$, the probability density ratio for observation i upon an exposure shift of δ. We aim to ensure that all observations have a ratio below a specific threshold λ. To do this, we iteratively decrease δ by a small amount, ϵ, until $H(a_\delta, w)_i < \lambda$ for all observations i:

$$\forall i,\ H(a_\delta, w) = \frac{g_{n_{-k}}(a_{n_{-k}} - \delta \mid w)}{g_{n_{-k}}(a_{n_{-k}} \mid w)} \leq \lambda$$

Here, λ is a preset threshold, and δ is reduced until all clever covariate density ratios fall below λ. By default, in our SuperNOVA package, λ is set to 50, and ϵ to 10% of δ. This means that if any predicted conditional probabilities exceed the probability under observed exposure by a factor of 50, we reduce δ.

Finally, we account for data-adaptive δ during the pooling process. If δ is constant, the pooling is simply an average of the estimates across folds. However, for a data-adaptive δ, we average δ and pair it with the average estimates and the pooled variance calculations described previously.

Interpreting Shifts when Exposure is Discretized

In the case where the exposure variable is discretized into quantiles, we can still interpret the results in a continuous context. If we let $A_{\min}$ and $A_{\max}$ denote the minimum and maximum values of the continuous exposure, respectively, and n_{bins} denote the number of quantiles, then each quantile represents an interval of size $\frac{A_{\max}-A_{\min}}{n_{\text{bins}}}$ on the continuous scale.

$$A_{\text{quantile}} = A_{\min} + \left(\frac{A_{\max} - A_{\min}}{n_{\text{bins}}}\right) \cdot (q-1) \tag{3.1}$$

Here, q denotes the quantile rank, which ranges from 1 (for the smallest values of the exposure) to n_{bins} (for the largest values of the exposure). Each value of A_{quantile} represents the lower bound of the interval on the continuous scale that corresponds to that quantile.

For example, if we have an exposure variable ranging from 0 to 10, and we discretize it into 5 quantiles, then each quantile represents an interval of size $\frac{10-0}{5} = 2$ on the continuous scale. Thus, the first quantile represents the interval from 0 to 2, the second quantile represents the interval from 2 to 4, and so on. Despite using a discretized version of the exposure in the analysis, the interpretation can still be related back to the original continuous exposure scale. In this way, if the discretized approach is preferable, a pseudo-continuous interpretation is still possible.

3.5 Simulations

In this section, we demonstrate using simulations that our approach identifies the correct mediating pathways in a complex mixture of exposures and mediators and correctly estimates the natural direct, indirect and total effects for a given pathways using stochastic interventions.

Data-Generating Processes

We first construct a simple data-generating process (DGP) where Y is generated from a linear combination of an exposure and mediator. In this DGP we measure the asymptotical behavior of the in(direct) effect estimators keeping the pathway fixed (not data-adaptively discovering the pathway). We do simulations for both continuous and discrete exposures to

investigate the behavior of the estimator using numeric integration vs. simple weighted sums. In the second DGP, we generate multiple pathways to the outcome from multiple exposures and measure the estimators performance in data-adaptively identifying the correct pathways.

Simple Mediation Simulation

This DGP has the following characteristics, $O = W, A, Z, Y$. We call this the "DGP 1" moving forward which we use to investigate the asymptotic behavior of our estimates. The data-generating process involved the following steps:

1. Baseline covariates:
   - $W_1 \sim \mathcal{N}(20, 2^2)$: generated from a normal distribution with mean 20 and standard deviation 2.
   - $W_2, w_3 \sim \text{Binomial}(1, 0.5)$: generated from binomial distributions with size 1 and probability 0.5.
   - $W_4 \sim \mathcal{N}(30, 3^2)$: generated from a normal distribution with mean 30 and standard deviation 3.
   - $W_5 \sim \text{Poisson}(1.2)$: generated from a Poisson distribution with rate 1.2.
2. The exposure A was generated from a normal distribution, conditional on the covariate W_1:
$$A \sim \mathcal{N}(1 + 0.5W_1, 1^2)$$
3. The exposure A was shifted by an amount δ (in this example $\delta = 1$), producing A_δ:
$$A_\delta = A + \delta$$
4. The mediator Z was generated from a normal distribution, conditional on the exposure A and the covariate W_1:
$$Z \sim \mathcal{N}(2 \cdot A + W_1, 1^2)$$
5. The mediator Z was also shifted given a shift in A, producing Z_{A_δ}:
$$Z_{A_\delta} \sim \mathcal{N}(2 \cdot A_\delta + W_1, 1^2)$$
6. The outcome Y, Y_{A_δ} (Y given a shift in only A), and $Y_{A_\delta, Z_{A_\delta}}$ (Y given a shift in A and Z) was generated as a linear function of the exposure A and the mediator Z:
$$Y = 10 \cdot Z + 40 \cdot A + \epsilon$$
$$Y_{A_\delta} = 10 \cdot Z + 40 \cdot A_\delta + \epsilon$$
$$Y_{A_\delta, Z_{A_\delta}} = 10 \cdot Z_{A_\delta} + 40 \cdot A_\delta + \epsilon$$

We use this simulation to test NOVAPathways estimation of the total, direct and indirect effects. Our approach was to keep things relatively straightforward, keeping the DGP a linear process to test the asymptotic behavior of the estimator when the functional forms are correctly specified (GLMs are included in each Super Learner that model the true underlying function).

Complicated Mediation Simulation

We now want to create a more complicated scenario where there are many correlated exposures and some go through mediators to drive the outcome. In this simulation, we want to test NOVAPathways in discovering the correct paths. This data-generating process (DGP) has the following characteristics, $(O = (W, A, Z, Y))$, we call this moving forward "DGP 2". The exposures $A = (A_1, A_2, A_3, A_4, A_5)$ are generated and have potential indirect (through $Z = (Z_1, Z_2, Z_3, Z_4, Z_5)$) effects on Y. Even though there are a total of 25 possible mediating paths due to 5 exposures and 5 mediating variables, only the exposures A_1, A_2 have actual direct and indirect (through Z_1, Z_2 respectively) effects on Y. Our goal is to test the proportion of times across the simulation that the correct paths among the 25 potential ones are discovered. The data-generating process involved the following steps:

1. Baseline covariates:
    - $W_1 \sim \mathcal{N}(20, 2^2)$: generated from a normal distribution with mean 20 and standard deviation 2.
    - $W_2, W_3 \sim \text{Binomial}(1, 0.5)$: generated from binomial distributions with size 1 and probability 0.5.
    - $W_4 \sim \mathcal{N}(30, 3^2)$: generated from a normal distribution with mean 30 and standard deviation 3.
    - $W_5 \sim \text{Poisson}(1.2)$: generated from a Poisson distribution with rate 1.2.
2. Five exposure variables $A = (A_1, A_2, A_3, A_4, A_5)$ are generated from a multivariate normal distribution, conditional on the covariates W:
    - $A_1 \sim \mathcal{N}(1 + 0.5 \cdot W_1, \Sigma)$
    - $A_2 \sim \mathcal{N}(2 \cdot W_2 \cdot W_3, \Sigma)$
    - $A_3 \sim \mathcal{N}(1.5 \cdot W_4/20 \cdot W_1/3, \Sigma)$
    - $A_4 \sim \mathcal{N}(3 \cdot W_4/2 \cdot W_2/3, \Sigma)$
    - $A_5 \sim \mathcal{N}(2 \cdot W_5, \Sigma)$

    The exposures are correlated as per the following correlation matrix, Σ, which represents common scenarios in air pollution where particulate matter and gaseous pollutants show high intra-group correlation but lower inter-group correlation:

$$\Sigma = \begin{bmatrix} 1 & 0.8 & 0.3 & 0.3 & 0.2 \\ 0.8 & 1 & 0.3 & 0.3 & 0.2 \\ 0.3 & 0.3 & 1 & 0.8 & 0.2 \\ 0.3 & 0.3 & 0.8 & 1 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.2 & 1 \end{bmatrix}$$

3. Five mediators $Z = (Z_1, Z_2, Z_3, Z_4, Z_5)$ are generated from normal distributions, conditional on the exposures A and covariates W:
   - $Z_1 \sim \mathcal{N}(2 \cdot A_1 + W_1, 1^2)$
   - $Z_2 \sim \mathcal{N}(2 \cdot A_2 + W_2, 1^2)$
   - $Z_3 \sim \mathcal{N}(5 \cdot A_3 \cdot A_4 + W_3, 1^2)$
   - $Z_4 \sim \mathcal{N}(3 \cdot A_4 \cdot W_4, 1^2)$
   - $Z_5 \sim \mathcal{N}(4 \cdot A_5 \cdot W_5, 1^2)$
4. The outcome Y is generated as a linear function of Z_1, A_1, W_3, A_2, and Z_2:

$$Y = 10 \cdot Z_1 + 40 \cdot A_1 + 15 \cdot W_3 - 6 \cdot A_2 + 7 \cdot Z_2$$

Calculating Ground-Truth

We numerically approximated the natural direct effect (NDE), natural indirect effect (NIE), and total effect (ATE) of the exposure A on the outcome Y to high precision using 100000 samples from our DGP. In our DGP which assesses estimation (simple DGP), δ is equal to 1.

1. The NDE was calculated as the mean difference in Y when shifting the exposure A while keeping the mediator Z constant:

$$\text{NDE} = \mathbb{E}(Y_{A_\delta, Z} - Y)$$

2. The NIE was calculated as the mean difference in Y when shifting both the exposure A and the mediator Z:

$$\text{NIE} = \mathbb{E}(Y_{A_\delta, Z_{A_\delta}} - Y_{A_\delta})$$

3. The ATE was calculated as the sum of NDE and NIE:

$$\text{ATE} = \text{NDE} + \text{NIE}$$

Additionally, we conducted the same analysis using a discrete exposure that has been split into quantiles (10) after step 2. and compute the quantile-based NDE, NIE, and total effect estimates.

Evaluating Performance

We assessed the asymptotic convergence to the true exposure relationships used in the DGP, as well as the convergence to the true in(direct) effects and total effects for these exposure-mediator pathways, in each simulation. To do so, we followed the following steps:

1. We generated a random sample of size n, which we divided into K equal-sized estimation samples of size $n_k = n/K$, each with a corresponding parameter generating sample of size $n - n_k$.

2. At each iteration, we used the parameter generating sample to define the mediation pathway(s) and create the estimators for the nuisance parameters used for $\theta(\delta)_d$ and $\theta(\delta)_t$. We then use the estimation sample to obtain the causal parameter estimate using generating equations and TMLE. We repeated this process for all folds.

3. At each iteration, we output the stochastic shift estimates given the pooled one-step and TMLE estimation.

4. For the simple DGP, we use the var_sets parameter in SuperNOVA to bypass the data-adaptive discovery of mediating paths and simply examine performance of the one $A - Z$ pathway. For the complicated DGP we use the discover_only parameter to do only pathway discovery and skip estimation. In the complicated DGP we report the proportion of iterations NOVAPathways identifies the correct two pathways out of the possible twenty-five pathways.

To evaluate the performance of our approach, we calculated several metrics for each iteration, including bias, variance, MSE, confidence interval (CI) coverage, and the proportion of instances in which the true meditating pathways were identified. To visually inspect if the rate of convergence was at least as fast as $\sqrt{n}$, we show projections of a $\sqrt{n}$ consistent estimator starting from the initial bias. For brevity, we focus on the absolute bias and confidence interval coverage. We calculated these performance metrics at each iteration, performing 50 iterations for each sample size n = (250, 500, 1000, 1500, 2000, 2500, 3000). We used SuperNOVA with 10-fold cross-validation and default learner stacks for each nuisance parameter and data-adaptive parameter. Additionally, the quantile threshold was set to 0 to include all basis functions used in the final best fitting model. To ensure our estimator has a sampling distribution that is normal, we standardize the bias by dividing by the standard deviation of the estimate at each sample size and plot the density distributions for the direct, indirect and total effects.

Default Estimators

SuperNOVA has two density estimating methods that come built into the package. The haldensify estimator [36] can be used for conditional density estimation of $g_n = p(A|W)$, $e_n = p(A|W, Z)$ and $r_n = p(Z|W)$. Haldensify is a flexible, data-adaptive approach that

employs a histogram-based technique to estimate densities. The maximum interaction degree is set by the user as is the number of bins to discretize the outcome. Haldensify works by constructing a histogram of the data and employing a multivariate step function to estimate the density, which makes it computationally efficient and suitable for a wide range of applications.

As an alternative to haldensify, the SuperNOVA package also offers the option to use Super Learner for conditional density estimation. The default Super Learner stack includes a diverse set of learners, such as glm [62], elastic net [27], random forest [111], and xgboost [13]. We create estimators based on homoscedastic errors (HOSE) and heteroscedastic errors (HESE). For the simulations presented in this paper, we have opted to use Super Learner. This way we can investigate the behavior of the estimator when the true function or an algorithm that approximates the true function is included in the Super Learner library.

Additionally, we need an estimator for $\bar{Q} = E(Y|A, W)$. SuperNOVA provides default algorithms to be used in a Super Learner [59] that are both fast and flexible. For our data-adaptive procedure, we include learners from the packages earth [67], polspline [81], and hal9001 [16]. The results from each of these packages can be formed into a model matrix, on which we can fit an ANOVA to obtain the resulting linear model of basis functions.

In the case where A is discrete, $g_n = p(A|W)$ and $e_n = p(A|W, Z)$ are instead Super Learners built from categorical outcome estimators such as neural networks, random forest and polspline.

Results

Do Target Parameters Estimated by NOVAPathways Converge to Truth at $1/\sqrt{n}$ for Continuous Exposures?

An important aspect of our estimator's performance is its convergence rate. In the context of our simulation (DGP 1 with one exposure-mediator pathway), the convergence rate signifies how quickly the estimator approaches the true parameter value as the sample size increases. Ideally, we want estimates for the total effect, direct effect and indirect effect to show convergence to the truth at $\sqrt{n}$ using a DGP that, although simple, at least includes confounding and relationships that feasible could be observed in a real-world analysis setting.

Figure 3.2 exhibits the absolute bias and the anticipated rate of convergence for a $\sqrt{n}$ consistent estimator, given the initial bias, when the exposure is truly continuous. It shows the bias as the sample size increases to 3000. Observing the estimates from the integration method, the bias is generally lower but exhibits a non-convergent behavior when reaching a sample size of 3000, particularly for Natural Direct Effect (NDE) and Natural Indirect Effect (NIE). Although the pseudo-regression approach displays greater consistency, the bias remains considerably high, hindering proper coverage.

Coverage, illustrated in **Figure 3.3**, refers to the proportion of iterations for each sample size where confidence intervals contain the true value. For both methodologies—integration and pseudo-regression—the estimated coverage for NDE and NIE does not achieve the desired

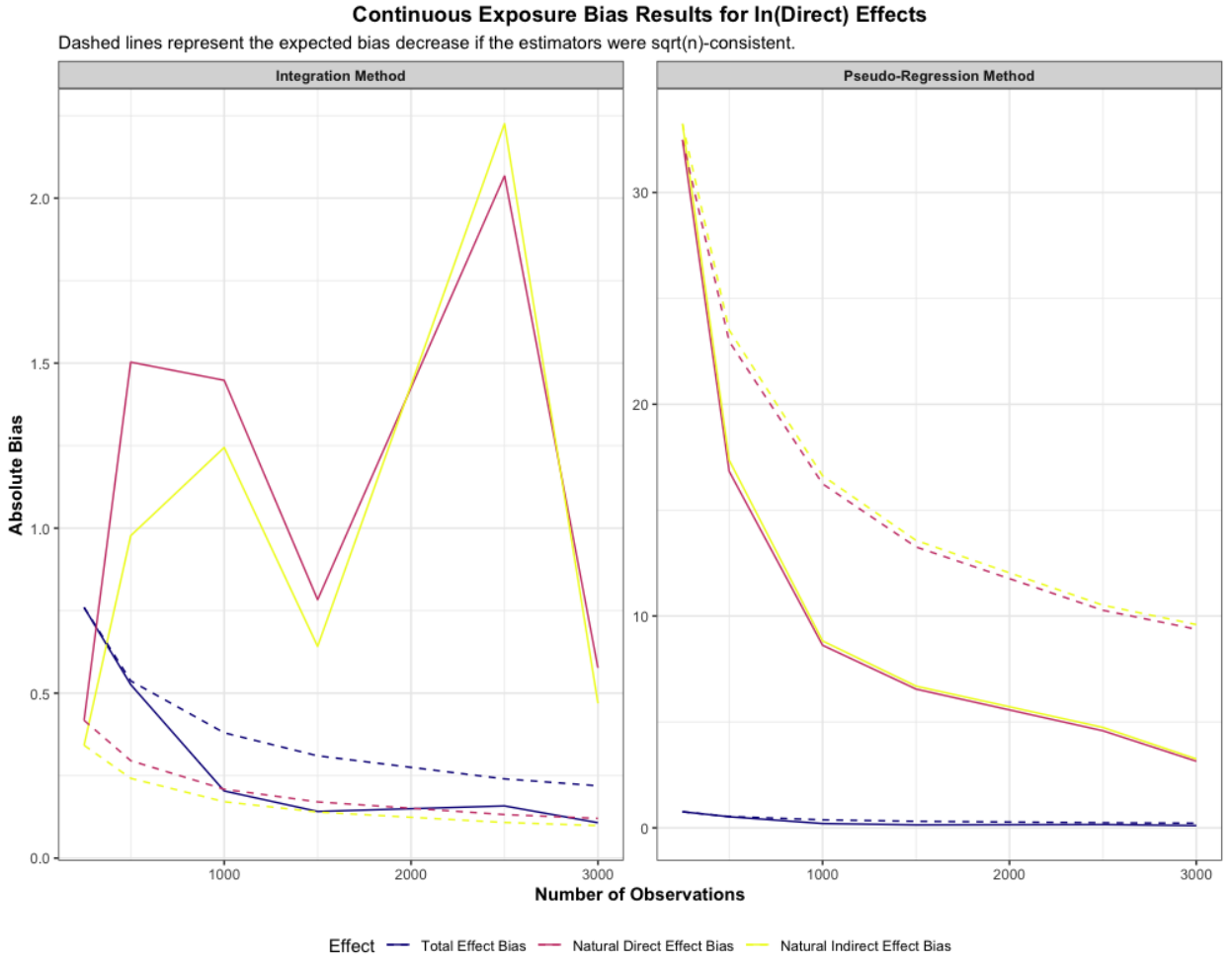


Figure 3.2: Absolute Bias and Expected $\sqrt{n}$ Convergence Across Sample Sizes for Total, Natural Direct and Natural Indirect Effects when Exposure is Continuous in DGP 1

95% level. This shortfall is likely attributable to the bias in estimates induced by numeric integration, a necessary procedure for estimating the nuisance parameters in the case of a continuous exposure.

Therefore, continuous exposures do not demonstrate the expected $\sqrt{n}$ convergence. This behavior implies that our estimator falls short of the necessary criteria to qualify as asymptotically normal. The departure from asymptotic normality may partly stem from approximations made during the numerical integration required for our estimation process. Alternatively, coding inaccuracies could be at play. Theoretically, the estimator should function correctly, so these anomalies warrant further investigation. Additionally, a sample size of 3000 may still be too small to assess for normality.

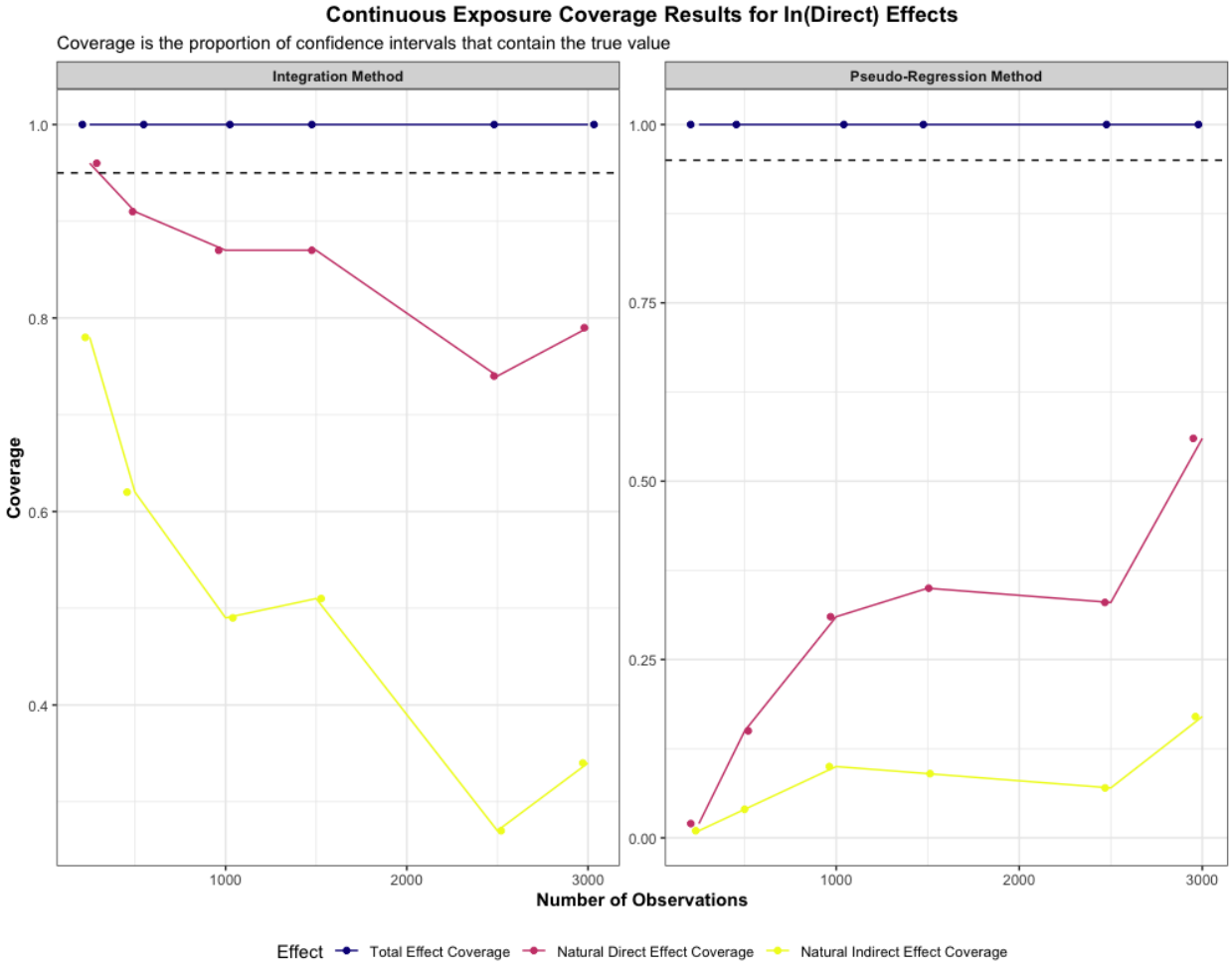


Figure 3.3: Confidence Interval Coverage for Total, Direct and Indirect Estimates using Integration and Pseudo-Regression in DGP 1

Do Target Parameters Estimated by NOVAPathways Converge to Truth at $1/\sqrt{n}$ for Quantized Exposures?

We also evaluated the performance of the NOVAPathways estimator under our DGP 1 scenario where the exposure variable is quantized. Like for the truly continuous exposure, the main aspects of the evaluation are the rate of convergence and the coverage of the confidence intervals, as these metrics represent the robustness and reliability of the estimator.

In contrast to the results for continuous exposures, we observe satisfactory performance of the estimator for quantized exposures. The bias for the estimated NDE, NIE, and total effect demonstrates a clear trend of convergence towards zero with increasing sample size, for both integration and pseudo-regression methods. **Figure 3.4** shows the absolute bias and

expected $\sqrt{n}$ convergence given initial bias as sample size increases. This pattern is more prominent for the integration method, with bias levels generally being much lower than in pseudo-regression. For instance, for NDE, the absolute bias using the integration method decreases from 0.668 at a sample size of 250 to 0.0621 at a sample size of 3000.

Coverage of confidence intervals for these estimates also shows a desirable pattern. Considering the average coverage across all sample sizes, the coverage for NDE reaches an average of 95.6 % for the pseudo-regression method and remains at 100% for the integration method. For NIE, the pseudo-regression method provides an average coverage of 85%, while the integration method provides a higher average coverage of 96%. **Figure 3.5** shows the proportion of confidence intervals that contain the true value for each approach at increasing sample size. Finally, for the total effect, the average coverage across all sample sizes reaches 100%. It's worth noting that for both NDE and NIE, the pseudo-regression method exhibits lower coverage than the integration method.

In summary, our results demonstrate that when the exposure variable is quantized, the NOVAPathways estimator exhibits desirable characteristics of a reliable estimator. It provides a rate of convergence that meets the $1/\sqrt{n}$ standard, and the confidence intervals demonstrate appropriate coverage. These results confirm the robustness of NOVAPathways when applied to quantized exposure variables, and underline the necessity of having appropriately quantized exposure variables in order to achieve reliable and valid results.

NOVAPathways Correctly Identifies Mediating Pathwways in a Realistic Complex Mixed Exposure-Mediator Situation

In DGP 2, despite having 25 potential mediating pathways in the complex exposure mixture-mediation simulation, NOVAPathways consistently identified the two true pathways ($A_1 - Z_1$ and $A_2 - Z_2$) with a frequency of 100%, across various sample sizes ranging from 250 to 3000 observations. Furthermore, direct effects of A_1 and A_2 on the outcome Y were also consistently identified across all scenarios, reinforcing the robustness of our detection method. **Figure 3.6** shows the frequency each pathways was detected for each sample size. Note that, only pathways detected are reported.

However, it is noteworthy that there were some instances of incorrectly identified pathways, as seen from the non-zero frequencies of pathways such as $A_1 - Z_2$, $A_2 - Z_1$, and others, which could be attributed to the high correlation between the exposures. While these false discoveries present opportunities for methodological refinement, the consistently correct identification of the true pathways underpins the effectiveness of our methodology in the presence of multiple mediators and exposures, which is a common scenario in air pollution research. Of note is that, the incorrect pathways were identified in very few folds and in such cases the analyst would report the inconsistency of such a finding.

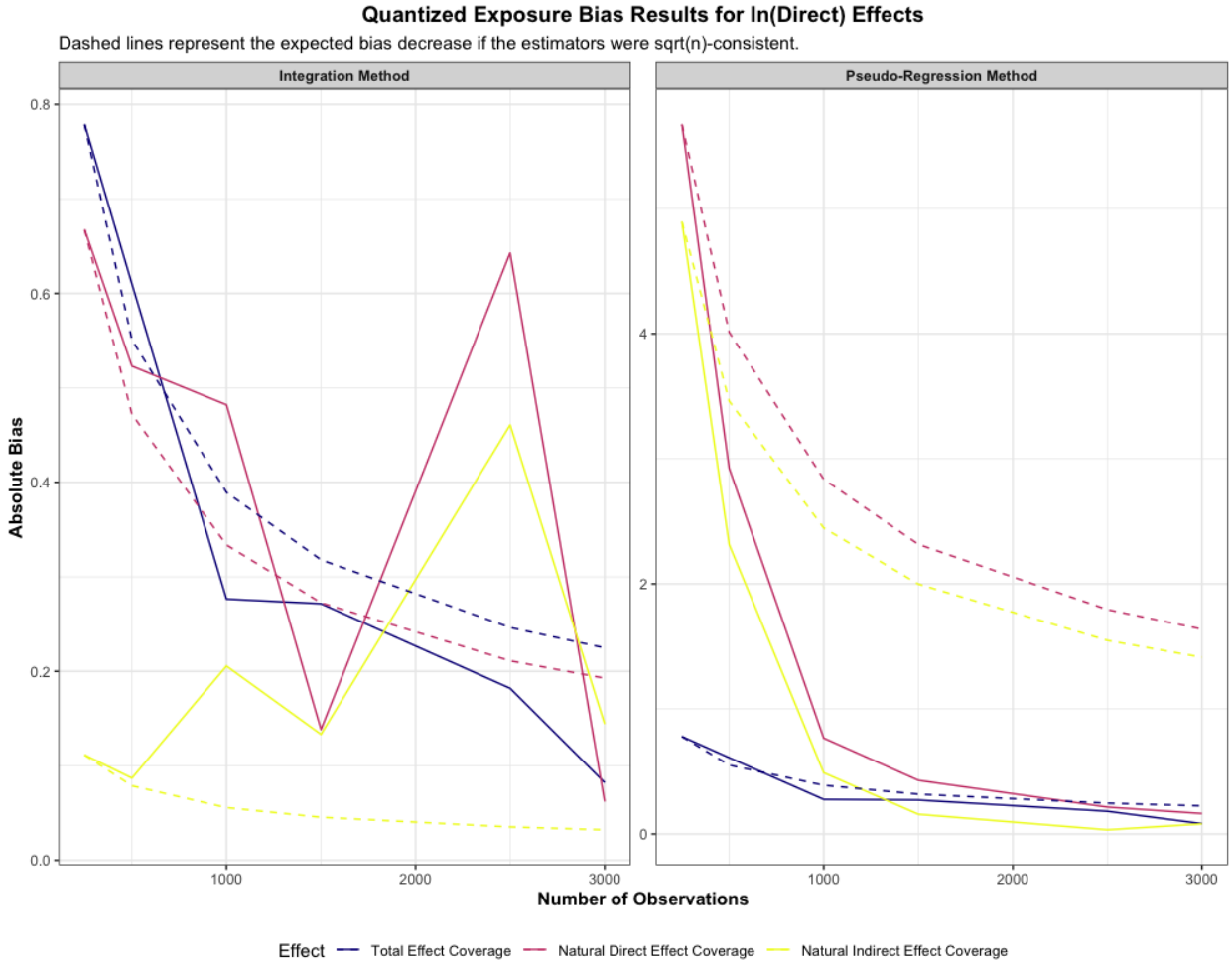


Figure 3.4: Absolute Bias and Expected $\sqrt{n}$ Convergence Across Sample Sizes for Total, Natural Direct and Natural Indirect Effects when Exposure is Quantized in DGP 1

Assessing the Validity of NOVAPathways's Inference through Simulations

The fundamental premise of a robust inference is the verification of the estimator's normal sampling distribution, centered at zero and progressively narrowing with increasing sample size. This premise is tested in the context of the NOVAPathway estimator for natural direct, indirect, and total effects. We illustrate the empirical distribution of the standardized bias, defined as the difference between the estimated and true values from the data-generating process, normalized by the standard deviation of the estimates across iterations. The assessment is conducted using 50 iterations per sample size and visualized as a probability density distribution in **Figure 3.7**.

In **Figure 3.7**, we observe the convergence of the sampling distribution to a mean-zero

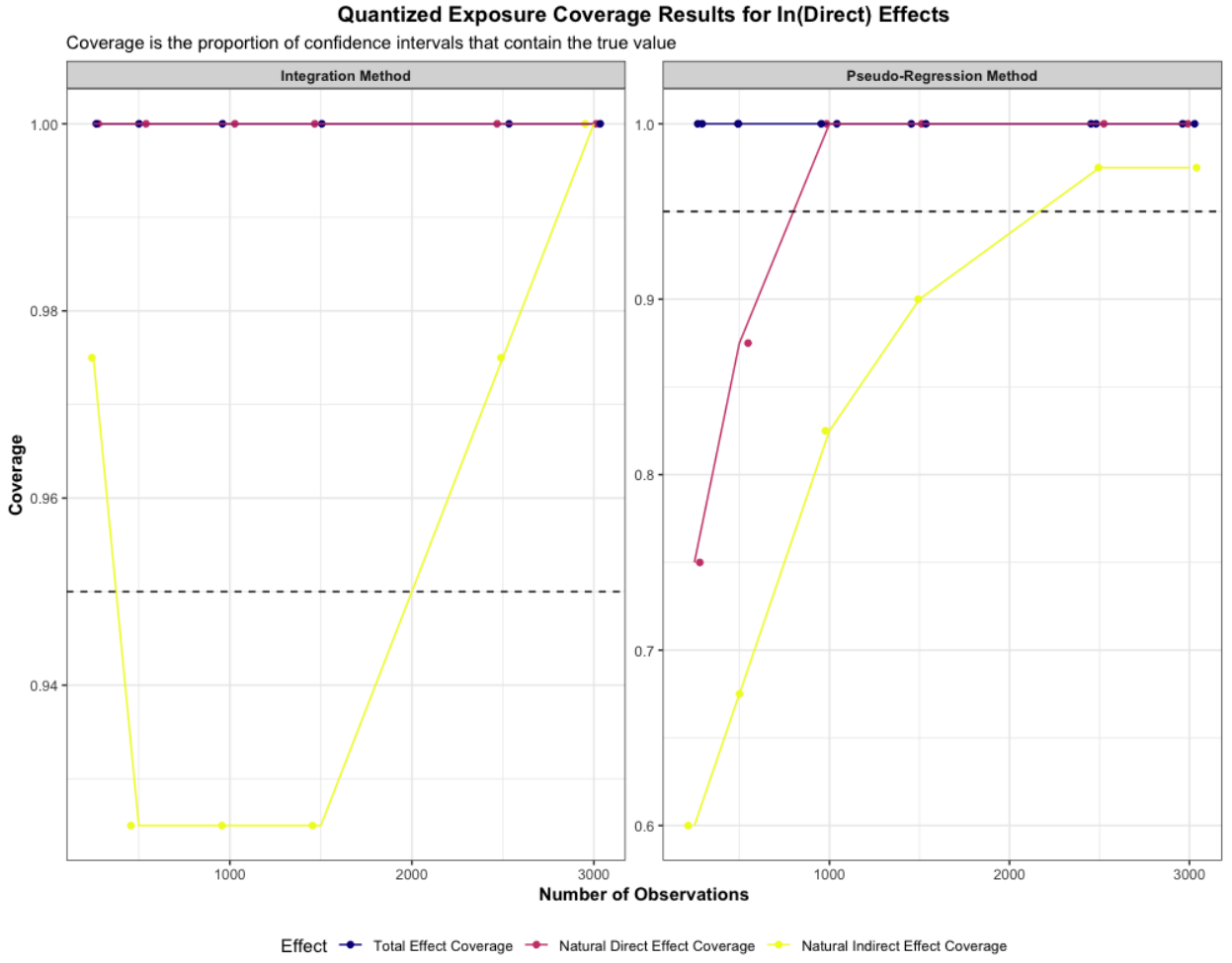


Figure 3.5: Confidence Interval Coverage for Total, Direct and Indirect Estimates using Integration and Pseudo-Regression in DGP 1

normal as sample size escalates. This phenomenon is evident across all types of effect estimates. The total effect, calculated via one-step, remains consistent regardless of whether the natural direct effect (NDE) is computed using integration or pseudo-regression methods. The NDE for the integration method is concentrated more closely around zero, albeit exhibiting greater tail variability, while the pseudo-regression counterpart maintains a smoother, centered distribution. The natural indirect effect (NIE) demonstrates the widest dispersion, although still centered around zero. Notably, the pseudo-regression method achieves a slightly narrower distribution around zero for NIE.

All plots in **Figure 3.7** exhibit normal or near-normal distributions centered at zero that contract with an increase in sample size. This characteristic is crucial for the validity of

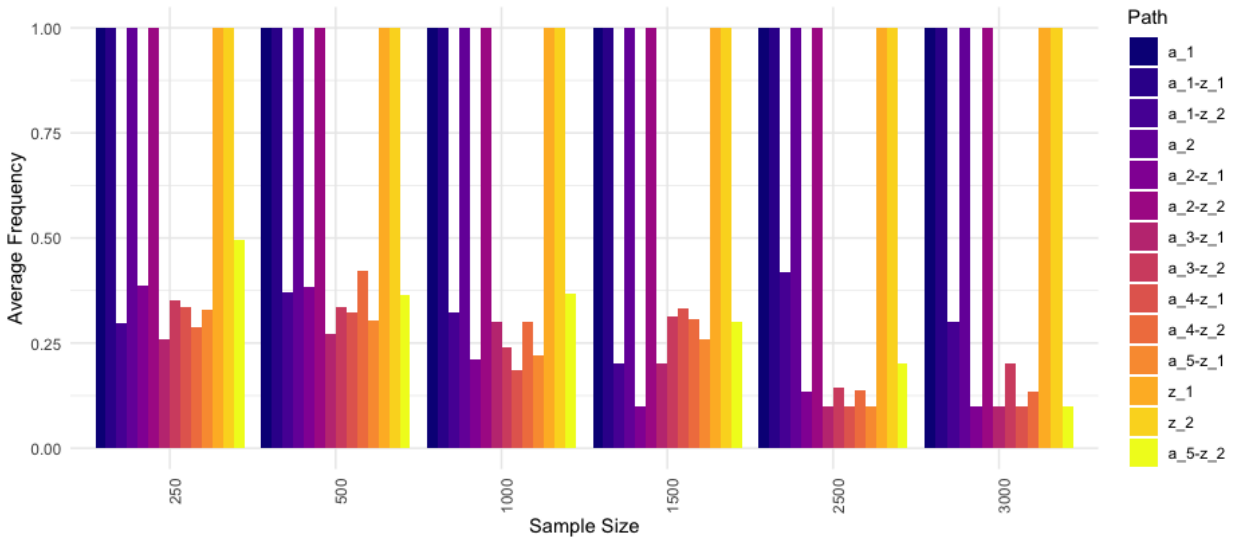


Figure 3.6: Average Frequency Each Path Was Detected in the Mixed Exposure-Mediator Simulation in DGP 2

confidence interval construction and underscores the reliability of our estimator. As such, the simulation results affirm the soundness of NOVAPathways's inference methodology.

3.6 Applications

NHANES Data

Data Description

To provide a motivating example for the application of NOVAPathways we extracted data from the 2001-2002 cycle of the National Health and Nutrition Examination Survey (NHANES). The NHANES program, managed by the Centers for Disease Control and Prevention (CDC), is a comprehensive set of studies designed to assess the health and nutritional status of adults and children in the United States [124]. These studies employ a combination of interviews and physical examinations to capture a broad array of health information. NHANES data is particularly suitable for motivating the use of NOVAPathways due to its representative sample of the U.S. population (specifically for pollution exposure), broad collection of health-related variables, and its open availability. This enables us to make our analysis transparent and easily replicable, fostering open science practices and facilitating methodological testing [109]. For these purposes, all code for data cleaning and curation for this motivating analysis example using NHANES is included in the SuperNOVA package which uses the NOVAPathways method.

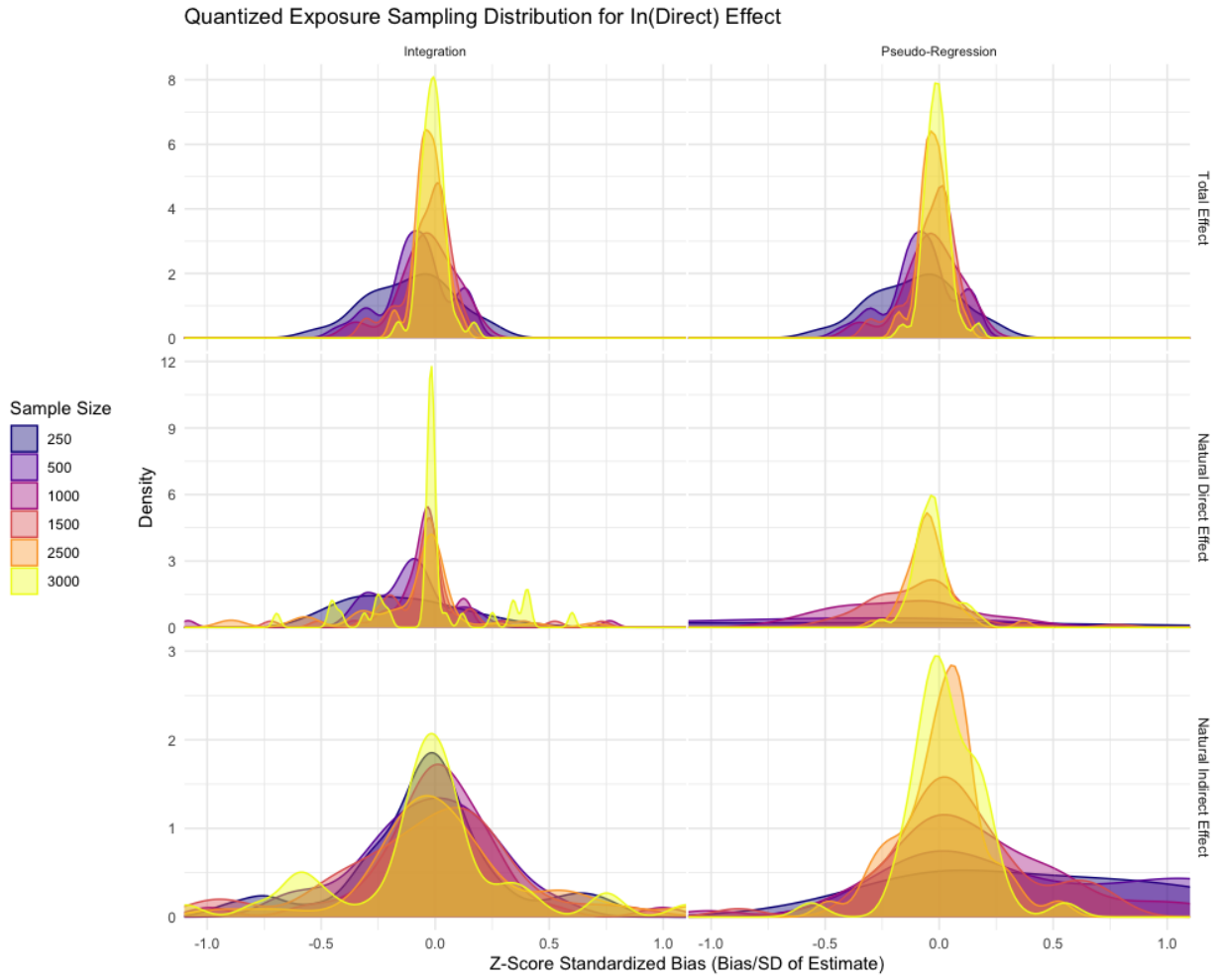


Figure 3.7: Bias Standardized by Standard Error of Estimates for the Natural In(Direct) Effects and Total Effect in DGP 1

One significant challenge of using cross-sectional datasets like NHANES is the potential for reverse causality, wherein the outcomes/mediators may influence the exposures rather than vice versa. This characteristic violates the temporal assumption required for traditional causal inference [41]. However, the use of NHANES data in our study is not primarily to establish causal relationships but rather to provide a real-world demonstration of the capabilities of our method, NOVAPathways.

The NHANES data provides a large number of well-measured toxic metal exposures, biomarkers for potential mediating pathways, and covariates. This comprehensive dataset offers an invaluable opportunity to determine if NOVAPathways identifies consistent mediating pathways in high-dimensional data and delivers interpretable direct, indirect, and total effect

results based on data-adaptively determined pathways from stochastic shift interventions. Our chosen example focuses on the association of a mixture of toxic metals on asthma, both directly and possibly indirectly through biomarkers for inflammation, oxidative stress, and immune function.

The decision to investigate this specific association is informed by a growing body of evidence suggesting a correlation between heavy metal exposure and the onset of asthma, particularly in children. For instance, a cross-sectional, population-based study leveraging NHANES data from 2007-2012 investigated the associations between heavy metal exposure and childhood asthma or wheezing [113]. Their analysis found that higher blood lead concentrations were associated with higher odds of active asthma in children aged 2-15 years, with a particularly pronounced effect in the 6-11 years age group. Moreover, the study also noted associations between blood lead concentrations and the incidence of wheezing, an asthma symptom. Although their results indicated a lower risk of wheezing with higher blood mercury concentrations, the overall evidence points towards a noteworthy connection between heavy metal exposure and asthma.

Given these findings, our exploration seeks to further illuminate potential mediating pathways between toxic metal exposure and asthma. This could provide vital insights into how and why such an association exists, potentially guiding the development of more targeted interventions for this prevalent health issue. By doing so, we hope to contribute to the broader understanding of environmental health risks, particularly those tied to heavy metal exposure.

Our choice of the 2001-2002 NHANES data cycle was informed by the fact that this cycle included all relevant variables necessary for a comprehensive investigation into the associations between toxic metal exposures, inflammation, immune function, oxidative stress, and the prevalence of asthma [3]. This particular NHANES cycle collected exhaustive data on these variables, offering a unique opportunity to conduct our investigation within a representative sample of the U.S. population. The exposure variables are several toxic metals, including barium, cadmium, cobalt, cesium, molybdenum, lead, antimony, thallium, and tungsten measured in urine. Covariates included are variables that could potentially confound the relationship between the exposure (metals) and the outcome (asthma). They include demographic variables (age, gender, race, education level), health behavior variables (average daily physical activity, muscle training, vigorous intensity in the last 30 days), substance use (cotinine, an indicator of nicotine exposure, and alcohol), body mass index, family poverty ratio, exam weight, interview weight, and caffeine intake. The mediators are the variables representing potential biological pathways through which the metal exposures could affect the outcome. They include various biomarkers for cell aging (mean telomere length, standard deviation of telomere length), immune function (white blood cell count, monocyte and neutrophil percentages), inflammation (C-reactive protein), and antioxidant nutrients (alpha and beta carotene, vitamins A and E, trans lycopene, lutein and zeaxanthin). These variables represent potential mediators of the relationship between heavy metal exposure and asthma. A binary indicator for asthma is the outcome.

The original NHANES 2001-2002 dataset consisted of 11,039 participants, with 4,260

individuals providing blood samples and consenting to DNA analysis [124]. After applying our exclusion criteria, such as missing environmental chemical analysis data, missing key covariate data, and insufficient stored samples for telomere length estimation, our final study sample was comprised of 1,344 participants.

Our data cleaning and curation techniques were relatively basic as our main goal is to demonstrate our proposed methodology and software, not provide a thorough analysis. Nontheless, data cleaning and curation were undertaken to ensure the integrity of distributions in our dataset was retained while allowing us to not lose too many observations due to missingness. We first omitted observations with missing values in the outcome variable (asthma) and in the crucial exposure variables (toxic metals). We then retained columns where less than 20% of the data was missing. This balance allowed us to maximize the use of available data while avoiding the potential bias from imputing excessive missing values.

Next, we imputed missing data in the remaining variables through suitable methods: mean imputation for numeric variables and mode imputation for categorical ones [42]. This strategy helped ensure the final dataset maintained the original distributions and variable relationships to the greatest extent possible.

We then quantized the metal exposure data to address the methodological issue with our proposed method when the exposure is fully continuous. As shown, continuous exposures necessitate numeric integration in the calculations of the mediation effects. However, this approach lead to approximations that are not precise enough, inducing asymptotic bias and resulting in poor confidence interval coverage. To avoid this issue, we quantized the continuous exposure data, transforming each exposure into a categorical variable with equal frequency bins (deciles in our case). This transformation allows a shift delta = 1 to represent an increase in decile, and we can calculate each nuisance function as a simple weighted sum rather than a numeric integration. By doing this, we have observed improved asymptotic behavior of our estimators and accurate confidence interval coverage.

The selection of toxic metals as exposures in our study was informed by prior literature demonstrating the potential link between toxic metal exposure, oxidative stress, inflammation, and immune function—all factors implicated in the etiology of asthma [33, 103]. Several studies have shown that exposure to toxic metals can lead to oxidative stress, which in turn can trigger inflammatory responses and modulate immune function [23, 52, 80]. These processes can potentially contribute to the onset or exacerbation of asthma, hence our interest in exploring these relationships in this study. By investigating these associations within the NHANES 2001-2002 dataset, we aim to show that semi-parametric methods utilizing efficient estimators and data-adaptive target parameters can yield a deeper understanding of the complex interplay between environmental exposures, molecular biomarkers, and disease outcomes.

Through this process, we seek to illustrate the utility of our SuperNOVA software which incorporates the NOVAPathways mediation methodology. Our results, which are presented in subsequent tables, offer a comprehensive view of the information that SuperNOVA can generate from a provided dataset.

Consistent Findings for Toxic Metal Exposure on Asthma Through Inflammatory, Oxidative Stress and Immune Function Mediators

Because NOVAPathways data-adaptively discovers exposure-mediator pathways, it's best to report first any notable consistencies across the multiple folds. Cesium, as an exposure, was found in 80% of the folds, demonstrating the greatest consistency among all exposures investigated. This highlights the potential relevance of cesium in our model, warranting further investigations into its role and impact on asthma in future studies.

When considering the mediators, monocyte percentage and vitamin E emerged as the most consistent across the folds, being detected in 80% and 60% of the folds, respectively. The consistent appearance of monocyte percentage, a key indicator of immune system activation, underscores the possible involvement of immune modulation in the effect of toxic metal exposure on asthma. This observation aligns with previous studies that have reported a link between elevated monocyte counts and increased asthma risk, potentially due to the role of monocytes in driving inflammation and airway remodeling processes in asthma [78].

Similarly, the recurring detection of vitamin E may imply a role for antioxidant mechanisms in modulating the exposure-asthma relationship. Vitamin E is a known antioxidant and anti-inflammatory agent, and its deficiency has been associated with higher incidence and severity of asthma. Evidence from prior research also suggests that higher vitamin E levels could provide some protective effect against pollutants or other harmful environmental exposures related to asthma [107]. Our findings are, therefore, in line with the existing body of research, suggesting potential roles for both immune system activation and antioxidant mechanisms in the relationship between toxic metal exposure and asthma.

Furthermore, we found the exposure-mediator pairs of cesium-monocyte percentage and tungsten-monocyte percentage in 60% of the folds. The pairings of specific exposures with monocyte percentage suggest potential pathways where these elements could influence asthma pathogenesis through immune mechanisms. Cesium is typically released into the environment from nuclear power plants and during the production and detonation of nuclear weapons, and exposure can also occur from consuming food and water contaminated with cesium. Tungsten is often used in industrial applications like metalworking, mining, and construction. Reducing human exposure to these metals might involve tighter regulations on nuclear energy production and emissions standards for industries using tungsten, as well as efforts to reduce contamination of agricultural areas near such sites.

Meanwhile, the lead-vitamin E pair appeared in 50% of the folds, alluding to another potential pathway via oxidative stress mechanisms. Lead exposure is a well-known public health issue, and significant steps have been taken to reduce lead in house paint, gasoline, and plumbing. Nevertheless, exposure still occurs, often from legacy sources such as older homes with leaded paint or pipes. Continued efforts to remediate these sources and educate the public on how to reduce their exposure to lead can be instrumental in reducing the burden of lead-associated health issues, including asthma.

Given these results, we next report the fold specific and pooled results of Cesium, the Cesium-Monocyte Percentage pathway, the Lead-Vitamin E pathway, and the Tungsten-

Monocyte Percentage pathway.

Results for Cesium, Lead and Tungsten Through Monocyte Percentage and Vitamin E

We first examined the potential impact of Cesium exposure on the likelihood of developing asthma, independent of mediation. In 8 out of 10 folds, Cesium consistently appeared, implying a possible influence in the disease's progression. Here we use a decile shift increment in Cesium exposure and observe the expected probability of asthma given this shift compared to the observed probability of asthma. Here, a decile increase is equivalent to a rise of 4.182,/L on the Cesium continuous scale.

While the results varied slightly across the folds, the effect generally leaned towards the positive. In the pooled analysis, a decile increase in Cesium corresponded to a 0.012 increase in the asthma probability. However, this result didn't achieve statistical significance at the conventional 0.05 level (p-value = 0.17). While the findings do not conclusively establish a relationship between cesium and asthma, the consistent results hint at a potential correlation warranting further investigation. **Table 3.1** presents the total effects of a decile shift in Cesium on the likelihood of asthma.

Psi	Variance	SE	Lower CI	Upper CI	P-value	Fold	Type	Variables	N	Delta
0.02	0.00	0.01	-0.01	0.04	0.14	1	Indiv Shift	cesium	135.00	1.00
0.02	0.00	0.05	-0.07	0.12	0.64	4	Indiv Shift	cesium	135.00	1.00
0.04	0.00	0.03	-0.03	0.10	0.29	5	Indiv Shift	cesium	135.00	1.00
-0.00	0.00	0.02	-0.04	0.04	0.92	6	Indiv Shift	cesium	134.00	1.00
0.00	0.00	0.01	-0.02	0.02	0.85	7	Indiv Shift	cesium	134.00	1.00
0.00	0.00	0.01	-0.01	0.01	0.88	8	Indiv Shift	cesium	134.00	1.00
0.02	0.00	0.02	-0.03	0.06	0.42	9	Indiv Shift	cesium	134.00	1.00
0.00	0.00	0.01	-0.01	0.02	0.74	10	Indiv Shift	cesium	133.00	1.00
0.01	0.00	0.01	-0.01	0.03	0.17	Pooled TMLE	Indiv Shift	cesium	1074.00	1.00

Table 3.1: Results for Association Between a Decile Shift in Cesium and Probability of Asthma

We subsequently examined how much of this effect passed through the monocyte percentage as opposed to not. **Table 3.2** displays the fold-specific and pooled results for the NDE, NIE, and total effect of Cesium on asthma, via monocyte percentage, using both pseudo-regression and double integration methods to construct the estimator. With pseudo-regression estimates, we observed an NDE of 0.61 (-0.66 - 1.88) and an NIE through monocyte percentage of -0.60 (-1.86 - 0.66). Both results were not significant. Hence, the positive NDE and negative NIE estimate, we would expect both to be positive given a positive total effect, in this case, both estimates include 0 in the confidence interval. However, despite the absence of traditional statistical significance, the persistence of cesium through the monocyte percentage across the majority of the folds suggests a potential influence of this pathway on asthma.

For Lead and Tungsten, a similar process was followed. In this instance, a decile increase in Lead and Tungsten corresponded to a rise of 1.3,/L and 0.285,/L on the respective continuous scales. A pathway for Tungsten through monocyte percentage was found in 60%. These results are provided in **Table ??**. Like Cesium-monocyte percentage, although these pathways were found in a majority of folds, the effects were not significant. A one decile increase in Tungston is associated with a -0.005 (-0.013 - 0.004) decrease in the probability of asthma (p-value = 0.28). Results for the NDE using pseudo-regression are the same as the total effect indicating no indirect effect through monocyte percentage.

Lastly, we give results for Lead on asthma through vitamin E. **Table 3.4** shows these results. Again, the total effect, NDE and NIE are not significant given a decile increase in lead although this pathway was identified in 50% of the folds. It should be noted that, across the results the effects are quite small and the pathways are not found across all the folds. If this were a true analysis being published by the analyst, these measures should be reported. The fact that, the path discovery procedure does pick up mediation in 60% of the folds (for Cesium through monocyte percentage) suggests that there is signal but perhaps the impact is weak, or another measure of immune function captures the mediation better.

Through an analysis of the NHANES dataset, we highlight the proficiency of NOVA-Pathways in identifying mediating pathways within high-dimensional data contexts. The dataset in focus included 9 exposures and 12 mediators, thus theoretically encompassing 108 potential pathways. These pathways could mediate the effects of toxic metals via proxies of inflammation, oxidative stress, and immune function.

Using flexible basis estimators, NOVAPathways successfully discerned the most influential pathways and provided estimates associated with a one-decile increment in exposure. While none of the effects reached the threshold of statistical significance, some exhibited borderline significance.

Our comparison of Natural Direct Effects (NDE) estimates from pseudo-regression and integration procedures revealed similar trends. The stability of estimates across the folds and a decreased variance for the pooled results, as anticipated, reaffirmed the advantage of pooling estimates across folds for precision.

We acknowledge the potential limitations of our method, as we discretized exposures prior to implementing NOVAPathways. This more rudimentary representation of exposures, although simplifying the data, might make pathway discovery more challenging.

Nevertheless, our primary objective was not the pinpoint accuracy of a causal inference but rather a demonstration of the potential output from NOVAPathways. We sought consistency of results across folds and the provision of interpretable estimates for NDE, NIE, and total effects. In conclusion, this example underscores NOVAPathways' utility in navigating complex associations within high-dimensional data, offering a useful tool for analysts working with multiple exposures and potential mediators.

3.7 Software

The accessibility and application of statistical software that executes semi-parametric methods which respect data-generating processes found in real-world data is pivotal for ensuring consistent and reproducible outcomes across research studies. SuperNOVA, an open-source R package, attempts to address this need by facilitating the evaluation of causal effects from mixed exposures using asymptotically linear estimators, which now includes the NOVAPathways method for mediation. These estimators are proven to converge to the true estimand at $\sqrt{n}$ given estimates of nuisance parameters converge at $n^{1/4}$. Its ability to handle both continuous and discretized exposures addresses a notable limitation of its predecessor, the medshift package [37] developed by Ivan Diaz and Nima Hejazi, which only supports binary exposures. We also offer some additional functionality compared to the longitudinal modified treatment policies approach and packege [22, 108] by data-adaptively finding mediating pathways in cross-sectional data. While continuous exposures are accommodated in SuperNOVA, caution is warranted due to the potential for bias introduced by numerical integration, which we have shown.

At the heart of SuperNOVA, with its integrated NOVAPathways, is Super Learning, a machine learning technique employed via the SL3 package [17]. This methodology allows SuperNOVA to adaptively identify mediating pathways using ensembles of basis-function estimators, improving the adaptability and efficiency of the software to find pathways even in complex exposure settings. Likewise, Super Learning is used for the estimation of each nuisance function.

Comparison with existing software illustrates the potential for SuperNOVA to enhance the accuracy and flexibility of mixed exposure-mediator research. Many environmental health studies that have performed mediation analyses have used packages such as medflex [94] and mediation [97] which are largely reliant on parametric assumptions. For instance, these packages make strong assumptions about functional form, and they often assume no interactions between the exposure and mediator, which can lead to biased estimates of direct and indirect effects. In contrast, SuperNOVA's semi-parametric approach relaxes these assumptions, potentially resulting in more accurate and consistent estimates. Additionally, no method or package currently exists which can identify pathways and make valid inference on these pathways in the presence of high-dimensional data.

SuperNOVA's design allows for both sequential and parallel computing, leveraging the parallel processing capabilities offered by the furrr package [104]. Its computational efficiency expands its suitability for use on personal computers, which can be crucial in resource-limited research settings. Additionally, in the context where the analyst has a pre-defined pathways they want to test, the path discovery section of NOVAPathways can be skipped and the direct, indirect and total effects can directly be estimated using the cross-validation procedure. Conversely, if the analyst is instead interested in only finding the most relevant exposure-mediator paths to guide future study develop, this approach is also available.

Additional features of SuperNOVA include a comprehensive vignette, a detailed exposition of the underlying semi-parametric theory, and comparisons to existing methods. The package

also offers the NHANES mixed metal exposure data for reproducibility purposes, coding notebooks illustrating the application of the software, and interpretative summaries of SuperNOVA output. SuperNOVA is regularly updated, available on GitHub (https://github.com/blind-contours/SuperNOVA), and aims to equip researchers with robust tools to advance the quality of research in mixed exposure and environmental health.

3.8 Limitations

Even as we have made a concerted effort to apply rigorous methodology in this study, following [21] there are several limitations to consider which influenced our results, particularly when the exposure is truly continuous.

Firstly, we used Monte Carlo integration methods, which are inherently stochastic. This could have introduced some level of bias in our estimates. We sought to minimize this by implementing four times the sample size for the number of Monte Carlo samples. However, in high-dimensional or complex model scenarios, such adjustments may not fully eradicate the error.

Furthermore, data variability, particularly in the density estimation, could have contributed to bias introduction. Specifically, when density values hover at the extremes - either exceedingly low or high - the subsequent variance in the estimator may inflate the bias.

Our proposed mediation method for continuous exposures also struggled with potential issues regarding integration boundaries. Even though we were cautious in setting these boundaries (the range of the exposure), the region of integration might have covered areas where the functions integrated were not well-behaved. This could have added to the bias.

Moreover, instabilities in the numerical computations could have subtly influenced our findings. Despite the power of contemporary computational tools, they are not entirely devoid of errors. Instances of round-off or truncation errors could subtly impact the results.

Likely, this issue with continuous exposures arises as a cumulative effect of the aforementioned limitations. Nevertheless, our results have demonstrated that when exposure is quantized into a discrete form, thereby bypassing numeric integration, our estimator exhibits the expected asymptotic behavior. Moreover, it provides valid confidence intervals for inference - results that can be interpreted continuously.

In relation to positivity, violations of this principle are often an unavoidable reality in many contexts. Nonetheless, our suggested approach optimizes the situation by considering smaller shifts. These shifts are based on the ratio of exposure densities, which contrast the density under shift to the observed density when there is no shift. Similar to our methodology for path discovery, this strategy is heuristic in nature. It attempts to strike a balance between ease of comprehension and implementation while effectively achieving the intended objective.

3.9 Discussion

In this study, we introduce a novel approach for the estimation of natural direct, indirect, and total effects, facilitated through data-adaptive identification of mediating pathways in high-dimensional data. This breakthrough addresses a significant gap in current analytical methods, particularly when dealing with data that comprises numerous exposures and mediators, which is a common occurrence in environmental omics data.

Our approach first fits a very large statistical model to the exposure-mediator-covariate space and treats the basis functions used in this model as a data-adaptive target parameter. This is done in two stages to discover the mediating pathways, the first step discerns the mediating pathways by determining which exposures influence the mediators and subsequently identifying the mediators that impact the outcome. The discovery process yields a set of exposure-mediators, termed pathways.

With these pathways fixed, we estimate the average change in the outcome under stochastic shift interventions on exposures, which are further partitioned into direct and indirect effects. We use and extend the methodology first proposed by [21]. We use the same efficient influence function for the expected change in outcome given a stochastic shift intervention on the exposure holding the mediator at observed values. We explore numeric integration required for nuisance function estimation and build software for mediation when the exposure is continuous or discrete. The resulting estimates, derived within a cross-validated framework paired with general estimating equations and targeted learning, are asymptotically unbiased with the lowest possible variance, subject to the fulfillment of the unconfoundedness and positivity assumptions. Our proposed method delivers valid confidence intervals, unfettered by the number of exposures, covariates, or the intricacy of the data-generating process, provided the exposures are binned into an arbitrary set of categories. As shown, the numeric integration required for exposures that are modeled truly as continuous induces bias in the estimator which prevents the estimator from converging at the required $\sqrt{n}$ rate, which prohibits our ability to construct valid confidence intervals.

However, we acknowledge the method's limitations, primarily its requirement for binned exposures and the computational demands of density estimation. Furthermore, interpretation can be challenging in instances where findings are inconsistent. To enhance the reliability and consistency of the data, we recommend reporting the number of folds in which estimates occur and running NOVAPathways with a high number of folds so a majority of data is used for path discovery in each fold.

Notwithstanding these constraints, both our simulations and real-world data applications underscore the robustness and interpretability of our approach, particularly when exposures are binned, which still have valid continuous interpretations. Our NOVAPathways method provides the research community with a statistical machine wherein, the researcher simply puts in a vector of exposures, mediators, covariates, an outcome, estimators used in the Super Learner of each nuisance parameter, and deltas for each respective exposure. The researcher is then provided a table of proportions for each pathway found in the folds and tables providing direct, indirect and total effects for each pathway.

To support the adoption of semi-parametric methods such as the one we propose, we have made NOVAPathways available via the SuperNOVA R package on GitHub. We believe that by equipping researchers with tools that are not only robust but also flexible, we are inching closer towards solving complex questions in environmental health research.

Fold	Parameter	Psi	Variance	SE	Lower CI	Upper CI	P-Value
Fold 1	NDE-Pseudo-Reg	0.05	0.00	0.06	-0.07	0.17	0.38
Fold 1	NDE-Double-Int	0.05	0.00	0.06	-0.07	0.17	0.41
Fold 1	NIE-Pseudo-Reg	-0.04	0.00	0.06	-0.16	0.07	0.46
Fold 1	NIE-Double-Int	-0.04	0.00	0.06	-0.16	0.08	0.50
Fold 1	Total Effect	0.01	0.00	0.01	-0.01	0.03	0.41
Fold 4	NDE-Pseudo-Reg	2.66	7.64	2.76	-2.76	8.08	0.34
Fold 4	NDE-Double-Int	2.66	7.64	2.76	-2.76	8.07	0.34
Fold 4	NIE-Pseudo-Reg	-2.62	7.66	2.77	-8.05	2.80	0.34
Fold 4	NIE-Double-Int	-2.62	7.66	2.77	-8.04	2.80	0.34
Fold 4	Total Effect	0.04	0.00	0.05	-0.06	0.13	0.45
Fold 5	NDE-Pseudo-Reg	0.92	7.34	2.71	-4.39	6.23	0.73
Fold 5	NDE-Double-Int	0.92	7.34	2.71	-4.39	6.22	0.74
Fold 5	NIE-Pseudo-Reg	-0.88	7.16	2.67	-6.12	4.36	0.74
Fold 5	NIE-Double-Int	-0.87	7.16	2.67	-6.12	4.37	0.74
Fold 5	Total Effect	0.04	0.00	0.04	-0.03	0.11	0.28
Fold 6	NDE-Pseudo-Reg	-0.02	0.00	0.04	-0.10	0.07	0.72
Fold 6	NDE-Double-Int	-0.02	0.00	0.04	-0.11	0.07	0.64
Fold 6	NIE-Pseudo-Reg	-0.02	0.00	0.06	-0.14	0.10	0.73
Fold 6	NIE-Double-Int	-0.01	0.00	0.06	-0.13	0.10	0.80
Fold 6	Total Effect	-0.04	0.00	0.03	-0.09	0.02	0.18
Fold 9	NDE-Pseudo-Reg	0.03	0.00	0.02	-0.02	0.07	0.26
Fold 9	NDE-Double-Int	0.02	0.00	0.02	-0.02	0.07	0.30
Fold 9	NIE-Pseudo-Reg	-0.02	0.00	0.03	-0.07	0.04	0.57
Fold 9	NIE-Double-Int	-0.01	0.00	0.03	-0.07	0.04	0.62
Fold 9	Total Effect	0.01	0.00	0.02	-0.03	0.04	0.64
Fold 10	NDE-Pseudo-Reg	-0.01	0.00	0.01	-0.04	0.02	0.67
Fold 10	NDE-Double-Int	-0.01	0.00	0.01	-0.04	0.02	0.47
Fold 10	NIE-Pseudo-Reg	0.01	0.00	0.02	-0.04	0.05	0.72
Fold 10	NIE-Double-Int	0.01	0.00	0.02	-0.03	0.06	0.59
Fold 10	Total Effect	0.00	0.00	0.01	-0.03	0.03	0.89
Pooled	NDE-Pseudo-Reg	0.61	0.42	0.65	-0.66	1.88	0.35
Pooled	NDE-Integrated	0.60	0.42	0.65	-0.66	1.87	0.35
Pooled	NIE-Pseudo-Reg	-0.60	0.41	0.64	-1.86	0.66	0.35
Pooled	NIE-Integrated	-0.59	0.41	0.64	-1.86	0.67	0.35
Pooled	Total-Pooled-TMLE	0.01	0.00	0.01	-0.01	0.03	0.40

Table 3.2: NDE and NIE of Cesium on Asthma Through Monocyte Percentage Across the Folds

Fold	Parameter	Psi	Variance	SE	Lower CI	Upper CI	P-Value
Fold 1	NDE-Pseudo-Reg	-0.02	0.00	0.03	-0.08	0.03	0.43
Fold 1	NDE-Double-Int	-0.02	0.00	0.03	-0.08	0.03	0.41
Fold 1	NIE-Pseudo-Reg	0.01	0.00	0.03	-0.06	0.08	0.72
Fold 1	NIE-Double-Int	0.01	0.00	0.03	-0.05	0.08	0.69
Fold 1	Total Effect	-0.01	0.00	0.01	-0.04	0.02	0.48
Fold 2	NDE-Pseudo-Reg	-0.02	0.00	0.02	-0.06	0.01	0.17
Fold 2	NDE-Double-Int	0.03	0.00	0.03	-0.03	0.09	0.35
Fold 2	NIE-Pseudo-Reg	0.03	0.00	0.03	-0.03	0.08	0.33
Fold 2	NIE-Double-Int	-0.03	0.00	0.04	-0.10	0.05	0.47
Fold 2	Total Effect	0.00	0.00	0.01	-0.02	0.03	0.84
Fold 3	NDE-Pseudo-Reg	0.03	0.01	0.08	-0.12	0.18	0.72
Fold 3	NDE-Double-Int	0.03	0.01	0.08	-0.13	0.18	0.74
Fold 3	NIE-Pseudo-Reg	-0.02	0.01	0.07	-0.17	0.12	0.77
Fold 3	NIE-Double-Int	-0.02	0.01	0.07	-0.16	0.12	0.78
Fold 3	Total Effect	0.01	0.00	0.01	-0.02	0.03	0.61
Fold 6	NDE-Pseudo-Reg	-0.01	0.00	0.01	-0.03	0.02	0.61
Fold 6	NDE-Double-Int	-0.01	0.00	0.01	-0.03	0.01	0.52
Fold 6	NIE-Pseudo-Reg	-0.01	0.00	0.04	-0.07	0.07	0.89
Fold 6	NIE-Double-Int	-0.00	0.00	0.04	-0.07	0.07	0.93
Fold 6	Total Effect	-0.01	0.00	0.03	-0.07	0.04	0.70
Fold 9	NDE-Pseudo-Reg	-0.06	0.00	0.05	-0.16	0.03	0.19
Fold 9	NDE-Double-Int	-0.06	0.00	0.05	-0.16	0.03	0.18
Fold 9	NIE-Pseudo-Reg	0.04	0.00	0.06	-0.07	0.15	0.44
Fold 9	NIE-Double-Int	0.04	0.00	0.06	-0.07	0.15	0.43
Fold 9	Total Effect	-0.02	0.00	0.02	-0.05	0.01	0.21
Pooled	NDE-Pseudo-Reg	-0.02	0.00	0.02	-0.06	0.02	0.37
Pooled	NDE-Integrated	-0.01	0.00	0.02	-0.05	0.03	0.70
Pooled	NIE-Pseudo-Reg	0.01	0.00	0.02	-0.03	0.05	0.61
Pooled	NIE-Integrated	0.00	0.00	0.02	-0.04	0.04	0.95
Pooled	Total-Pooled-TMLE	-0.01	0.00	0.01	-0.02	0.01	0.41

Table 3.4: NDE and NIE of Lead on Asthma Through Vitamin E Across the Folds

Chapter 4

Open Source Causal Inference Software

The modern era of scientific research presents both exciting opportunities and unprecedented challenges. Vast, complex datasets, innovative computational methods, and evolving statistical models have revolutionized the landscape of scientific discovery. However, these advances have also introduced significant complications that undermine the traditional pillars of scientific inquiry—namely, transparency, reproducibility, and verifiability [46].

Over 70% of biologists, for example, struggle to reproduce their own findings, let alone those of others [47]. Such a lack of reproducibility not only inhibits scientific progress and squanders valuable resources but also erodes public trust in research. Various inefficiencies such as poorly formulated research questions, inadequate research design and methods, and flawed publication practices have led to the alarming squander of approximately 85% of research investment—around $200 billion in 2010 [12].

While these issues affect every branch of scientific research, they pose significant challenges in epidemiology and biostatistics where research questions are frequently governed by extrinsic factors such as funding, available data, and statistical methods, rather than their inherent significance or relevance. This leads to compromised research design and quality, and fosters a research culture that prioritizes selective reporting and post-hoc rationalization over rigorous scientific inquiry.

This is where the concept of a data-adaptive estimand becomes crucial. As a one-shot statistical method, it respects the unknown complexities of high-dimensional data and acknowledges the limitations of the researcher [54]. The estimand is designed to identify latent relationships in the data flexibly and robustly, limiting the opportunities for subjective tinkering with data and analysis. The resulting transparency and consistency can help mitigate the issues related to reproducibility.

However, open and reproducible code alone is not sufficient if the critical elements of discovery and estimation are conducted behind a curtain. In a rigorous scientific process, the software must facilitate not only the estimation but also the discovery. A data-adaptive estimand, therefore, is a framework that enables the integration of discovery and estimation within the same reproducible and transparent process.

But even within this framework, there is still room for human input. In fact, the human

component is indispensable in determining what kind of estimand to pursue. It forces the researcher to focus on the policy relevance and actionability of the estimand rather than its publishability.

By fostering a shift in focus from publishability to policy relevance, we pave the way for a scientific culture that prioritizes real-world impact over academic recognition. Open-source software platforms like CVtreeMLE and SuperNOVA play a pivotal role in enabling this paradigm shift. These tools encourage transparency, ease citation, record run parameters, and detail the machine learning libraries used in estimation.

Adopting such tools, and prioritizing reproducibility and policy-relevance, enhances the reliability of research and facilitates the rapid translation of findings into public policy. It contributes to the advancement of public health by reducing human bias in model selection and producing estimates directly relevant to public policy.

The following sections will provide an in-depth introduction to CVtreeMLE and SuperNOVA software packages. Each was developed to implement the data-adaptive estimand methods described above. Written for the R programming language, their open-source code is freely available on GitHub, supplemented with comprehensive documentation, brief introductions to the underlying theory, unit tests, example data, and user feedback [79].

By harnessing these tools and championing the ethos of reproducibility, transparency, and policy-relevance, we can ensure that scientific research lives up to its foundational principles and delivers tangible benefits to society.

4.1 The `CVtreeMLE` Package

Summary

Statistical causal inference of mixed exposures has been limited by reliance on parametric models and, until recently, 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. Marginal methods for mixture variable selection such as ridge/lasso regression are biased by linear assumptions and the interactions modeled are chosen by the user. Clustering methods such as principal component regression lose both interpretability and valid inference. Newer mixture methods such as quantile g-computation [50] are biased by linear/additive assumptions. More flexible methods such as Bayesian kernel machine regression (BKMR) [8] are sensitive to the choice of tuning parameters, are computationally taxing and lack an interpretable and robust summary statistic of dose-response relationships. No methods currently exist which finds the best flexible model to adjust for covariates while applying a non-parametric model that targets for interactions in a mixture and delivers valid inference for a target parameter.

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. Our target audience are those analysts who would normally use a potentially biased GLM based model for a mixed exposure. Instead, we hope to provide users with a non-parametric statistical machine where users simply specify the exposures, covariates and outcome, `CVtreeMLE` then determines if a best fitting decision tree exists and delivers interpretable results.

Although users do not need strong knowledge of the underlying theory, `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. `CVtreeMLE` uses V-fold cross-validation and partitions the full data into parameter-generating samples and estimation samples. For example, when V=10, integers 1-10 are randomly assigned to each observation with equal probability. In fold 1, observations assigned to 1 are used in the estimation sample and all other observations are used in the parameter-generating sample. This process rotates through the data until all the folds are complete. In the parameter-generating sample, decision trees are applied to a mixed exposure to obtain rules and estimators are created for our statistical target parameter. 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 guaranteed to converge to the truth (under assumptions) at a particular rate as sample size increases. Additionally, `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 led 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 [[119, 50, 11]. In many instances, 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 semi-parametric statistical theory into well-documented functional software is a formidable challenge. Such implementation requires understanding of causal inference, semi-parametric 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 many research scenarios, the analyst is interested in causal inference for an *a priori* specified treatment or exposure. This is because when a single exposure/treatment is measured the analyst is interested in understanding how this exposure/treatment impacts an outcome, controlling for covariates. 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. This is because the conditional outcome given every combination of exposures is not measured. Furthermore, it is likely that, only certain exposures within a mixture have marginal or interacting effects on an outcome. In such a setting, new methods are needed for statistical learning from data that go beyond the usual requirement that the estimand is *a priori* defined in order to allow for proper statistical inference [44].

In the case of mixtures, it is necessary to map a set of continuous mixture components into a lower dimensional representation of exposure using a pre-determined algorithm, and then estimate a target parameter on this more interpretable exposure. 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 subspace (delineated by the rule) compared to those unexposed, or the average treatment effect (ATE) for the mixed exposure. Decision trees have been used as a data-adaptive parameter to explore and estimate heterogeneous treatment effects of a binary treatment [2]. Using a so-called "honest" approach, this method estimates the treatment effect in subpopulations based on covariates in a left-out sample. This approach is limited by not making use of the full data and not data-adaptively selecting the best decision tree.

Advancements in using decision trees as a data-adaptive parameter that solve both these issues and guarantees nominal confidence interval coverage under certain assumptions are needed. Under normal assumptions of conditional independence (A is independent of Y

given W) and positivity (enough experimentation in the data) identifiability of the ATE causal parameter is obtained from the observed data via the statistical functional for a data adaptively determined exposure. This is because, 1. by using Super Learner as our estimator, we are asymptotically guaranteed to select the correct functional for the underlying joint distribution thereby removing bias due to model error and 2. by using TMLE we debias our initial counterfactual for the ATE to target the parameter of interest. The remaining potential bias is therefore due to aggregated data and not the statistical method.

`CVtreeMLE`'s Scope

Building on prior work related to data-adaptive parameters [44] and CV-TMLE [99], chapter 27. `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 generalized 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 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 `R` [79], 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. It integrates with the [‘sl3‘ package](https://github.com/tlverse/sl3) [59] to allow for ensemble machine learning to be leveraged in the estimation of nuisance parameters.

Availability

The `CVtreeMLE` package has been made publicly available [via GitHub](https://github.com/blind-contours/CVtreeMLE). Use of the `CVtreeMLE` package has been extensively documented in the package's ‘README‘ and a vignette.

Code Demonstration

Loading Necessary Packages

```
# The following libraries are needed for our computations and
    visualizations:
library(CVtreeMLE)
library(sl3)
library(pre)
library(partykit)
library(kableExtra)
library(ggplot2)
```

Data Simulation

```
# We use the simulate_mixture_cube function to generate synthetic
    data:
sim_data <- simulate_mixture_cube(
  n_obs = 800,
  splits = c(0.99, 2.0, 2.5),
  mins = c(0, 0, 0),
  maxs = c(3, 4, 5),
  subspace_assoc_strength_betas = c(
    0, 0, 0, 0,
    0, 0, 6, 0
  )
)
```

The 'simulate_mixture_cube' function generates 800 observations across three exposures. The splits parameter determines the thresholds for creating subregions within each variable, much like partitioning a Rubik's Cube into smaller cubes. Each variable is split into two regions at the specified cut-off points, resulting in eight distinct regions or subspaces.

To understand the structure of our simulated data, we can examine the first few rows:

```
# Preview of the simulated data:
head(sim_data)

      age         bmi         sex         M1          M2         M3
                    y
0.01651435 -0.4227082 -1.0221195 1.7594922 0.03442708 2.7936966
    -0.9910446
0.19072911  0.4842019  0.9771362 0.1961772 2.34932053 1.3962661
    1.1728384
-0.18790449  0.4828171 -1.0221195 0.4488381 0.04331044 2.6834768
    -1.2116326
```

```
-0.19596384 -1.1133632 -1.0221195 0.1387679 2.78777587 0.6990761
    -1.2167154
0.26243848  0.6081797  0.9771362 1.6475103 1.33051234 1.6460804
    1.2452379
-1.32782405 -1.0698419 -1.0221195 1.4097762 0.00406810 2.8495084
    -2.3520191
```

The ‘subspace_assoc_strength_betas‘ parameter assigns outcome values to specific regions. In this example, we have set the outcome to 6 for the seventh region (where M2 and M3 are above their split points, and M1 is below its split point), and to 0 for all other regions.

The indices of ‘subspace_assoc_strength_betas‘ correspond to the following regions:

1. All mixtures are lower than specified thresholds.
2. M1 is higher, but M2 and M3 are lower.
3. M2 is higher, but M1 and M3 are lower.
4. M1 and M2 are higher, and M3 is lower.
5. M3 is higher, and M1 and M2 are lower.
6. M1 and M3 are higher, and M2 is lower.
7. M2 and M3 are higher, and M1 is lower.
8. All mixtures are higher than thresholds.

Running CVtreeMLE

Next, we pass the simulated data and variable names to the ‘CVtreeMLE‘ function:

```
# Run CVtreeTMLE:
sim_results <- CVtreeMLE(
  data = sim_data,
  w = c("age", "sex", "bmi"),
  a = c(paste("M", seq(3), sep = "")),
  y = "y",
  n_folds = 5,
  parallel_cv = TRUE,
  seed = 2333,
  parallel_type = "multi_session",
  family = "continuous",
  num_cores = 6
)
```

The `CVtreeMLE` function is executed using default estimators for all parameters, unless user-defined estimators are passed. The default estimators are designed to be non-parametric and computationally efficient, and include techniques like random forests, XGBoost, elastic net, and GLMs. By adjusting the 'num_cores' parameter, users can improve computation speed.

Analyzing Pooled TMLE Results

We can examine the pooled TMLE results to verify if 'CVtreeMLE' consistently identified the correct rule across all folds:

```
# Pooled TMLE results:
mixture_results <- sim_results$`Pooled TMLE Mixture Results`
consistent_results <- mixture_results %>%
   dplyr::filter(Proportion_Folds == 1.0)
consistent_results

  Mixture ATE Standard Error Lower CI Upper CI P-value P-value Adj
         Vars  RMSE

      Union_Rule Proportion_Folds
1       3.259          0.158    2.949    3.568       0           0
       M1-M2 2.128                              M1 >= 0.002 & M1 <=
   0.966 & M2 >= 1.336 & M2 <= 3.968                1
2       5.935          0.037    5.862    6.007       0           0
   M1-M2-M3 1.069 M1 >= 0.002 & M1 <= 0.989 & M2 >= 1.966 & M2 <=
   3.968 & M3 >= 2.436 & M3 <= 4.99                 1
```

The 'ATE' column in the result represents the Average Treatment Effect. In this case, the ATE for the second rule is 5.94 (with 95% confidence interval from 5.84 to 6.03), which is close to the true ATE of 6 used in the data generation.

We can also assess the stability of the estimates and rules by inspecting the v-fold specific results:

```
# v-fold specific results:
mixture_v_results <- sim_results$`V-Specific Mix Results`
mixture_v_results$`M1-M2-M3`

   ate    se lower_ci upper_ci p_val p_val_adj  rmse

       mix_rule   fold variables
1 5.893 0.066   5.7630   6.0230     0         0 1.184
                                              M3 > 2.468 & M2 >
   1.975 & M1 <= 0.986      1  M1-M2-M3
```

```
7 2 5.946 0.043   5.8610   6.0300     0          0 0.997
                                                 M3 > 2.481 & M1 <=
    0.995 & M2 > 1.975      2  M1-M2-M3
8 3 5.946 0.109   5.7320   6.1600     0          0 1.178
                                                 M2 > 2.006 & M3 >
    2.408 & M1 <= 0.985      3  M1-M2-M3
9 4 5.940 0.114   5.7160   6.1630     0          0 1.300
                                                 M3 > 2.481 & M2 >
    1.966 & M1 <= 0.986      4  M1-M2-M3
10 5 5.948 0.071   5.8090   6.0880     0          0 1.113
                                                 M3 > 2.481 & M2 >
    1.975 & M1 <= 0.989      5  M1-M2-M3
11 6 5.935 0.190   5.5627   6.3077     0          0 1.089 M1 >= 0.002 &
    M1 <= 0.989 & M2 >= 1.966 & M2 <= 3.968 & M3 >= 2.436 & M3 <=
    4.99 Pooled  M1-M2-M3
12 >
```

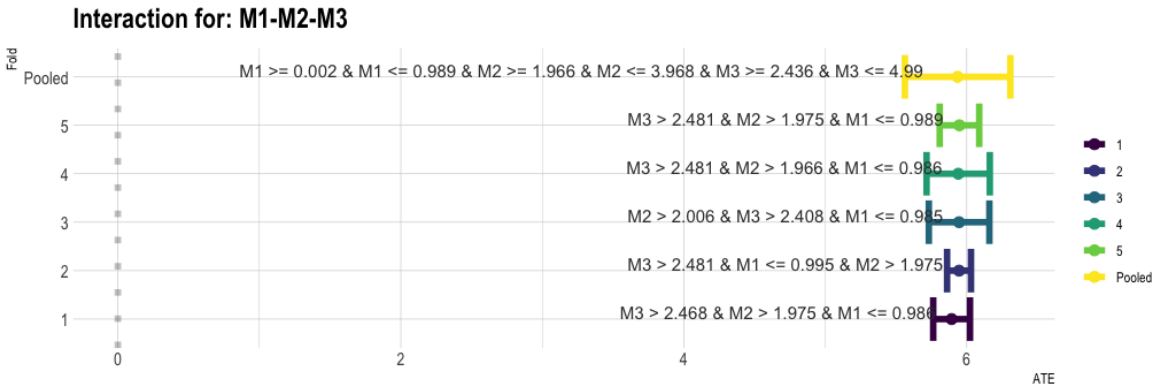
The v-fold specific results include a pooled estimate, which is a weighted average of the fold-specific ATEs and the harmonic mean of the variances. This is similar to meta-analysis approaches.

Visualizing the Results

Finally, we can visualize our findings:

```
# Creating the plot:
mixture_plots <- plot_mixture_results(
  v_intxn_results = sim_results$`V-Specific Mix Results`,
  hjust = 1.05
)
mixture_plots$`M1-M2-M3`
```

This function returns a list of plots, each corresponding to one of the interactions identified.



4.2 The SuperNOVA Package

Summary

Environmental epidemiology studies aim to understand the impact of mixed exposures on health outcomes while adjusting for covariates. However, traditional statistical methods make simplistic assumptions that may not be applicable to public policy decisions. Researchers are ultimately interested in answering causal questions, such as the impact of reducing toxic chemical exposures on adverse health outcomes like cancer. For example, in the case of PFAS, a class of chemicals measured simultaneously in blood samples, identifying the shifts that result in the greatest reduction in thyroid cancer rates can help more directly inform policy decisions on PFAS. In mixtures, nonlinear and non-additive relationships call for new statistical methods to estimate such modified exposure policies.

To address these limitations, the open-source SuperNOVA package has been developed to use data-adaptive machine learning methods for identifying variable sets that have the most explanatory power on an outcome of interest. This package applies non-parametric definitions of interaction and effect modification to these variable sets in a mixed exposure, enabling researchers to explore modified treatment policies using stochastic interventions and answer causal questions.

The SuperNOVA software implements the data-adaptive discovery of variable sets and estimation using optimal estimators for stochastic interventions described in our paper "Semi-Parametric Identification and Estimation of Interaction and Effect Modification in Mixed Exposures using Stochastic Interventions" [64].

Statement of Need

Reliable and accurate estimation of treatment effects is essential in public health and medical research. However, traditional parametric models have limitations, especially when dealing

with complex exposure scenarios like mixed exposures or treatments. Semi-parametric statistical methods are necessary to provide unbiased estimates and consistent findings, but they are not always accessible to researchers.

The open-source R package SuperNOVA addresses this need by offering a powerful and interpretable framework for estimating non-parametric definitions of interaction and effect modification target parameters. This software reduces the risk of model bias and can help drive faster public health decisions by removing human bias due to model selection. SuperNOVA provides a solution to the limitations of traditional parametric models, enabling researchers to adopt these new methods more easily and achieve more consistent findings in public health and medical research.

Background

The package `SuperNOVA` was developed to address the limitations of traditional statistical methods in environmental epidemiology studies. These traditional methods often make overly simplistic assumptions, such as linear and additive relationships, and the resulting statistical quantities may not be directly applicable to public policy decisions. `SuperNOVA` addresses these limitations by using data-adaptive machine learning methods to identify the variables and variable sets that have the most explanatory power on an outcome of interest. In the variable set discovery, the package builds a discrete Super Learner [17] which is a library of machine learning estimators that uses cross-validation to select the best fitting estimator. This Super Learner is composed of flexible basis function estimators, the best of which is analyzed using ANOVA style analysis to determine the variables that contribute most to the model fit through an F-statistic for basis functions.

The variable sets used in the basis functions drive the target parameters estimated. In the event of basis functions for an individual exposure A, the effects of an individual shift are estimated, for basis function with A and W (a baseline covariate), the effect modification parameter is estimated, which is an individual shift in a covariate region and if two exposures are included in a basis function A_1, A_2 the interaction target parameter is estimated, which is the expected outcome under dual shift of both exposures compared to the sum of expected outcomes given individual shifts independently. For each target parameter we use ensemble machine learning to ascertain the expected outcome under a shift and we use cross-validated targeted maximum likelihood estimation [44] to debias our initial estimates thereby creating an asymptotically unbiased estimator with minimum variance. When we say shift, we mean a stochastic shift [48].

In this framework we calculate the average outcome after shifting the exposure. A stochastic intervention changes the function that defines the exposure A and its conditional density $g(A \mid W)$ with a candidate density $g_{A_\delta}(A \mid W)$. The new density defines how the exposure is modified by a random draw from $g_{A_\delta}(A \mid W)$. This can include static interventions, where all mass is placed on a single value, such as the average treatment effect, where all observations either receive or don't receive treatment or a shift such as in pollution or drug where we increase all exposure by say 100 parts per million to a chemical such as PFAS

and observe the change in outcomes. Stochastic interventions give rise to a counterfactual outcome $Y_{A_\delta} := f_Y(A_\delta, W, U_Y)$, which is obtained by replacing the natural value of the exposure with a shifted value. The degree of shift δ describes the reduction in exposure, based on the individual's baseline characteristics W. We can evaluate the causal effect of the intervention by finding the counterfactual mean of the outcome under the modified distribution, $\psi_{0,\delta} = E_{P_0}^{A_\delta} Y_{A_\delta}$.

In this way, `SuperNOVA` allows analysts to explore modified treatment policies and ask causal questions (under assumptions) about the impact of mixed exposures on health outcomes. `SuperNOVA` uses V-fold cross-validation procedures to avoid over-fitting and incorrect model assumptions by creating parameter generating samples wherein the variable sets are determined and estimators for nuisance parameters are trained, an estimation sample is then used to estimate the target parameters of interest [121]. Additionally, to avoid positivity violations (user inputs a shift amount that there isn't enough experimentation in the data to estimate) the shift amount can also be input as a data-adaptive parameter which finds the maximum shift possible for each exposure.

`SuperNOVA`'s Scope

The `SuperNOVA` software package is built for the R language and implements our proposed methodology for estimating modified treatment policies in environmental epidemiology studies for data-adaptively identified variable sets. It is specifically designed to estimate the effects of mixed exposures on health outcomes, while adjusting for covariates and potential confounders.

As input, `SuperNOVA` takes in variable sets A (exposures), W (covariates), Y (outcome) and a vector of deltas for each exposure in A, representing the degree of shift in each exposure if it is identified as predictive of the outcome. The output of `SuperNOVA` is a dose-response analysis for variable sets data-adaptively identified in the mixed exposure, estimating the expected outcome under a change in exposure compared to the observed outcome under the observed exposure. Using these shift parameters, users are provided with estimates of non-parametric definitions of interaction and effect modification that are directly informative for public health policy. `SuperNOVA` is a valuable tool for researchers in many fields who need an interpretable and robust statistical approach to answer modified treatment policy questions, estimates are interpreted as the expected outcome if an exposure was changed by the respective delta.

`SuperNOVA` is designed to provide analysts with both V-fold specific and pooled results for stochastic intervention causal effects. It integrates with the [‘sl3‘ package] [17] to allow for ensemble machine learning to be leveraged in the estimation of nuisance parameters.

Availability

The `SuperNOVA` package has been made publicly available [via GitHub](https://github.com/blind-contours/SuperNOVA). Use of the ‘SuperNOVA‘package has been extensively documented in the package's ‘README‘ and a vignette.

Code Demonstration

```
library(SuperNOVA)
library(devtools)
library(kableExtra)
library(sl3)

set.seed(429153)
# simulate simple data
n_obs <- 100000
```

Data Simulation

```
sim_out <- simulate_data(n_obs = n_obs)
data <- sim_out$data
```

Sample Data and Run SuperNOVA

```
data_sample <- data[sample(nrow(data), 4000), ]

w <- data_sample[, c("W1", "W2", "W3")]
a <- data_sample[, c("M1", "M2", "M3", "M4")]
y <- data_sample$Y

deltas <- list("M1" = 1, "M2" = 1, "M3" = 1, "M4" = 1)

ptm <- proc.time()
sim_results <- SuperNOVA(
  w = w,
  a = a,
  y = y,
  delta = deltas,
  n_folds = 3,
  num_cores = 6,
  outcome_type = "continuous",
  quantile_thresh = 0,
  seed = 294580
)
proc.time() - ptm

# Extract the results from the returned object:
basis_in_folds <- sim_results$`Basis Fold Proportions`
```

```
indiv_shift_results <- sim_results$`Indiv Shift Results`
em_results <- sim_results$`Effect Mod Results`
joint_shift_results <- sim_results$`Joint Shift Results`
```

Let's first look at the variable relationships used in the folds:

```

 M1 M1M4   M3 M3W3   M4   W3
0.33 1.00 0.67 1.00 1.00 1.00
```

In this example, y is generated from $M1$, $M4$ and interaction between these two variables and $M3$ which is modified by $W3$. The impact of $M1$ is small and here we only find it in 1 of 3 folds. The impact of the interaction is strong , we see it in all the folds, as is the impact of $M4$. This example run was only done with 3 fold CV and so results are not stable as if it was done with 10 fold CV, we did this simply for computational time.

Let's look at the results for the marginal impact of $M4$. The truth for this effect is:

```
sim_out$m4_effect
[1] 10.3882
```

```
indiv_shift_results$M4

Condition      Psi  Variance        SE Lower CI Upper CI
   P-value        Fold        Type Variables    N Delta
        M4 10.43028 0.2861536 0.5349332   9.3818  11.4787
           1.135361e-84           1 Indiv Shift        M4 1334     1
        M4 10.40188 0.2930203 0.5413135   9.3409  11.4628
           2.719726e-82           2 Indiv Shift        M4 1333     1
        M4 10.42346 0.7148800 0.8455058   8.7663  12.0806
           6.395340e-35           3 Indiv Shift        M4 1333     1
        M4 10.46028 0.1846333 0.4296898   9.6181  11.3025
           6.744836e-131 Pooled TMLE Indiv Shift        M4 4000     1
```

This table shows that, for a 1 unit increase in $M4$ the outcome Y increases by 10.4, this finding is consistent across all the folds. The pooled estimate has the smalles variance, utilizing data across the folds, as expected.

Let's next look at the effect modification found between $M3$ and $W3$.

```
em_results$M3W3
                        Condition       Psi  Variance       SE
                           Lower_CI Upper_CI      P_value        Fold
                                 Type Variables    N Delta
Level 1 Shift Diff in W3 <= 0 -5.316240 76.097437 8.723384 -22.4138
    11.7813 5.422435e-01           1 Effect Mod      M3W3 1334     1
Level 0 Shift Diff in W3 <= 0  1.371439  2.378236 1.542153  -1.6511
      4.3940 3.738412e-01           1 Effect Mod      M3W3 1334     1
```

```
Level 1 Shift Diff in W3 <= 0 -5.263901 36.093815 6.007813 -17.0390
      6.5112 3.809344e-01            2 Effect Mod      M3W3 1333     1
Level 0 Shift Diff in W3 <= 0  2.102544  2.664307 1.632270  -1.0966
      5.3017 1.977077e-01            2 Effect Mod      M3W3 1333     1
Level 1 Shift Diff in W3 <= 0 11.567295 10.834292 3.291549   5.1160
    18.0186 4.410127e-04            3 Effect Mod      M3W3 1333     1
Level 0 Shift Diff in W3 <= 0 18.794371  6.395027 2.528839  13.8379
    23.7508 1.069555e-13            3 Effect Mod      M3W3 1333     1
Level 1 Shift Diff in W3 <= 0  3.934214  5.951079 2.439483  -0.8471
      8.7155 1.068044e-01 Pooled TMLE Effect Mod      M3W3 4000     1
Level 0 Shift Diff in W3 <= 0 12.103738  1.412560 1.188512   9.7743
    14.4332 2.338840e-24 Pooled TMLE Effect Mod      M3W3 4000     1
```

The covariate $W3$ here is binary and so we get the effect for a shift in $M3$ when $W3$ is 1 and 0. If the effect modifier is not binary, then a partition is found and results are given at each level of the partition in the covariate space. Here we see a larger impact in $M3$ on the outcome when the effect modifier is low compared to high. This matches our ground-truth where the effect is 1 when the modifier is 0 and 11 when it is 1. So we have proper coverage and correctly identify the modifier in this simulated data. The true effects look like:

```
> sim_out$effect_mod
$`Level 0 Shift Diff in W3 <= 0`
[1] 10.99079

$`Level 1 Shift Diff in W3 <= 0`
[1] 1
```

Lastly, let's look at the interaction effect. The true effect is:

```
> sim_out$m1_effect
[1] 1.600192
> sim_out$m4_effect
[1] 10.3882
> sim_out$m14_effect
[1] 12.38839
> sim_out$m14_intxn
[1] 0.4
> 1.600192 + 10.3882
[1] 11.98839
```

Above, the marginal effect of $M1$ is 1.6, $M4$ is 10.4, the additive effect of these is then 12, the actual effect of a joint shift is 12.4 so Ψ is 0.4, the difference of the joint effect compared to the additive effect.

The output from `SuperNOVA` looks like:

```
```

```
> joint_shift_results$M1M4
   Condition          Psi   Variance        SE Lower CI Upper CI
      P-value         Fold        Type Variables    N Delta M1 Delta
      M4
M1  1.46336129  2.5445985 1.5951798  -1.6631   4.5899 2.466049e-01
              1 Interaction     M1&M4 1334         1         1
M4 10.54131272  0.3050387 0.5523031   9.4588  11.6238 1.147076e-45
              1 Interaction     M1&M4 1334         1         1
M1&M4 12.47959870  0.3060496 0.5532175  11.3953  13.5639
   3.506850e-63           1 Interaction     M1&M4 1334         1
          1
Psi  0.47492469  2.5464908 1.5957728  -2.6527   3.6026 7.069482e-01
              1 Interaction     M1&M4 1334         1         1
M1  8.61817807 10.2244664 3.1975719   2.3511  14.8853 1.438919e-06
              2 Interaction     M1&M4 1333         1         1
M4 10.39947363  0.2938431 0.5420729   9.3370  11.4619 2.671335e-45
              2 Interaction     M1&M4 1333         1         1
M1&M4 12.38345613  0.2947550 0.5429135  11.3194  13.4475
   2.188112e-63           2 Interaction     M1&M4 1333         1
          1
Psi -6.63419557 10.2504833 3.2016376 -12.9093  -0.3591 2.091672e-04
              2 Interaction     M1&M4 1333         1         1
M1  1.35150469  1.4025618 1.1842980  -0.9697   3.6727 2.142730e-01
              3 Interaction     M1&M4 1333         1         1
M4 10.36554589  0.7165151 0.8464721   8.7065  12.0246 1.922329e-29
              3 Interaction     M1&M4 1333         1         1
M1&M4 11.37912283  0.7675117 0.8760774   9.6620  13.0962
   5.245696e-34           3 Interaction     M1&M4 1333         1
          1
Psi -0.33792775  1.3974915 1.1821555  -2.6549   1.9791 7.559496e-01
              3 Interaction     M1&M4 1333         1         1
M1  0.05600595  0.7490726 0.8654898  -1.6403   1.7523 9.519956e-01
   Pooled TMLE Interaction     M1&M4 4000         1         1
M4 10.44655789  0.1847928 0.4298753   9.6040  11.2891 3.730056e-57
   Pooled TMLE Interaction     M1&M4 4000         1         1
M1&M4 11.98522952  0.1923595 0.4385880  11.1256  12.8448
   3.336193e-73 Pooled TMLE Interaction     M1&M4 4000         1
          1
Psi  1.48266568  0.7440697 0.8625947  -0.2080   3.1733 1.104011e-01
   Pooled TMLE Interaction     M1&M4 4000         1         1
```

Our pooled Ψ estimate for the interaction covers the truth, as does our estimates for the marginal impact of $M1$, $M4$ and the joint shift.

This simulation function comes with `SuperNOVA` so users can test performance on this

simulation and ensure results cover the truth.

Chapter 5

Future Investigations

5.1 Mediation Analysis for CVtreeMLE

The method CVtreeMLE is designed to transform complex, mixed exposures, including multiple continuous variables, into a binary indicator. This indicator represents a specific region of exposure space, as defined by a tree-based rule. This rule identifies combinations of exposure levels that most effectively influence a certain outcome. The aim of CVtreeMLE is to calculate the Average Regional Exposure Effect (ARE), equivalent to an Average Treatment Effect (ATE), which estimates the mean difference in outcomes between individuals within the identified region versus those outside it.

While ARE measures the total impact of the exposure region on the outcome, understanding the specific causal pathways behind this effect is often valuable. For example, exposure to particular levels of environmental toxins may cause changes in biomarkers leading to disease. Identifying the exposure levels that trigger these changes can provide insights into disease mechanisms and inform intervention strategies.

By converting a complex mixture of exposures into a binary regional exposure vector, we can break down the total regional effect into direct and indirect effects, thus highlighting mediating pathways. Here, direct effects show the exposure's influence on the outcome without any mediators, while indirect effects represent the influence of exposure on the outcome through mediators, a concept articulated in the NOVAPathways approach.

While this methodology is similar to the NOVAPathways approach (decomposing a total effect into natural effects), a notable divergence exists as our exposure is defined in binary terms. Let's clarify the mathematical representation of this: if we denote Y as the outcome, A as the exposure (the binary regional exposure indicator, in our context), Z as the mediator, and W as the covariates, we can express the direct effect as $E[Y(1, Z(0)) - Y(0, Z(0))]$ and the indirect effect as $E[Y(Z(1), 1) - E(Y(Z(0), 1)]$. It's important to stress that we're not working with a singular exposure A, but rather regions of the exposure space, denoted as A_{region}. This adjustment necessitates a reinterpretation of $g(A|W)$, which previously represented the probability of exposure given the covariates, as the likelihood of exposure to a

region given the covariates, i.e., $g(A_{region}|W)$. Thus, we are estimating the natural direct and indirect effects being exposed to a combination of exposure levels has on a disease outcome.

Moving forward, we utilize the asymptotically linear estimator put forth by [122]. Zheng's work advances the field by building on the efficient scores (under a nonparametric model) for various natural effect parameters and formulating general robustness conditions. This work also presents an estimating equation-based estimator utilizing the efficient score, a concept expanded upon in the work of Tchetgen Tchetgen and Shpitser [96].

By harnessing the targeted maximum likelihood framework described by [122], we construct a semiparametric efficient, multiply robust, substitution estimator for the natural direct effect. This estimator satisfies the efficient score equation, as derived by Tchetgen Tchetgen and Shpitser [96], for the Average Regional Exposure Effect.

It's worth noting that this estimator was originally coded for the tmle3mediate package [39], developed collaboratively by myself, Nima Hejazi, and James Duncan, with the aim of providing researchers a package for mediation analysis with binary exposure. Building on this, I extend this work to account for the ARE case, integrating data-adaptive target parameters and leveraging the Cross-Validation Targeted Maximum Likelihood Estimation (CV-TMLE) method. With additional testing, the ARE can be decomposed into (in)direct effects, therefore researchers can include a set of mediators in CVtreeMLE. For regions in the exposures, estimates can then be given for the effect of exposure to certain levels of exposures and if this effect goes through certain mediating variables.

Estimating Natural Direct Effects in CVtreeMLE

The estimation process for the Natural Direct Effect (NDE) begins by creating $\overline{Q}_{Y,n}$, an estimate of the conditional mean of the outcome given Z, A, and W. With this estimate, we can predict $\overline{Q}_Y(Z, 1, W)$ (setting $A = 1$) and $\overline{Q}Y(Z, 0, W)$ (setting $A = 0$). The difference, $\overline{Q}$diff, helps us understand variations in the conditional mean of Y across contrasts of A.

We construct a targeted maximum likelihood (TML) estimator for the NDE by treating $\overline{Q}$diff as a nuisance parameter. In this procedure, we regress its estimate $\overline{Q}\text{diff}, n$ on baseline covariates W, only considering observations in the control condition (i.e., those where $A = 0$ is observed). The goal is to remove part of the marginal impact of Z on $\overline{Q}_{\text{diff}}$, as the covariates W precede the mediators Z in time.

Estimating Natural Indirect Effects in CVtreeMLE

The process of deriving and estimating the natural indirect effect (NIE) mirrors that of the natural direct effect (NDE). The NIE represents the influence of a variable A on an outcome Y solely through a mediator variable Z. This influence is quantified as the difference between the conditional mean of Y given $A = 1$ and $Z(1)$ and the conditional mean of Y given $A = 1$ and $Z(0)$.

Similar to the NDE, we can replace $q_Z(Z|A, W)$ with $e(A|Z, W)$ in the estimation process, which sidesteps the need to estimate a potentially multivariate conditional density. However,

unlike the NDE, the estimation of the NIE involves a two-step regression process. First, the conditional mean of Y given Z and W when $A = 1$ is regressed on W among the treated units ($A = 1$). Then, the same value is regressed on W among the control units ($A = 0$). The average difference between these two steps is a valid estimator of the NIE, demonstrating the treatment's marginal effect on the conditional mean of Y given $A = 1$ and Z, through its effect on Z.

Integration into CVtreeMLE

The above approach has been preliminarily coded into CVtreeMLE. Our goal is to code these new estimators, needed for mediation of the ARE and deliver both fold specific and pooled effects. Additional testing is required and this will result in another publication, extending work from the original ARE proposed in CVtreeMLE.

5.2 Stochastic Interventions in the Context of CVtreeMLE

A natural extension to our CVtreeMLE method, and an avenue for future research, lies in the incorporation of a more realistic representation of exposure regulations. The current iteration of CVtreeMLE translates a mixture of continuous exposures into a binary regional exposure indicator, providing an average regional exposure effect (ARE). This approach assumes homogeneity of the exposure distribution within a specified region. However, this does not accurately reflect real-world scenarios.

For instance, consider a policy regulation enforcing a limit on lead levels below 15 ppb (regulations established by the EPA). In our current model, any observation within this region, be it 14 or 0.1, is treated as equally likely. When estimating the counterfactual, we essentially estimate the effect if all observations were moved into this region. In practical terms, however, following the implementation of such a regulation, there is likely to be a clustering of data points just below the threshold, as entities try to adhere to the regulation in the least disruptive or costly way.

To better reflect this situation, we propose a modification to our estimand that incorporates this heterogeneity. Rather than deterministically setting our binary exposure, A, to 1 or 0, we could data-adaptively identify thresholds in the mixture that optimally explain disease outcomes, similar to our existing CVtreeMLE method. Our estimates, however, would be based on stochastic shifts towards this threshold.

For individuals already within the identified region, no shift occurs, while those outside the region experience a shift to the minimum or maximum value within the data-adaptively identified regulation. This approach results in a concentration of data points around the threshold level, representing a more realistic regulatory scenario.

To formulate this mathematically, following the example set by Díaz and van der Laan, we could define a stochastic intervention that, given an initial exposure A, generates a new exposure A^* based on a conditional distribution $g(A^*|A)$.

Our goal would then be to estimate the interventional causal effect of a shift towards the threshold, quantified by the contrast between the expectation of the outcome under this new exposure level and the observed outcome. This revised approach would offer a more accurate reflection of real-world dynamics in the context of exposure regulations.

5.3 Interaction Mediation

While the primary purpose of NOVAPathways is to estimate direct and indirect effects of data-adaptively identified pathways involving a single exposure and mediator, there can be situations where understanding the compounded impact of multiple exposures on a mediating pathway becomes crucial. These instances prompt an examination of whether the collective impact of exposures is more than the sum of their individual effects—introducing the concept of interaction through a mediating biological system.

Consider a hypothetical scenario involving two endocrine disrupting compounds (EDCs): compound A and compound B. These compounds, when studied within our stochastic intervention framework, permit us to estimate the mediation for a joint exposure defined as $\psi(\delta_1, \delta_2)$, mediation of each individual exposure $\psi(\delta_1)$ and $\psi(\delta_2)$, and $\psi(1, 2)$ the outcome under observed exposures. These equations represent the interaction indirect effect (IIE) and interaction direct effect (IDE) when estimated for both EDCs operating simultaneously.

$$\psi(\delta_1, \delta_2) = \underbrace{E[Y(A_{\delta_1}, A_{\delta_2}, Z(A_{\delta_1}, A_{\delta_2})) - Y(A_{\delta_1}, A_{\delta_2}, Z)]}_{IIE} + \underbrace{E[Y(A_{\delta_1}, A_{\delta_2}, Z) - Y(A_1, A_2, Z)]}_{IDE}$$

Here, the IIE signifies the effect attributable to a change in the mediator Z resulting from a joint change in both exposures (A). Conversely, the IDE captures the outcome's change resulting from a joint change in both exposures while the mediator remains fixed.

To examine the influence of each exposure individually, we calculate the direct and indirect effects given a shift in each compound:

$$\psi(\delta_1) = \underbrace{E[Y(A_{\delta_1}, A_2, Z(A_{\delta_1})) - Y(A_{\delta_1}, A_2, Z)]}_{IIE} + \underbrace{E[Y(A_{\delta_1}, A_2, Z) - Y(A_1, A_2, Z)]}_{IDE}$$

and:

$$\psi(\delta_2) = \underbrace{E[Y(A_{\delta_2}, A_1, Z(A_{\delta_2})) - Y(A_{\delta_2}, A_1, Z)]}_{IIE} + \underbrace{E[Y(A_{\delta_2}, A_1, Z) - Y(A_1, A_2, Z)]}_{IDE}$$

The observed outcome under observed exposures is simply represented as:

$$\psi(1,2) = E[Y(A_1, A_2)]$$

Subsequently, we can denote:
Total effect as:

$$\psi(\delta_1, \delta_2) - \psi(\delta_1) - \psi(\delta_2) + \psi(1,2)$$

Indirect effect as:

$$\psi(\delta_1, \delta_2)_{IIE} - \psi(\delta_1)_{IIE} - \psi(\delta_2)_{IIE} + \psi(1,2)$$

Direct effect as:

$$\psi(\delta_1, \delta_2)_{IDE} - \psi(\delta_1)_{IDE} - \psi(\delta_2)_{IDE} + \psi(1,2)$$

Employing these metrics allows us to disaggregate individual and joint effects of the two EDCs. The indirect effect measures the impact of concurrent exposures through a shared receptor pathway, whereas the direct effect uncovers the influence of these compounds outside this shared pathway.

The scope of this methodology transcends EDCs, offering essential insights into synergistic effects of concurrent exposures across various research areas. By presenting a structured way to evaluate the collective impact of simultaneous exposures, we facilitate deeper comprehension of multifactorial health risks, aiding in formulating more effective intervention strategies. Future research will seek to apply and expand this novel statistical framework for mediation of interactions across a diverse range of scenarios to capture the intricacies of concurrent exposures and their aggregate effects on health outcomes. This estimation can be built into the SuperNOVA infrastructure where, if two exposures are found to operate through a mediator in the pathway discover section, estimation of this mediation for an interaction can be estimated.

5.4 Concluding Remarks and Vision for the Future

The culmination of the endeavors and insights that have been gathered throughout this dissertation calls to mind the familiar adage: "Data is the new oil." However, with the ever-growing influx of data, the question arises: How can we harness this oil to power a reliable machine—one that delivers trustworthy, reproducible results while also being easily understood and handled by its users? The crux of this dissertation lies in addressing this question, adopting an airplane analogy: Like a passenger who trusts the airplane to take them to their destination without delving into the technical intricacies of its operation, the objective is to create statistical machinery where data goes in, and results come out, with the data itself guiding the process. Throughout the chapters of this dissertation, we navigated through a sequence of decisions used in this machine:

1. What are we aiming to uncover within the data—thresholds, variable sets, pathways?
2. What algorithmic strategy would be best suited to find these elements for us?
3. Given the outputs from 1 and 2, what do we aim to achieve with this information? Are we looking to reduce exposure, determine safe levels, what's our estimand?
4. What suite of algorithms are selected for estimating the nuisance parameters for this estimand?
5. How does the cross-validation function, and do we need to consider adjustments for repeated measures on the same individual within the CV framework? Is there a longitudinal structure?
6. How do we pool results across the folds and assess for reliability of the data-adaptive target parameter?

This sequence of choices, with changes at various points, forms the backbone of the presented dissertation. Despite adopting diverse approaches, each chapter respects the overarching problem-solving strategy while acknowledging the intricacies of estimating effects in high-dimensional data, where details about the exposures, mediators, and interactions might be scarce.

Envisioning the future, we imagine a user-friendly interface that guides users to describe the data generating process (DGP) for their data using a Directed Acyclic Graph (DAG), choose their research question, and make the aforementioned decisions. Once these choices are made, the user hits 'Go,' and the statistical machine takes over. This system becomes a unified platform for causal inference, even for simpler models, like a single binary exposure, utilizing CV-TMLE and other relevant methods.

The future emanates from the existing work, currently scattered across various packages, to construct a singular, object-oriented package with GUI interfaces. The analyst merely needs to drag variable names to different parts of the DAG, choose their interest—mediation or interaction search, select the algorithms that compose each Super Learner, and hit 'Go.' The process then remains largely hands-off, with findings, parameters of the statistical machine, and code all published for consistency.

If we can streamline this statistical research process through a common platform, we can promote transparency, reproducibility, and trust in statistical research outcomes. Such an ecosystem could transform the current landscape of biostatistics and epidemiology, ensuring a more robust, accessible, and efficient approach to data analysis. In this way, rather than the researchers spending time tinkering with models, they can instead focus on what questions are most pertinent to public health and how to mathematically formalize them. Once an estimand is established which respects the DGP as close as possible following the statistical roadmap, then the user can use the platform to make estimation.

This vision succinctly captures the essence of this dissertation, serving as both a compendium of our endeavors and a roadmap to future explorations in this field. We stand on

the tarmac of a grand journey, with our 'statistical airplane' prepped for take-off, symbolizing the readiness of our methodology. However, this take-off into the future of statistics will require more than just mechanical readiness; it will necessitate a firm commitment to our core mission: the generation of reliable, reproducible results that can inform and influence public policy.

The status quo in our field awaits a paradigm shift—a statistical revolution. This transformative change rests upon embracing data-adaptive target parameters as a new standard for statistical analysis. With these in place, we can foster an environment of trust, ensuring that our research outcomes are not only reproducible but also trustworthy, thereby instilling confidence in our stakeholders and effectuating tangible societal impact.

This dissertation has meticulously laid the groundwork, paving the way for this necessary revolution. The blueprint is now in our hands, and it's incumbent upon us - the collective force of statisticians, epidemiologists, environmental scientists, and policy makers - to construct this "aircraft". We're at the helm of this transformation, entrusted with the mission to steer our field into a future where statistical analysis becomes the bastion of robustness, transparency, and, above all, trust.

Bibliography

[1] Nur Adila Adnan et al. "Comparison of joint effect of acute and chronic toxicity for combined assessment of heavy metals on Photobacterium sp.NAA-MIE". In: *International Journal of Environmental Research and Public Health* 18.12 (2021). ISSN: 16604601. DOI: 10.3390/ijerph18126644.

[2] Susan Athey and Guido Imbens. "Recursive partitioning for heterogeneous causal effects". In: *Proceedings of the National Academy of Sciences of the United States of America* 113.27 (2016), pp. 7353–7360. ISSN: 10916490. DOI: 10.1073/pnas.1510489113. arXiv: 1504.01132.

[3] Mahdi Balali-Mood et al. "Toxic Mechanisms of Five Heavy Metals: Mercury, Lead, Chromium, Cadmium, and Arsenic". In: *Frontiers in Pharmacology* 12.April (2021), pp. 1–19. ISSN: 16639812. DOI: 10.3389/fphar.2021.643972.

[4] Julia A. Bauer et al. "Associations of a metal mixture measured in multiple biomarkers with IQ: Evidence from italian adolescents living near ferroalloy industry". In: *Environmental Health Perspectives* 128.9 (2020), pp. 097002–1–097002–12. ISSN: 15529924. DOI: 10.1289/EHP6803.

[5] David Benkeser and Mark Van Der Laan. "The Highly Adaptive Lasso Estimator". In: *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 2016, pp. 689–696. DOI: 10.1109/DSAA.2016.93.

[6] Peter J Bickel. "On adaptive estimation". In: *The Annals of Statistics* (1982), pp. 647–671.

[7] Cécile Billionnet et al. "Quantitative assessments of indoor air pollution and respiratory health in a population-based sample of French dwellings". In: *Environmental Research* 111.3 (2011), pp. 425–434. ISSN: 00139351. DOI: 10.1016/j.envres.2011.02.008.

[8] Jennifer F. Bobb et al. "Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures". In: *Biostatistics* 16.3 (2014), pp. 493–508. ISSN: 14684357. DOI: 10.1093/biostatistics/kxu058.

[9] L. Breiman et al. "Classification and Regression Trees". In: 1984.

[10] Benfeng Cao et al. "U-shaped association between plasma cobalt levels and type 2 diabetes". In: *Chemosphere* 267.12 (2021), pp. 1876–1881. ISSN: 18791298. DOI: 10.1016/j.chemosphere.2020.129224.

[11] Danielle J Carlin et al. "Unraveling the health effects of environmental mixtures: an NIEHS priority". In: *Environmental health perspectives* 121.1 (2013), A6–A8. DOI: 10.1289/ehp.1206187.

[12] Iain Chalmers and Paul Glasziou. "Avoidable waste in the production and reporting of research evidence". In: *The Lancet* 374.9683 (2009), pp. 86–89. ISSN: 01406736. DOI: 10.1016/S0140-6736(09)60329-9. URL: http://dx.doi.org/10.1016/S0140-6736(09)60329-9.

[13] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

[14] Victor Chernozhukov et al. *Double/debiased machine learning for treatment and structural parameters*. 2018.

[15] Cassidy Clarity et al. "Associations between polyfluoroalkyl substance and organophosphate flame retardant exposures and telomere length in a cohort of women firefighters and office workers in San Francisco". In: *Environmental Health: A Global Access Science Source* 20.1 (2021), pp. 1–14. ISSN: 1476069X. DOI: 10.1186/s12940-021-00778-z.

[16] Jeremy R Coyle et al. *hal9001: The scalable highly adaptive lasso*. R package version 0.4.3. 2022. DOI: 10.5281/zenodo.3558313. URL: https://github.com/tlverse/hal9001.

[17] Jeremy R Coyle et al. *sl3: Modern pipelines for machine learning and Super Learning*. https://github.com/tlverse/sl3. R package version 1.1.0. 2018.

[18] Frank De Vocht, Nicola Cherry, and Jon Wakefield. "A Bayesian mixture modeling approach for assessing the effects of correlated exposures in case-control studies". In: *Journal of Exposure Science and Environmental Epidemiology* 22.4 (2012), pp. 352–360. ISSN: 15590631. DOI: 10.1038/jes.2012.22.

[19] Ivan Díaz and Mark J van der Laan. "Population intervention causal effects based on stochastic interventions". In: *Biometrics* 68.2 (2012), pp. 541–549.

[20] Ivan Díaz and Mark J van der Laan. "Stochastic treatment regimes". In: *Targeted Learning in Data Science*. Springer, 2018, pp. 219–232.

[21] Iván Díaz and Nima S. Hejazi. "Causal Mediation Analysis for Stochastic Interventions". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.3 (Feb. 2020), pp. 661–683. ISSN: 1369-7412. DOI: 10.1111/rssb.12362. eprint: https://academic.oup.com/jrsssb/article-pdf/82/3/661/49323651/rssb12362-sup-0001-supinfo.pdf. URL: https://doi.org/10.1111/rssb.12362.

[22] Iván Díaz et al. "Non-parametric causal effects based on longitudinal modified treatment policies". In: *Journal of the American Statistical Association* (2021). DOI: 10.1080/01621459.2021.1955691.

[23] Abbas Esmaeilzadeh et al. "Role of oxidative stress in respiratory diseases: from molecular mechanisms to therapeutic approaches". In: *Respiratory Research* 22 (2021), p. 210. DOI: 10.1186/s12931-021-01801-4. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8330548/.

[24] Kristen M. Fedak et al. "Applying the Bradford Hill criteria in the 21st century: How data integration has changed causal inference in molecular epidemiology". In: *Emerging Themes in Epidemiology* 12.1 (2015), pp. 1–9. ISSN: 17427622. DOI: 10.1186/s12982-015-0037-4.

[25] Marjolein Fokkema. "Fitting Prediction Rule Ensembles with R Package pre". In: *Journal of Statistical Software* 92.12 (2020), pp. 1–30. DOI: 10.18637/jss.v092.i12.

[26] Marjolein Fokkema. "Fitting prediction rule ensembles with R package pre". In: *Journal of Statistical Software* 92.12 (2020). ISSN: 15487660. DOI: 10.18637/jss.v092.i12. arXiv: 1707.07149.

[27] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. DOI: 10.18637/jss.v033.i01. URL: https://www.jstatsoft.org/v33/i01/.

[28] Jerome H. Friedman and Bogdan E. Popescu. "Predictive learning via rule ensembles". In: *Annals of Applied Statistics* 2.3 (2008), pp. 916–954. ISSN: 19326157. DOI: 10.1214/07-AOAS148.

[29] Miguel García-Villarino et al. "Exposure to metal mixture and growth indicators at 4–5 years. A study in the INMA-Asturias cohort". In: *Environmental Research* 204 (2022). ISSN: 10960953. DOI: 10.1016/j.envres.2021.112375.

[30] Elizabeth A Gibson et al. "An overview of methods to address distinct research questions on environmental mixtures: an application to persistent organic pollutants and leukocyte telomere length". In: *Environmental Health* 18 (2019), pp. 1–16.

[31] Adam N. Glynn and Kevin M. Quinn. "An introduction to the augmented inverse propensity weighted estimator". In: *Political Analysis* 18.1 (2009), pp. 36–56. ISSN: 10471987. DOI: 10.1093/pan/mpp036.

[32] Arthur S Goldberger. "Structural equation methods in the social sciences". In: *Econometrica: Journal of the Econometric Society* (1972), pp. 979–1001.

[33] Katarzyna Grzela et al. "Oxidative Stress and Bronchial Asthma in Children—Causes or Consequences?" In: *Frontiers in Pediatrics* 5 (2017), p. 129. DOI: 10.3389/fped.2017.00129. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5523023/.

[34] Sebastian Haneuse and Andrea Rotnitzky. "Estimation of the effect of interventions that modify the received treatment". In: *Statistics in Medicine* (2013).

[35] Trevor Hastie and Robert Tibshirani. *Generalized additive models*. Wiley Online Library, 1990.

[36] Nima S Hejazi, David Benkeser, and Mark J van der Laan. *haldensify: Highly adaptive lasso conditional density estimation*. R package version 0.2.3. 2022. DOI: 10.5281/zenodo.3698329. URL: https://github.com/nhejazi/haldensify.

[37] Nima S Hejazi and Iván Dıaz. *medshift: Causal mediation analysis for stochastic interventions*. R package version 0.1.4. 2020. URL: https://github.com/nhejazi/medshift.

[38] Nima S Hejazi et al. "Nonparametric causal mediation analysis for stochastic interventional (in)direct effects". In: *Biostatistics* (2022), pp. 1–22. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxac002. arXiv: 2009.06203.

[39] Nima S Hejazi et al. *tmle3mediate: Targeted Learning for Causal Mediation Analysis*. R package version 0.0.3. 2021. URL: https://github.com/tlverse/tmle3mediate.

[40] Nima S. Hejazi et al. "Efficient nonparametric inference on the effects of stochastic interventions under two-phase sampling, with applications to vaccine efficacy trials". In: *Biometrics* 77.4 (2021), pp. 1241–1253. ISSN: 15410420. DOI: 10.1111/biom.13375. arXiv: 2003.13771.

[41] Miguel A Hernan and James M Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2010.

[42] Nicholas J Horton and Ken P Kleinman. "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models". In: *The American Statistician* 61.1 (2007), pp. 79–90.

[43] Torsten Hothorn and Achim Zeileis. "partykit: A Modular Toolkit for Recursive Partytioning in R". In: *Journal of Machine Learning Research* 16 (2015), pp. 3905–3909. URL: https://jmlr.org/papers/v16/hothorn15a.html.

[44] Alan E. Hubbard, Sara Kherad-Pajouh, and Mark J. Van Der Laan. "Statistical Inference for Data Adaptive Target Parameters". In: *International Journal of Biostatistics* 12.1 (2016), pp. 3–19. ISSN: 15574679. DOI: 10.1515/ijb-2015-0013.

[45] "Individual and joint effects of metal exposure on metabolic syndrome among Chinese adults". In: *Chemosphere* 287.P3 (2022), p. 132295. ISSN: 18791298. DOI: 10.1016/j.chemosphere.2021.132295. URL: https://doi.org/10.1016/j.chemosphere.2021.132295.

[46] John P.A. Ioannidis. "How to Make More Published Research True". In: *PLoS Medicine* 11.10 (2014). ISSN: 15491676. DOI: 10.1371/journal.pmed.1001747.

[47] John P.A. Ioannidis. "Why most published research findings are false". In: *Getting to Good: Research Integrity in the Biomedical Sciences* 2.8 (2018), pp. 2–8. ISSN: 15491277. DOI: 10.1371/journal.pmed.0020124.

[48] Iván Díaz Muñoz and Mark van der Laan*. "Population Intervention Causal Effects Based on Stochastic Interventions". In: *Biometrics*. 68.2 (2012), pp. 541–549. ISSN: 15378276. DOI: 10.1111/j.1541-0420.2011.01685.x.Population. arXiv: NIHMS150003. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf.

[49] Tomoyuki Kawada. "RE: Evaluating additive versus interactive effects of copper and cadmium on life history". In: *Environmental Science and Pollution Research* 28.41 (2021), p. 58816. ISSN: 16147499. DOI: 10.1007/s11356-021-16371-3.

[50] Alexander P. Keil et al. "A quantile-based g-computation approach to addressing the effects of exposure mixtures". In: *arXiv* 128.April (2019), pp. 1–10. ISSN: 23318422. DOI: 10.1097/01.ee9.0000606120.58494.9d. arXiv: 1902.04200.

[51] Edward H Kennedy. "Nonparametric causal effects based on incremental propensity score interventions". In: *Journal of the American Statistical Association* (2018), pp. 1–12.

[52] Zhiyang Kong, Chunhong Liu, and Opeyemi Joshua Olatunji. "Asperuloside attenuates cadmium-induced toxicity by inhibiting oxidative stress, inflammation, fibrosis and apoptosis in rats". In: *Scientific Reports* 13.1 (2023), p. 5698. ISSN: 2045-2322. DOI: 10.1038/s41598-023-29504-0. URL: https://doi.org/10.1038/s41598-023-29504-0.

[53] Andreas Kortenkamp. "Ten years of mixing cocktails: A review of combination effects of endocrine-disrupting chemicals". In: *Environmental Health Perspectives* 115.SUPPL1 (2007), pp. 98–105. ISSN: 00916765. DOI: 10.1289/ehp.9357.

[54] Neha Kulkarni et al. "Reproducible bioinformatics project: A community for reproducible bioinformatics analysis pipelines". In: *BMC Bioinformatics* 19.Suppl 10 (2018). ISSN: 14712105. DOI: 10.1186/s12859-018-2296-x.

[55] Sören R Künzel et al. "Metalearners for estimating heterogeneous treatment effects using machine learning". In: *Proceedings of the national academy of sciences* 116.10 (2019), pp. 4156–4165.

[56] M.J. van der Laan and S. Rose. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer Series in Statistics. Springer International Publishing, 2018. ISBN: 9783319653044. URL: https://books.google.com/books?id=vKFTDwAAQBAJ.

[57] M.J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Series in Statistics. Springer New York, 2011. ISBN: 9781441997821. URL: https://books.google.com/books?id=RGnSX5aCAgQC.

[58] Mark J. van der Laan and Daniel Rubin. In: *The International Journal of Biostatistics* 2.1 (2006). DOI: doi:10.2202/1557-4679.1043. URL: https://doi.org/10.2202/1557-4679.1043.

[59] J. van der Laan Mark, Polley Eric C, and Hubbard Alan E. "Super Learner". In: *Statistical Applications in Genetics and Molecular Biology* 6.1 (2007), pp. 1–23. URL: https://EconPapers.repec.org/RePEc:bpj:sagmbi:v:6:y:2007:i:1:n:25.

[60] Xuefeng Lai et al. "Individual and joint associations of co-exposure to multiple plasma metals with telomere length among middle-aged and older Chinese in the Dongfeng-Tongji cohort". In: *Environmental Research* 214.P3 (2022), p. 114031. ISSN: 10960953. DOI: 10.1016/j.envres.2022.114031. URL: https://doi.org/10.1016/j.envres.2022.114031.

[61] Miguel Angel Luque-Fernandez et al. "Targeted maximum likelihood estimation for a binary treatment: A tutorial". In: *Statistics in Medicine* 37.16 (2018), pp. 2530–2546. ISSN: 10970258. DOI: 10.1002/sim.7628.

[62] Ian C. Marschner. "glm2: Fitting generalized linear models with convergence problems". In: *The R Journal* 3 (2011), pp. 12–15.

[63] David McCoy, Alan Hubbard, and Mark Van der Laan. "CVtreeMLE: Efficient Estimation of Mixed Exposures using Data Adaptive Decision Trees and Cross-Validated Targeted Maximum Likelihood Estimation in R". In: *Journal of Open Source Software* 8.82 (2023), p. 4181. DOI: 10.21105/joss.04181. URL: https://doi.org/10.21105/joss.04181.

[64] David Mccoy et al. "SuperNOVA: Semi-Parametric Identification and Estimation of Interaction and Effect Modification in Mixed Exposures using Stochastic Interventions in R". In: *Journal of Open Source Software* 0 (2023), pp. 1–4. ISSN: 2475-9066. arXiv: arXiv:2305.01849v1.

[65] Archana J. McEligot et al. "Logistic lasso regression for dietary intakes and breast cancer". In: *Nutrients* 12.9 (2020), pp. 1–14. ISSN: 20726643. DOI: 10.3390/nu12092652.

[66] Xia Meng et al. "Short term associations of ambient nitrogen dioxide with daily total, cardiovascular, and respiratory mortality: Multilocation analysis in 398 cities". In: *The BMJ* 372.2 (2021). ISSN: 17561833. DOI: 10.1136/bmj.n534.

[67] S. Milborrow. Derived from mda:mars by T. Hastie and R. Tibshirani. *earth: Multivariate Adaptive Regression Splines*. R package. 2011. URL: http://CRAN.R-project.org/package=earth.

[68] Susanna D Mitro et al. "Cross-sectional associations between exposure to persistent organic pollutants and leukocyte telomere length among US adults in NHANES, 2001–2002". In: *Environmental health perspectives* 124.5 (2016), pp. 651–658.

[69] "Molybdenum and cadmium co-induce oxidative stress and apoptosis through mitochondria-mediated pathway in duck renal tubular epithelial cells". In: *Journal of Hazardous Materials* 383.August 2019 (2020), p. 121157. ISSN: 18733336. DOI: 10.1016/j.jhazmat.2019.121157. URL: https://doi.org/10.1016/j.jhazmat.2019.121157.

[70] National Institute of Environmental Health Sciences (NIEHS). "2018-2023 Strategic Plan". In: *National Institutes of Health* (2018). URL: https://www.niehs.nih.gov/about/strategicplan/strategicplan20182023_508.pdf.

[71] Ana Navas-Acien et al. "Blood DNA Methylation and Incident Coronary Heart Disease: Evidence from the Strong Heart Study". In: *JAMA Cardiology* 6.11 (2021), pp. 1237–1246. ISSN: 23806591. DOI: 10.1001/jamacardio.2021.2704.

[72] Austin Nichols. "Causal Inference with Observational Data". In: *The Stata Journal* 7.4 (2007), pp. 507–541. DOI: 10.1177/1536867X0800700403.

[73] Joshua R Nugent and Laura B Balzer. "A demonstration of Modified Treatment Policies to evaluate shifts in mobility and COVID-19 case rates in U.S. counties". In: *American Journal of Epidemiology* (Jan. 2023). kwad005. ISSN: 0002-9262. DOI: 10.1093/aje/kwad005. eprint: https://academic.oup.com/aje/advance-article-pdf/doi/10.1093/aje/kwad005/48583062/kwad005.pdf. URL: https://doi.org/10.1093/aje/kwad005.

[74] JUDEA PEARL. "Causal diagrams for empirical research". In: *Biometrika* 82.4 (Dec. 1995), pp. 669–688. ISSN: 0006-3444. DOI: 10.1093/biomet/82.4.669. eprint: https://academic.oup.com/biomet/article-pdf/82/4/669/698263/82-4-669.pdf. URL: https://doi.org/10.1093/biomet/82.4.669.

[75] Judea Pearl. *Causal inference in statistics : a primer*. eng. Chichester, West Sussex: Wiley, 2016 - 2016. ISBN: 9781119186854.

[76] Judea Pearl. "Direct and Indirect Effects". In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. UAI'01. Seattle, Washington: Morgan Kaufmann Publishers Inc., 2001, pp. 411–420. ISBN: 1558608001.

[77] Jing Qin. "Inferences for case-control and semiparametric two-sample density ratio models". In: *Biometrika* 85.3 (1998), pp. 619–630.

[78] Quang Luu Quoc et al. "Contribution of monocyte and macrophage extracellular traps to neutrophilic airway inflammation in severe asthma". In: *Allergology International* xxxx (2023). ISSN: 14401592. DOI: 10.1016/j.alit.2023.06.004. URL: https://doi.org/10.1016/j.alit.2023.06.004.

[79] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL: https://www.R-project.org/.

[80] Irfan Rahman and I. M. Adcock. "Oxidative stress and redox regulation of lung inflammation in COPD". In: *European Respiratory Journal* 28.1 (2006), pp. 219–242. ISSN: 09031936. DOI: 10.1183/09031936.06.00053805.

[81] B. D. Ripley and W. Venables. *polspline: Polynomial Spline Routines*. R package version 1.1.26. 2021. URL: https://CRAN.R-project.org/package=polspline.

[82] Steven Roberts and Michael A. Martin. "Using supervised principal components analysis to assess multiple pollutant effects". In: *Environmental Health Perspectives* 114.12 (2006), pp. 1877–1882. ISSN: 00916765. DOI: 10.1289/ehp.9226.

[83] James M Robins. "A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect". In: *Mathematical Modelling* 7 (1986), pp. 1393–1512.

[84] James M Robins and Sander Greenland. "Identifiability and exchangeability for direct and indirect effects". In: *Epidemiology* 3.0 (1992), pp. 143–155.

[85] James M Robins, Miguel A Hernan, and Uwe Siebert. "Effects of multiple interventions". In: *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors* 1 (2004), pp. 2191–2230.

[86] James M Robins and Thomas S Richardson. "Alternative graphical causal models and the identification of direct effects". In: *Causality and psychopathology: Finding the determinants of disorders and their cures*. 2010, pp. 103–158.

[87] James M Robins et al. "New statistical approaches to semiparametric regression with application to air pollution research". In: *Research report (Health Effects Institute)* 175 (2013), pp. 3–129.

[88] James M. Robins and Sander Greenland. "Identifiability and Exchangeability for Direct and Indirect Effects". In: *Epidemiology* 3 (1992), pp. 143–155.

[89] Donald B Rubin. "Estimating Causal Effects of Treatments in Randomized & Nonrandomized Studies". In: *Journal of Educational Psychology* (1974). URL: http://www.eric.ed.gov/ERICWebPortal/detail?accno=EJ118470.

[90] Donald B. Rubin. "Estimating causal effects from large data sets using propensity scores". In: *Matched Sampling for Causal Effects* (2006), pp. 443–453. ISSN: 0003-4819. DOI: 10.1017/CBO9780511810725.035.

[91] S. Safe. "Toxicology, structure-function relationship, and human and environmental health impacts of polychlorinated biphenyls: Progress and problems". In: *Environmental Health Perspectives* 100 (1993), pp. 259–268. ISSN: 00916765. DOI: 10.1289/ehp.93100259.

[92] Anton Schick. "On asymptotically efficient estimation in semiparametric models". In: *The Annals of Statistics* (1986), pp. 1139–1151.

[93] Matthew J. Smith et al. "Introduction to computational causal inference using reproducible Stata, R, and Python code: A tutorial". In: *Statistics in Medicine* 41.2 (2022), pp. 407–432. ISSN: 10970258. DOI: 10.1002/sim.9234.

[94] J Steen et al. "medflex: An R Package for Flexible Mediation Analysis using Natural Effect Models". In: *Journal of Statistical Software* 76 (11 2017), pp. 1–46. DOI: 10.18637/jss.v076.i11.

[95] James H Stock. "Nonparametric policy analysis". In: *Journal of the American Statistical Association* 84.406 (1989), pp. 567–575.

[96] Eric J. Tchetgen and Ilya Shpitser. "Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness and sensitivity analysis". In: *Annals of Statistics* 40.3 (2012), pp. 1816–1845. ISSN: 00905364. DOI: 10.1214/12-AOS990.

[97] Dustin Tingley et al. "mediation: R Package for Causal Mediation Analysis". In: *Journal of Statistical Software* 59.5 (2014), pp. 1–38. URL: http://www.jstatsoft.org/v59/i05/.

[98] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. DOI: 10.1017/CBO9780511802256.

[99] Mark J van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011. DOI: 10.1007/978-1-4419-9782-1. URL: https://doi.org/10.1007/978-1-4419-9782-1.

[100] Tyler J. Vanderweele. "On the distinction between interaction and effect modification". In: *Epidemiology* 20.6 (2009), pp. 863–871. ISSN: 10443983. DOI: 10.1097/EDE.0b013e3181ba333c.

[101] Samantha L. VanEtten et al. "Telomeres as targets for the toxicity of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) and polychlorinated biphenyls (PCBs) in rats". In: *Toxicology and Applied Pharmacology* 408.October (2020), p. 115264. ISSN: 10960333. DOI: 10.1016/j.taap.2020.115264. URL: https://doi.org/10.1016/j.taap.2020.115264.

[102] Stijn Vansteelandt and Tyler J VanderWeele. "Natural direct and indirect effects on the exposed: effect decomposition under weaker assumptions". In: *Biometrics* 68.4 (2012), pp. 1019–1027.

[103] Daniela Vargas Vargas et al. "Metals and Metalloids in Asthma: A Role for Environmental Exposure?" In: *Frontiers in Pharmacology* 12 (2021), p. 643972. DOI: 10.3389/fphar.2021.643972. URL: https://www.frontiersin.org/articles/10.3389/fphar.2021.643972/full.

[104] Davis Vaughan and Matt Dancho. *furrr: Apply Mapping Functions in Parallel using Futures*. https://github.com/DavisVaughan/furrr, https://furrr.futureverse.org/. 2022.

[105] Stefan Wager and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1228–1242.

[106] Bin Wang et al. "Exposure to acrylamide and reduced heart rate variability: The mediating role of transforming growth factor-β". In: *Journal of Hazardous Materials* 395.March (2020). ISSN: 18733336. DOI: 10.1016/j.jhazmat.2020.122677.

[107] T. Whyand et al. "Pollution and respiratory disease: Can diet or supplements help? A review". In: *Respiratory Research* 19.1 (2018), pp. 1–14. ISSN: 1465993X. DOI: 10.1186/s12931-018-0785-0.

[108] Nicholas Williams and Iván Díaz. *lmtp: Non-parametric Causal Effects of Feasible Interventions Based on Modified Treatment Policies*. R package version 1.3.1. 2020. DOI: 10.5281/zenodo.3874931. URL: https://github.com/nt-williams/lmtp.

[109] Greg Wilson et al. "Good enough practices in scientific computing". In: *PLoS computational biology* 13.6 (2017).

[110] Christopher Winship and Stephen L. Morgan. "The estimation of causal effects from observational data". In: *Annual Review of Sociology* 25.May (1999), pp. 659–707. ISSN: 03600572. DOI: 10.1146/annurev.soc.25.1.659.

[111] Marvin N. Wright and Andreas Ziegler. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software* 77.1 (2017), pp. 1–17. DOI: 10.18637/jss.v077.i01.

[112] Sewall Wright. "The Method of Path Coefficients". In: *The Annals of Mathematical Statistics* 5.3 (1934), pp. 161–215. DOI: 10.1214/aoms/1177732676. URL: https://doi.org/10.1214/aoms/1177732676.

[113] Keh Gong Wu et al. "Associations between environmental heavy metal exposure and childhood asthma: A population-based study". In: *Journal of Microbiology, Immunology and Infection* 52.2 (2019), pp. 352–362. ISSN: 19959133. DOI: 10.1016/j.jmii.2018.08.001. URL: https://doi.org/10.1016/j.jmii.2018.08.001.

[114] Fang Xia et al. "Association between urinary metals and leukocyte telomere length involving an artificial neural network prediction: Findings based on NHANES 1999–2002". In: *Frontiers in Public Health* 10 (2022). ISSN: 22962565. DOI: 10.3389/fpubh.2022.963138.

[115] Lili Xiao et al. "Cadmium exposure, fasting blood glucose changes, and type 2 diabetes mellitus: A longitudinal prospective study in China". In: *Environmental Research* 192.June 2020 (2021), p. 110259. ISSN: 10960953. DOI: 10.1016/j.envres.2020.110259. URL: https://doi.org/10.1016/j.envres.2020.110259.

[116] Xin Wang, Bhramar Mukherjee and Sung Kyun Park. "Associations of Cumulative Exposure to Heavy Metal Mixtures with Obesity and Its Comorbidities Among U.S. Adults in NHANES 2003-2014". In: *Environment International* 121.Pt 1 (2018), pp. 683–694. DOI: 10.1016/j.envint.2018.09.035.Associations.

[117] Tao Xu et al. "Associations of urinary carbon disulfide metabolite with oxidative stress, plasma glucose and risk of diabetes among urban adults in China". In: *Environmental Pollution* 272 (2021), p. 115959. ISSN: 18736424. DOI: 10.1016/j.envpol.2020.115959. URL: https://doi.org/10.1016/j.envpol.2020.115959.

[118] Maayan Yitshak-Sade et al. "Estimating the combined effects of natural and built environmental exposures on birthweight among urban residents in massachusetts". In: *International Journal of Environmental Research and Public Health* 17.23 (2020), pp. 1–16. ISSN: 16604601. DOI: 10.3390/ijerph17238805.

[119] Linling Yu et al. "A review of practical statistical methods used in epidemiological studies to estimate the health effects of multi-pollutant mixture". In: *Environmental Pollution* 306.January (2022). ISSN: 18736424. DOI: 10.1016/j.envpol.2022.119356.

[120] Yuqing Zhang et al. "Association between exposure to a mixture of phenols, pesticides, and phthalates and obesity: Comparison of three statistical models". In: *Environment International* 123.July 2018 (2019), pp. 325–336. ISSN: 18736750. DOI: 10.1016/j.envint.2018.11.076.

[121] Wenjing Zheng and MJ van der Laan. "Asymptotic theory for cross-validated targeted maximum likelihood estimation". In: *U.C. Berkeley Division of Biostatistics Working Paper Series* 273 (2010). URL: http://biostats.bepress.com/ucbbiostat/paper273/.

[122] Wenjing Zheng and Mark J. Van Der Laan. "Targeted maximum likelihood estimation of natural direct effects". In: *International Journal of Biostatistics* 8.1 (2012). ISSN: 15574679. DOI: 10.2202/1557-4679.1361.

[123] Yun Zhou et al. "Urinary polycyclic aromatic hydrocarbon metabolites and altered lung function in Wuhan, China". In: *American Journal of Respiratory and Critical Care Medicine* 193.8 (2016), pp. 835–846. ISSN: 15354970. DOI: 10.1164/rccm.201412-2279OC.

[124] George Zipf et al. "National health and nutrition examination survey: Plan and operations, 1999–2010". In: *Vital and health statistics. Series 1, Programs and collection procedures* 56 (2013), pp. 1–37.

[125] Paul N. Zivich and Alexander Breskin. "Machine Learning for Causal Inference: On the Use of Cross-fit Estimators". In: *Epidemiology* 32.3 (2021), pp. 393–401. ISSN: 15315487. DOI: 10.1097/EDE.0000000000001332. arXiv: 2004.10337.

[126] Ami R. Zota et al. "Associations of cadmium and lead exposure with leukocyte telomere length: Findings from National Health And Nutrition Examination Survey, 1999-2002". In: *American Journal of Epidemiology* 181.2 (2015), pp. 127–136. ISSN: 14766256. DOI: 10.1093/aje/kwu293.