Simulations

Past simulation studies

Here, we preface our simulation study with a brief overview of examples in the literature which compare the efficacy of various methods using simulations. @taylor_statistical_2016 conclude that, in general for exposure mixture studies, no single method consistently outperforms others across all situations and, importantly, that a method should be chosen based on the question of interest. Thus, for each study, we highlight not only the findings, but also the data-generating scenarios and the identified question of interest.

@lazarevic_performance_2020 compare the performance of a broad range of methods for accurate variable selection of important exposures. They simulated exposure data using a multivariate *t*-copula based on real-world data and the response by specifying a regression relationship with only a subset of truly significant exposures and a normal error term. Two correlation structures were considered — one with the original Spearman correlation matrix and one with the values halved — as well as two signal-to-noise ratios — one with an R^2 for the true model at 10% and one at 30%. They found that BKMR, along with three other flexible regression methods that allow for nonlinearity, provided more accurate variable selection results compared to two machine learning methods. Moreover, they observed that, in general, low signal-to-noise ratios had a stronger impact on performance than did increasing multicollinearity.

@hoskovec_model_2021 compare Bayesian methods, including BKMR, while considering 4 research questions: accurate estimation, selection of important exposures, exclusion of unimportant exposures, and identification of interactions. They use observed exposure and covariate data to simulate response data using regression relationships; they considered three exposure-response scenarios of varying complexity and included two-way multiplicative interaction terms. For each simulated dataset, they randomly assigned exposures to be active components of the mixture to incorporate variability in the data. Overall, they found that Bayesian methods outperformed traditional linear regressions, and that BKMR performed best when the exposure-response function takes on a complex form.

Most recently, @pesenti_comparative_2023 compare BKMR, BSR, and the Bayesian Least Absolute Shrinkage and Selection Operator (LASSO) for variable selection. Data were generated using a multivariate normal with moderate and strong correlation structures specified manually by the researchers. They found that, in situations with additivity and linearity, Bayesian LASSO was appropriate. Across the other scenarios, BKMR generally performed best, while BSR selected exposures with high heterogeneity when the sample size was smaller due to the influence of the degrees of freedom, d, tuning parameter. Notably, multicollinearity did not generally lead to spurious variable selection.

Finally, we briefly comment on studies by @sun_statistical_2013 and @barrera-gomez_systematic_2017, whose explicit goal is to compare methods for identifying interactions. Both studies generate exposure data using the correlation structure from an existing dataset; @sun_statistical_2013 uses a multivariate lognormal, while @barrera-gomez_systematic_2017 uses a multivariate normal. Both only consider two-way, multiplicative interactions. While neither of these studies consider the methods used in this thesis, they find that, in general, models that formally allow for interaction effects perform better than models that only allow for univariate additive effects.

Methods

The goal of our simulation study is to provide guidance on the choice between BSR and BKMR for characterizing a diverse range of complex interactions between predictors. In particular, we aim to extend findings from previous simulation studies by considering a more comprehensive range of interaction types, including different effect sizes, non-multiplicative interactions, and three-way interactions. We also explore interactions between exposures and categorical covariates, a previously understudied form of interaction in exposure mixture studies.

MADRES data

In order to make our simulations comparable to real-world exposure mixture studies, we based our simulation data on the Maternal And Developmental Risks from Environmental and Social Stressors (MADRES) pregnancy cohort. The MADRES cohort is an ongoing, prospective pregnancy cohort of predominantly lower-income, Hispanic women in Los Angeles, California, which began in 2015 [@bastain_study_2019]. Urine samples were collected by participants at their first visit, and questionnaires were administered during their first visit, with follow-ups at the first, second, and third trimesters. See @bastain_study_2019 for further details on study design.

@howe_prenatal_2020 previously examined the effect of prenatal metal mixtures of birth weight (BW) for gestational age (GA) in this cohort. They used BKMR to identify associations between metal mixtures and BW for GA, as well as BSR to conduct inference on interactions between metals. Briefly, using BKMR, they found that, of the metals in the mixture, Hg and Ni were most strongly associated with BW for GA. Moreover, BKMR results suggested that a potential interaction between Hg and Ni; however, when run through BSR, the PIP for this interaction was extremely small, despite being the highest of all two-way interactions.

Data from the study by @howe_prenatal_2020 were obtained from publicly available data in the Human Health Exposure Resource (HHEAR) Data Repository, which has been approved under Icahn School of Medicine at Mount Sinai IRB Protocol #16-00947. The Digital Object Identifiers associated with the urinary trace element data and epidemiological data are 10.36043/1945_159 and 10.36043/1945_177, respectively.

We followed the approach by @howe_prenatal_2020 for preparing the data for analysis. This resulted in retaining 10 metals in analysis: arsenic (As), cadmium (Cd), cobalt (Co), mercury (Hg), nickel (Ni), molybdenum (Mo), lead (Pb), antimony (Sb), tin (Sn), and thallium (Tl). @howe_prenatal_2020 used speciated As, but this was not available in HHEAR, so we used total As. Metals were expressed in μg/g and natural log transformed to reduce right-skewness and then standardized to keep values scale-free. Among the full range of covariates considered by @howe_prenatal_2020, we used the subset of 4 that were available in HHEAR: any smoke exposure during pregnancy, maternal prepregnancy body mass index (BMI), maternal age during firt trimester, and maternal race by ethnicity and birth place. Race by ethnicity and birth place was collapsed into the following categories: non-Hispanic white, non-Hispanic black, non-Hispanic other, Hispanic born in the US, and Hispanic born outside the US. We observed 8 missing values for BMI in the data from HHEAR, which were not reported by @howe_prenatal_2020. We mean imputed these missing values. Our final analytic dataset included 252 participants, which was 10 fewer than in @howe_prenatal_2020, likely due to small discrepancies in their dataset and the one made available in HHEAR.

- The data used for simulations in this thesis was supported by the National Institute of Environmental Health Sciences of the National Institutes of Health under Award Numbers U2CES026555 and U2CES026553. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institutes of Health.

Simulating predictor data

define predictor = both exposure and covariate

We simulated exposure and covariate data using a multivariate t-copula fit on the 252 participants in the MADRES cohort. We used copulas as they can preserve both the correlation structure and marginal distributions from the observed data, allowing us to replicate conditions in a real-world scenario.

First, we briefly introduce copulas in the context of their use in this simulation, based on the presentation in @nelsen_introduction_2006. Copulas are joint cumulative distribution functions (CDFs) defined on the unit cube $[0,1]^n$ that capture the dependence between n uniformly distributed marginals. Sklar's theorem allows us to apply copulas to our observed data. Sklar's theorem states that, if $H(x_1, \dots x_n)$ is a joint CDF of the marginal CDFs $F_1(x_1), \dots, F_n(x_n)$, then there exists a copula C such that, for all $(x_1, \dots, x_n)$ in $(X_1, \dots, X_n)$,

$$H(x_1, \dots x_n) = C(F_1(x_1), \dots, F_n(x_n)).$$

Note that, by the probability integral transform, or the universality of the uniform, the CDFs $F_1(x_1), \dots, F_n(x_n)$ are distributed uniformly.

We have two categorical covariates in our data: smoke exposure and race by ethnicity and birthplace.

In our case, we transformed the observed predictor values to the uniform distribution based on their empirical marginal CDFs, a process called generating "pseudo-random" samples. Then, we fit the copula and back-transformed the data to its original marginal structure using the empirical marginal CDFs. The Spearman's rho correlation structure remains consistent through these transformations, as the ranking of the data is not affected.

Various families of copulas have been described, each of which specifies a different shape for the dependence structure. We fit the set of multivariate copulas used by @lazarevic_performance_2020 in their simulation study, which included the Gaussian, t, Gumbel, Frank, Clayton, and Joe copulas. We selected the ___ copula based on _____.

show correlation and marginal distribution of original and simulations

Simulation predictor-response relationships

- simulate outcome using a formula w/ different types of interaction
- change effect size (three levels? — NOTE, this is diff from signal:noise ratio)
- change nature of interaction (mathematical formulation, two- or three-way, b/t just exposures or b/t exposures and covariates)
- take into account collinear structure (interactions b/t correlated exposures or uncorrelated exposures?)
- change sample size to inform study design

Base case:

$$\begin{aligned} Y_i = \mathrm{Hg}_i + \frac{3}{1+\exp(-4\mathrm{Ni}_i)} + \frac{1.5}{1+\exp(-4\mathrm{Sn}_i)} - \mathrm{Sb}_i^2 + 0.5\mathrm{Sb}_i \\ + \mathrm{age} + 0.5\mathrm{bmi} - \mathrm{race}_{\mathrm{oth}} - \mathrm{race}_{\mathrm{hisp.us}} - 1.5\mathrm{race}_{\mathrm{hisp.non}} - \mathrm{smoke} + \varepsilon_i, \end{aligned}$$

This resulted in a total of 41 scenarios.

- see appendix for surfaces

Models

Software: @bobb_statistical_2018 on CRAN, @antonelli_estimating_2020 on GitHub

Models compared, specify the parameters for each (justify them!)

- MLR
- MLR with known form of interactions specified (oracle method)
- BKMR with component-wise
- BSR

check convergence with trace plots

Table 1: Specification of interaction terms in simulations.

	Effect size	
	Small	Large
Univariately significant		
Multiplicative	0.3Hg$*$Ni	0.6Hg$*$Ni
Polynomial	0.1Hg$*$(Ni$-$1)^2	0.2Hg$*$(Ni$-$1)^2
Univariately insignificant		
Multiplicative	0.3Cd$*$As	0.6Cd$*$As
Polynomial	0.1Cd$*$(As$-$1)^2	0.2Cd$*$(As$-$1)^2
Highly correlated		
Multiplicative	0.3Hg$*$Co	0.6Hg$*$Co
Polynomial	0.1Hg$*$(Co$-$1)^2	0.2Hg$*$(Co$-$1)^2
Three-way interaction		
Multiplicative	0.3Hg$*$Ni$*$Tl	0.6Hg$*$Ni$*$Tl
Polynomial	0.1Hg$*$(Ni$-$1)^2$*$Tl	0.2Hg$*$(Ni$-$1)^2$*$Tl

Model assessment

- use median probability model threshold — marginal PIP of at least 0.5
- how many times is interaction picked up?
  - sensitivity and false discovery rate
- testing MSE
- potentially explore mpower package

Results

- example output from representative model
- figures + tables w/ model performance