# Simulations

## Past simulation studies

@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. Here, we preface our simulation study with a brief overview of examples in the literature which compare the efficacy of various methods using simulations. 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 the findings from @pesenti_comparative_2023 by considering different forms of interactions, including interactions generated with non-multiplicative interaction terms, with different effect sizes. We also explore interactions between exposures and categorical covariates, a previously understudied goal of exposure mixture studies with important public health implications.

Table 1: Specification of interaction terms in simulation

| | Effect size | |
| --- | --- | --- |
| | Small | Large |
| **Univariately significant** | | |
| Multiplicative | $0.5\text{Hg}*\text{Ni}$ | $\text{Hg}*\text{Ni}$ |
| Polynomial | $\text{Hg}*(\text{Ni}-1)^2$ | $2\text{Hg}*(\text{Ni}-1)^2$ |
| **Univariately insignificant** | | |
| Multiplicative | $0.5\text{Cd}*\text{As}$ | $\text{Cd}*\text{As}$ |
| Polynomial | $\text{Cd}*(\text{As}-1)^2$ | $2\text{Cd}*(\text{As}-1)^2$ |
| **Highly correlated** | | |
| Multiplicative | $0.5\text{Hg}*\text{Co}$ | $\text{Hg}*\text{Co}$ |
| Polynomial | $\text{Hg}*(\text{Co}-1)^2$ | $2\text{Hg}*(\text{Co}-1)^2$ |
| **Three-way interaction** | | |
| Multiplicative | $0.5\text{Hg}*\text{Ni}*\text{Tl}$ | $\text{Hg}*\text{Ni}*\text{Tl}$ |
| Polynomial | $\text{Hg}*(\text{Ni}-1)^2*\text{Tl}$ | $2\text{Hg}*(\text{Ni}-1)^2*\text{Tl}$ |

**MADRES data**

- @bastain_study_2019 original study design paper
- @howe_prenatal_2020 applied BKMR and BSR to this data, what did they find?

In order to make our simulations comparable to real-world exposure mixture studies, we based our simulation data on the cohort in the MADRES study.

We selected this study because @howe_prenatal_2020

**Simulating data**

- use original predictor values in MADRES data. allows us to explore these methods on predictors with realistic correlation structure OR use original correlation structure + multivariate t-copula
- 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:

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

- 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

**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