Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Evaluating multi-pollutant methods for PFAS mixture effects on immunometabolic health

Jonas Meijerink

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Quantitative Epidemiology

SUPERVISOR :
Prof. dr. Christel FAES

SUPERVISOR :
dr. Bianca COX

2024
2025

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Evaluating multi-pollutant methods for PFAS mixture effects on immunometabolic health

Jonas Meijerink

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Quantitative Epidemiology

SUPERVISOR :

Prof. dr. Christel FAES

SUPERVISOR :

dr. Bianca COX

Acknowledgements

First and foremost, thanks to everyone who didn't run away when I started talking about my master's thesis... I am deeply thankful to my internal supervisor Prof. dr. Christel Faes, for her calm guidance and continued belief even in uncertain periods. My external supervisor, Dr. Bianca Cox, deserves special thanks for answering all my questions, sharing valuable data and insights from VITO (Flemish Institute for Technology Research). A heartfelt thank you to my friends, Lode, Lore, and Lynn, for their support and companionship throughout these five years of studying together. Many thanks to my Dad, Sara and Sarah for their input and suggestions on improving the manuscript. Finally, I'm grateful to my family for all their support. After all, this thesis didn't write itself!

Contents

Abstract

Background: Humans are constantly exposed to various chemicals. The impact of these pollutants on human health is often unknown or poorly characterised. This has led to a growing interest in the field of environmental epidemiology. This master's thesis investigates statistical methods for correlated environmental exposures, with a particular emphasis on per- and polyfluoroalkyl substances (PFAS) in human biomonitoring studies. We address two main questions: (1) how to identify individual pollutants associated with a health outcome and (2) how to estimate the joint effect of a pollutant mixture.

Methods: The statistical methods considered include frequentist shrinkage approaches (Ridge, LASSO, Elastic Net), Bayesian shrinkage models using different priors (Laplace, spike and slab, horseshoe), Bayesian model averaging with Bayesian adaptive sampling and G-computation. Additionally, we considered methods specifically developed for mixture analysis, such as weighted quantile sum (WQS) regression and Bayesian kernel machine regression (BKMR). The methods are evaluated through a literature review, along with simulation studies and a case study.

Results: For individual effect estimation, we found that multicollinearity poses the main challenge. Ridge regression was best suited in terms of power due to its grouping property and ability to handle multicollinearity. Bayesian shrinkage regression offered improved interpretability via posterior distributions and reduced bias, though at the cost of wider credible intervals. For WQS regression, we found that using standardised continuous exposures improves statistical power without compromising robustness. We demonstrated that highly collinear exposures have a reduced impact on joint effect standard errors from ordinary least squares (OLS) due to variance cancellation.

Conclusion: For individual effect estimation, Bayesian shrinkage methods have shown promise due to their ability to quantify uncertainty, minimal shrinkage of relevant coefficients and interpretability. Additionally, these methods provide a flexible framework that can be extended to accommodate different outcome distributions, clustering, spatial dependencies and non-linear associations. For joint effect estimation, traditional OLS regression performed well. The more flexible models may demonstrate improved performance in larger sample sizes and indicate potential areas for future research.

Keywords: human biomonitoring, PFAS exposure, multi-pollutant methods, shrinkage methods, weighted quantile sum regression, Bayesian kernel machine regression, individual and joint effect estimation

List of Figures

List of Tables

List of abbreviations and frequently used symbols

Abbreviations

ADEMP	Aims Data-generating mechanism Estimand Methods Performance measures
BART	Bayesian additive regression trees
BAS	Bayesian adaptive sampling
BIP	Bootstrap inclusion probability
BMA	Bayesian model averaging
BKMR	Bayesian kernel machine regression
bs	Bootstrap
CI	Confidence or credible interval
GLM	Generalised linear model
HBM	Human biomonitoring
HBM4EU	Human Biomonitoring for Europe
LASSO	Least absolute shrinkage and selection operator
LOQ	Limit of quantification
MCMC	Markov chain Monte carlo
MLE	Maximum likelihood estimation
MSE	Mean squared error
NOEC	No observed effect concentration
OLS	Ordinary least squares
PARC	Partnership for the Assessment of Risks from Chemicals
PFAS	Per- and polyfluoroalkyl substances
PFBA	Linear perfluorobutanoic acid
PFDA	Linear perfluorodecanoic acid
PFHXS (total)	Linear + branched perfluorohexanesulfonic acid
PFNA	Linear perfluorononanoic acid
PFOA (total)	Linear + branched perfluorooctanoic acid
PFOS	Linear perfluorooctanesulfonic acid
PFOS (branched)	Branched perfluorooctanesulfonic acid
PIP	Posterior inclusion probability
PLS	Partial least squares
Q1	25th percentile
Q2	50th percentile
Q3	75th percentile
rh	Repeated holdout

rs	Random subset
RSS	Residual sum of squares
SE	Standard error
UPLC-MS/MS	Ultra-performance liquid chromatography coupled to tandem mass spectrometry
VITO	Flemish Institute for Technology Research
WQS	Weighted quantile sum
WQS_{bs}	Bootstrap weighted quantile sum
WQS_{rs}	Random subset weighted quantile sum
WQS_{rh}	Repeated holdout weighted quantile sum

Frequently used symbols

n	Sample size
p	Total number of predictors ($p \equiv c + z$)
c	Number of exposures
z	Number of additional covariates
$\mathbf{X}$	Design matrix
$\mathbf{A}$, $\boldsymbol{a}$	Exposure matrix or vector
$\mathbf{Z}$, $\boldsymbol{z}$	Additional covariate matrix or vector
$\mathbf{Y}$, Y	Random outcome vector or scalar
$\boldsymbol{\beta}$	Unknown coefficient vector
$\boldsymbol{\epsilon}, \epsilon$	Random error vector or scalar
σ^2	Variance
$\boldsymbol{I}$	Unit matrix
l	Eigenvalues
$(\cdot)^\top$	Transpose
$\mathbb{E}(\cdot)$	Expectation
$\mathrm{Tr}(\cdot)$	Trace
$\mathrm{Var}(\cdot)$	Variance
$\mathrm{Cov}(\cdot)$	Covariance
$\tau, \lambda, \pi, r, \delta$	Shrinkage or variable selection hyperparameters
ϵ_0^2, c_0^2	Spike variance, slab variance
$\mathbf{q}, q$	Quantile vector or scalar (e.g., quartiles $0, 1, 2, ...$)
$g(\cdot)$	Link function in generalised linear model
β_0	Intercept
β_1	Regression coefficient for the weighted quantile sum
$\boldsymbol{w}, w$	Weight vector or scalar in weighted quantile sum
$\boldsymbol{\varphi}$	Unknown coefficient vector of the additional covariates
wqs	Weighted quantile sum scalar
wqs_{bs}	Bootstrap weighted quantile sum scalar
B	Number of bootstrap samples
B^*	Number of bootstrap samples following directional homogeneity
d	Number of randomly selected exposures in random subset WQS regression
S	Total variable subsets in random subset WQS regression
R	Total partitions in repeated holdout WQS regression
$\hat{Y}^{\boldsymbol{a}}$	Predicted counterfactual outcome
Δ	Quantity or estimator of interest

$\boldsymbol{\gamma}$	Indicator vector
Γ	Model space
$\mathcal{M}_{\boldsymbol{\gamma}}$	Specific model configuration by $\boldsymbol{\gamma}$
$\boldsymbol{\beta}_{\boldsymbol{\gamma}}$	Unknown coefficient vector by $\boldsymbol{\gamma}$
$\mathbf{X}_{\boldsymbol{\gamma}}$	Design matrix formed by columns of $\mathbf{X}$ based on $\boldsymbol{\gamma}$
$h(\cdot)$	High-dimensional exposure-response function

Chapter 1

Introduction

1.1 Background

Our environment has been severely polluted by many chemicals, creating a lasting legacy of contamination in our air, water and soil. Furthermore, many of these pollutants are also used in personal care products or household consumer items. As a result, these chemicals can enter the human body via inhalation, dermal absorption or ingestion (HBM4EU, 2021). The effects of these chemicals on human health are often unknown or poorly characterised. As a result, there has been a growing interest in the field of environmental epidemiology, dealing with these chemicals. This shift has led to changes in statistical methodology to address the complexities of chemical pollutants and answering specific research questions.

In the past, researchers have typically focused on studying the effects of a single chemical in relation to a specific health outcome. They conducted statistical analyses using classical methods with one exposure. This approach is often referred to as a *single-pollutant* method. Silva et al. (2002) argued that this approach can lead to significant underestimation of risk. They demonstrated that estrogenic agents can interact together to generate significant effects even when combined at concentrations below their NOEC (no observed effect concentration). Many researchers argue for switching to the analysis of chemical mixtures, also called *multi-pollutant* methods (Dominici et al., 2010). This refers to multiple chemicals or exposures that may interact additively, synergistically or antagonistically (Braun et al., 2016). In contrast to *single-pollutant* models, these methodologies also take into account the potential confounding effects posed by other chemicals included in the mixture. Therefore, this approach provides a richer understanding of how different pollutants affect a health outcome.

A chemical mixture presents various challenges. The main statistical issue is the potential of multicollinearity due to high correlations among chemicals in the mixture. Many traditional statistical methods suffer from this, as high correlation between exposures leads to inflated standard errors. Therefore, it is essential to design a statistical method that effectively addresses this problem. Apart from dealing with multicollinearity, epidemiologists want to use these models to answer specific research questions. Braun et al. (2016) described three broad questions related to chemical mixtures:

- What are the health effects of individual chemicals within a mixture?
- What are the interactions between chemicals within a mixture?

- What is the health effect of cumulative chemical exposure?

Given the challenges associated with multicollinearity, along with the presence of three key questions, statisticians and epidemiologists have developed various novel statistical methodologies. This thesis will describe some of the most commonly used *multi-pollutant* methods found in the literature. The objective is to evaluate the assumptions, robustness, and interpretability associated with the specified methods. This evaluation will be based on a particular case study provided by VITO (Flemish Institute for Technological Research).

1.2 Societal relevance and stakeholder awareness

As a science-to-technology partner, VITO supports companies, governments and society in their sustainability transition. VITO Health conducts research to understand the harmful effects of environmental factors on human health. By analysing data, VITO provides insights into these impacts at both individual and population levels. Moreover, they offer targeted policy advice on substances of concern (VITO, 2025).

VITO Health is a crucial partner in the PARC (Partnership for the Assessment of Risks from Chemicals) seven-year initiative under Horizon Europe, FLEHS (Flemish Environment and Health Study) and has previously served as co-coordinator for HMB4EU (Human Biomonitoring for Europe) (Marx-Stoelting et al., 2023; Gilles et al., 2022). Significant efforts have been made during these projects to develop statistical guidelines addressing general issues related to human biomonitoring data. The application of *multi-pollutant* methods within HBM4EU has been limited. The objective is that these methods become the standard approach for exposure–effect analyses in future human biomonitoring studies within PARC. VITO Health acknowledges that a deeper exploration is needed to fully understand their implications and effectiveness. As a result of their keen interest and active involvement, VITO has emerged as the principal requesting party for this master's thesis. Their involvement reinforces the importance of collaboration, highlighting VITO's commitment to advancing knowledge and innovation within the field.

Further insights into these methodologies are crucial for VITO, given their implications for public health. A better understanding of these methods will yield more robust findings and facilitate clearer communication with the general population. The findings should guide safer chemical policies and increase public awareness. The methods presented are applicable to a diverse range of pollutants, including per- and polyfluoroalkyl substances (PFAS), phthalates, atmospheric contaminants (air pollutants), heavy metals and pesticides. In this manner, a better understanding of advanced statistical techniques benefits research institutes as VITO, companies, governments and society as a whole.

1.3 Research question

This master thesis will address the following questions:

- How do different statistical methods estimate the overall (joint) effect of PFAS mixtures on immune-related health outcomes?
- How do these methods identify key PFAS compounds within the mixture and estimate their individual contributions to the health outcome?
- What are the assumptions, strengths and limitations of these statistical methods when applied to PFAS mixture data and how do they compare in terms of performance and interpretability?

The questions will be addressed through a literature review, a simulation study and a case study provided by VITO. Moreover, we are encouraged to adopt a broader perspective than the methodology employed by VITO and to critically evaluate certain assumptions that are frequently presented in the literature.

1.4 Ethical considerations

This master's thesis did not involve any human participation, nor did it address clinical issues or involve the collection of personal data. Internal data provided by VITO was utilised as a case study throughout this thesis. This case study titled “Teenager study HBM - 3M site” was approved by the *Committee for Medical Ethics at UZA/UAntwerpen.* The personal data acquired from the human biomonitoring study was received in an anonymous format. It should be noted that in a statistical context, case studies are not intended to yield generalisable results. They aim to demonstrate how a given approach can be implemented, interpreted and adapted to address similar problems, rather than providing answers to specific research questions with certainty. Furthermore, this thesis underscores the scientific integrity and transparency, in alignment with the principles set forth by Hasselt University and VITO. In accordance with these principles, AI-based tools were employed for specific tasks. Details of the specific AI tools and tasks are provided in Appendix A.

1.5 Structure of the thesis

The remainder of this thesis is organised as follows. Chapter 2 introduces human biomonitoring studies and describes the dataset. Chapter 3 addresses the challenge of multicollinearity and introduces both frequentist and Bayesian shrinkage methods. In Chapter 4, we explore methods particularly used for mixture analysis, such as weighted quantile sum (WQS) regression and Bayesian kernel machine regression (BKMR). Additionally, we explore more flexible techniques for estimating mixture effects using G-computation. Chapter 5 presents a simulation study comparing these methods in terms of individual and joint effect estimation. It also addresses the WQS quantile choice and joint effect estimation using ordinary least squares (OLS). Finally, Chapter 6 details the results of the case study, while the overall findings of this thesis are discussed in Chapter 7 and a conclusion is given in Chapter 8.

Chapter 2

Data

2.1 Human biomonitoring studies

In our daily lives, individuals are exposed to a variety of chemicals. Furthermore, our environment has been severely polluted by these contaminants. Consequently, these chemicals can enter the human body through inhalation, dermal absorption or ingestion (HBM4EU, 2021). To evaluate potential risks, it is crucial to gain a deep understanding of the various forms of exposure and the adverse effects they may pose to our health. Human biomonitoring (HBM) is a methodology utilised to assess the concentrations of chemicals present within the human body. This is accomplished through the collection of biological specimens such as blood, urine or breast milk. These specimens reflect the multiple pathways that individuals are exposed to, such as diet, consumer products or environment (Gilles et al., 2022). The collected specimens are analysed in a laboratory to determine the exact concentrations of various pollutants. These results are then studied using statistical techniques to uncover the complex relationships between these pollutants and effect biomarkers (e.g., immune parameters) or potential health outcomes (e.g., asthma or allergies).

2.2 Case study

This master's thesis is based on VITO's research regarding human biomonitoring studies. Specifically, it focuses on the exposure of humans to PFAS (per- and polyfluoroalkyl substances) and the associated immunometabolic health effects in teenagers. PFAS refers to a large group ($>$ 6 000) of human-made chemicals, encompassing a wide range of molecular sizes from small to very large. The substances are defined by their strong bonds between carbon and fluorine atoms. These bonds make PFAS highly resistant to degradation, which has resulted in their classification as "forever chemicals". Since the 1950s, PFAS have been used worldwide to make consumer products resistant to water, oil, grease and prevent staining (HBM4EU, 2021).

VITO has conducted multiple human biomonitoring (HBM) studies in the last 20 years. This thesis will use the data from the "Teenager HBM Study - 3M site" (Consortium UAntwerpen, VITO, PIH, UHasselt and VUB, 2023). 3M has been a major producer of PFAS in the past. As PFAS is persistent, over the years, the environment surrounding the factory has become significantly contaminated. One of the main objectives of this study was to investigate the environmental health implications for young people living near the

3M production site. The dataset consists of a sample size of 303 teenagers living within a radius of 5km from the 3M site. A full description of the study population, collection of samples, questionnaires and measurements of PFAS in blood can be found in Consortium UAntwerpen, VITO, PIH, UHasselt and VUB (2023). The original dataset includes diverse information, but this thesis will primarily focus on the measured concentrations of PFAS in relation to immune-related effect biomarkers.

Figure 2.1 gives an overview of all PFAS compounds that were measured in the participants' blood serum and the percentage that was above the limit of quantification (LOQ). This limit refers to the lowest concentration at which the exposure can be reliably quantified. This is often referred to as left-censoring of the exposure, which is a statistical research field on its own. Since this thesis does not focus on complex methods for handling the LOQ, a single random imputation from a censored log-normal distribution was applied by VITO Health to variables with a detection rate of at least 60%. For exposure-effect analysis, variables below 60% are often dichotomised ($<$ LOQ versus $\geq$LOQ). These recommendations were specified in the statistical analysis plan for the PARC project (Hassen et al., 2023). The dichotomised exposures were excluded in this thesis, allowing for a focus on continuous exposures as required by some of the *multi-pollutant* methods. The following exposures are used in the case study in Chapter 6: linear perfluorobutanoic acid (PFBA), linear + branched perfluorooctanoic acid (PFOA total), linear perfluorononanoic acid (PFNA), linear perfluorodecanoic acid (PFDA), linear + branched perfluorohexanesulfonic acid (PFHXS total), linear perfluorooctanesulfonic acid (PFOS) and branched perfluorooctanesulfonic acid (PFOS branched).

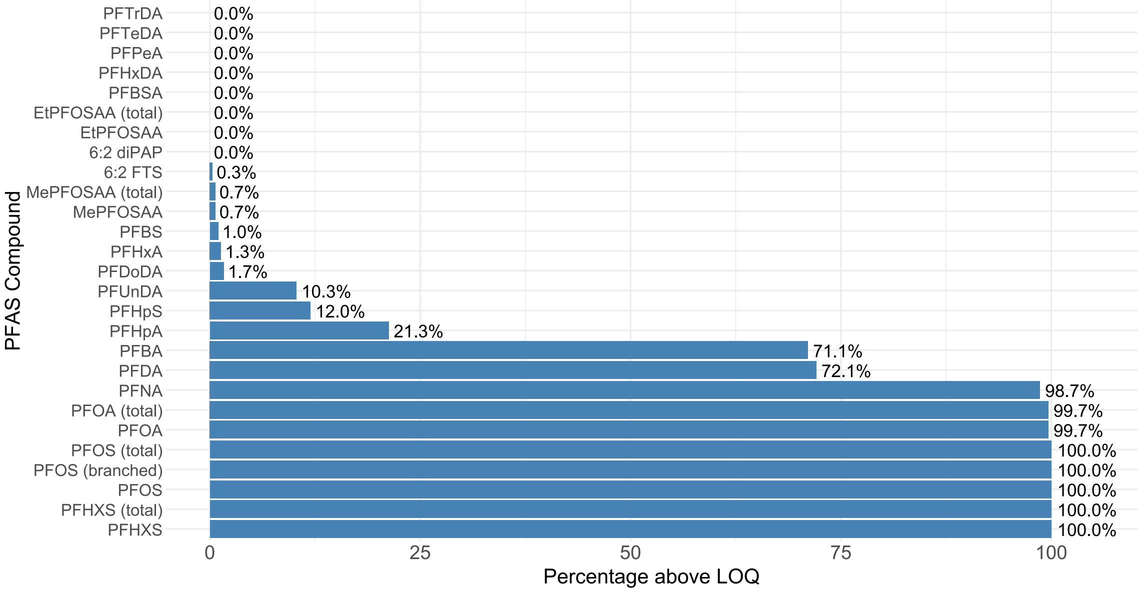


Figure 2.1: *Percentage of participants (excluding missing values) with values above the LOQ for various PFAS compounds. Total refers to the combined sum of linear and branched PFAS variants. If there are no specific indications in brackets, it refers to the linear variant.*

Table E.1 gives the relevant descriptive statistics for the exposures expressed in μg/L serum. To address the right-skewness typically observed in environmental exposure data, we applied log transformations to the exposure variables before analysis. Chemical compounds, such as PFAS, are often highly correlated, as discussed in the introduction. In the upcoming chapters, we will delve into this topic further. Figure 2.2 illustrates the Spearman correlation across the various compounds, providing insight into the complexity and strength of these correlations. This will serve as a guide for the discussions in the subsequent chapters.

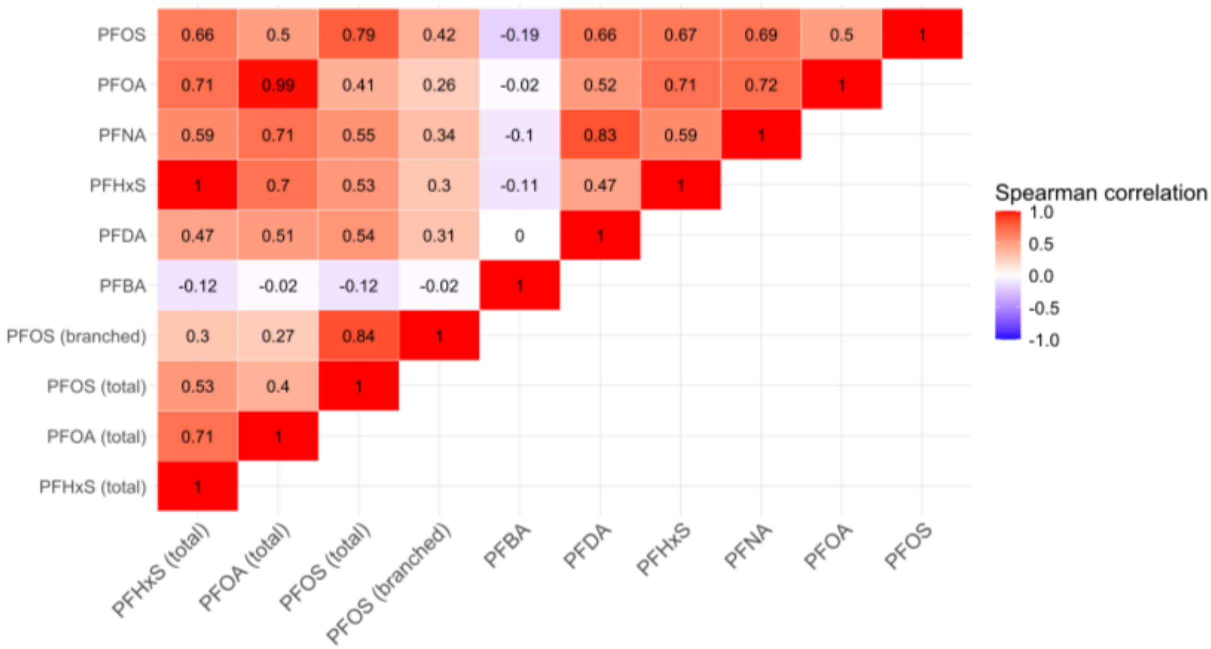


Figure 2.2: *Pairwise Spearman correlations between PFAS concentrations (after single imputation and log-transformation on the exposure). Total refers to the combined sum of linear and branched PFAS variants. If there are no specific indications in brackets, it refers to the linear variant.*

The health outcome of interest related to PFAS exposure is the number of immune cells in the blood. An effective immune response relies on balanced coordination between the innate and adaptive immune systems. Suppression of the immune system can increase the risk of infections, while over-activation may lead to allergic reactions or autoimmune diseases. Leukocytes (white blood cells) are key components of the immune system and play a crucial role in defence against infections and cancer development (Consortium UAntwerpen, VITO, PIH, UHasselt and VUB, 2023). Therefore, it serves as an important biomarker of immune system activity. Research conducted in animal studies has demonstrated that elevated exposure to PFAS leads to a reduction in leukocyte counts (Ehrlich et al., 2023).

The second outcome of interest is the ratio of CD4+ (helper T cells) to CD8+ (cytotoxic T cells). This ratio reflects immune balance and can serve as a sensitive biomarker for immune dysregulation. The majority of epidemiological studies indicate that there is no significant association between PFAS exposure and the CD4+/CD8+ T-cell ratio (Ehrlich et al., 2023). Appendix E contains visualisations of the log-transformed outcome distribution and the standardised log-transformed exposure distributions, as well as a summary table of the exposures, covariates and outcomes used in the case study presented later. The complete analysis will be presented in Chapter 6, as the methodology must first be introduced to answer the research question of interest.

Chapter 3

Shrinkage methods

This initial chapter on methodology outlines statistical shrinkage methods that are widely recognised and utilised, grounded in a robust theoretical framework. These methods are often considered as a baseline for comparison with new approaches that (sometimes) lack this theoretical foundation. Within this chapter, both frequentist and Bayesian shrinkage methods will be discussed. The goal is to use these models to deal with multicollinearity and assess the effect of individual chemicals within a mixture on a health outcome.

Let n denote the number of observations, c the number of exposure variables in the exposure matrix $\mathbf{A}$, z the number of additional covariates in the matrix $\mathbf{Z}$ and define $p \equiv c + z$ as the total number of predictors. In the upcoming sections, we will, for simplicity and without loss of generality, treat exposures and additional covariates together in the notation as an $(n \times p)$ design matrix $\mathbf{X} = [\mathbf{A}\ \mathbf{Z}]$ of rank p. In practice, these are often separated in the analysis, as for shrinkage methods, we typically do not penalise the additional covariates.

The issue of multicollinearity is best understood in a linear model context. Hoerl and Kennard (1970) considered a standard model for multiple linear regression

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.1}$$

where it is assumed that $\mathbf{Y}$ is a $(n \times 1)$ random outcome vector, $\boldsymbol{\beta}$ is the $(p \times 1)$ dimensional unknown parameter vector, $\boldsymbol{\epsilon}$ is a $(n \times 1)$ random error vector, $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \sigma^2 \boldsymbol{I}_n$. The ordinary least squares (OLS) estimator is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} \quad \text{with} \quad \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} \tag{3.2}$$

which represents the individual effects in a chemical mixture. Multicollinearity has to do with the ill-conditioning of the Gram matrix $\mathbf{X}^\top\mathbf{X}$. This can be best understood by examining the total variance of $\hat{\boldsymbol{\beta}}$:

$$\text{Total Var}(\hat{\boldsymbol{\beta}}) = \mathrm{Tr}\big(\mathrm{Var}(\hat{\boldsymbol{\beta}})\big) = \sigma^2\ \mathrm{Tr}\big((\mathbf{X}^\top\mathbf{X})^{-1}\big) = \sigma^2 \sum_{i=1}^{p} \frac{1}{l_i} \tag{3.3}$$

where $l_{max} = l_1 \geq l_2 \geq ... \geq l_p = l_{min} > 0$ are the eigenvalues of $\mathbf{X}^\top\mathbf{X}$. This follows from the fact that the Gram matrix $\mathbf{X}^\top\mathbf{X}$ is semi-positive definite. When two columns in $\mathbf{X}$ have an approximate linear relationship and are thus highly correlated, l_i will be small. This will lead to an inflation of the total variance as indicated in equation (3.3). This phenomenon is known as multicollinearity (Hoerl and Kennard, 1970; Thas, 2023).

The OLS estimator is affected by potentially high correlations among predictors. While it can still provide unbiased estimates, the standard errors of the regression parameters may be inflated. This inflation makes it difficult to identify which chemical within a mixture has a significant effect on a health outcome. For these particular settings, researchers have developed shrinkage methods. These techniques introduce minimal bias while effectively decreasing standard errors, leading to improvements in mean square error (MSE). The following sections will shortly introduce various shrinkage methods that are well-known.

3.1 Frequentist shrinkage methods

3.1.1 Ridge regression

In the context of non-orthogonal regression problems, Hoerl and Kennard (1970) introduced a biased estimator known as the ridge regression estimator, defined as:

$$\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^\top \mathbf{X} + k\boldsymbol{I}_p)^{-1}\mathbf{X}^\top \mathbf{Y} \tag{3.4}$$

This estimator is the solution of minimising the residual sum of squares (RSS), subject to a constraint on the L2 or Euclidean norm of the coefficient vector. This approach is equivalent to penalising the RSS, where $k \geq 0$ can be interpreted as the Lagrange multiplier or penalty parameter.

If $k = 0$, the OLS estimator (3.2) is obtained. As k increases, the variance of $\hat{\boldsymbol{\beta}}_{ridge}$ decreases because the influence of small eigenvalues is suppressed. At the same time, a larger k results in a stronger penalty, causing the coefficients to shrink towards zero. Hoerl and Kennard (1970) showed that the MSE of $\hat{\boldsymbol{\beta}}_{ridge}$, decomposes into the total variance of the ridge estimator and squared bias introduced by the ridge penalty. This illustrates the fundamental bias-variance trade-off in ridge regression. Hoerl and Kennard (1970) present an explicit formulation for the standard error of the ridge coefficient. Nevertheless, it is common practice to employ bootstrapping as an alternative method. Bootstrap percentile-based confidence intervals will be utilised in this thesis.

In conclusion, ridge regression can be effectively applied to chemical mixtures to prevent the inflation of standard errors. Although the shrinkage introduced by ridge regression adds some bias, it often leads to a reduction in the MSE of the individual components within the mixture. Consequently, the selection of the ridge penalty parameter k is crucial, as it determines the balance between bias and variance. In this thesis, 5-fold cross-validation will be utilised for all frequentist shrinkage methods to choose the largest k whose MSE is within one standard error of the minimum.

3.1.2 LASSO regression

Ridge regression, as previously described, addresses the issues associated with the OLS estimator by shrinking the coefficients towards zero. Tibshirani (1996) highlighted a second issue with OLS, which has not been discussed yet: the interpretation. When dealing with a large number of predictors, it is often the goal to identify a smaller subset of those predictors that have the strongest effects. This is not the case for Ridge regression. In contrast, it does not set coefficients to zero and may not yield an easily interpretable model. To accommodate for this, Tibshirani (1996) proposed a method called LASSO or "least absolute shrinkage and selection operator". The coefficients $\hat{\boldsymbol{\beta}}_{lasso}$ are obtained by minimising the RSS, subject to a non-differentiable constraint expressed by the L1 or Manhattan norm of the coefficient vector. Due to the continuous shrinking operation of this norm, it can produce coefficients that are exactly zero. This leads to variable selection, which was not the case in Ridge regression.

In contrast to Ridge regression, LASSO does not have an explicit formula for calculating the standard error due to the non-differentiable characteristics of the L1 penalty. Therefore, Tibshirani (1996) proposed the use of bootstrapping for calculating the standard error (SE) and percentile-based confidence intervals. Alternatively, various standard error estimators have been proposed in the literature. Kyung et al. (2010) give a comprehensive overview of the available methodologies. However, Kyung et al. (2010) argue that Bayesian methods are often more accessible and practical. This topic will be discussed in detail in Section 3.2.

LASSO can thus be useful in a situation where there are many chemical pollutants and people believe that only a few of these chemicals have significant effects. However, this is not the case in our study presented in Chapter 2, in which $n \gg p$ and the predictors tend to be highly correlated. Tibshirani (1996) demonstrated that in these cases, ridge regression often outperforms LASSO in terms of prediction performance. Moreover, in the context of mixture analysis, the use of LASSO may result in problematic variable selection as well. If there is a group of pollutants among which the pairwise correlations are high, then the LASSO tends to select only one variable from that group (Zou and Hastie, 2005). There are several alternative versions of LASSO available that offer slight improvements. A discussion of these variations is beyond the scope of this context.

3.1.3 Elastic Net regression

Considering the limitations outlined in the previous paragraph, Zou and Hastie (2005) proposed Elastic Net regression. Similar to Ridge and LASSO regression, Elastic Net is a penalised regression technique. It can be seen as a penalised least squares method where the penalty is a convex combination of the LASSO and ridge penalty $((1-\alpha)||\boldsymbol{\beta}||_1+\alpha||\boldsymbol{\beta}||_2^2)$ (Zou and Hastie, 2005). The Elastic Net faces the same standard error challenges as the LASSO. When $\alpha = 1$, Ridge regression is obtained, whereas when $\alpha = 0$, LASSO regression is achieved. Within VITO, it is common practice to select a value of $\alpha = 0.5$ to balance sparsity and shrinkage, particularly in the presence of correlated predictors. While tuning α via cross-validation can offer improved model performance and more stable variable selection in some settings, we follow the internal convention and fix $\alpha = 0.5$.

The primary reason for using Elastic Net in the context of chemical mixtures is its grouping effect, which means that identical predictors receive the same coefficients (Zou and Hastie, 2005). While this holds for both Ridge and Elastic Net, it does not hold for LASSO. For LASSO, the sum of the two coefficients can be arbitrarily split between the two and does not even have a unique solution. Carrico et al. (2015) argue that this grouping effect can also be problematic. The grouping effect would assign similar coefficients to highly correlated predictors even when one of the predictors does not necessarily have an association with the health outcome. Nevertheless, the Elastic Net approach is often seen as a baseline method for mixture analysis and will serve as a comparison point.

3.2 Bayesian shrinkage methods

Techniques such as Ridge, LASSO and Elastic Net are mainly used for prediction, selection and shrinkage. However, our main objective is to identify a specific chemical within a mixture and to conduct statistical inference as well. Since there are challenges in calculating standard errors or bootstrap approximation are required for the frequentist shrinkage methods, Bayesian shrinkage methods could be a suitable alternative. The Bayesian approach utilises priors as a replacement for the previously imposed penalties. A detailed discussion of this methodology will be provided in the subsequent paragraphs.

3.2.1 Bayesian LASSO regression

Consider a linear model where the outcome follows a normal distribution using previous notation.

$$\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2 \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n) \tag{3.5}$$

Tibshirani (1996) noted that LASSO estimates can be interpreted as posterior mode estimates under the assumption that the regression parameters follow independent identically distributed Laplace priors. Park and Casella (2008) suggest a hierarchical representation of the full Gaussian model where the Laplace distribution is represented as a scale mixture of normals with an exponential mixing density.

$$\begin{aligned} \beta_i \mid \tau_i^2,\ \sigma^2 &\sim \mathrm{N}\left(0,\sigma^2\tau_i^2\right) \\ \tau_i^2 &\sim \mathrm{Exp}\left(\frac{\lambda^2}{2}\right) \\ \pi(\sigma) &\propto \frac{1}{\sigma} \\ \lambda^2 &\sim \mathrm{Gamma}(0.1,0.1) \end{aligned} \tag{3.6}$$

with $i = 1,...,c$ the index for each chemical in the mixture. For σ^2 Park and Casella (2008) recommend an improper prior or an inverse-gamma prior as it would also maintain conjugacy. For choosing the LASSO parameter λ they proposed a class of gamma priors on λ^2.

In conclusion, Bayesian LASSO can serve as an alternative to construct credible intervals for the parameters. The name Bayesian LASSO may be misleading. Park and Casella (2008) showed in an example that Bayesian LASSO appears to be a compromise between frequentist LASSO and Ridge regression. It tends to pull weak signals faster to zero than frequentist ridge regression. In contrast to the frequentist LASSO, it will not set coefficients exactly to zero as it allows posterior mass near zero. This aspect is frequently regarded as a critique of the Bayesian LASSO approach. The subsequent methods introduced will further address this concern.

3.2.2 Bayesian spike and slab regression

The spike and slab prior is a very popular shrinkage and variable selection prior. Piironen and Vehtari (2017) referred to it as the “golden standard” for sparse Bayesian estimation. It was first introduced by Lempers (1971); Mitchell and Beauchamp (1988); George and McCulloch (1993). The spike and slab prior is defined as a mixture of two normal distributions with different variances. It can be defined for (3.5) using previous notations as

$$\begin{aligned} \beta_i \mid \lambda_i, c_0, \epsilon_0 &\sim \lambda_i \cdot \mathrm{N}(0, c_0^2) + (1 - \lambda_i) \cdot \mathrm{N}(0, \epsilon_0^2) \\ \lambda_i &\sim \mathrm{Ber}(\pi) \\ \pi &\sim \mathrm{Beta}(1, 1) \end{aligned} \tag{3.7}$$

with $\epsilon_0 \ll c_0$. The "spike" variance is defined as $\epsilon_0 \approx 0$, which ensures that when $\lambda_i = 0$ the coefficient β_i will be close to zero. Alternatively, the "slab" variance is defined as $c_0 \gg 0$ such that when $\lambda_i = 1$, the coefficient β_i will be away from zero. The selection of these hyperparameters can often be challenging. George and McCulloch (1993) argue that from a subjective Bayesian perspective, c_0 should be sufficiently large to support the strong signals as seen in Figure 3.1. The hyperparameter tuning is commonly regarded as a limitation of the spike and slab prior.

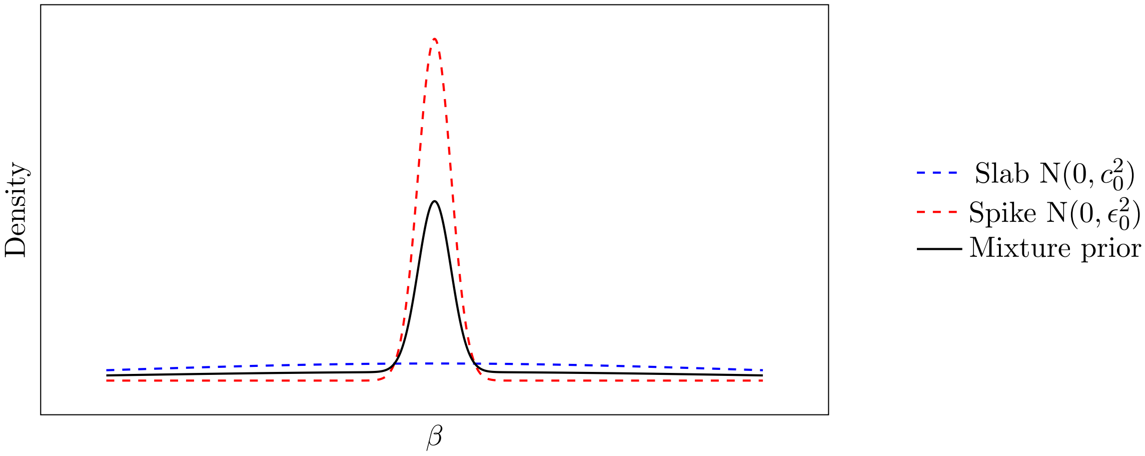


Figure 3.1: *Spike and slab prior with* $\pi = 0.5,\ c_0 = 0.1,\ \epsilon_0 = 0.005$.

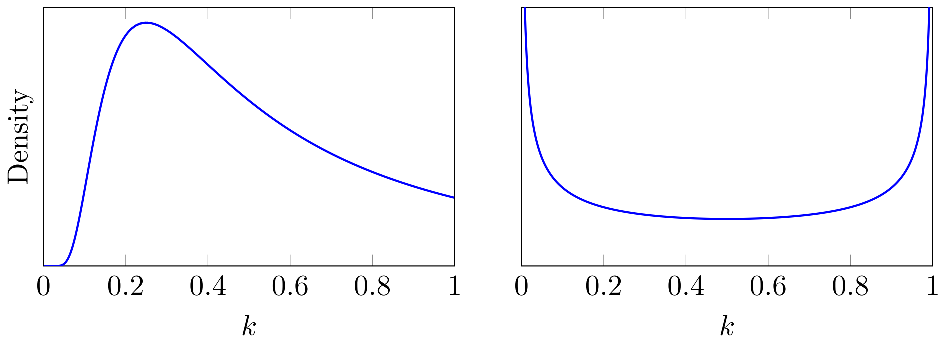
3.2.3 Bayesian horseshoe regression

The Bayesian LASSO has difficulties with adapting to situations with both strong and weak signals. On the other hand, the spike and slab often suffers from sensitivity to hyperparameter tuning. To address these limitations, Carvalho et al. (2009, 2010) introduced a novel approach to sparsity, in which the parameter $\boldsymbol{\beta}$ is believed to be sparse. They called this the horseshoe prior. The horseshoe prior is defined as a scale mixture of normals:

$$\begin{aligned} \beta_i \mid \lambda_i,\ \tau &\sim \mathrm{N}(0, \lambda_i^2 \tau^2) \\ \lambda_i &\sim \mathrm{C}^+(0, 1) \\ \tau &\sim \mathrm{C}^+(0, 1) \end{aligned} \tag{3.8}$$

with $\mathrm{C}^+(0, 1)$ the half-Cauchy distribution. Carvalho et al. (2009, 2010) refer to λ_i as the local shrinkage parameter and τ the global shrinkage parameter.

To better understand the Horseshoe prior, Carvalho et al. (2009, 2010) introduced the shrinkage coefficient $k_i = \frac{1}{1+\lambda_i^2}$, using fixed values for σ and τ set to 1. The shrinkage coefficient quantifies the extent to which the observed data influences the posterior mean of β_i. When $k \approx 1$, the posterior mean is heavily influenced and pulled towards zero, indicating strong shrinkage. Conversely, when $k \approx 0$, the posterior mean remains largely unshrunken. The right panel in Figure 3.2 illustrates the horseshoe-shaped shrinkage profile. This characteristic arises because the half-Cauchy prior on λ_i results in a Beta $\left(\frac{1}{2}, \frac{1}{2}\right)$ distribution for k_i. This shows that the Horseshoe prior favours weak or strong signals as it has a high peak near 0 and 1. The flat density in the middle reflects that moderate shrinkage is discouraged. In contrast, the left panel on Figure 3.2 representing the LASSO prior shrinkage profile, has high density in the middle. This reflects the tendency of LASSO to result in the over-shrinkage of large values, while simultaneously under-shrinking observations characterised by noise (Carvalho et al., 2009).



Laplace (LASSO)

Horseshoe

Figure 3.2: *Density functions of k_i for selected shrinkage priors (up to constants) based on Table 1 and Figure 2 in Carvalho et al. (2010).*

Extensions such as the regularised horseshoe have been proposed by Piironen and Vehtari (2017). These are particularly advantageous for data with moderate signals, as the regularised horseshoe prior is less aggressive. This is an important topic to consider for future research. In conclusion, the horseshoe prior is particularly well-suited for chemical mixture modelling, as it effectively shrinks the impact of irrelevant exposures to near-zero levels while preserving the truly important chemicals.

Chapter 4

Multi-pollutant methods

The methods outlined in Chapter 3 are frequently utilised in a variety of contexts beyond chemical mixtures. Recent advancements have introduced new techniques specifically designed to address the unique complexities and questions associated with chemical mixtures. A literature review conducted by Yu et al. (2022) indicated that between 2018 and 2022, methods as weighted quantile sum (WQS) regression and Bayesian kernel machine regression (BKMR) have gained significant popularity. In addition to the aforementioned methods, this discussion will also encompass Bayesian model averaging (BMA) and G-computation. These techniques have been frequently cited in the literature and have been employed in multi-pollutant analysis by VITO in prior studies.

4.1 Weighted quantile sum regression

Recently, innovative methodologies have been proposed to tackle the complexities associated with chemical mixtures. One notable strategy involves the empirical development of a weighted sum. This idea is better known as weighted quantile sum (WQS) regression. First, the exposure values are quantised and combined into a unidirectional weighted sum, reducing dimensionality and avoiding multicollinearity. Secondly, the significance of the WQS is determined, providing a single overall effect estimate of the mixture and the weights are interpreted as the relative importance (Carrico et al., 2015; Tanner et al., 2019). The discussion will first focus on the (bootstrap) WQS regression, followed by improvements aimed at stabilising weight estimation.

4.1.1 WQS regression

Let c represent the number of correlated components or exposures, which are scored into quantiles denoted by $\mathbf{q} = (q_1, ..., q_c)$ (e.g., for quartiles $q_i = 0, 1, 2$ or 3), where $i = 1, ..., c$. This quantisation helps mitigate the influence of extreme exposure values, accounts for potential non-linear relationships and allows for straightforward interpretation as the effect of a one-quantile increase in exposure. Nonetheless, this assumption has faced criticism, which is examined further in the simulation study presented in Section 5.1.

The original weighted index model by Christensen et al. (2013) is defined as

$$g(\mathbb{E}[Y|\mathbf{q},\mathbf{z}]) = \beta_0 + \beta_1 \left(\sum_{i=1}^{c} w_i q_i \right) + \mathbf{z}^\top \boldsymbol{\varphi} \tag{4.1}$$

where g is a monotonic, differentiable link function as in a generalised linear model (GLM), β_0 is the intercept, β_1 is the regression coefficient for the weighted quantile sum, w_i is the weight associated with the i^{th} component, $\mathbf{z}$ is the vector of additional covariates and $\boldsymbol{\varphi}$ is the vector of parameters associated with the covariates. For estimation of the parameters, the dataset is first divided into a training and validation set. Next, the weights w_i are estimated using the training set with constraints

$$\sum_{i=1}^{c} w_i = 1 \quad \text{and} \quad 0 \leq w_i \leq 1 \ \ \forall i \in \{1, ..., c\} \tag{4.2}$$

The estimation is performed by maximising the likelihood of the (non-)linear regression model (4.1) while enforcing the constraints (4.2). The estimated weights $\hat{w}_i$ are now used to define the WQS

$$wqs = \sum_{i=1}^{c} \hat{w}_i q_i \quad . \tag{4.3}$$

In the final step model (4.4) is used to estimate the effect, β_1, of the WQS on the outcome using maximum likelihood estimation (MLE) on the validation set.

$$g(\mathbb{E}[Y|wqs,\mathbf{z}]) = \beta_0 + \beta_1 wqs + \mathbf{z}^\top \boldsymbol{\varphi} \tag{4.4}$$

The parameter β_1 is interpreted as the joint effect if all exposures simultaneously increase by one quantile. If this effect is found to be significant, the weights $\hat{w}_i$ can be interpreted as the relative importance of a specific individual chemical.

4.1.2 Bootstrap WQS regression

To stabilise the weight estimates, Carrico et al. (2015) proposed a bootstrap step. A fixed number B of bootstrap samples from the training dataset is used to estimate the unknown weights w_i that maximise the likelihood for the model (4.1), as previously described. After estimation, a post hoc constraint is applied to retain only the weights associated with bootstrap samples in which the estimated coefficient $\hat{\beta}_1$ has the same sign. This constraint enforces directional homogeneity, meaning all exposures in the index are assumed to influence the outcome in the same direction (either positive or negative). As a result, a set of B^* estimated directional weights is obtained and the weighted quantile sum index is then calculated as

$$wqs_{bs} = \sum_{i=1}^{c} \bar{w}_i q_i \quad \text{with} \quad \bar{w}_i = \frac{1}{B^*} \sum_{b=1}^{B^*} \hat{w}_{i(b)} f(\hat{\beta}_{1(b)}) \tag{4.5}$$

with $f(\cdot)$ a pre-specified “signal function”. The signal function is specifically designed to assign greater weight to samples with higher signals. The wqs in model (4.4) is then replaced by wqs_{bs} and the estimation of β_1 is done in the same way.

4.1.3 Random subset WQS regression

In high-dimensional mixtures with highly collinear predictors or when the number of predictors exceed the number of subjects, bootstrap WQS regression may fail. Therefore, Curtin et al. (2019) proposed a novel implementation, random subset WQS (WQS_{rs}) regression. The idea is to select a random subset of the total predictor set. These subsets of predictors are more decorrelated and thus the ill-conditioning in the variable selection algorithm is improved.

Let d be the fixed number of randomly selected predictors and S the total number of subsets. For each subset $s = 1, ..., S$, randomly select d exposures out of the total c. Next, the weights for each subset will be estimated using the same constraints as in (4.2) and the model previously applied in (4.1). After estimation, a post hoc constraint is applied to retain only the weights associated with subsets in which the estimated coefficient β_1 has the same sign. Afterwards, the unique weights are averaged across all subsets to determine the final variable weights used in the calculation of the WQS_{rs} index. This is then similarly employed in a model on the validation data, as done previously.

Curtin et al. (2019) showed in their simulation study the behaviour of this method for $34, 59$ or 472 exposures. They concluded that for smaller mixtures, there are relatively few combinations of random subsets, making WQS_{bs} potentially advantageous. However, in cases with larger predictor sets or when the number of predictors exceeds the number of subjects, WQS_{rs} should be implemented instead of WQS_{bs}.

4.1.4 Repeated holdout validation WQS regression

The previously discussed bootstrap WQS regression applications divided the data into a single training set and a validation set. In finite study samples, this reduces statistical power and may result in unrepresentative partitions and unstable estimates (Tanner et al., 2019). Therefore, Tanner et al. (2019) proposed a repeated holdout validation. First, the dataset is R times randomly partitioned (with replacement) into a training and validation set. Next, bootstrap WQS regression is done on each set. Across the R sets, the mean is used as the final estimate

$$\hat{\beta}_{rh} = \frac{1}{R}\sum_{r=1}^{R}\hat{\beta}_{1r} \tag{4.6}$$

For coefficient inference, the 95% confidence intervals are based on the standard deviation of the simulated sampling distribution.

The study conducted by Tanner et al. (2019) did not present any simulation results; instead, it focused solely on an empirical case study with 26 predictors. This strengthens their belief that WQS_{rh} can produce more stable WQS index estimates compared to WQS_{bs}. The reason for this is that in smaller sample sizes, extreme individual weights are averaged out, unlike in single partitions. However, it should be noted that this comes at a large computational cost. A repeated holdout with 100 partitions will take 100 times longer to run as compared to bootstrap WQS regression.

The final advantage of repeated holdout validation is that it allows for the characterisation of weight uncertainty. As the WQS weights are constrained to be non-negative and sum to one, classical statistical inference (e.g., hypothesis tests for individual weights $H_0 : w_i = 0$) is not applicable. Carrico et al. (2015) proposed a specific cut-off point: when the weights fall below this threshold, these components are identified as "bad actors" or not selected. It is important to note that this cut-off point should be smaller when there are many predictors and larger when there are only a few. This choice of their cut-off point is arbitrary. Given that inference based on WQS weights lacks a formal theoretical foundation, we focus on interpreting the overall mixture effect and examining the relative magnitudes of weights to explore which exposures may be more influential.

4.2 G-computation for joint effect estimation

WQS regression requires strong assumptions about directional homogeneity and the linear, additive effects of individual exposures. Additionally, little theoretical statistical evidence exists about internal validity, such as bias, consistency and confidence interval coverage. Therefore, Keil et al. (2020) introduced a novel method called quantile g-computation, without such strong assumptions. It relies on the G-computation principle from causal inference and allows for flexible techniques to estimate the joint effect. In contrast to the work of Keil et al. (2020), which concentrated on both joint and individual effect estimation, we will focus only on joint effect estimation through the application of G-computation. This approach aims to effectively address the potential synergistic effects associated with PFAS exposures. The specific terminology related to causal inference and identifiability assumptions are introduced in Appendix B.

Historically, the parametric G-formula or G-computation was introduced in a series of articles by Robins (1986). It is sometimes also referred to as "standardisation". Vansteelandt and Keiding (2011) argued that this term is more familiar to epidemiologists and better captures the essence of G-computation for point exposures. The idea is to model the outcome Y and use that model to predict counterfactual outcomes across the entire population. Snowden et al. (2011) provide guidelines for implementing this idea. It involves the following steps:

1. Select a model for the outcome Y on the exposure $\boldsymbol{a}$ and covariates $\mathbf{z}$. This model is commonly referred to as the *Q-model* and is typically a regression-type model, expressed as $\mathbb{E}[Y|\boldsymbol{a}, \mathbf{z}]$. However, G-computation also allows for the use of more flexible methods, such as non-parametric and machine learning techniques, to estimate the Q-model. These methods relax the parametric assumptions of the Q-model, which can help reduce bias from model misspecification.

2. Use the model fit in step 1 to predict outcomes $\hat{Y}^{\boldsymbol{a}}$, reflecting counterfactuals, for each observation under two hypothetical scenarios while covariates remain at their observed values.

3. Finally, average the predictions under the two hypothetical scenarios and take the difference. A reasonable hypothetical scenario in the context of chemical mixtures could involve making predictions at the first and third quartile values. The average difference between these predictions would represent the marginal effect of an interquartile range increase in all predictors.

Under the *identifiability assumptions*, and additionally assuming no model misspecification or measurement error, the estimates are considered unbiased for the causal effect. In practice, some degree of mismeasurement of most variables, mismeasured models or potential confounding is unavoidable. (Hernan and Robins, 2025; Snowden et al., 2011)

4.3 Bayesian adaptive sampling for variable selection and model averaging

Previously described models typically assume linear additive relationships among exposures, although some can be easily modified to incorporate non-linear or interaction effects. It is essential to evaluate whether the inclusion of an interaction effect or predictor improves the model's fit. This evaluation process, which involves examining various combinations of predictors, is referred to as model selection. A limitation of this process is that the final estimates derived from the selected model do not account for the uncertainty inherent in the selection process. To address this concern, Hoeting et al. (1999) proposed the method of Bayesian model averaging (BMA).

4.3.1 Bayesian model averaging

The key idea of model averaging is to make weighted predictions based on the model's fit for a specific quantity Δ of interest, in our case, the regression coefficient. In the context of Bayesian model averaging, the weights are chosen to be the posterior model probability. Consider a model $\mathcal{M}_{\boldsymbol{\gamma}}$, with $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_c) \in \{0,1\}^c \equiv \Gamma$, the elements in $\boldsymbol{\gamma}$ are indicators representing whether or not a specific predictor is included in the design matrix $\boldsymbol{X}_{\boldsymbol{\gamma}}$. The Gaussian linear model can now be defined as

$$\boldsymbol{Y} \mid \beta_0, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \mathcal{M}_{\boldsymbol{\gamma}} \sim \mathrm{N}(\boldsymbol{I}_n\beta_0 + \boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\ ,\ \boldsymbol{I}_n\sigma^2) \tag{4.7}$$

using previous notation. Based on this formulation, each model is assigned a posterior model probability. This defines the fit to the data and prior model probability. The posterior probability of a model $\mathcal{M}_{\boldsymbol{\gamma}'}$ is computed by Bayes' theorem

$$P(\mathcal{M}_{\boldsymbol{\gamma}'}|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}'})P(\mathcal{M}_{\boldsymbol{\gamma}'})}{\sum_{\boldsymbol{\gamma}\in\Gamma} P(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}})P(\mathcal{M}_{\boldsymbol{\gamma}})} \tag{4.8}$$

with $P(\mathcal{M}_{\boldsymbol{\gamma}})$ the prior model probability and $P(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}})$ proportional to the marginal likelihood of $\mathcal{M}_{\boldsymbol{\gamma}'}$ obtained by integrating the joint likelihood with respect to the prior distribution over all parameters. The last step involves calculating the quantity of interest as a weighted average, weighted by the posterior model probabilities (Hoeting et al., 1999; Clyde et al., 2011).

$$P(\Delta|\boldsymbol{Y}) = \sum_{\boldsymbol{\gamma}\in\Gamma} P(\Delta|\mathcal{M}_{\boldsymbol{\gamma}}, \boldsymbol{Y})P(\mathcal{M}_{\boldsymbol{\gamma}}|\boldsymbol{Y}) \tag{4.9}$$

As implemented by Clyde et al. (2011), each Bayesian linear model will include a different combination of exposure variables. When only main effects were considered, the total model space consists of 2^c possible models, where c is the number of exposures. In this framework, regression coefficients for the exposures are interpreted as weighted averages across all considered models, with weights given by the posterior model probabilities. This does not resolve the issue of multicollinearity.

The different combinations of exposure variables in each model allow us to interpret the posterior inclusion probability (PIP). The PIP serves as a measure of variable importance, defined as the sum of posterior model probabilities for all models that include a specific variable. A high PIP indicates strong evidence that the variable is important for explaining the outcome, while a low PIP suggests limited support for its inclusion.

4.3.2 Bayesian adaptive sampling

In situations where the number of predictors is substantial, the space of models Γ may become excessively large to analyse. For instance, in a scenario involving 25 predictors, considering only the main effects would result in $2^{25} = 33\ 554\ 432$ potential combinations of predictors that could be incorporated into the model. Therefore, Clyde et al. (2011) proposed a Bayesian adaptive sampling (BAS) without replacement algorithm. Unlike other algorithms, this method guarantees that it will enumerate the complete space of models if computational resources permit. When this is not possible, BAS uses a stochastic sampling algorithm.

The model space Γ is structured as a binary tree. In this framework, each level of the tree reflects a decision to either include ($\gamma_j = 1$) or exclude ($\gamma_j = 0$) the j-th predictor. Consequently, every model corresponds to a distinct path within the binary tree, resulting in a total of 2^c paths. For a binary tree, the distribution can be expressed as

$$f(\boldsymbol{\gamma}|\boldsymbol{\rho}) = \prod_{j=1}^{c} f(\gamma_j|\boldsymbol{\gamma}_{<j}) = \prod_{j=1}^{c} (\rho_{j|<j})^{\gamma_j}(1-\rho_{j|<j})^{1-\gamma_j} \tag{4.10}$$

where $\boldsymbol{\gamma}_{<j}$ indicates the subset of inclusion indicators, $\rho_{j|<j} \equiv f(\gamma_j = 1|\boldsymbol{\gamma}_{<j})$ and $\boldsymbol{\rho}$ the collection of all $\{\rho_{j|<j}\}$. When a model is sampled using formula (4.10), the equation will first be updated with a new value of $\boldsymbol{\rho}$. This update guarantees that all previously sampled models will have a probability of zero. Once this update is done, a new model can be sampled. Clyde et al. (2011) proposed updating $\rho_{j|<j}$ using π_j, the marginal posterior inclusion probability for predictor j. They recommend starting with an estimate and iteratively updating it using sampled models:

$$\hat{\pi}_j^{(t)} = \frac{\sum_{\boldsymbol{\gamma}\in S_t} p(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}})\gamma_j}{\sum_{\boldsymbol{\gamma}\in S_t} p(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}})} \tag{4.11}$$

with S_t the set of models that have been sampled at step t. To ensure all models can be sampled, they advise bounding ρ away from 0 and 1 (Clyde et al., 2011). In conclusion, Bayesian adaptive sampling is often employed when the model space is extensive.

4.4 Bayesian kernel machine regression

The previous models rely on parametric functional forms. Therefore, in settings with complex mixtures, which are often the case in reality, these models are easily misspecified. Therefore, Bobb et al. (2014) proposed a novel, flexible technique called Bayesian kernel machine regression (BKMR). This approach models the health outcome as a smooth kernel function while adjusting for confounding variables.

The focus in this section will be on semi-parametric models, denoted as

$$Y = h(\boldsymbol{a}) + \mathbf{z}^{\top}\boldsymbol{\varphi} + \epsilon \tag{4.12}$$

with $\boldsymbol{a}$ a vector of c continuous exposures, $h : \mathbb{R}^c \to \mathbb{R}$ a high-dimensional exposure response function and ϵ assumed independent and follows $\mathrm{N}(0, \sigma^2)$. The remaining notation is similar to WQS regression model (4.1). In particular, the exposure-response function $h(\cdot)$ will be of interest. First, the connection with kernels will be made. Afterwards, the focus will shift to Bayesian variable selection within the exposure-response function.

4.4.1 Kernel function

The function $h(\boldsymbol{a})$ is expressed in terms of a kernel function $K(\cdot, \cdot)$ that defines similarity between input vectors such that

$$h(\boldsymbol{a}) = \sum_{i=1}^{n} \alpha_i K(\boldsymbol{a}_i, \boldsymbol{a}) \tag{4.13}$$

with $\{\alpha_i\}_{i=1}^{n}$ the corresponding set of weights. The number of weights α_i is related to the number of data points n. As a result, the dimension of the feature space does not affect the computational complexity (Cristianini and Shawe-Taylor, 2000). Commonly used kernels include the *dth polynomial kernel* and the *Gaussian kernel*. This chapter will primarily focus on the *Gaussian kernel*, which is defined as

$$K(\boldsymbol{a}_1, \boldsymbol{a}_2) = \exp\left(-\frac{\|\boldsymbol{a}_1 - \boldsymbol{a}_2\|^2}{\rho}\right) \tag{4.14}$$

with ρ a tuning parameter and $\|\cdot\|$ the L2 norm or Euclidean norm. From now on, a Gaussian kernel function will be used to represent $h(\cdot)$ in model (4.12). Liu et al. (2007) showed the Bayesian representation of model (4.12), they treat $h(\boldsymbol{a})$ as a random vector with a Gaussian process prior with mean 0 and covariance $cov(h(\boldsymbol{a_1}), h(\boldsymbol{a_2})) = \tau K(\boldsymbol{a}_1, \boldsymbol{a}_2)$. Thus the Bayesian formulation is

$$Y \mid \boldsymbol{\varphi}, \mathbf{z}, h(\boldsymbol{a}), \sigma^2 \sim N(\mathbf{z}^{\top}\boldsymbol{\varphi} + h(\boldsymbol{a})\ ,\ \sigma^2) \text{ with } h(\cdot) \sim \mathcal{GP}(0\ ,\ \tau K(\cdot,\cdot)) \text{ and } \boldsymbol{\varphi} \propto 1 \tag{4.15}$$

4.4.2 Bayesian kernel machine regression

Bobb et al. (2014) expanded the ideas of Liu et al. (2007) to a Bayesian semi-parametric framework that also allows for variable selection. They proposed *component-wise* variable selection and *hierarchical* variable selection. In this thesis, only the first option will be of interest, as our motivating example makes it difficult to partition the components into groups. To allow for *component-wise* variable selection Bobb et al. (2014) proposed an augmented Gaussian kernel function defined as

$$K(\boldsymbol{a_1}, \boldsymbol{a}_2; \boldsymbol{r}) = \exp\left(-\sum_{i=1}^{c} r_i (a_{1i} - a_{2i})^2\right) \tag{4.16}$$

The *component-wise* selection is now introduced by using a spike and slab prior on $\boldsymbol{r}$

$$r_i \mid \delta_i \sim \delta_i f_1(r_i) + (1 - \delta_i) P_0 \text{ with } i = 1, ..., c \text{ and } \delta_i \sim \mathrm{Ber}(\pi) \tag{4.17}$$

where $f_1(\cdot)$ is a pdf with support on $\mathbb{R}^+$ and P_0 density with point mass at 0. The posterior mean of δ_i can be interpreted as the inclusion probability for a specific component i.

Chapter 5

Simulation results

Simulation studies are experiments that generate data through pseudo-random sampling from known distributions. A key strength of these studies is the ability to understand the behaviour of methods since the "truth" can be derived from the data-generating process. This allows for the investigation of properties such as bias or robustness against model misspecifications. Morris et al. (2019) provide guidelines on how these experiments should be best set up. Their "ADEMP" structure will be closely followed in the Appendix D, which includes all additional details on the simulation procedures. In Appendix A, we introduce the software and packages used for each method.

All simulation studies are inspired by "Teenager HBM Study - 3M site", presented in Chapter 2. The models introduced in Chapter 3 and Chapter 4 will be compared in a moderate sample size within an epidemiological framework. First, we will study the impact of quantiles instead of continuous exposure values using WQS. The second simulation study studies the effect of the exposure correlation structure on joint effect estimation using OLS. Finally, a realistic exposure response setting will be considered where all methods will be compared based on 4 performance measures. In addition, a simulation study on the grouping property will also be considered.

5.1 Continuous vs quantised exposures in WQS

Carrico et al. (2015) were the first to propose the implementation of quantiles in WQS regression. They argued that this approach provides a clear interpretation of the joint effect. Additionally, they pointed out that estimating weights without bounds on the components can lead to extreme values having influence that grows with the weights. Nonetheless, they acknowledge that adopting quantiles may result in a loss of information. While quartiles are often used, one may be interested in understanding the differences in bias associated with weights between quartile exposure, decile exposure or standardised continuous exposures. Since Carrico et al. (2015) did not explore this topic in their paper, we will conduct a simulation study based on three different data-generating mechanisms.

The simulation study evaluates how continuous versus quantised (quartiles or deciles) exposures affect the stability of weight estimation and the power of the joint effect in a repeated holdout WQS regression model. A detailed description of the rationale behind the methodological choices and simulation procedure can be found in Appendix D.1. Data for 300 individuals are simulated under three scenarios:

(a) Exposures drawn from a multivariate t-distribution with two degrees of freedom, characterised by heavy tails with either no correlation or high correlation among the exposures.

(b) Exposures from a multivariate normal distribution with added exposure-driven outliers by multiplicative inflation on 5 randomly selected values with either no correlation or high correlation among the exposures.

(c) Log-transformed exposure profiles (rows) resampled from the case study dataset to preserve joint distributions and including confounders.

Outcomes are simulated using an additive linear continuous WQS formulation with standardised exposures, which includes known weights and noise. Confounders are incorporated only in scenario (c). The estimand of interest is the average exposure weight and joint effect calculated from the repeated holdout splits. Performance is evaluated by measuring the absolute bias in estimated weights and the power of the joint effect for both continuous and quantised exposures.

Figure 5.1 presents the results obtained from the three distinct exposure scenarios. The exposures were standardised prior to conducting the analysis. This is important because the weights in WQS regression sum to 1 and would otherwise adjust for the scale of the exposure. In all three scenarios analysed, X_3 or PFBA demonstrates the highest level of systematic bias, resulting in a consistent underestimation of its true effect, independent of how the exposure was simulated. A similar pattern is observed with X_5 or PFOA (total). It should be noted that these two components collectively contribute to 65% of the overall effect in the simulation procedure. The opposite effect is observed for X_2, X_6 and X_4 where there is a consistent overestimation of the true effect. These three chemicals accounted for 8% of the joint effect in the simulation procedure. For the first two scenarios, we also distinguished between simulations with approximately no correlation or high correlation among the exposures. The bias on the components is similar, but there is more variability when the correlation is higher. In conclusion, the findings seem to suggest that strong effects are underestimated and small effects overestimated. Similar results were found in terms of relative bias or when using a single repeated holdout split (results not shown).

When comparing quartiles, deciles and continuous exposures, there is little difference in biases across the scenarios. The weights based on continuous exposures in a repeated holdout WQS model seem to be robust against exposures with heavy tails. This is in contrast with Carrico et al. (2015), as they emphasised that having no bounds on the weights could result in extreme values. They discussed this in relation to a WQS regression model with a single repeated holdout split, as originally proposed in their first paper. In contrast, we used 100 repeated holdout splits as recommended by Tanner et al. (2019), which improves the stability of the weight estimates. As a result, we may not see differences in the weight estimation between quantised exposures and continuous exposures.

The right panel of Figure 5.1 shows the power of the joint effect estimate across the three exposure scenarios. In scenarios (a) and (c), continuous exposures tend to have greater statistical power. In scenario (b), the three exposure types have similar behaviour. When examining the correlation, WQS regression demonstrates greater power for the joint effect when the correlation among the exposures is strong. This difference in power will therefore be the topic of interest in the next simulation study.

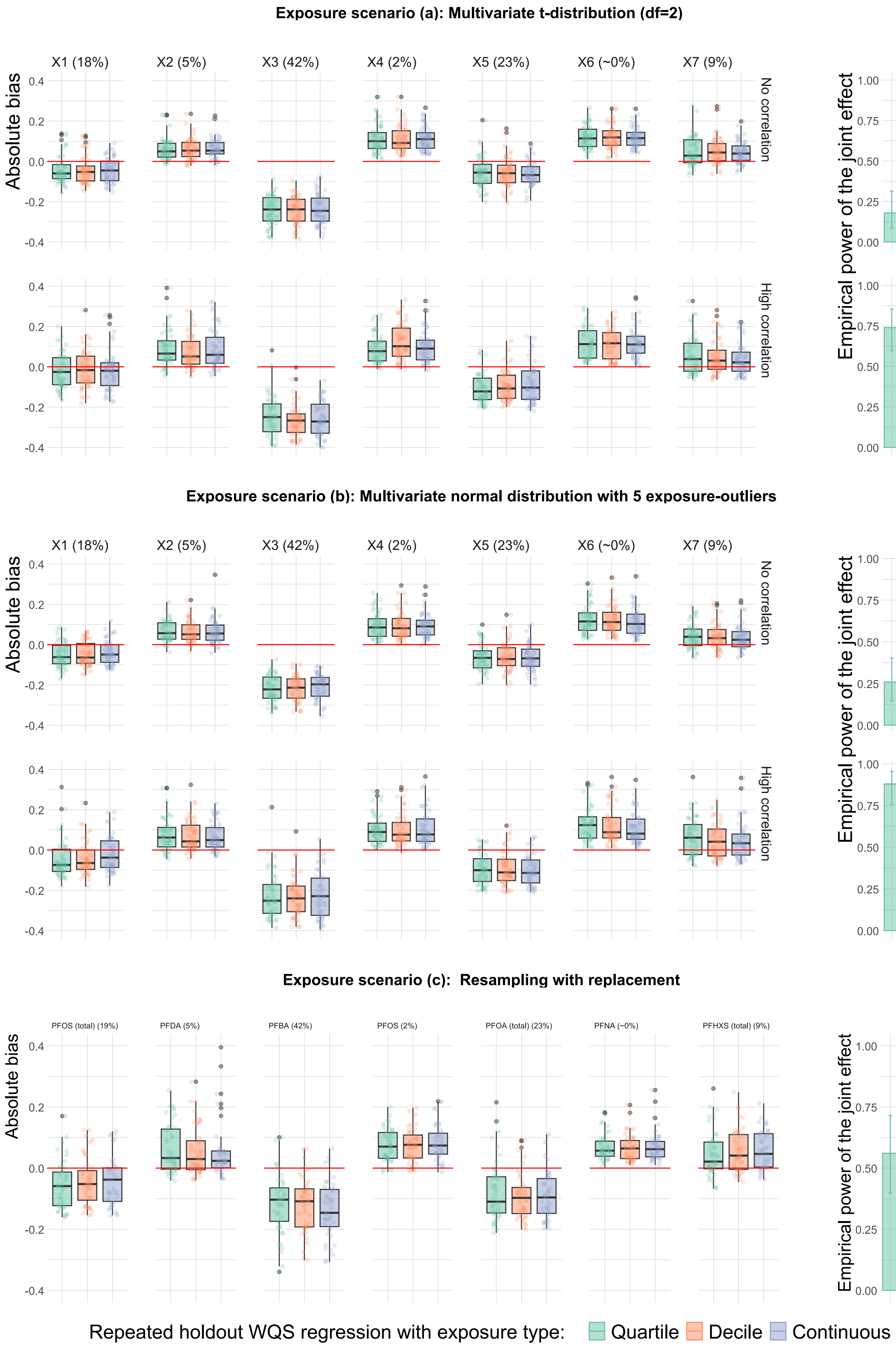


Figure 5.1: *The results are derived from repeated holdout WQS regression, which employed 100 holdout splits, with 20 bootstrap steps for each split. This entire process was repeated 50 times for each scenario; more details can be found in Appendix D.1. The left panel: The boxplot displays the estimated absolute bias of the weights for the various exposures. The right panel: The bar plot illustrates the empirical power of the joint effect with 95% CI.*

5.2 Joint effect estimation using OLS

In the previous simulation study using WQS, we identified a pattern suggesting that the statistical power associated with the joint mixture effect tends to be greater in the presence of high correlation. This simulation study investigates how high correlations among exposures affect the standard errors, power and Type I error of joint effects in multiple linear regression. The individual effects will also be presented for the purpose of comparison. Exposure data are generated from a multivariate normal distribution with equal pairwise correlation $\rho = 0, ..., 0.9$. The outcome is simulated with a linear additive relationship to the predictors. Three effect scenarios are considered: (a) one active exposure effect of -0.2, (b) multiple large/small effects summing to -0.2 and (c) no exposure effect. A correctly specified linear model is fitted in each simulation and the joint effect is computed as the sum of the coefficients. Performance is evaluated using the mean SE, empirical power and Type I error rate over 1 000 simulations. Exact details can be found in Appendix D.2.

The left panels shown in Figure 5.2 present the average SE as a function of the pairwise correlations. The SE for individual exposure effects remain small at low correlation levels but increases rapidly as correlations become higher. This is the well-known phenomenon of multicollinearity. In contrast, the SE of the joint effect decreases as the correlation increases. This trend is consistent across the three effect scenarios considered. The right panels in Figure 5.2 for cases (a) and (b) demonstrate the relationship between power and increasing pairwise correlations. The observed pattern is, as expected, consistent with the SE findings. Specifically, as the correlation increases, the power of the joint effect increases, while the power of the individual effects decreases. Empirical Type I error rates for both the sum of exposures and individual exposures are consistently close to the nominal level of approximately 0.05 and demonstrate stability across varying correlation levels. This observation indicates effective control of error rates, irrespective of the degree of correlation.

In conclusion, the joint effect estimation does not seem to suffer as much as the individual exposures from high correlations. This phenomenon was also observed by Carrico et al. (2015) and Keil et al. (2020) in the context of WQS regression and G-computation. Keil (2020) gave an intuitive explanation for this. Consider the simple example of two highly correlated predictors in Figure C.1. When Exposure A increases by one unit, Exposure B often increases as well due to their high correlation. As a result, there are few instances where A increases while B does not, making it challenging to estimate the effect of A while keeping B constant. In contrast, when considering the joint effect, we observe a diagonal movement along the data cloud. Increasing both A and B together is supported by the data, allowing the model to estimate their joint effect with more certainty.

The intuition above was given by Keil (2020). However, no theoretical explanation was given. In the introduction of Chapter 3, we showed why individual effects suffer from high correlation. In Appendix C.2, a theoretical approach is given to understand this in the context of joint effects. The key point is that in the context of two predictors with positive correlations, the covariance becomes highly negative. As a result, even if the sum of the individual standard errors is large, the standard error of the joint effect can be small due to the negative covariance.

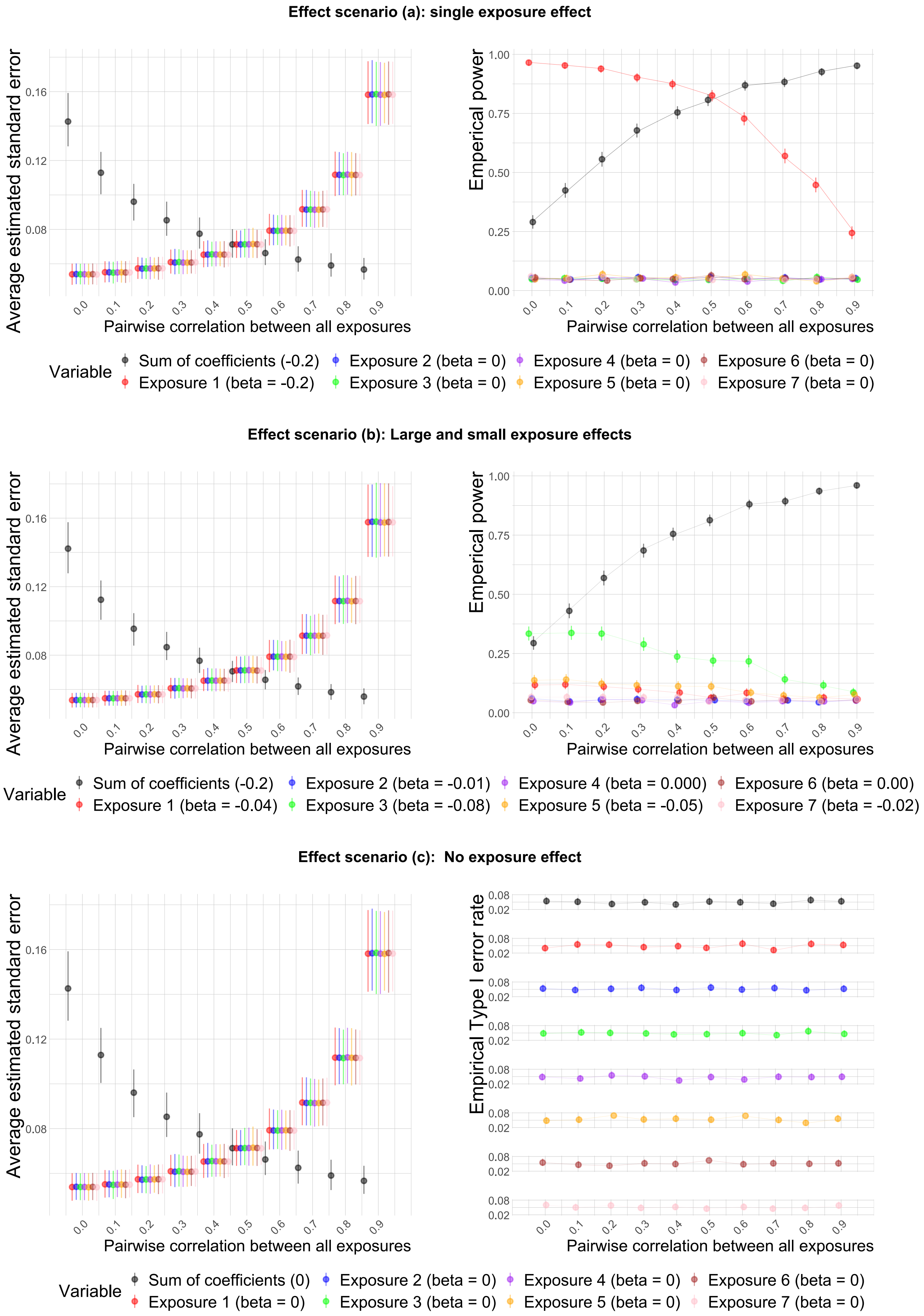


Figure 5.2: *Standard errors, power and Type I error rates for individual exposures and their joint effect across increasing pairwise correlation levels. Each point represents results from* 1 000 *simulations. See Appendix Appendix D.2 for full methodological details. Left panels: Mean standard errors with 95% simulation intervals. Right panels: Empirical power (for scenarios a and b) and Type I error rates (for scenario c), shown with 95% CI.*

5.3 Evaluating methods on realistic exposure mixtures

The goal is to evaluate the behaviour of the methods introduced in Chapter 3 and Chapter 4. Evaluation will be done based on different performance measures for estimating both joint and individual mixture effects, utilising realistic exposure data from the case study described in Chapter 2. The full description and all simulation setup details are given in Appendix D.3.1. We give here the intuition behind the simulation study. We generate 7 exposures by sampling $n = 300$ complete log-transformed exposure profiles with replacement from the case study dataset. This is done to preserve the original joint distribution and associated covariates. The strength of the correlations is shown in Figure 2.2, excluding the two components that have a correlation close to 1 and PFOS (branched). The outcomes are simulated using a weighted sum model (D.14) with standardised continuous exposures and a Gaussian outcome. The weighted sum coefficient and error variance are chosen such that we have a moderate signal-to-noise scenario, which reflects typical conditions in environmental health research. We assume directional homogeneity. Two formulations of the weighted sum are discussed:

- *Linear additive effect*: All exposures have an assigned weight greater than zero, with components having large, moderate or small effect weights. Weights were informed by the case study. This scenario represents rather a dense setting than a sparse setting.
- *Synergistic effect*: Secondly, the weighted sum is selected to incorporate certain interactions representing synergistic effects. In the literature, this is often regarded as more realistic. The interactions represent 35% of the total effect.

This setup evaluates method performance under both additive and synergistic exposure scenarios common in environmental health research.

5.3.1 Linear additive exposures

Let us begin by comparing the joint effect estimates in Figure C.2. As mentioned in the previous section, a multiple pollutant linear model used for joint effect estimation does not necessarily suffer from multicollinearity. This can be observed in terms of confidence interval (CI) width, OLS performs well in comparison to other methods. We note that the frequentist shrinkage methods exhibit high bias and result in the largest CI width. As these methods are developed for stabilising individual effect estimates, they perform very poorly in the context of joint effects. In contrast, the Bayesian shrinkage methods are less biased and demonstrate smaller widths for the CIs.

The G-computation linear model is unbiased and essentially identical to the multiple linear model, given that the underlying true mechanisms are linear and additive. For the CI width, bootstrapping was employed, resulting in a slightly larger CI width compared to the multiple pollutant linear model. In contrast, the G-computation methods using random forest and BKMR are more biased. The random forest G-computation produces a small CI. Conversely, BKMR yields a larger CI due to its flexibility. The various types of weighted index models are slightly biased but have the smallest confidence interval width. In weighted index models, a notable difference exists between using one repeated holdout (rh) versus using 100 repeated holdouts. The models that incorporate 100 repeated holdouts provide more stable estimates, leading to smaller CIs. This stability arises because the CIs are based on simulations from the 100 repeated holdouts.

Secondly, the individual effects are worth examining. Figure C.3 presents inclusion probabilities that allow us to compare the relative importance of the different effects. For all methods, the strongest effect (PFBA) shows the highest poster inclusion probability (PIP) or bootstrap inclusion probability (BIP). However, the relative importance of the other effects is less clearly defined. This uncertainty may stem from high pairwise correlations, making it challenging to accurately assign the correct effects to their respective components. This issue will be further investigated in subsection 5.3.3.

The performance measures for the exposures related to the individual effect of PFOA (total) are shown in Figure 5.3. This effect was selected due to its strong correlation and moderate influence on the outcome. First, note that the single pollutant linear model is biased due to high correlation with other components. Additionally, the multiple pollutant linear model reveals a high relative CI width, indicating potential issues with multicollinearity. In frequentist shrinkage approaches, the effects tend to be biased downwards, which is a common characteristic of these methods. This results in the smallest overall CIs. Specifically, Ridge regression performs best in terms of achieving the smallest CI width and, consequently, the highest power. On the other hand, Bayesian shrinkage methods exhibit less bias. Bayesian model averaging yields similar results to a multiple linear model, indicating it potentially suffers as well from multicollinearity. Lastly, the repeated holdout weighted index models perform well in terms of CI width. The bias across the three variants (single holdout, random subset, repeated holdout) was similar, but less variability was observed for repeated holdout. As noted earlier, these models are only comparable in relation to other components in the weighted sum, and thus statistical power is not reported. In conclusion, these findings are observed in settings with dense effects. Therefore, methods like BMA, and spike and slab regression may perform better in sparser settings. This will be investigated in subsection 5.3.3.

5.3.2 Linear synergistic exposures

In the case of a linear additive effect, a multiple pollutant linear model is the most effective for estimating joint effects. However, we need to consider how realistic this scenario is in practice. Therefore, we also compare methods under a linear synergistic scenario. It is important to note that a significant contribution from interaction effects should favour the performance of random forest and BKMR models, as other methods are misspecified in this context. Given our belief that a moderately weak interaction effect is realistic, we also present the results for this scenario. In Figure C.4, we observe that the multiple pollutant linear model and BMA are clearly biased due to their misspecifications. The random forest and BKMR now demonstrate comparable levels of bias when compared to other methods. However, we refrain from offering further interpretations, as the conclusions drawn heavily depend on the simulation procedures used.

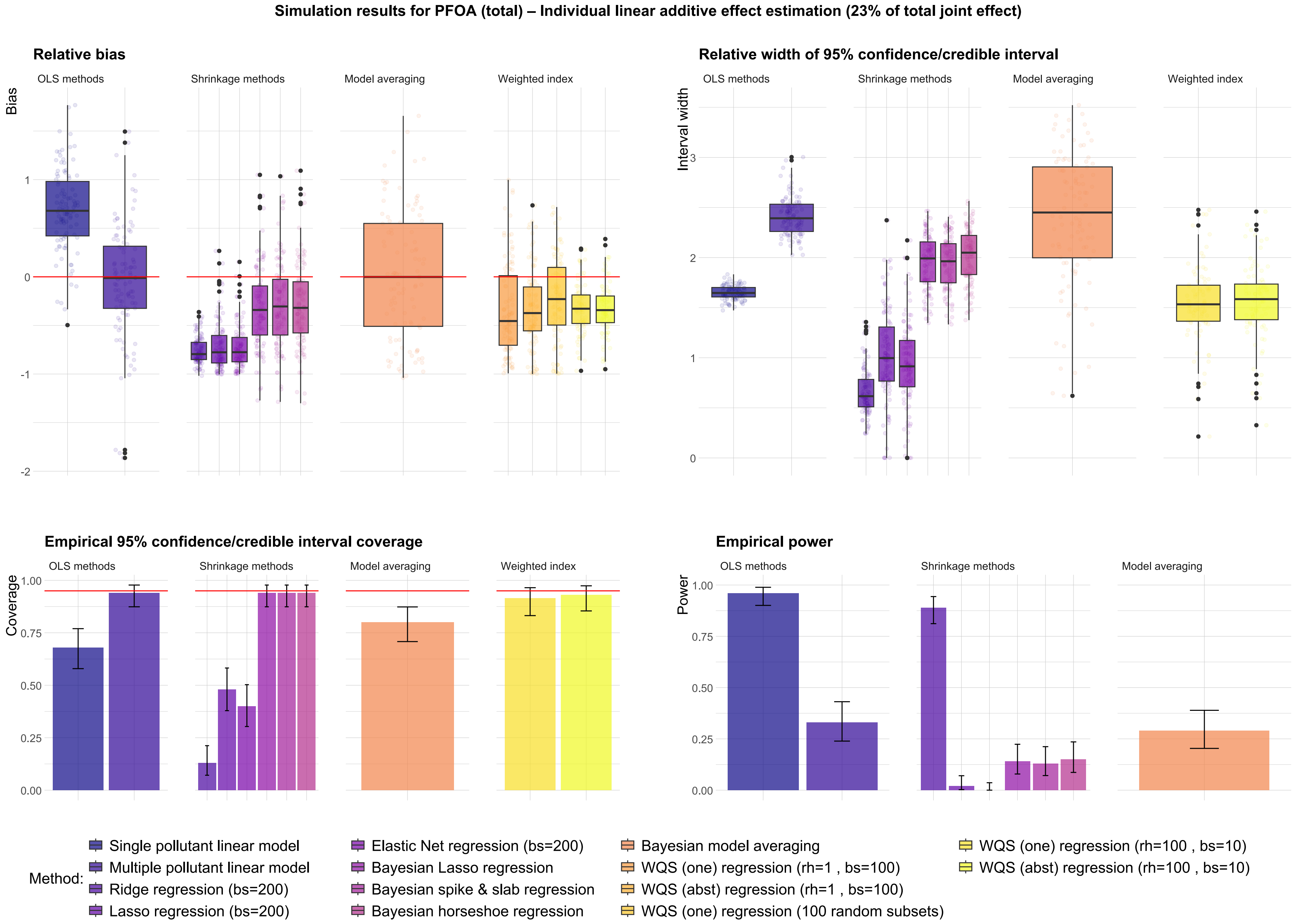


Figure 5.3: *Comparison of different methods based on relative bias, CI width, CI coverage and power for PFOA (total). All exposures had a linear additive effect on the outcome. Simulations are based on a sample size of $n = 300$ repeated $n_{sim} = 100$ times; more details can be found in Appendix D.3.1. "One" is the simple average of weights and "abst" uses the absolute t-statistic as a weight for averaging.*

5.3.3 Evaluating grouping behaviour

In Chapter 3, the grouping effect was discussed in the context of Ridge, LASSO and Elastic Net regression. Here, we further investigate the ability of various methods to exhibit the grouping property, defined as the tendency to assign similar coefficients to highly correlated predictors. In light of the critique by Carrico et al. (2015), who questioned the appropriateness of this property in certain settings, we also examine a scenario in which only one of the two highly correlated predictors has a true effect on the outcome.

To reflect the real-world structure of the case study data, we preserve the joint distribution of the two highly correlated exposures, PFNA, PFDA and the additional covariates by sampling entire exposure profiles (rows) with replacement. In contrast, the additional five exposures are sampled independently, breaking any correlation among them and added as uncorrelated nuisance parameters. We simulate outcomes using standardised exposures and a weighted continuous sum formulation, considering two scenarios:

- Equal effects, where PFNA and PFDA each contribute equally to the outcome
- Unequal effects, where only PFDA is assigned an effect while PFNA acts as a highly correlated nuisance variable

This design provides a framework to evaluate the capacity of each method to appropriately attribute individual effect sizes to correlated exposures, depending on whether their true contributions are equal or unequal. Moreover, this can be seen as a sparse setting in which there are only two true effects (equal effects) or one true effect (unequal effects). All simulation details can be found in Appendix D.3.2.

Correlated predictors of equal effect

Let us start by comparing the BIP for LASSO and Elastic Net in Figure C.5. Both methods exhibit high BIPs (close to 1) for the two correlated predictors (PFDA & PFNA). However, inclusion probability alone does not fully reflect the strength or stability of the estimated effects. In contrast, when we look at the associated power in Figure C.6, we do see a difference. Ridge and Elastic Net have similar power and bias for both components, indicating that they assign similar coefficient values. LASSO shows lower power overall and notably a larger discrepancy in power between PFDA and PFNA. This difference can be attributed to LASSO's tendency to select only one variable from a set of highly correlated predictors.

Based on the relative bias and power in Figure C.6, the Bayesian shrinkage methods seem to assign similar coefficients to both effects. This also holds for the PIP of the spike and slab regression in Figure C.5, which is close to 1 for both components. Similarly, multiple pollutant linear regression assigns equal coefficients to both predictors. This is also true for WQS regression. However, the true effect is clearly underestimated in this case. This underestimation arises from the constraint that the weights must be non-negative and sum to one, which forces the model to assign small weights to the nuisance components, thereby biasing down the true large effect.

BMA seems to have instability shown by the large variability in estimates in the top left plot in Figure C.6. This was also observed by the PIP in Figure C.5. For PFNA, this is almost one, while for PFDA, there is clearly more variability. In the case of BKMR, the PIP in Figure C.5 shows large variability. The model may alternate between assigning moderate PIP to one variable or to its correlated neighbour. This leads to PIPs that fluctuate between simulations. Finally, as expected, the single-pollutant method exhibits bias because it estimates an indirect effect for both components.

In conclusion, Ridge, Elastic Net and Bayesian shrinkage methods performed well, consistently assigning biased stable effect estimates to both correlated predictors. In contrast, BMA and BKMR showed poorer performance. Large uncertainty was observed in the estimated effects, suggesting difficulties in reliably attributing effects to both predictors.

Correlated predictors of unequal effect

We can start by comparing the PIP and BIP shown in Figure C.7 for the various variable selection methods. For PFDA, all the inclusion probabilities are close to 1, which aligns with the true underlying process. However, the component PFNA, which is highly correlated with PFDA but does not have an effect on the outcome, is particularly interesting. For PFNA, Elastic Net regression behaves quite differently, showing a median inclusion probability of around 0.5. This can be explained by the grouping effect mentioned earlier. In contrast, the other methods tend to have low inclusion probabilities, indicating their effectiveness in handling this specific situation. When we compare the nuisance parameters, we find that they also exhibit low probabilities.

Figure C.8 presents all estimates and allows for comparison with methods that do not explicitly perform variable selection. First, it is important to note that absolute bias was assessed since PFNA has no effect. Therefore, we should be cautious when comparing this to the weighted index methods, as they operate on a different scale. As expected, the single pollutant estimate for PFNA is highly biased due to the strong correlations among the two components, which resulted in high power for both components. In contrast, the shrinkage methods demonstrate high power for PFDA, with only ridge regression showing high power for PFNA. This issue arises from the grouping property, which is problematic in this case. Model averaging does not encounter this issue. Weighted index models, on the other hand, appear to underestimate the true effect of PFDA while incorrectly assigning an effect to PFNA that does not exist. The model may spread the weight across correlated predictors, leading to an underestimation of truly important variables. This was also observed by Carrico et al. (2015).

In conclusion, the methods that performed poorly in the previous setting (equal effects) show improved performance here. In particular, BMA and BKMR yield more stable results. The Bayesian shrinkage methods performed well in both settings. Therefore, in the next chapter, Bayesian horseshoe regression will be applied for individual effect estimation along with Ridge regression due to its high statistical power. In addition, repeated holdout WQS regression and BKMR will be included for illustrative purposes. To estimate joint effects, OLS regression will be used, as it demonstrated good performance for linear additive or weak interaction effects. A random forest-based G-computation approach will also be explored, as it performed well in the synergistic simulation study.

Chapter 6

A case study: PFAS exposure and immunometabolic health

The “Teenager HBM Study - 3M site” was introduced in Chapter 2. In this chapter, we primarily focused on the characteristics of the exposures to inform our discussion on shrinkage and multi-pollutant methods. Our current objective is to analyse two immune biomarkers using some of the models introduced earlier and translate the results into interpretable and meaningful effects.

The guidelines from the original study conducted on this dataset will be followed. The original dataset included 303 adolescents aged 12 to 17 years living within 5 kilometres of the 3M site in Zwijndrecht. Consortium UAntwerpen, VITO, PIH, UHasselt and VUB (2023) gave an overview of exclusion parameters. Adolescents taking growth hormones ($n = 7$), medication for thyroid disease ($n = 1$), diabetes medication ($n = 2$), kidney disease medication ($n = 1$) or intake of a high dose of cortisone ($n = 1$) were excluded. This resulted in a sample size of 289. Among the remaining 289 adolescents, a complete case analysis was performed due to missing values in PFAS exposures ($n = 2$), confounding variables ($n = 29$) or outcome ($n = 2$ or 3). Exposure values below the LOQ were not considered missing as they were imputed (see Section 2.2). This resulted in a final sample size for analysis of 259 for the CD4+/CD8+ T-cell ratio and 260 for the leukocyte counts.

The study conducted by Consortium UAntwerpen, VITO, PIH, UHasselt and VUB (2023) provides a comprehensive list of confounding variables to be included in the analysis. These variables are: gender (binary), age (trichotomous), financial stability with the income (trichotomous), BMI (trichotomous), birth weight (binary) and exposure to tobacco smoke at home, elsewhere or through own smoking (binary). The exact levels and distribution are given in Table E.1. The two outcomes, CD4+/CD8+ T-cell ratio and leukocyte counts, introduced in Section 2.2 were log-transformed to ensure approximate normality (see Figure E.2 for the distribution). Scatter plots of the log-transformed outcome versus each exposure, stratified by intervals of a second exposure, reveal no substantial deviations from parallel trends, suggesting limited evidence of potential interaction effects. Additionally, a LASSO regression model was employed that included all two-way interactions. We calculated the bootstrap inclusion probabilities (BIP) to determine which two-way interactions could enhance the model’s fit. All BIP values were below 0.5, indicating that any potential interaction effects may be weak or obscured by noise (see Table E.2). This suggests that interactions between exposures are minimal and can be ignored when estimating individual exposure effects.

First, we will begin by identifying the components that significantly influence the outcome using Ridge regression and Bayesian horseshoe regression. The general structure for these models is defined as follows:

$$\log(y_i) = \beta_0 + \sum_{j=1}^{7} \beta_j \left(\frac{\log(a_{ij}) - \mu_{\log(\mathbf{a}_j)}}{\sigma_{\log(\mathbf{a}_j)}} \right) + \mathbf{z}_i^\top \boldsymbol{\varphi} + \epsilon_i \tag{6.1}$$

where y_i is the outcome for adolescent $i = 1, ..., 259$ or 260, β_0 is the intercept, β_j is the effect for each exposure $j = 1, ..., 7$, $\mu_{\log(\mathbf{a}_j)}$ and $\sigma_{\log(\mathbf{a}_j)}$ represent respectively the mean and standard deviation of the log transformed exposure j, $\mathbf{z}_i$ represents additional covariates with coefficients $\boldsymbol{\varphi}$ and $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ is the error term. The evaluation of model assumptions and the convergence of MCMC, when applicable, were conducted initially. For a detailed account of the procedures utilised, we refer to Appendix E. Next, effect sizes were expressed as the average multiplicative change in the outcome Y per interquartile fold change in exposure concentration. This means the ratio of the expected outcome (see (E.1)) when the exposure is at the 75th percentile (Q3) compared to the 25th percentile (Q1). All exposures and joint effects will be described with their 95% confidence or credible intervals. Next, exposure effects will be interpreted using PIP from BKMR and by interpreting relative weights from repeated holdout WQS regression. Following this, to assess the joint effect, we will employ a multiple pollutant linear model and a random forest G-computation approach.

6.1 Leukocyte count

Figure 6.1 visualises the influence of various exposures on leukocyte counts based on Ridge and Bayesian horseshoe regression. Their counterparts that do not perform shrinkage, OLS or a flat prior, have also been added to facilitate interpretation. PFOS, PFDA, PFHXS (total) and branched PFOS do not have a significant impact on the outcome, as their 95% confidence or credible intervals include one. In contrast, PFNA, PFBA and PFOA (total) exhibit borderline significance, showing weak evidence of an effect. An increase in PFBA from 0.1 μg/L (Q1) to 0.19 μg/L (Q3) is associated with an average leukocyte count decrease of 3.0% ([0.0 : 5.7]; 95%CI) for ridge regression and 2.2% ([-0.4 : 7.5]; 95%CI) for horseshoe regression. Note that the effect sizes are biased downwards and therefore should be interpreted with caution. These findings, based on a model with an R^2 of approximately 0.3, suggest cautious interpretation due to modest explanatory power and the potential presence of unmodelled factors.

Next, two repeated holdout WQS regression models were fit, one with a negative association with the outcome and one with a positive. Carrico et al. (2015) advise first testing the joint effect, as weights cannot be interpreted otherwise. Fitting both a positive and a negative WQS model to the same data involves conducting two hypothesis tests. To address this issue, a Bonferroni adjustment was applied, resulting in an insignificant test outcome. Finally, a BKMR model with variable selection was employed. All PIPs were found to be (close to) zero. Therefore, as an illustration, we refitted the model without variable selection, which will increase the risk of overfitting. We added Figure E.4 to visualise the multiplicative effect on the leukocyte count for three exposures, identified by Ridge regression. These plots provide a view of the direction, but show significant uncertainty and cannot be used to draw conclusions from.

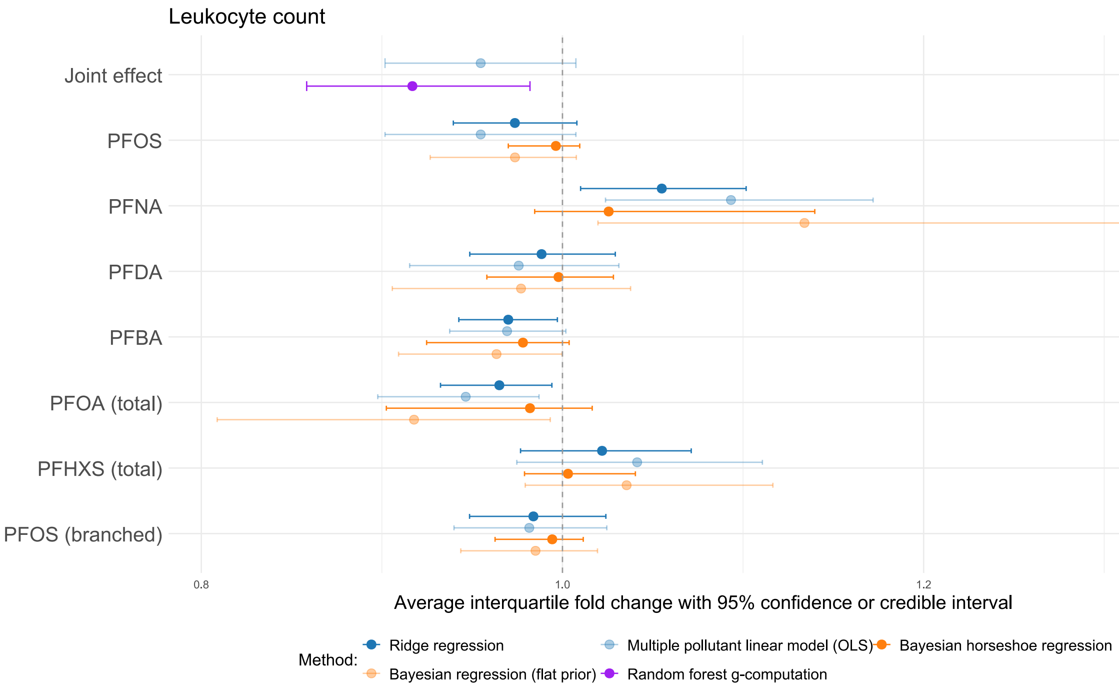


Figure 6.1: *Average interquartile fold change effect for each PFAS exposure on the number of leukocytes. Confidence intervals were calculated as percentiles of 2000 bootstrap samples and equal-tailed credible intervals were constructed based on the posterior distributions.*

Finally, G-computation with random forest was used to estimate the population-averaged multiplicative effect of a shift in all exposures simultaneously from low to high concentrations. This was quantified as the multiplicative effect of whether an adolescent had PFAS exposure levels at the 25th percentile against the 75th percentile. This corresponds to an increase in PFOS (branched) from 2.6 μg/L to 6.8 μg/L, PFHXS (total) 0.38 μg/L to 0.88 μg/L, PFOA (total) 0.92 μg/L to 1.5 μg/L, PFBA 0.095 μg/L to 0.19 μg/L, PFDA 0.096 μg/L to 0.20 μg/L, PFNA 0.19 μg/L to 0.33 μg/L and PFOS 1.4 μg/L to 5.4 μg/L.

Increasing all exposures from Q1 to Q3 resulted in an average population-specific decrease in leukocyte count of 8.3% with corresponding 95%CI [1.8% : 14.2%] based on percentiles from 2000 bootstrap samples. G-computation has the objective of obtaining causal effects instead of associations. Therefore, in Appendix B we checked the validity of the different assumptions. It is clear that some of these assumptions cannot be fully verified. Furthermore, the cross-sectional design raises concerns about temporal ordering, complicating causal attribution. As a result, we interpret the findings as associations, recognising that unverified assumptions and study design limitations may prevent definitive causal inference.

The joint effect was also calculated using a multiple pollutant additive linear model (OLS). An increase in all exposures from Q1 to Q3 resulted in an average decrease of 4.5% in leukocyte count with 95%CI [-0.7% : 9.8%] based on percentiles from 2 000 bootstrap samples. It is important to note that G-computation estimates the effect conditional on the study population's covariate distribution. In contrast, the joint effect of a multiple pollutant additive linear model does not depend on the covariate distribution in the sense that it describes the effect of an exposure at any fixed level of a covariate. Therefore, we should be careful with comparing these two effect estimates.

6.2 CD4+/CD8+ T-cell ratio

Figure 6.2 visualises the influence of various exposures on CD4+/CD8+ ratio based on Ridge and Bayesian horseshoe regression. Their counterparts that do not perform shrinkage, OLS or a flat prior, have also been added to facilitate interpretation. The analysis was done analogously to the leukocyte counts. None of the exposures have a significant impact on the outcome, as their 95% confidence or credible intervals include one. Next, two repeated holdout WQS regression models were fit, one with a negative association with the outcome and one with a positive. Both effects were found to be insignificant. Therefore, the weights are not interpreted. Finally, a BKMR model with variable selection was employed to balance model flexibility and prevent overfitting. All PIPs were below 0.1, suggesting that none of the exposures are strong predictors.

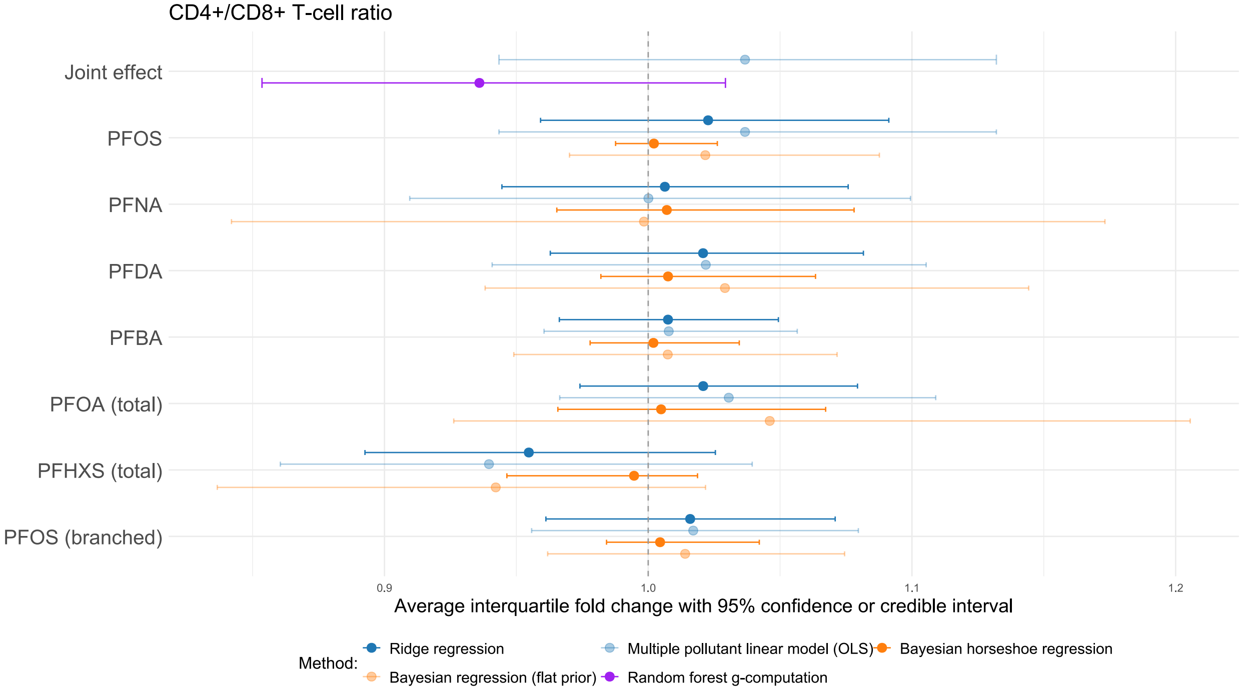


Figure 6.2: *Average interquartile fold change effect for each PFAS exposure on the CD4+/CD8+ T-cell ratio. Confidence intervals were calculated as percentiles of 2000 bootstrap samples and equal-tailed credible intervals were constructed based on the posterior distributions.*

A random forest-based G-computation approach was used, similar to the analysis conducted on leukocytes. When all PFAS exposure levels were simultaneously increased from their 25th to 75th percentiles, there was an average population-specific decrease of 6.4% in the CD4+/CD8+ T-cell ratio. However, this estimate came with a high degree of uncertainty, as indicated by a 95% confidence interval ranging from -2.9% to 14.6%, based on 2 000 bootstrap samples. In contrast, a traditional additive linear multiple pollutant regression model showed an average increase of 3.7%, with a 95% bootstrap-confidence interval of -5.7% to 13.2%. Overall, both approaches yielded imprecise and directionally inconsistent estimates, suggesting no clear evidence of an association.

Chapter 7

Discussion

This master's thesis began by addressing the problem of multicollinearity, a common issue in environmental health research. This problem was not observed in the past as analyses typically relied on *single-pollutant* models. However, as the field is shifting toward *multi-pollutant* approaches, the issue of multicollinearity has become a key challenge. We discussed this specifically in the context of human biomonitoring studies, where multiple chemicals, such as PFAS, tend to be moderately to strongly correlated. In this context, we want to distinguish between two types of questions. The first type involves assessing the effect of a single chemical on a particular health outcome while controlling for confounding variables and other pollutants. The second type focuses on estimating the overall impact of all pollutants on a specific health outcome. We started by concentrating on statistical models that isolate individual effects. Traditionally, penalised regression techniques have been employed in these contexts, such as Ridge, LASSO and Elastic Net regression.

LASSO is known to perform well in a sparse setting or if the number of predictors is large relative to the sample size ($n \approx p$ or $n < p$) (Tibshirani, 1996). This is not the case in human biomonitoring studies as they typically have $5, ..., 25$ exposures and a sample size of about $250, ..., 1000$. Moreover, LASSO does not have the grouping property (Zou and Hastie, 2005). In contrast, Ridge regression does have the grouping property and is designed to specifically address the problem of multicollinearity (Hoerl and Kennard, 1970). However, this comes at the cost of bias. Due to the strong correlation structure, we observed that the bias was relatively large in our simulation studies. Therefore, the interpretation of the coefficients may not be straightforward. In our case study, we addressed this issue by examining the coefficient paths across a range of penalty values. This allowed us to select models with only moderate shrinkage. In this way, we accepted a small amount of bias in exchange for greater stability and better interpretation.

The use of bootstrap inclusion probabilities (BIPs) in Elastic Net models can provide an alternative measure of variable importance to classical coefficients. In a simulation study involving two highly correlated predictors with equal effects on the outcome, both variables consistently received BIPs close to 1. In contrast, when only one of the two predictors had a true effect, the variable without an effect often received different BIPs across simulations. This was not the case for LASSO regression, as it does not have the grouping property. This illustrates the fundamental problem of two highly correlated predictors. From a statistical perspective, there is no model that can distinguish between such predictors, as they have nearly identical information. In the context of human biomonitoring studies, it seems more plausible that both exposures have the same effects on the outcome, as many of these chemicals share the same molecular properties. Therefore, models that exhibit the grouping property, such as Ridge regression, are more appropriate in our context.

The second type of models discussed were Bayesian shrinkage methods, which use a prior distribution to allow estimates to shrink toward zero. One of the advantages over frequentist shrinkage techniques is the posterior distribution for each parameter. This allows us to reflect on the uncertainty in the form of credible intervals. This is less straightforward when using LASSO or Elastic Net, where bootstrap confidence intervals may fail. Three different shrinkage priors were discussed: the Laplace (LASSO) prior, the spike & slab prior and the horseshoe prior. In comparison with frequentist shrinkage methods, we observed less bias, which facilitates interpretation but at the cost of wider credible intervals. The specific prior choice was less important in our simulation studies. We observed only minor differences between them in terms of power, bias, CI coverage and width.

From a theoretical perspective, each prior has its advantage. In our context, the Horseshoe prior seems most appropriate. It applies adaptive shrinkage, allowing strong signals to remain unshrunk while heavily shrinking noise. Beyond the specific choice of the prior, the main advantage lies in the flexibility of the Bayesian framework. It is particularly valuable for future research, as it allows for extensions. These include modifying the outcome distribution, incorporating random or spatial effects, or using spline-based models.

Thirdly, a highly popular method known as weighted quantile sum (WQS) regression has been introduced. This method has gained significant popularity over the past five years (Yu et al., 2022). Our focus was mainly on repeated holdout validation WQS regression, which has demonstrated more stability from a theoretical perspective, but also based on the simulation study. One of the main questions that often remains unanswered is the impact of using quantiles. The argument previously given in single partition WQS regression is that quantiles stabilise the influence of outliers or heavy-tailed skewed exposure distributions. Our simulation study looked at three potential exposure distributions. In terms of estimated weights, we observed little difference between quartiles, deciles or continuous exposures across the three distributions. However, with respect to power for detecting an overall effect, there was a small improvement when using continuous exposures. Thus, the repeated holdout validation procedure in WQS regression appears to be relatively robust to outliers and heavy-tailed distributions. Therefore, we argue in favour of using standardised log-transformed continuous exposures, as this simplifies comparison with other methods and yields a small gain in statistical power.

A second important remark concerns the constraint imposed on the weights in WQS regression. While the primary goal of WQS is to identify key pollutants that contribute most to the overall mixture effect, the estimated weights can only be interpreted relative to one another. No formal statistical inference is available for individual weights. Carrico et al. (2015) recommended applying a threshold to determine whether a weight is considered influential. However, this cutoff point or threshold is arbitrarily chosen and lacks a strong theoretical foundation. Moreover, the simulation study revealed an underestimation of large effects and an overestimation of small effects. This discrepancy can be attributed to the constraints imposed on the weights.

Fourthly, the idea of Bayesian model averaging with Bayesian adaptive sampling was introduced. The Bayesian adaptive sampling algorithm is a powerful tool when the model space is rather large. In our context, the model space was computationally tractable and thus there is no need for such an algorithm. Moreover, in our simulation study, Bayesian model averaging based on Bayesian linear models performed poorly in terms of credible interval width. We also observed substantial uncertainty in PIPs, particularly in the scenario with two highly correlated predictors of equal effect. This increased uncertainty likely comes from the model uncertainty in BMA, which averages over models that include or exclude each predictor. Given our aim to retain all predictors in the model and the limited number of exposures considered, BMA as presented by Clyde et al. (2011) is not well-suited for this setting.

The final model used for estimating individual effects was Bayesian kernel machine regression (BKMR). This method has gained popularity due to its ability to account for non-linearities and interactions. However, our relatively small sample size limited our ability to fully benefit from these advantages, increasing the risk of overfitting. As a consequence, BKMR showed large uncertainty in PIPs. Although its high flexibility is a strength, it also reduces interpretability. In our case study, we tried to analyse exposure effects using graphical tools. While these plots offered a general sense of direction, the substantial uncertainty around the estimates hindered our ability to draw conclusions.

The second question concerns estimating the overall effect of a mixture. This means the effect on the outcome if all exposures increase simultaneously. We began by demonstrating that the joint effect is less affected by multicollinearity than the individual effect estimates. This occurs due to variance cancellation, supported both by a simple theoretical illustration and by simulation results. In our simulations, the multiple pollutant linear model (OLS) was shown to be unbiased and had narrower confidence intervals than many alternative methods. Based on this, penalised methods appear unnecessary in this context, as they often produce similar or even wider intervals with large bias.

The method of greater interest was the G-computation approach, a technique from causal inference used to estimate causal effects under hypothetical interventions. Its main advantage lies in the ability to use flexible modelling techniques that naturally account for non-linearities and interactions, thus capturing potential synergistic effects. However, this flexibility comes at the cost of possible overfitting. In our simulation study, random forest G-computation showed substantial bias. Random forest produced narrow confidence intervals, while BKMR yielded wide intervals, reflecting considerable uncertainty. Although the goal of G-computation is to estimate causal effects, this was not achieved in our case study. Several causal assumptions could not be fully verified. Therefore, we refrain from making causal claims and instead interpret the results as associational estimates.

Finally, a case study was conducted to reveal associations between PFAS exposures and two immune biomarkers. For leukocyte counts, both Ridge and Horseshoe regression identified PFBA as borderline significant. Based on Ridge regression, an increase in PFBA from 0.1 μg/L (Q1) to 0.19 μg/L (Q3) is associated with an average leukocyte count decrease of 3.0% ([0.0 : 5.7]; 95%CI). The joint effect, estimated using G-computation, was found to be significant. Increasing all exposures from Q1 to Q3 resulted in an average population-specific decrease in leukocyte count of 8.3% with corresponding 95%CI [1.8% : 14.2%]. However, as the individual exposure effects varied in positive and negative directions, these may partially cancel each other out. Therefore, the joint effect should be interpreted with caution but highlights an important direction for future research. No associations were found for the CD4+/CD8+ T-cell ratio, possibly due to the limited sample size and the substantial unexplained variability.

In conclusion, Table 7.1 provides a comprehensive overview of all the methods discussed in this master's thesis. It outlines the type of method and indicates which questions each method is particularly suited for, distinguishing between individual and joint effects. The advantages and disadvantages presented are drawn from theoretical arguments outlined in the literature review in Chapter 3 and Chapter 4, as well as from the results of simulation studies discussed in Chapter 5.

Limitations

Let us begin by discussing the limitations of the simulation study. First, the simulation design was closely aligned with the case study in terms of sample size and the number of exposures. While this alignment increased the relevance of the findings, it also restricted the generalisability of the results. In larger sample size settings, more flexible methods, such as Bayesian kernel machine regression (BKMR), may perform better and provide more stable estimates. Additionally, some human biomonitoring studies involve a greater number of exposures, which could affect the relative performance of the methods, particularly those designed for high-dimensional settings.

We limited the data-generating mechanism to a linear additive effect with the assumption of directional homogeneity. While this choice benefits less flexible methods and allows for straightforward interpretation, it does not capture the complexity often seen in real-world exposure scenarios. Finally, each simulation scenario was repeated only a limited number of times due to the high computational cost of Bayesian methods and bootstrapping. We were also unable to conduct full convergence diagnostics for all Bayesian methods in every run. Although we manually inspected the convergence in a subset of runs, we cannot ensure that all Bayesian models converged adequately across all iterations.

Several methodological limitations of the case study should be acknowledged. First, exposure values below the limit of quantification were imputed without considering the correlation structure between exposures. Second, single imputation was used rather than multiple imputation, limiting our ability to reflect uncertainty in the imputed values. We also limited ourselves to exposures with sufficient data (60% $>$ LOQ) and ignored all other measured PFAS exposures. Third, missing data in covariates were handled using a complete case analysis, which assumes data are missing completely at random. Fourth, some adolescents lived in the same household ($n = 41$), introducing potential clustering effects that were not accounted for in the analysis. Finally, the cross-sectional design of the study restricts causal interpretation, as exposures and outcomes were measured simultaneously.

Table 7.1: *Overview of statistical methods for mixture analysis*

Type of method	**Method**	**Effect type**	**Pros**	**Cons**	**Interpretation**
Ordinary least squares	Single pollutant linear model	Individual	Easy to interpret effect estimates if implemented linear additively	Highly biased due to omission of correlated causal pollutants	Unadjusted associations for other pollutants
	Multi pollutant linear model			Large uncertainty due to highly correlated pollutants	Associations conditional on other pollutants
Shrinkage methods	Ridge regression	Individual	Handles multicollinearity well and has the grouping property	Careful penalty tuning to avoid over-shrinkage	Biased downwards associations conditional on other pollutants
	LASSO regression		Handles sparse settings and performs variable selection	Does not have grouping property and no uncertainty	Biased downwards associations conditional on other pollutants, BIP as a measure of variable importance
	Elastic Net regression		Variable selection and grouping property	No uncertainty quantification	
	Bayesian LASSO prior		Uncertainty quantification via posterior	Computationally slower and risk of over/under-shrinkage	Biased downwards associations conditional on other pollutants
	Horseshoe prior		Uncertainty quantification via posterior, performs adaptive shrinkage	Computationally slower	
	Spike and slab prior		Uncertainty quantification via posterior, works well in sparse settings	Computationally slower and sensitive to hyperpriors	Biased downwards associations conditional on other pollutants, PIP as measure of variable importance
G-computation	Linear model	Joint	Performs well in small sample sizes	Requires manual inclusion of interactions or non-linearities	Population-specific average causal effect under two exposure scenarios
	Random forest		Captures complex interactions and non-linearities	Risk of overfitting in smaller sample sizes	
Model averaging	Bayesian model averaging with Bayesian adaptive sampling	Individual	Computationally efficient for many pollutants (> 25)	Large uncertainty due to highly correlated predictors	Model-averaged associations across subsets of pollutants, PIP expresses how likely a pollutant is to be included in a model
Weighted index models	Bootstrap WQS regression	Individual Joint	Computationally efficient	No theoretical framework for weight inference, assumes directional homogeneity	Joint exposure effect as a weighted sum in a particular direction, with weights indicating each chemical's relative importance
	Random subset WQS regression		Computationally efficient for large number of predictors		
	Repeated holdout WQS regression		Provides weights uncertainty and performs well in smaller sample sizes	Computationally demanding, assumes directional homogeneity, no theoretical framework for weight inference	
Kernel regression	Bayesian kernel machine regression	Individual Joint	Captures complex interactions and non-linearities	Computationally demanding and large uncertainty in small sample sizes	PIPs indicate variable importance, effects visualised trough plots

Future research

The previously discussed limitations offer several directions for future research. First, the current simulation study could be extended to more exposures ($p \approx 25$), a larger sample size ($n \approx 1000$) and more complex exposure-response structures. In order to achieve this, we will need to utilise high-performance computing resources. The case study could be improved by adopting a more rigorous approach to account for censoring in the exposure data. A more robust strategy would involve multiple imputation, incorporating both the correlation structure among exposures and the relationship with covariates. Alternatively, a fully Bayesian approach could be implemented, treating censored exposure values as unknown parameters and assigning them informative priors to appropriately reflect the uncertainty below the LOQ. A sensitivity analysis evaluating extreme scenarios where censored values are imputed as either zero or at the LOQ may help define a plausible range for effect estimates. The same issue applies to missingness in the covariates, where imputation methods are more appropriate than relying on complete case analysis. As within-household correlation was ignored, we could easily account for it in a Bayesian context by adding a random effect. However, in the case of the other methodologies, it is less clear how to extend the methods.

The previously discussed extensions are most naturally implemented within a Bayesian framework and relate closely to the case study. Looking ahead, key priorities include adapting models to accommodate other outcome distributions and incorporating binary exposures. This is particularly relevant when a substantial proportion of exposure measurements fall below the LOQ. Furthermore, given the substantial amount of unexplained variability observed, accounting for spatial clustering or intra-household correlation could enhance model fit. A spatial component may capture the influence of unmeasured environmental factors, such as industrial pollution, water quality or air pollution. However, implementing such spatial models typically requires a sufficient sample size to support reliable inference. Finally, alternative methods such as Bayesian additive regression trees (BART), regularised horseshoe regression or partial least squares (PLS) could be explored in this context (Chipman et al., 2010; Wold et al., 1984; Piironen and Vehtari, 2017).

This thesis presented a case study concentrating on immune-related biomarkers as outcome. However, the ultimate aim is to perform a full mediation analysis. At VITO Health and PARC, the focus is to explore how immune and inflammation parameters mediate the relationship between PFAS and various health outcomes. Achieving this objective necessitates the development of methodologies capable of accommodating multiple pollutants and multiple mediators in order to investigate these complex causal pathways. This master's thesis marks the initial phase of this research, as it explores statistical methods for characterising the relationship between multiple PFAS exposures and immune biomarkers.

Chapter 8

Conclusion

This master's thesis explored statistical methods for estimating both individual and joint effects of correlated environmental exposures, with a particular focus on PFAS in human biomonitoring studies. We addressed two main questions: (1) how to identify individual pollutants in a mixture and (2) how to estimate the overall effect of a pollutant mixture on a health outcome.

For individual effect estimation, we found that multicollinearity poses the main challenge. Among frequentist shrinkage methods, Ridge regression was best suited for this context due to its grouping property and ability to handle multicollinearity. Bayesian shrinkage methods offered improved interpretability via posterior distributions and reduced bias, though at the cost of wider credible intervals. Among the priors considered, the horseshoe prior was most appropriate in our setting due to its adaptive shrinkage properties. We also assessed specific methods designed for analysing chemical mixtures. The two most popular methods are weighted quantile sum (WQS) regression and Bayesian kernel machine regression (BKMR). WQS was originally designed for quantile-transformed exposures. In contrast, our results suggest that using continuous exposures improves statistical power without compromising robustness. However, WQS is not optimal for individual effect estimation as it lacks formal inferential support. BKMR did not perform well in our setting, potentially due to the small sample size.

Regarding joint effect estimation, we demonstrated that multicollinearity has a reduced impact on joint standard errors due to variance cancellation. G-computation in combination with a flexible method offers the potential to capture synergistic effects within mixtures. Our simulation results indicated that G-computation with flexible learners like random forest may suffer from overfitting, especially in small samples. Additionally, in the case study, the necessary assumptions for causal interpretation of the G-computation results were not met.

In summary, this thesis offers an overview of modern methods for environmental mixture analysis, highlighting the trade-offs between interpretability, flexibility and statistical performance.

Appendix A

Software details & AI tools

The software used for data analysis, simulation results and data visualisation was R version 4.4.2 for Windows (R Core Team, 2024). Table A.1 provides an overview of the different packages utilised in this thesis apart from the base functions in R. All R code used in this thesis is available on GitHub at this link. I acknowledge using Grammarly (Grammarly Inc., 2025) for grammar and rephrasing suggestions, as well as ChatGPT (OpenAI, 2023) for assistance in rephrasing and troubleshooting coding. All outputs were critically evaluated, reviewed and edited by me.

Table A.1: *An overview of the various packages utilised in this thesis.*

	Package (version)	Citation
Data manipulation	readxl (1.4.3)	Wickham and Bryan (2023)
	dplyr (1.1.4)	Wickham et al. (2023)
	tidyverse (2.0.0)	Wickham et al. (2019)
	reshape2 (1.4.4)	Wickham (2007)
	caret (7.0-1)	Kuhn and Max (2008)
Data visualisation	ggplot2 (3.5.1)	Wickham (2016)
	hrbrthemes (0.8.7)	Rudis (2024)
	viridis (0.6.5)	Garnier et al. (2024)
	patchwork (1.30)	Pedersen (2025)
	ggpubr (0.6.0)	Kassambara (2025)
Simulation study		
Parallel computing	foreach (1.5.2)	Microsoft and Weston (2022b)
	doParallel (1.0.17)	Microsoft and Weston (2022a)
Data-generating mechanism	MASS (7.3-64)	Venables and Ripley (2002)
	mvtnorm (1.3-3)	Genz and Bretz (2009)
Multi-pollutant methods	gWQS (3.0.5)	Renzetti et al. (2023)
	BAS (1.7.5)	Clyde (2024)
	nimble (1.3.0)	de Valpine et al. (2017)
	coda (0.19-4.1)	Plummer et al. (2006)
	MCMCvis (0.16.3)	Youngflesh (2018)
	glmnet (4.1-8)	Friedman et al. (2010); Tay et al. (2023)
	bkmr (0.2.2)	Bobb et al. (2018)
	randomForest (4.7-1.2)	Breiman et al. (2022)

Appendix B

Causal assumptions

B.1 Counterfactuals and causal inference

Let us start by introducing some essential concepts of causal inference based on the textbook of Hernan and Robins (2025). Causal inference represents a special case of the more general process of scientific reasoning, one in which we try to separate causation from association. To make our understanding of causality suitable for statistical analysis, we will first introduce some specific notation.

Consider the following setting where we have, for simplicity, a binary single exposure A (1: exposed, 0: unexposed) and a continuous outcome variable Y. A key question that a researcher might ask is: What is the causal effect of exposure A on the outcome Y? To answer this question using causal inference, it is necessary to reconstruct a hypothetical framework in which each individual could have been either exposed (1) or not exposed (0). Let $Y^{a=1}$ be the outcome that would have been observed if an individual was exposed and $Y^{a=0}$ if the individual was not exposed. These variables are known as *potential outcomes* or *counterfactual outcomes*. We can now provide a formal definition of a causal effect for an individual: the exposure A has a causal effect on an individual's outcome Y if

$$Y^{a=1} \neq Y^{a=0} \tag{B.1}$$

for that individual. This is known as the *sharp causal null hypothesis* (Hernan and Robins, 2025).

An individual cannot be both exposed and unexposed simultaneously. This issue is often referred to as the *fundamental problem of causal inference*: we can never observe both $Y^{a=1}$ and $Y^{a=0}$ for the same individual. Therefore, it is too ambitious to draw any conclusions about causal effects at the individual level. A more realistic goal is to concentrate on the population-level or average causal effect. This hypothesis is defined as

$$\mathbb{E}[Y^{a=1}] \neq \mathbb{E}[Y^{a=0}] \quad . \tag{B.2}$$

Note that the average causal effect is always equal to the average of the individual causal effects, as it holds that

$$\mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}] = \mathbb{E}[Y^{a=1} - Y^{a=0}] \quad . \tag{B.3}$$

For clarity, we will henceforth refer to "average causal effects" simply as "causal effects".

However, the task remains to select quantities from the observed data that serve as reasonable estimates for the hypothetical quantities $\mathbb{E}[Y^{a=1}]$ and $\mathbb{E}[Y^{a=0}]$.

Let us first simplify this question to a randomised experiment, later on to an observational study. In this case, it can be shown by design (randomised experiment) that the following statement holds.

$$\begin{aligned}\mathbb{E}[Y^{a=1} - Y^{a=0}] =& \mathbb{E}[Y^{a=1}] - \mathbb{E}[Y^{a=0}] && \text{(B.4)}\\ \overset{(a)}{=}& \mathbb{E}[Y^{a=1}|A = 1] - \mathbb{E}[Y^{a=0}|A = 0] && \text{(B.5)}\\ \overset{(b)}{=}& \mathbb{E}[Y|A = 1] - \mathbb{E}[Y|A = 0] && \text{(B.6)}\end{aligned}$$

Step (a), known as *mean exchangeability* or $Y^a \perp\!\!\!\perp A$, states that the exposed group and unexposed group would have experienced the same average counterfactual outcome if they were (not) exposed (either $a = 0$ or $a = 1$). Randomisation is expected to produce *exchangeability*. This means that the exposure allocation is not associated with the mean of the counterfactual outcomes. Step (b), known as *consistency*, states that an individual who was (not) exposed has observed outcome Y equal to his counterfactual outcome $Y = Y^{a=1}$ ($Y = Y^{a=0}$). In conclusion, this means that under these assumptions, the causal effect can be estimated as the difference of the conditional means. (Hernan and Robins, 2025; Thas, 2023)

However, as mentioned before, we are dealing with an observational study and thus the exposure is not randomised among the individuals. The ideal randomised scenario above does not hold, so what do we do? When randomisation is not possible, we need to mimic the conditions of a randomised trial. We call these *identifiability assumptions* in observational studies:

- *Conditional exchangeability* or $Y^a \perp\!\!\!\perp A \mid Z$: This means that, within levels of measured covariates Z, the exposed and unexposed are exchangeable, just as they would be in a randomised trial. The key question is whether Z is the only predictor that is distributed unevenly between the exposed and unexposed groups. Unfortunately, that question must remain unanswered, so we must hope that our expert knowledge guides us correctly to collect enough data so that the assumption is at least approximately true.

- *Positivity* or $0 < P(A = a \mid Z) < 1$: Conditional on covariates Z, there is a probability greater than zero of being assigned to each of the exposure levels. We did not emphasise *positivity* in experimental studies as it is often assumed in those studies.

- *Consistency* or $Y^a = Y$: The definition remains the same as what was previously outlined for experimental studies. To assess *consistency*, we need two things: a clear definition of the counterfactual Y^a and a clear link between counterfactuals and observed outcomes. We must ensure that individuals classified as exposed actually were exposed and likewise for the unexposed. A more detailed discussion can be found in Hernan and Robins (2025).

These assumptions are used to produce causal effects for observational studies using G-computation.

B.2 Assessing assumptions

The findings in Chapter 6 using G-computation should be interpreted cautiously if interpreted in a causal way. This is due to several key assumptions inherent to this approach. This discussion is inspired by Pelgrims et al. (2024). First, temporal ordering assumes that the PFAS exposure precedes the immune-related outcome and that confounders precede both. In cross-sectional studies, exposure, outcomes and additional covariates are measured simultaneously, which complicates the determination of whether PFAS exposure occurred before immune dysregulation or if immune conditions influenced PFAS levels (reverse causality).

Secondly, the assumption of conditional exchangeability is crucial. This was addressed by adjusting for confounders identified by experts, including gender, age, BMI, financial stability, birth weight and smoking status. However, there may still be several unmeasured confounding factors, such as genetic influences or dietary habits, which could significantly impact the association. Thirdly, the no-interference assumption states that the outcome of an individual is not affected by the exposures or outcomes of other individuals. This is plausible as immune dysregulation does not spread between individuals. For some adolescents living in the same household, shared PFAS exposure sources (eg, contaminated water) may correlate exposures among family members and (slightly) violate this assumption. The fourth assumption, positivity, states that for all combinations of covariates, there must be a non-zero probability of observing PFAS concentrations at both Q1 and Q3. This assumption is generally satisfied, although positivity may not hold in subgroups (e.g., birth weight and BMI).

The fifth assumption is the principle of consistency. In a medical context involving an intervention, adjusting treatment status is relatively straightforward. However, maintaining consistency can pose challenges when the exposure is attributable to an environmental pollutant (Pelgrims et al., 2024). Efforts to increase or decrease exposure to PFAS are less practical and may raise ethical concerns. Therefore, the hypothetical scenario is unlikely to occur as a real-world intervention. The sixth assumption is that there is no model misspecification. Random forest addresses misspecification by effectively capturing non-linear relationships and interactions without relying on specific functional forms. Lastly, we have the assumption of measurement error. Measurement error in PFAS levels is possible, even with precise lab methods like UPLC-MS/MS (ultra-performance liquid chromatography coupled to tandem mass spectrometry) (Consortium UAntwerpen, VITO, PIH, UHasselt and VUB, 2023). However, self-reported covariates (e.g., smoking) are potentially more subject to measurement error.

Appendix C

Additional simulation results and insights

C.1 Graphical visualisation of the individual and joint exposure effect

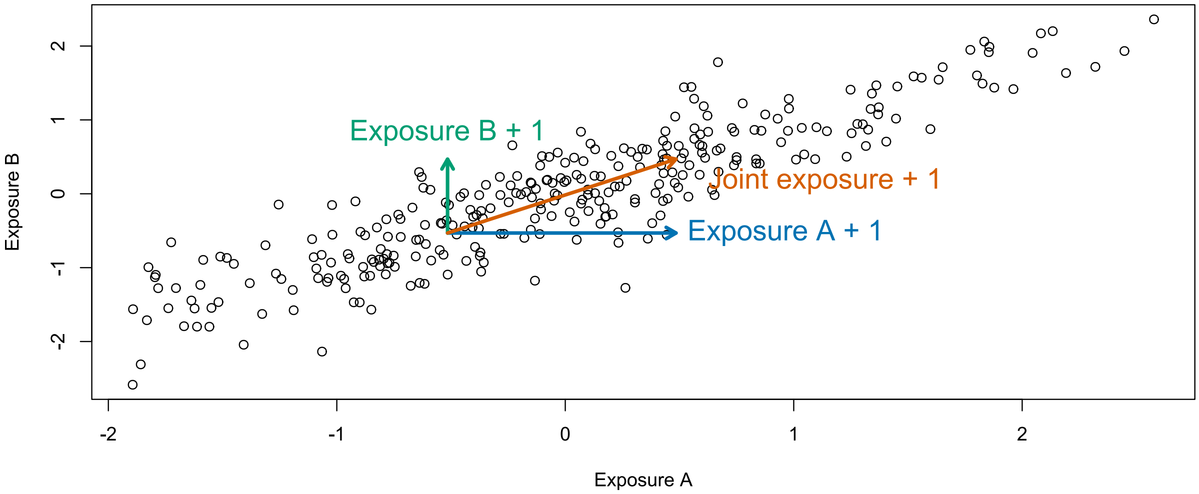


Figure C.1: *Scatterplot of two highly correlated exposures A and B with arrows illustrating a 1-unit increase in Exposure A (blue), Exposure B (green) and their joint increase (orange).*

C.2 Theoretical approach for the joint effect standard error

In the simulation study presented in Section 5.2, it was found that the joint effect appears unaffected by high correlations among the predictors. This observation can also be supported from a theoretical perspective. To simplify the discussion, let us focus on a scenario with two highly correlated predictors

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \boldsymbol{\epsilon} \tag{C.1}$$

with $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ centered, $\mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}$, $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \sigma^2 \boldsymbol{I}_n$ and $\text{Cor}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \rho \in (0, 1)$. We aim to show that the variance of the joint effect $\hat{\beta}_1 + \hat{\beta}_2$ is less than the sum of the individual effects using OLS.

$$\text{Var}(\hat{\beta}_1 + \hat{\beta}_2) < \text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) \tag{C.2}$$

Therefore, we need to prove that the covariance between the two parameters is negative. It is known from (3.2) that the covariance can be written as

$$\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \sigma^2 \Big\{ (\mathbf{X}'\mathbf{X})^{-1} \Big\}_{12} \tag{C.3}$$

with the subscript referring to the off-diagonal element and $\mathbf{X} = (\boldsymbol{x}_1 \ \boldsymbol{x}_2)$ the design matrix. The inverse of the Gramm matrix $\mathbf{X}'\mathbf{X}$ has the following from

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} \boldsymbol{x}_1'\boldsymbol{x}_1 & \boldsymbol{x}_1'\boldsymbol{x}_2 \\ \boldsymbol{x}_2'\boldsymbol{x}_1 & \boldsymbol{x}_2'\boldsymbol{x}_2 \end{pmatrix}^{-1} = \frac{1}{\boldsymbol{x}_1'\boldsymbol{x}_1\boldsymbol{x}_2'\boldsymbol{x}_2 - (\boldsymbol{x}_1'\boldsymbol{x}_2)^2} \begin{pmatrix} \boldsymbol{x}_2'\boldsymbol{x}_2 & -\boldsymbol{x}_1'\boldsymbol{x}_2 \\ -\boldsymbol{x}_2'\boldsymbol{x}_1 & \boldsymbol{x}_1'\boldsymbol{x}_1 \end{pmatrix} \tag{C.4}$$

Given the assumption that the vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are centered, the Pearson correlation can be written as

$$\rho = \frac{\text{Cov}(\boldsymbol{x}_1, \boldsymbol{x}_2)}{\sqrt{\text{Var}(\boldsymbol{x}_1)}\sqrt{\text{Var}(\boldsymbol{x}_2)}} = \frac{\frac{1}{n-1}\boldsymbol{x}_1'\boldsymbol{x}_2}{\sqrt{\frac{1}{n-1}\boldsymbol{x}_1'\boldsymbol{x}_1}\sqrt{\frac{1}{n-1}\boldsymbol{x}_2'\boldsymbol{x}_2}} = \frac{\boldsymbol{x}_1'\boldsymbol{x}_2}{\sqrt{\boldsymbol{x}_1'\boldsymbol{x}_1}\sqrt{\boldsymbol{x}_2'\boldsymbol{x}_2}} \tag{C.5}$$

By integrating (C.3) with (C.4) and implementing the substitution of (C.5), one can derive the following result:

$$\begin{aligned} \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \frac{\sigma^2(-\boldsymbol{x}_1'\boldsymbol{x}_2)}{\boldsymbol{x}_1'\boldsymbol{x}_1\boldsymbol{x}_2'\boldsymbol{x}_2 - (\boldsymbol{x}_1'\boldsymbol{x}_2)^2} &= \frac{-\sigma^2\rho\sqrt{\boldsymbol{x}_1'\boldsymbol{x}_1}\sqrt{\boldsymbol{x}_2'\boldsymbol{x}_2}}{\boldsymbol{x}_1'\boldsymbol{x}_1\boldsymbol{x}_2'\boldsymbol{x}_2 - (\boldsymbol{x}_1'\boldsymbol{x}_2)^2} && \text{(C.6)} \\ &= \frac{-\sigma^2\rho}{\sqrt{\boldsymbol{x}_1'\boldsymbol{x}_1}\sqrt{\boldsymbol{x}_2'\boldsymbol{x}_2} - \frac{(\boldsymbol{x}_1'\boldsymbol{x}_2)^2}{\sqrt{\boldsymbol{x}_1'\boldsymbol{x}_1}\sqrt{\boldsymbol{x}_2'\boldsymbol{x}_2}}} && \text{(C.7)} \\ &= \frac{-\sigma^2\rho}{\sqrt{\boldsymbol{x}_1'\boldsymbol{x}_1}\sqrt{\boldsymbol{x}_2'\boldsymbol{x}_2} - \rho\ \boldsymbol{x}_1'\boldsymbol{x}_2} && \text{(C.8)} \\ &= \frac{-\sigma^2\rho}{\sqrt{\boldsymbol{x}_1'\boldsymbol{x}_1}\sqrt{\boldsymbol{x}_2'\boldsymbol{x}_2}(1-\rho^2)} && \text{(C.9)} \end{aligned}$$

This now means that for $\rho \in (0, 1)$ the covariance is negative which implies that (C.2) is true. In conclusion, multicollinearity increases the standard errors of individual parameters, but decreases the standard errors of joint effects. Nonetheless, the inequality presented in (C.2) raises questions regarding its practical implications. First, a negative covariance does help reduce the variance of the joint effect, but it does not guarantee that the sum has low variance. While it helps reduce the variance, the individual variances and the strength of correlation will still play a crucial role.

C.3 Evaluating methods on realistic exposure mixtures

This section includes additional graphs referred to in the simulation discussion in Section 5.3.

C.3.1 Linear additive exposures

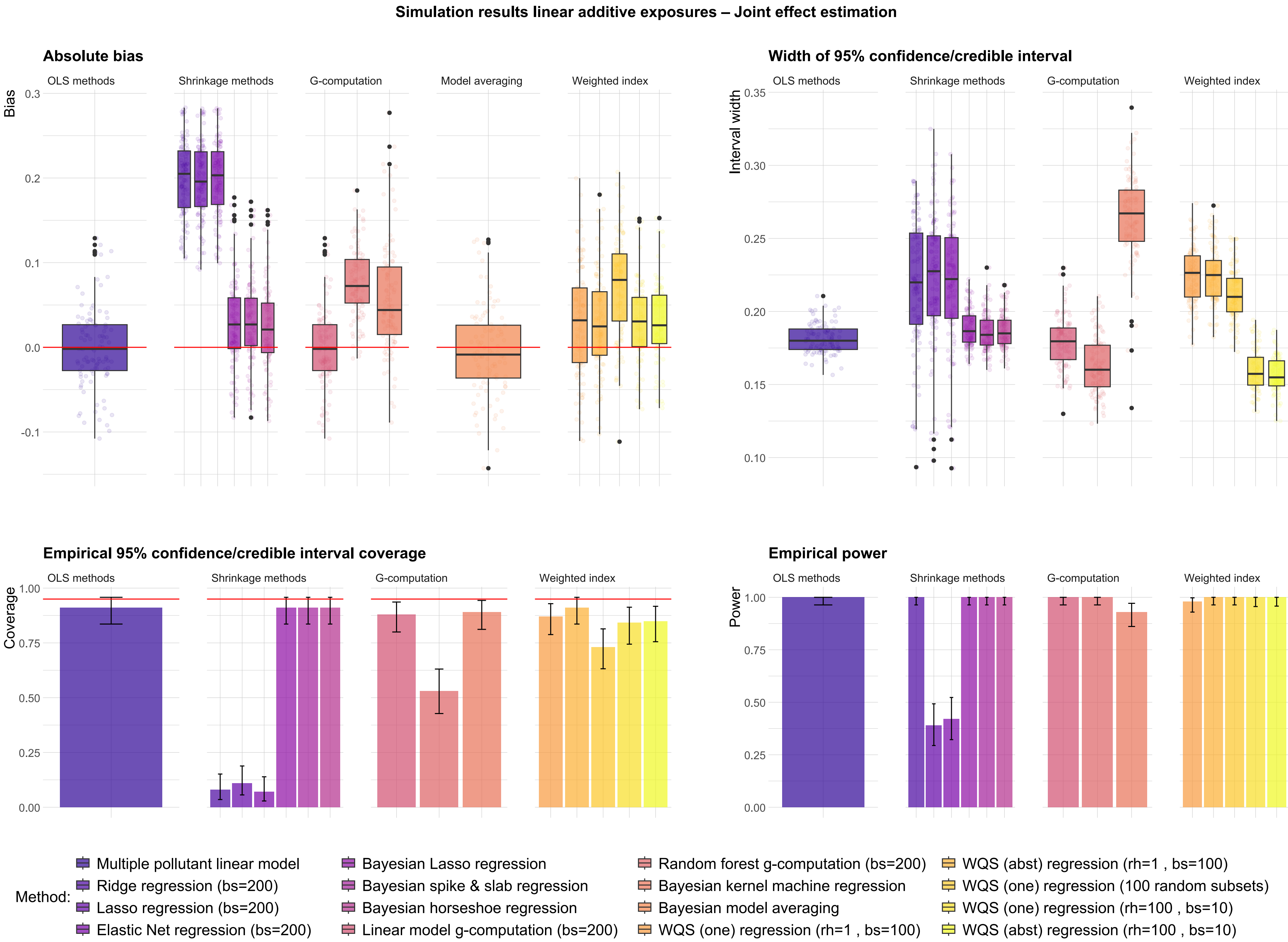


Figure C.2: *Linear additive exposure effects on the outcome were simulated to compare methods for estimating joint effects. Performance was evaluated based on absolute bias, confidence/credible interval (CI) width, CI coverage and power. Simulations used $n = 300$ observations and were repeated $n_{sim} = 100$ times; see Appendix D.3.1 for additional details. "One" is the simple average of weights and "abst" uses the absolute t-statistic as a weight for averaging.*

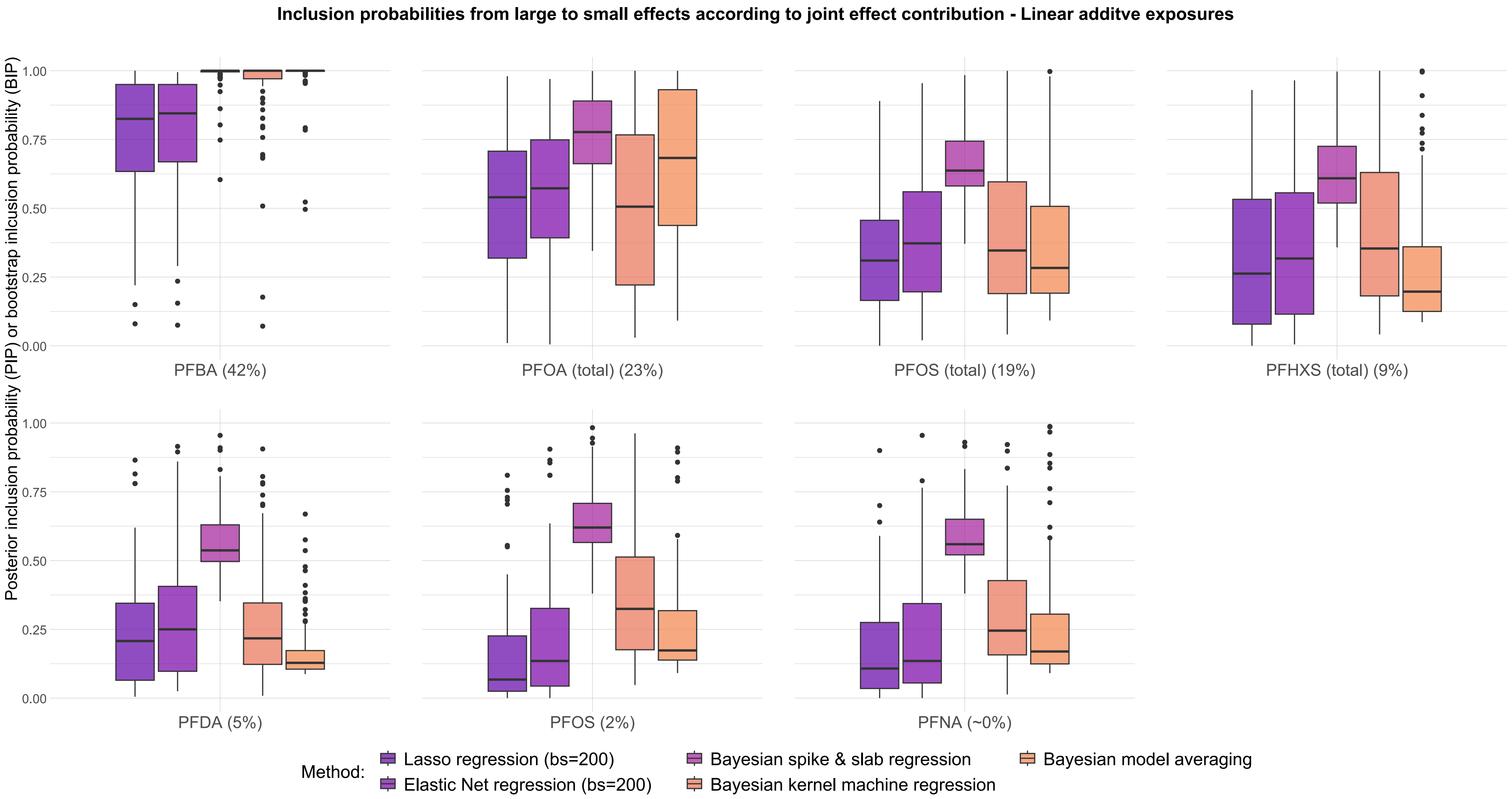


Figure C.3: *Posterior inclusion and bootstrap inclusion probabilities are shown for all variable selection methods. Brackets (%) indicate the contribution of each individual effect to the total effect, which was considered linear additive. Simulations were based on a sample size of $n = 300$ and repeated $n_{sim} = 100$ times; more details can be found in Appendix D.3.1.*

C.3.2 Synergistic effects

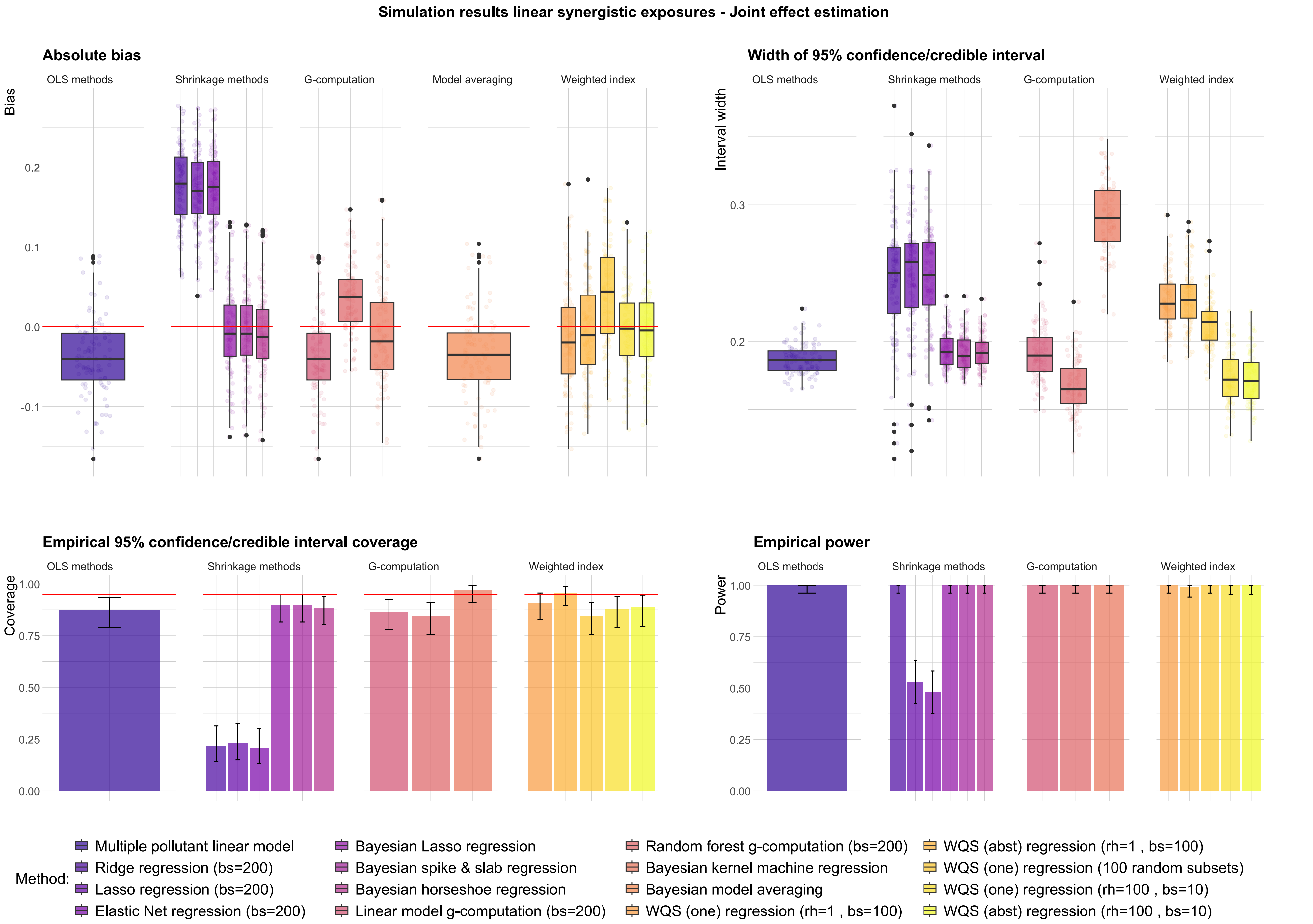


Figure C.4: *Synergistic and linear exposure effects on the outcome were simulated to compare methods for estimating joint effects. Performance was evaluated based on absolute bias, confidence/credible interval (CI) width, CI coverage and power. Simulations used $n = 300$ observations and were repeated $n_{sim} = 100$ times; see Appendix D.3.1 for additional details. "One" is the simple average of weights and "abst" uses the absolute t-statistic as a weight for averaging.*

C.4 Evaluating grouping behaviour

This section includes additional graphs referred to in the simulation discussion in subsection 5.3.3.

C.4.1 Correlated predictors of equal effect

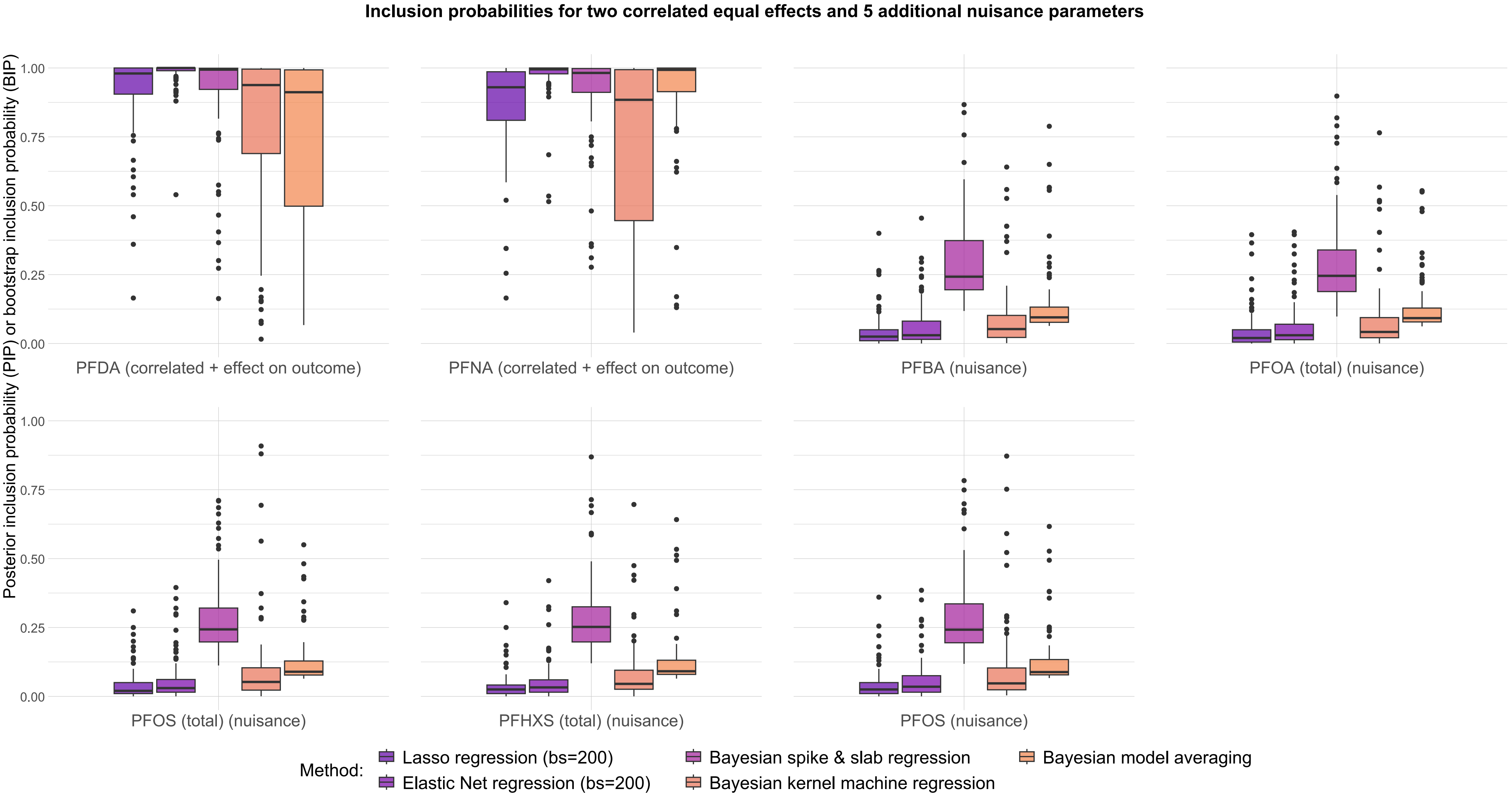


Figure C.5: *Posterior inclusion and bootstrap inclusion probabilities are shown for all variable selection methods. PFDA and PFNA were highly correlated, with both having the same effect on the outcome. All five additional parameters were not correlated and had no effect on the outcome; therefore, they were considered nuisances. Simulations were based on a sample size of* $n = 300$ *and repeated* $n_{sim} = 100$ *times; more details can be found in Appendix D.3.2.*

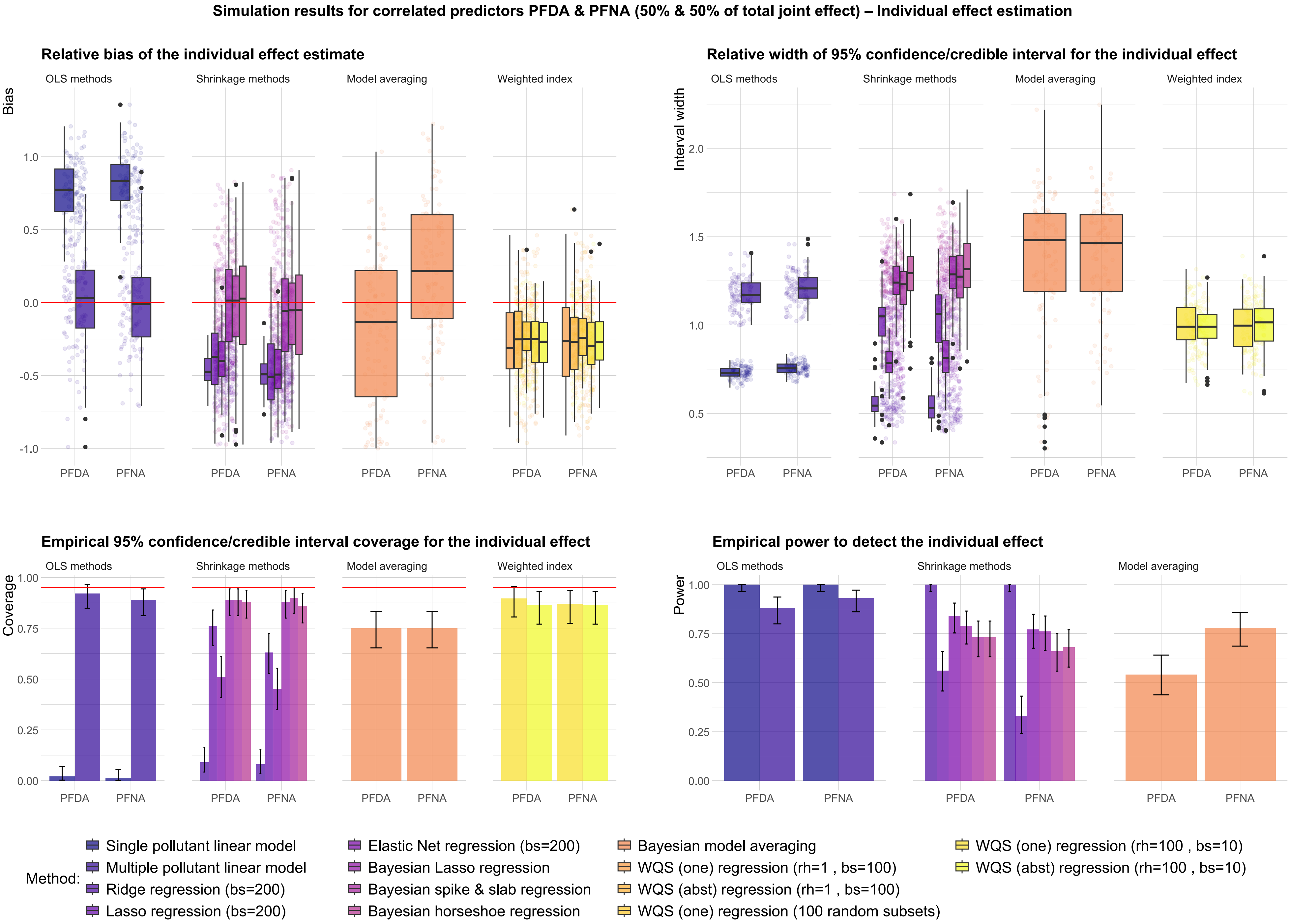


Figure C.6: *Comparison of different methods based on absolute bias, CI width, CI coverage and power for two individual effects that were highly correlated, with both correlated predictors having a true effect on the outcome. Simulations are based on a sample size of $n = 300$ repeated $n_{sim} = 100$ times; more details can be found in Appendix D.3.2. "One" is the simple average of weights and "abst" uses the absolute t-statistic as a weight for averaging.*

C.4.2 Correlated predictors of unequal effect

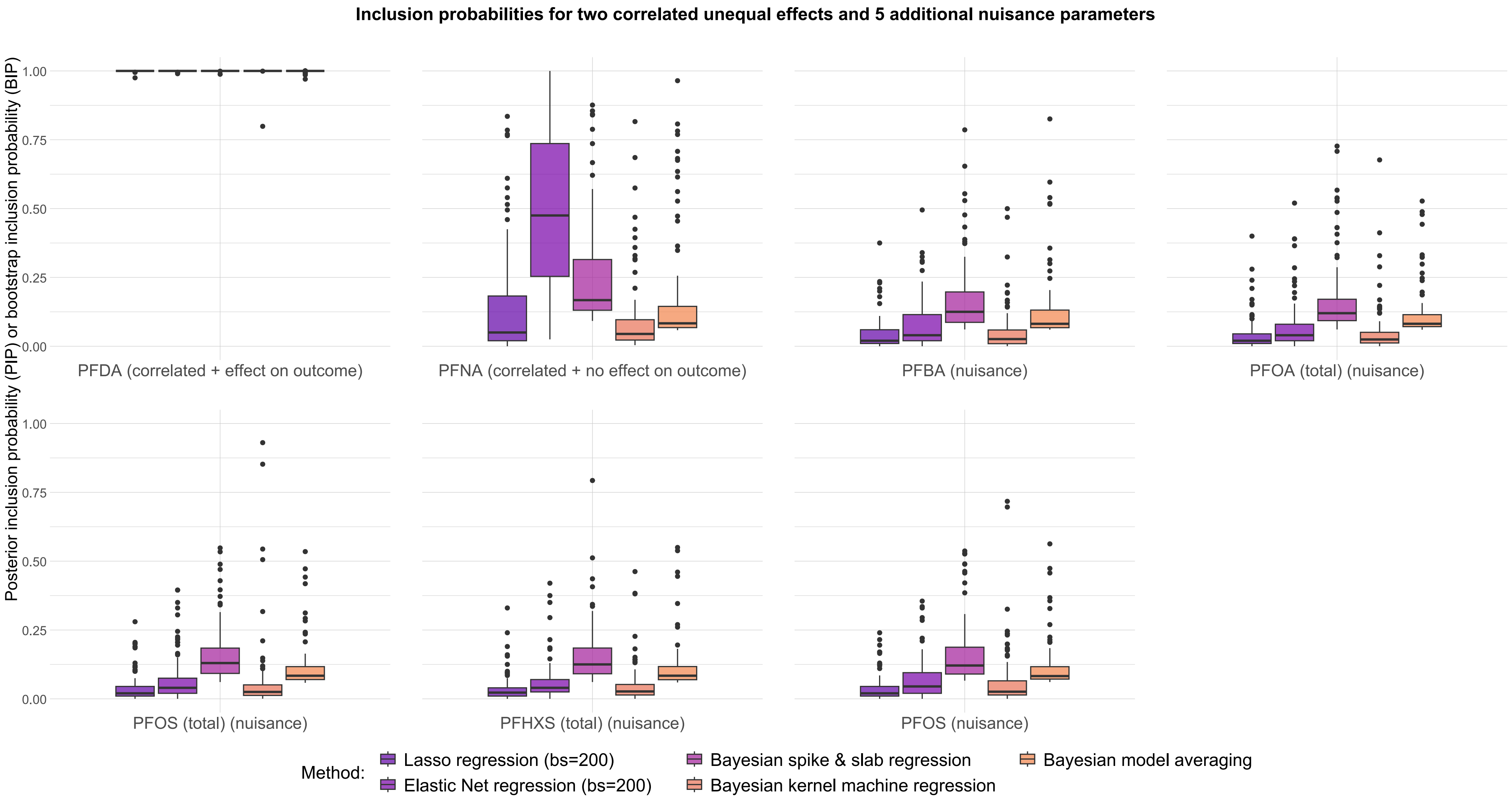


Figure C.7: *Posterior inclusion and bootstrap inclusion probabilities are shown for all variable selection methods. PFDA and PFNA were highly correlated, with only PFDA having a true effect on the outcome. All five additional parameters were not correlated and had no effect on the outcome; therefore, they were considered nuisances. Simulations were based on a sample size of* $n = 300$ *and repeated* $n_{sim} = 100$ *times; more details can be found in Appendix D.3.2.*

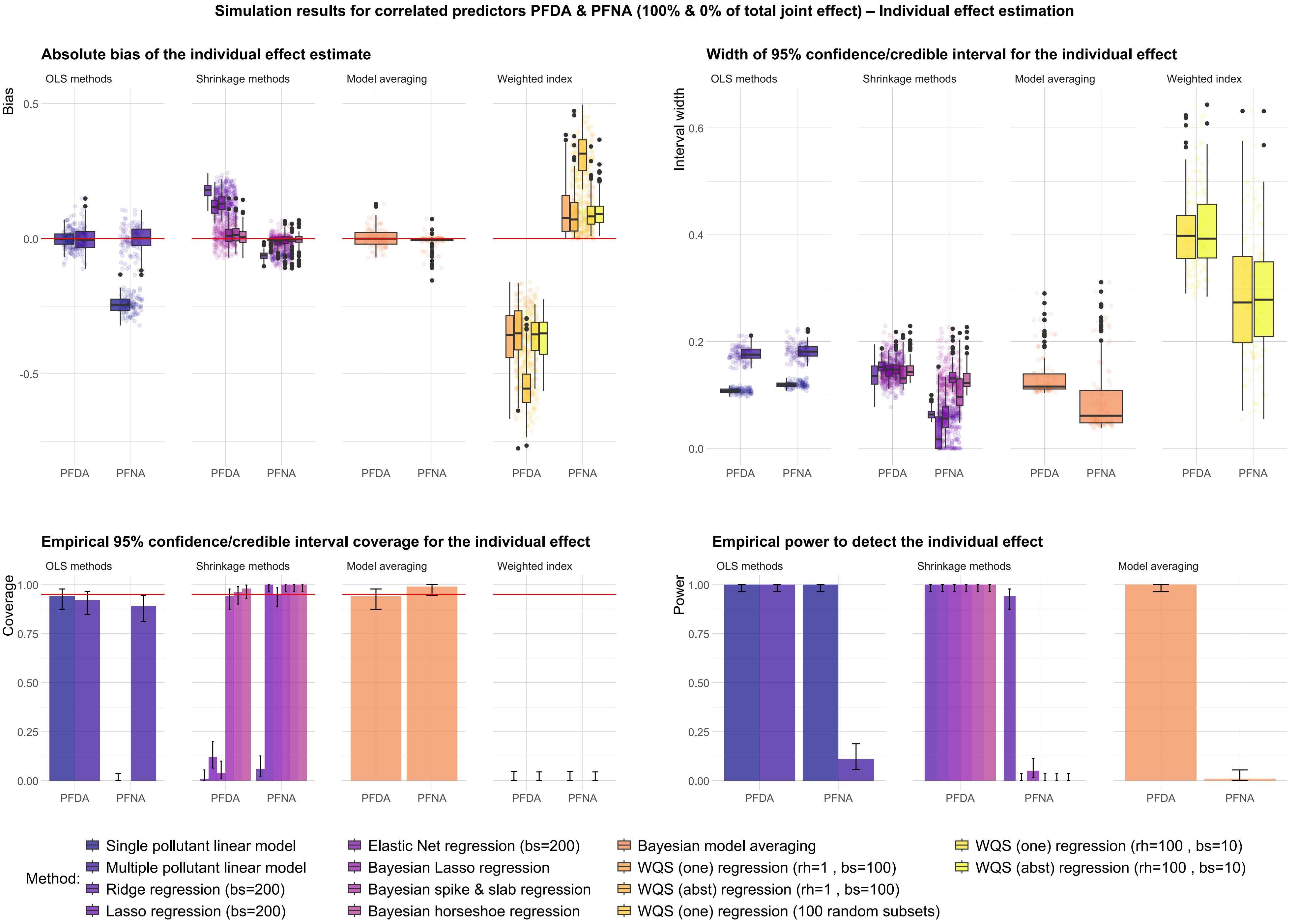


Figure C.8: *Comparison of different methods based on absolute bias, CI width, CI coverage and power for two individual effects that were highly correlated, with only PFDA having a true effect on the outcome. Simulations are based on a sample size of* $n = 300$ *repeated* $n_{sim} = 100$ *times; more details can be found in Appendix D.3.2. "One" is the simple average of weights and "abst" uses the absolute t-statistic as a weight for averaging.*

Appendix D

Details on simulation procedure

D.1 Continuous vs quantised exposures

We will now provide a detailed description of the simulation study presented in Section 5.1.

Aims: This simulation study aims to evaluate the effects of continuous versus quantised exposures ($Q = 4, 10$) on the stability of weight estimation and statistical power across various exposure distributions. The analysis employs repeated holdout WQS regression to achieve this objective.

Data-generating mechanisms: We consider three data-generating mechanisms. In all cases, data are simulated on $n = 300$ individuals, representing a typical sample size in a human biomonitoring study as described in Chapter 2. Consider the following three distributions for 7 exposures, notated in the exposure matrix $\boldsymbol{A}$:

(a) It is well established that the distribution of PFAS is typically right-skewed, characterised by long, high tails. While this right-skewness is often addressed through various transformation methods, the presence of high tails within the exposure may still introduce instability into the analysis (see Figure E.1). To account for this, we will sample the exposure from a multivariate t-distribution with 2 degrees of freedom:

$$\boldsymbol{A} \sim t_2(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{D.1}$$

The mean $\boldsymbol{\mu}$ of the distribution is zero, representing centred exposures. The scale matrix $\boldsymbol{\Sigma}$ is a unit matrix where the off-diagonal elements are set to 0 or 0.8 to simulate the extreme case of no correlation or high correlation.

(b) In addition to high tails, human biomonitoring studies often include exposure-driven outliers. To begin, we will simulate the exposure $\boldsymbol{A'}$ using a multivariate normal distribution

$$\boldsymbol{A'} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{D.2}$$

using previously described $\boldsymbol{\mu}, \boldsymbol{\Sigma}$. Next, 5 outliers will be added to $\boldsymbol{A'}$ at random. This is done by replacing the original exposure matrix $\boldsymbol{A'}$ by $\boldsymbol{A}$. For each selected index (j, k), the values in the exposure matrix will be increased multiplicatively by inflating the previous value with a factor that is uniformly distributed between 1.5 and 2.5.

This can mathematically be described as

$$\boldsymbol{A}_{j,k} = \begin{cases} \boldsymbol{A'}_{j,k} \cdot (1.5 + e_{j,k}) & \text{if} \;\; j \in \mathcal{I} \\ \boldsymbol{A'}_{j,k} \end{cases} \qquad \text{with } e_{j,k} \sim U[0,1] \tag{D.3}$$

with $\mathcal{I}$ a set of 5 indices randomly sampled (without replacement) and $k = 1, ..., 7$. The matrix $\boldsymbol{A}$ will be used as the exposure matrix. Figure D.1 visualises the outliers in a typical exposure distribution for this scenario.

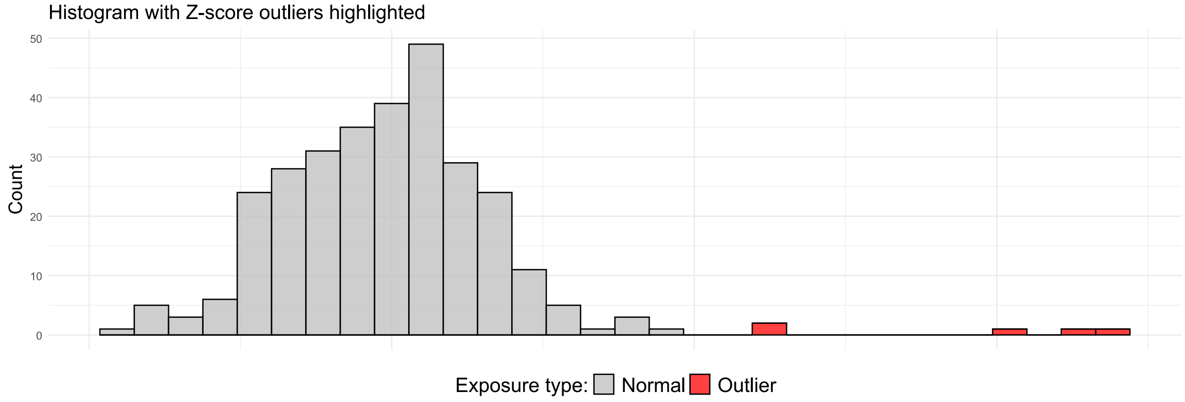


Figure D.1: *Histogram of exposure values with outliers identified using the Z-score method ($|Z| > 3$). Red bars indicate extreme values flagged as statistical outliers, while gray bars represent values within the normal range.*

(c) In the final data-generating process, our goal is to preserve the original distribution of the exposures in the case study. The exposure matrix $\boldsymbol{A}$ is constructed by sampling entire log-transformed exposure profiles (rows) with replacement from the original dataset. By selecting complete exposure vectors instead of individual exposure values, we retain the joint distribution and dependence structure of the mixture while introducing variability due to sampling. As we sample complete individual profiles, we also obtain the corresponding confounding variables, which will be stored in the $\boldsymbol{Z}$ matrix.

We simulate the outcome $\boldsymbol{Y}$ using the WQS formulation with the matrix $\boldsymbol{A} = [X_1 X_2 ... X_7]$, containing the exposures, being standardised. The simulation procedure is defined as

$$\begin{aligned} \boldsymbol{wqs} =& 0.1863 \cdot X_1 + 0.0515 \cdot X_2 + 0.4213 \cdot X_3 + 0.0207 \cdot X_4 + \\ & 0.2326 \cdot X_5 + 0.0024 \cdot X_6 + 0.0851 \cdot X_7 \\ \boldsymbol{Y} =& -1.4714 - 0.2 \cdot \boldsymbol{wqs} \;(\, + \,\boldsymbol{Z'}\boldsymbol{\varphi}\;) + \boldsymbol{\epsilon} \end{aligned} \tag{D.4}$$

with $\boldsymbol{\epsilon} \sim \mathrm{N}(0\ ,\ 0.8466)$ and the brackets indicate that only in the data-generating mechanism (c), confounding variables are included. All parameters are inspired by the case study, with an effect size of -0.2, approximately doubled from what was observed.

Estimands: The estimands of interest are the weights $\hat{\boldsymbol{w}}$ and joint effect $\hat{\beta}_1$ from the WQS formulation calculated by the mean across the repeated holdout splits for each exposure.

Methods: Each simulated dataset is analysed using repeated holdout WQS regression with 100 repeated holdouts, each with 20 bootstraps, with continuous, quartile or decile exposures. The exact model formulation corresponds to (D.4). This means that the model is correctly specified. Implementation was done through the R package *gWQS* with the constraint of directional homogeneity (negative direction) and without signal function (Renzetti et al., 2023).

Performance measures: The absolute bias will be evaluated for each exposure, and the power of the joint effect will be assessed across various models utilising $n_{sim} = 50$ simulations. The absolute bias is defined as follows

$$\text{Absolute bias} = \hat{w}_{ij} - \hat{w}_{True,\ j} \tag{D.5}$$

for each simulation where $i = 1, \ldots, n_{sim}$ and each predictor where $j = 1, \ldots, 7$. The power is approximated by

$$P(\text{reject } H_0|H_1) \approx \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \mathbb{I}(p_i < \alpha) \tag{D.6}$$

where $\alpha = 0.05$, $\mathbb{I}$ is an indicator function and p_i denotes the corresponding p-value for the two-sided null hypothesis $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. In the context of the repeated holdout WQS model, this is based on a Wald test statistic. A Clopper-Pearson interval will be employed to provide a 95% confidence interval around the estimated power.

D.2 Joint effect standard error

We will now provide a detailed description of the simulation study presented in Section 5.2.

Aims: The objective of this simulation study is to evaluate whether a traditional linear regression model (coefficients estimated through OLS) is affected by high correlations across the predictors when estimating the joint effect of a mixture.

Data-generating mechanisms: In order to assess the impact of high correlation on the SE, we will not use the exact exposure values derived from the case study. Instead, we will adopt a multivariate normal distribution with $n = 300$ observations, which enables us to maintain precise control over the correlation by means of the covariance matrix. The means of the exposures are zero, representing centred exposures. The covariance matrix Σ is created by using a compound symmetry correlation structure, where there is an equal correlation ρ between all exposures. The correlation ρ varies from 0 to 0.9 in increments of 0.1. This configuration does not entirely reflect the observed data, as we employ a uniform correlation among all predictors. The effect on the outcome is chosen to be linear and additive, as this simplifies the calculations of the joint effect SE. We consider three settings for the parameter $\boldsymbol{\beta}$:

(a) Only the first exposure has an effect, with all others set to zero. The sum of the coefficients equals the effect of the first exposure: In this scenario, $\boldsymbol{\beta}$ is considered to be:

$$\boldsymbol{\beta} = [-0.2\ ,\ 0\ ,\ 0\ ,\ 0\ ,\ 0\ ,\ 0\ ,\ 0] \tag{D.7}$$

(b) Multiple exposures contribute to the outcome, with varying effect sizes. The total joint effect remains -0.2. The vector $\boldsymbol{\beta}$ is considered to be approximately equal to:

$$\boldsymbol{\beta} = [-0.04\ ,\ -0.01\ ,\ -0.08\ ,\ 0.00\ ,\ -0.05\ ,\ 0.00\ ,\ -0.02] \tag{D.8}$$

(c) None of the exposures affect the outcome. This null scenario is used to examine the empirical Type I error rate. Thus, $\boldsymbol{\beta}$ is considered to be:

$$\boldsymbol{\beta} = [0\ ,\ 0\ ,\ 0\ ,\ 0\ ,\ 0\ ,\ 0\ ,\ 0] \tag{D.9}$$

Using $\boldsymbol{\beta}$ as defined in the respective cases, with standardised exposure matrix $\boldsymbol{X}$, the outcome will be generated as follows:

$$\boldsymbol{Y} = -1.4714 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{D.10}$$

with $\boldsymbol{\epsilon} \sim \mathrm{N}(0\ ,\ 0.8466)$.

Estimands: The estimand of interest is the SE of the individual exposure and joint exposure effect to contrast how these change over increasing correlation.

Methods: A multiple linear model is used to estimate the individual exposure estimates and their standard errors with the base *stats* package in R. The model is correctly specified according to (D.10). The joint effect of the mixture is then computed as the sum of the individual exposure estimates

$$\hat{\beta}_{joint} = \sum_{j=1}^{7} \hat{\beta}_j \tag{D.11}$$

with $\hat{\beta}_j$ the individual exposure effect. The SE of the joint effect $\hat{\beta}_{joint}$ is calculated using the covariance matrix from the fitted linear model. Note that there is no model specification.

Performance measures: The mean standard error and its empirical 95% interval (based on simulation quantiles) are computed for both individual and joint effects, based on $n_{sim} = 1000$ simulations. In cases (a) and (b), the power is approximated by

$$P(\text{reject } H_0|H_1) \approx \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \mathbb{I}(p_i < \alpha) \tag{D.12}$$

where $\alpha = 0.05$, $\mathbb{I}$ is an indicator function and p_i denotes the corresponding p-value for the two-sided null hypothesis $H_0 : \beta_j = 0$ against $H_1 : \beta_j \neq 0$ (equivalent for the joint effect). The p-value is based on a Wald test statistic. For case (c), the empirical Type I error rate is approximated by

$$P(\text{reject } H_0|H_0) \approx \frac{1}{n_{sim}} \sum_{i=1}^{n_{sim}} \mathbb{I}(p_i < \alpha) \tag{D.13}$$

with previously defined significance level and hypothesis. A Clopper-Pearson interval will be employed to provide a 95% confidence interval around the estimated power or Type I error rate.

D.3 Evaluating methods on realistic exposure mixtures

D.3.1 Linear additive or synergistic exposure effect

We will now present a comprehensive description of the simulation study outlined in Section 5.3. Furthermore, a detailed explanation of the precise implementation of all methodologies discussed in Chapter 3 and Chapter 4 will be provided.

Aims: The objective is to compare the methods discussed in Chapter 3 and Chapter 4 for handling the complexity of estimating joint mixture effects or individual effects within a mixture. The comparison will focus on absolute or relative bias, power and the (relative) width or coverage of confidence/credible intervals in a realistic setting.

Data-generating mechanisms: The objective is to conduct a comparison of various methods in a setting that accurately reflects real-world conditions. We aim to preserve the original distribution of exposures within the case study context presented in Chapter 2. To facilitate this, the exposure matrix $\boldsymbol{A}$ is constructed by sampling $n = 300$ entire log-transformed exposure profiles (rows) from the original dataset with replacement. By selecting complete exposure vectors rather than isolated exposure values, we maintain the joint distribution and dependence structure of the mixture while introducing variability through the sampling process. Additionally, as we sample these complete individual profiles, we simultaneously obtain the corresponding confounding variables, which will be stored as columns in the $\boldsymbol{Z}$ matrix.

We simulate the outcome $\boldsymbol{Y}$ using the WQS formulation with the matrix $\boldsymbol{A}$, containing the PFAS exposures, being standardised. The simulation procedure is defined as

$$\boldsymbol{Y} = -1.4714 - 0.3 \cdot \boldsymbol{wqs} \ + \ \boldsymbol{Z}'\boldsymbol{\varphi} \ + \boldsymbol{\epsilon} \tag{D.14}$$

with $\boldsymbol{\epsilon} \sim \mathrm{N}(0\ ,\ 0.21165)$. The intercept and coefficients for covariates such as age, gender and education were based on estimates from a real case study. It is defined as

$$\boldsymbol{\varphi} = [-0.016248\ ,\ 0.168078\ ,\ 0.103059\ ,\ 0.023381] \tag{D.15}$$

where the first value is a continuous confounder and the last three values correspond to binary confounders. The coefficient for the weighted continuous sum, still referred to as WQS, was set to -0.3, approximately triple the magnitude of the effect observed in the empirical data, in order to simulate a more pronounced signal. The error variance was chosen such that the resulting coefficient of determination is approximately $R^2 \approx 0.185$, representing a moderate signal-to-noise ratio that is typical in environmental health research. The assumption of directional homogeneity is made throughout this entire simulation study. The weighted sum is defined in two different ways:

- *Linear additive effect*: First, we will examine the simplest case, where the WQS is defined as the linear additive sum of all exposures in the mixture. It is important to note that the sum of the seven parameters equals one. The weights chosen below are based on the specific case study.

$$\begin{aligned}\boldsymbol{wqs} = & 0.1863 \cdot \mathrm{PFOS\ (total)} + 0.0515 \cdot \mathrm{PFDA} + 0.4213 \cdot \mathrm{PFBA} + 0.0207 \cdot \mathrm{PFOS} + \\ & 0.2326 \cdot \mathrm{PFOA\ (total)} + 0.0024 \cdot \mathrm{PFNA} + 0.0851 \cdot \mathrm{PFHXS\ (total)}\end{aligned} \tag{D.16}$$

- *Linear synergistic effect*: Secondly, the weighted sum is selected to incorporate certain interactions. While this may result in misspecification of various parametric models discussed later, it generates a synergistic effect. In the literature, this is often regarded as more realistic for chemical mixtures in contrast to previous additive effects.

$$\begin{aligned}\boldsymbol{wqs} =&0 \cdot \text{PFOS (total)} + 0 \cdot \text{PFDA} + 0.4213 \cdot \text{PFBA} + 0 \cdot \text{PFOS}+ \\ &0.2326 \cdot \text{PFOA (total)} + 0 \cdot \text{PFNA} + 0 \cdot \text{PFHXS (total)}+ \\ &0.173026 \cdot \text{PFDA} \cdot \text{PFNA} + 0.173026 \cdot \text{PFHXS (total)} \cdot \text{PFOS} \cdot \text{PFNA}\end{aligned} \tag{D.17}$$

Estimands: The estimand of interest is the estimated coefficient with its confidence/credible interval for both individual and joint exposure effects.

Methods: An overview is given of the exact implementation of all methods to estimate the individual exposure or joint exposure effects. Each time it is indicated in brackets if the model is used for individual or joint effect estimation. The choices made here are in line with Chapter 3 and Chapter 4. The models are implemented as follows:

Single pollutant linear model (individual): Seven separate linear regression models, each including one chemical exposure along with covariates age, gender and education level, are fitted. The functional form included additive linear exposures and covariates. For each model, the estimated effects of chemical exposure and their confidence intervals are extracted. Calculations were performed using OLS, with standard errors derived from OLS and Wald-type confidence intervals utilising the base *stats* package in R.

Multiple pollutant linear model (individual/joint): A multiple linear regression model including all seven chemical exposures, along with covariates age, gender and education level, was fitted. The functional form included additive linear exposures and covariates. The estimated effect for each chemical exposure was extracted, as well as the joint effect calculated by summing the individual coefficients. Calculations were conducted using OLS regression. Standard errors and Wald-type confidence intervals were obtained from the base *stats* package in R. To assess the joint effect, the OLS covariance matrix was utilised to calculate the standard errors and to compute the Wald-type confidence intervals.

Ridge regression (individual/joint): A ridge regression model including all seven chemical exposures along with covariates for age, gender, and education level was fitted using the *glmnet* package in R. The functional form included additive linear exposures and covariates. The analysis was conducted using 200 non-parametric bootstrap samples. For each bootstrap replicate, a ridge regression model was applied with the regularisation parameter (λ) selected via 5-fold cross-validation using the default grid in *glmnet*. Specifically, the λ value chosen was the largest value for which the cross-validated MSE was within one standard error of the minimum MSE. The confounding covariates were not penalised. The estimated coefficients for the seven chemical exposures were extracted and their joint effect was calculated as the sum of the individual exposure coefficients. Across the bootstrap replicates, percentile-based 95% confidence intervals were constructed for both individual and joint effects. The mean across the bootstrap samples was used as the final estimate.

LASSO regression (individual/joint): A LASSO regression model including all seven chemical exposures and covariates for age, gender and education level was fitted using the *glmnet* package in R. The functional form included additive linear exposures and covariates. The analysis was performed via non-parametric bootstrap resampling, repeated 200 times. For each bootstrap sample, LASSO regression was applied with the regularisation parameter (λ) selected via 5-fold cross-validation using the default grid in *glmnet*. Specifically, the λ value chosen was the largest value for which the cross-validated MSE was within one standard error of the minimum MSE. The confounding covariates were not penalised. The estimated coefficients for the seven chemical exposures were extracted from each fitted model and the joint effect was defined as the sum of the seven exposure-specific coefficients. For both the joint and individual effects, percentile-based 95% confidence intervals were computed from the bootstrap distribution. The mean across the bootstrap samples was used as the final estimate. The bootstrap inclusion probability (BIP) for each variable are determined by the bootstrap samples in which the variable is given a non-zero coefficient.

Elastic Net regression (individual/joint): An Elastic Net regression model including all seven chemical exposures and covariates for age, gender and education level was fitted using the *glmnet* package in R. The functional form included additive linear exposures and covariates. The analysis was performed via non-parametric bootstrap resampling, repeated 200 times. For each bootstrap sample, Elastic Net regression was applied with the regularisation parameter $\alpha = 0.5$ fixed and λ selected via 5-fold cross-validation using the default grid in *glmnet*. Specifically, the λ value chosen was the largest value for which the cross-validated MSE was within one standard error of the minimum MSE. The confounding covariates were not penalised. The estimated coefficients for the seven chemical exposures were extracted from each fitted model and the joint effect was defined as the sum of the seven exposure-specific coefficients. For both the joint and individual effects, percentile-based 95% confidence intervals were computed from the bootstrap distribution. The mean across the bootstrap samples was used as the final estimate. The bootstrap inclusion probabilities (BIP) for each variable are determined by the bootstrap samples in which the variable is given a non-zero coefficient.

Bayesian LASSO regression (individual/joint): A Bayesian regression model was fitted, incorporating all seven chemical exposures while adjusting for age, gender and education level. The model was implemented in R using the *nimble* package. The likelihood was assumed normal with the mean specified by a linear predictor including all exposures and covariates. All variables were modelled as additive and linear. The model used a Bayesian LASSO prior on the exposure coefficients, implemented as a scale mixture of normals with an exponential prior on local shrinkage parameters (see (3.6)). Covariate coefficients (age, gender and education level) were assigned weakly informative normal priors and were not subject to shrinkage. The residual standard deviation was assigned an improper Jeffreys prior via a normal prior on $log(\sigma)$. Two Markov chain Monte Carlo (MCMC) chains were run for 100 000 iterations with a burn-in of 5 000 and a thinning interval of 10, yielding 9 500 posterior samples per chain. Posterior summaries were computed for all model parameters. The joint effect of the chemical mixture was defined as the sum of the seven exposure-specific coefficients in each posterior sample. Final point estimates for both joint and individual effects were taken as the posterior means and 95% credible intervals were computed using the equal-tail quantiles of the posterior distributions. A sensitivity analysis using four different hyperpriors for λ^2 showed robust results with minor differences (results not shown).

Bayesian spike and slab regression (individual/joint): A Bayesian regression model was fitted, incorporating all seven chemical exposures while adjusting for age, gender and education level. The model was implemented in R using the *nimble* package. The likelihood assumed normally distributed outcomes with the mean specified by a linear predictor including all exposures and covariates. All variables were modelled as additive and linear. The model employed a Bayesian spike and slab prior on the exposure coefficients to enable variable selection (see (3.7)). The spike variance ϵ_0 was given a Gamma (5, 1000) prior (mean 0.005), concentrating probability mass near zero to encourage strong shrinkage of negligible effects. The slab variance c_0 was assigned a Gamma (2, 20) prior (mean 0.1), providing sufficient flexibility to accommodate moderate effect sizes. This configuration was selected based on the expected range of true effects, which varied from approximately -0.126 (largest) to -0.0006 (smallest) and was verified by examining the support of the resulting prior distribution for β_i. The inclusion probability π was assigned a Uniform$(0, 1)$ prior to reflect uncertainty in the number of active exposures. Covariate coefficients (age, gender and education level) were assigned weakly informative normal priors and were not subject to variable selection. The residual standard deviation was assigned an improper Jeffreys prior via a normal prior on $\log(\sigma)$. Two MCMC chains were run for 100 000 iterations with a burn-in of 5 000 and a thinning interval of 10, yielding 9 500 posterior samples per chain. Posterior summaries were computed for all model parameters. The joint effect of the chemical mixture was defined as the sum of the seven exposure-specific coefficients. Final point estimates for both joint and individual effects were taken as the posterior means and 95% credible intervals were computed using the equal-tail quantiles of the posterior distributions. The posterior inclusion probability (PIP) is defined as the proportion of posterior samples in which the variable's inclusion indicator, denoted as λ_i, is 1 (see (3.7)). In other words, it represents the probability that the coefficient is drawn from the "slab" component of the prior distribution instead of the spike around zero. Sensitivity analyses indicated that posterior estimates and posterior inclusion probabilities (PIP) for strong effects were stable across a range of spike ϵ_0 and slab c_0 hyperpriors. In contrast, variables with weaker effects (PIP < 0.6) showed moderate sensitivity to the slab variance, with PIP decreasing as the slab hyperprior was made more diffuse. The spike variance hyperprior had little impact on results within the tested range (results not shown).

Bayesian horseshoe regression (individual/joint): A Bayesian regression model was constructed to analyse all seven chemical exposures, while controlling for variables such as age, gender and education level. The model was implemented in R using the *nimble* package. The likelihood assumed normally distributed outcomes with the mean specified by a linear predictor including all exposures and covariates. All variables were modelled as additive and linear. The model employed a Bayesian horseshoe prior on the exposure coefficients (see (3.8)). Covariate coefficients (age, gender and education level) were assigned weakly informative normal priors and were not subject to shrinkage. The residual standard deviation was assigned an improper Jeffreys prior via a normal prior on $\log(\sigma)$. Two MCMC chains were run for 100 000 iterations, with a burn-in of 5 000 and thinning interval of 10, yielding 9 500 posterior samples per chain. Posterior summaries were computed for all model parameters. The joint effect of the chemical mixture was defined as the sum of the seven exposure-specific coefficients. Final point estimates for both joint and individual effects were taken as the posterior means and 95% credible intervals were computed using equal-tail quantiles of the posterior distributions. A sensitivity analysis using four different hyperpriors for λ_i showed robust results, completely unaffected (results not shown).

Linear model G-computation (joint): A linear regression model was fitted with the base *stats* package in R to estimate the joint effect of multiple chemical exposures on the outcome, adjusting for relevant covariates such as age, gender and education level. The model assumed an additive linear relationship with the continuous outcome. G-computation was used to estimate the joint effect of simultaneously increasing all exposures by one unit. This involved predicting outcomes under the observed exposure values and under a hypothetical scenario where all exposures were shifted upward by one unit, then averaging the difference in predicted outcomes. Uncertainty in the joint effect estimate was quantified using a non-parametric bootstrap procedure with 200 resamples. For each bootstrap sample, the G-computation procedure was repeated to generate a distribution of joint effect estimates. The 95% confidence interval was constructed from the empirical quantiles of this bootstrap distribution.

Random forest G-computation (joint): A random forest regression model was used to estimate the joint effect of simultaneously increasing all chemical exposures by one unit on the outcome, while adjusting for covariates including age, gender and education level. The model was fit using the *randomForest* package in R with 500 trees. G-computation was implemented by first predicting the outcome under the observed exposure values, then under a scenario where all exposures were shifted upward by one unit. The average difference between these predicted outcomes provided an estimate of the joint effect. To quantify uncertainty, a non-parametric bootstrap with 200 resamples was performed. For each bootstrap sample, the entire G-computation procedure was repeated to generate a distribution of joint effect estimates. The 95% confidence interval was derived from the empirical quantiles of this distribution.

Bayesian model averaging (individual/joint): Bayesian model averaging was applied to estimate the joint and individual effects of seven chemical exposures while adjusting for covariates including age, gender and education level. The analysis was conducted using the *bas* package in R with a Zellner–Siow Cauchy prior on the exposures. The model included all possible subsets of the seven exposures with covariates (age, gender and education indicators) forced into every model. Posterior model-averaged coefficients and their standard deviations were extracted. The joint effect of the mixture was calculated as the sum of the seven coefficients. For each exposure, approximate 95% credible intervals were computed assuming normality of the posterior distribution. Due to the construction of the package, it was not possible to obtain the posterior distribution of the sum; therefore, no credible intervals will be discussed. Posterior inclusion probability (PIP) is computed by summing the posterior probabilities of all models in the model space that include the variable of interest.

Bootstrap WQS regression (individual/joint): WQS regression was used to estimate the joint effect and weights, adjusting for covariates including age, gender and education level. The functional form is always additive and linear. In contrast to the literature, we used continuous exposures for the weighted sum to facilitate comparison with other methods. The model was implemented using the *gWQS* R package, assuming a Gaussian outcome. First, the data was split into a training set of 40% and a validation set of 60%. Bootstrapping was used to estimate the WQS weights, with the number of bootstrap samples set to 100. The joint effect was defined as the coefficient of the weighted sum in the final model and 95% Wald-type confidence intervals were computed based on the fitted model using OLS.

Random subset WQS regression (individual/joint): Random subset WQS regression was used to estimate the joint effect and weights, adjusting for covariates including age, gender and education level. The functional form is always additive and linear. In contrast to the literature, we used continuous exposures for the weighted sum to facilitate comparison with other methods. The model was implemented using the *gWQS* R package, assuming a Gaussian outcome. First, the data was split into a training set of 40% and a validation set of 60%. Bootstrapping was utilised to estimate WQS weights using different subsets of 3 out of 7 exposures, with the number of bootstrap samples set at 100. The joint effect was defined as the coefficient of the weighted sum in the final model and 95% Wald-type confidence intervals were computed based on the fitted model using OLS.

Repeated holdout WQS regression (individual/joint): Repeated holdout WQS regression was used to estimate the joint effect and weights, adjusting for covariates including age, gender and education level. The functional form is assumed to be additive and linear. In contrast to the literature, we used continuous exposures for the weighted sum to facilitate comparison with other methods. The model was implemented using the *gWQS* R package with a Gaussian outcome. The analysis used 100 repeated holdouts, where in each iteration the data were randomly split into training (40%) and validation (60%) sets. Within each training set, weights were estimated using 10 bootstrap samples and the joint effect was evaluated in the corresponding validation set. The final estimate of the joint effect and weights were defined as the average of coefficients across all holdout repetitions and 95% confidence intervals were computed based on the empirical distribution of these estimates.

Bayesian kernel machine regression (individual/joint): Bayesian kernel machine regression (BKMR) was applied to estimate the joint effect of the exposure mixture and to assess the relative importance of individual exposures, while adjusting for covariates including age, gender and education level. The model was implemented using the *bkmr* package in R, which provides a user-friendly interface using the MCMC algorithm with 10 000 iterations. It did not apply thinning and used a burn-in of 50%. The default settings from the *bkmr* packages are followed for the prior choices. A Gaussian likelihood for the outcome was assumed with normal, weakly informative priors for the confounding variables. For the residual variance modelled as precision σ^{-2}, both the shape and rate of the Gamma prior are set to 0.001. The notation used here refers to subsection 4.4.1. Component-wise variable selection (spike and slab prior) was used for the smoothness parameter r with a uniform prior on the inclusion probability δ_i, reflecting no strong prior belief about which variables are included. The slab $f_1(r)$ was defined as an inverse uniform prior with boundaries 0 and 100. For the kernel scale parameter $\lambda \equiv \tau\sigma^{-2}$, a Gamma prior is used with a mean of 10 and a standard deviation of 10. The posterior mean of δ will be interpreted to assess the relative importance and the concept of G-computation will be used for the joint effect. Posterior predictive distributions were obtained for each individual under both the observed exposure levels and a counterfactual scenario with increased exposures by one unit. For prediction counterfactuals, the first 50% of the posterior samples are dropped and only every 10 iterations are kept, resulting in 500 posterior samples. For each posterior draw, we computed the difference in predicted outcomes and averaged across individuals to obtain a draw from the posterior distribution of the average causal effect. The resulting distribution was summarised by its posterior mean and 95% equal-tail credible interval. The posterior inclusion probability (PIP) for each variable is computed as the proportion of posterior samples in which the corresponding indicator δ equals one.

Performance measures: The estimates were compared using several key metrics, such as the bias with $n_{sim} = 100$ simulations. For the joint effect, we considered the absolute bias defined as

$$\text{Absolute bias} = \hat{\Delta}_{ij} - \hat{\Delta}_{True,\ j} \tag{D.18}$$

for each simulation where $i = 1, \ldots, n_{sim}$ and each predictor where $j = 1, \ldots, 7$ with Δ estimand of interest for a specific model. To evaluate individual effects, we assessed the relative bias since the scales of the estimates differ across the models. It is defined as

$$\text{Relative bias} = \frac{\hat{\Delta}_{ij} - \hat{\Delta}_{True,\ j}}{\hat{\Delta}_{True,\ j}} \tag{D.19}$$

using previous notation. Next, the confidence/credible interval width was calculated for the joint effect and the relative confidence/credible width for the individual effects. This was simply defined as

$$\text{Relative CI width} = \frac{\text{CI width}}{\hat{\Delta}_{True,\ j}} \tag{D.20}$$

using previous notation. The third performance measure was the coverage of the different types of intervals for their true effect. A Clopper-Pearson interval was employed to provide a 95% confidence interval around the estimated coverage. Finally, the power was assessed and defined based on the inclusion of zero in the CI. A Clopper-Pearson interval was employed to provide a 95% confidence interval around the estimated power. It should be noted that not all performance measures could be calculated in the case of individual effect. For variable selection methods such as LASSO, Elastic Net, BMA, Spike and slab regression and BKMR, the posterior inclusion or bootstrap inclusion probabilities (PIP & BIP) are compared based on their ranking of the different individual effects.

D.3.2 Evaluating grouping behaviour

We will now provide a detailed description of the simulation study outlined in subsection 5.3.3.

Aims: The goal is to study the grouping effect described in Chapter 3 for all applicable methods introduced in Chapter 3 and Chapter 4.

Data-generating mechanisms: First, it should be noted that the previous simulation study Section 5.3 considered high correlations across all predictors. This makes it more difficult to study the grouping effect. We will now examine a specific scenario in which only two predictors have a high correlation, while the remaining predictors have correlations that are around zero. In line with our previous simulations, we will utilise a sample size of $n = 300$. As illustrated in Figure 2.2, we have decided to maintain the original distributions of PFNA and PFDA, given that these two variables have the highest pairwise correlation ($\rho = 0.83$). To facilitate this, the exposure matrix $\boldsymbol{A}$ is constructed by sampling entire log-transformed exposure profiles (rows) from the original dataset with replacement. Note that this is only done for the PFNA, PFDA and the additional covariates ($\boldsymbol{Z}$). In contrast, the remaining five exposures were sampled independently, with replacement. This means that the correlation structure among these exposures was not preserved, which resulted in pairwise correlations around zero.

We simulate the outcome $\boldsymbol{Y}$ using the WQS formulation with the matrix $\boldsymbol{A}$, containing the PFAS exposures, being standardised (see (D.14) for the choices and argumentation). Next, the WQS is defined with equal effect or unequal effects for the two correlated predictors. Consider the following two cases:

- *Equal effect*: First, the grouping property will be examined for the two highly correlated predictors PFNA and PFDA, both of which are assigned equal effects on the outcome. All other parameters are included as nuisance variables.

$$\begin{aligned} \boldsymbol{wqs} &= 0 \cdot \text{PFOS (total)} + 0.5 \cdot \text{PFDA} + 0 \cdot \text{PFBA} + 0 \cdot \text{PFOS} + \\ &\quad 0 \cdot \text{PFOA (total)} + 0.5 \cdot \text{PFNA} + 0 \cdot \text{PFHXS (total)} \end{aligned} \tag{D.21}$$

- *Unequal effect*: Secondly, the grouping property is evaluated in a scenario where only one of the two highly correlated predictors, PFDA, is assigned an effect on the outcome, while PFNA is given no effect. All other variables are included as nuisance variables.

$$\begin{aligned} \boldsymbol{wqs} &= 0 \cdot \text{PFOS (total)} + 1 \cdot \text{PFDA} + 0 \cdot \text{PFBA} + 0 \cdot \text{PFOS} + \\ &\quad 0 \cdot \text{PFOA (total)} + 0 \cdot \text{PFNA} + 0 \cdot \text{PFHXS (total)} \end{aligned} \tag{D.22}$$

Estimands: The estimand of interest is the estimated coefficient with its confidence/credible interval for the individual exposure effects.

Methods: We refer to the Methods in Appendix D.3.1 for a complete description of the methodologies used. Note that the focus is only on models estimating individual effects, as indicated in parentheses following each method.

Performance measures: Similar, for the performance measures we refer to Appendix D.3.1 which introduces all key metrics to compare the individual exposure effects based on $n_{sim} = 100$ simulations.

Appendix E

Case study: Descriptive statistics, results and implementation details

E.1 Descriptive statistics

Descriptive statistics in Table E.1 were computed after the exclusion of all individuals who met one or more of the specified exclusion criteria defined in Chapter 6.

Table E.1: *Characteristics of the study population, exposure to PFAS and outcomes. Exposure values are expressed in $\mu g/L$.*

	Variable name	**N (%)**	**$\geq$ LOQ (%)**	**Min**	**Q1**	**Median**	**Q3**	**Max**
Exposures	PFOS	287	100%	0.30	1.40	2.50	5.40	230.00
	PFOS (branched)	287	100%	0.42	2.60	4.20	6.80	23.00
	PFHXS (total)	287	100%	0.11	0.38	0.54	0.88	9.30
	PFOA (total)	287	100%	< LOQ	0.92	1.20	1.50	6.40
	PFBA	287	72%	< LOQ	< LOQ	0.15	0.19	0.89
	PFDA	287	73%	< LOQ	< LOQ	0.14	0.20	1.00
	PFNA	287	99%	< LOQ	0.19	0.26	0.33	0.89
Confounders	Gender	289						
	Male	143 (50%)						
	Female	146 (51%)						
	Age (years)	289						
	[12.5, 14.5]	117 (41%)						
	(14.5, 15.5]	101 (35%)						
	> 15.5	71 (25%)						
	Financial stability with the income	282						
	Struggling	7 (3%)						
	Getting by	89 (32%)						
	Comfortable living	186 (66%)						
	BMI	288						
	(severe) Underweight	26 (9%)						
	Normal weight	218 (76%)						
	(severe) Overweight	44 (15%)						
	Birth weight	269						
	< 2.5kg	19 (7%)						
	$\geq$ 2.5kg	250 (93%)						
	Exposure to tobacco smoke at home, elsewhere or through own smoking	289						
	No	234 (81%)						
	Yes	55 (19%)						
Outcome	Ratio CD4+ count over CD8+ count	286	100%	0.22	1.33	1.64	2.00	5.80
	Leukocyte count	287	100%	3230	4985	5850	7000	19760

N: refers to the total number of observations after applying the exclusion criteria and accounting for any missing values.
LOQ: lowest concentration of a substance that can be quantitatively measured

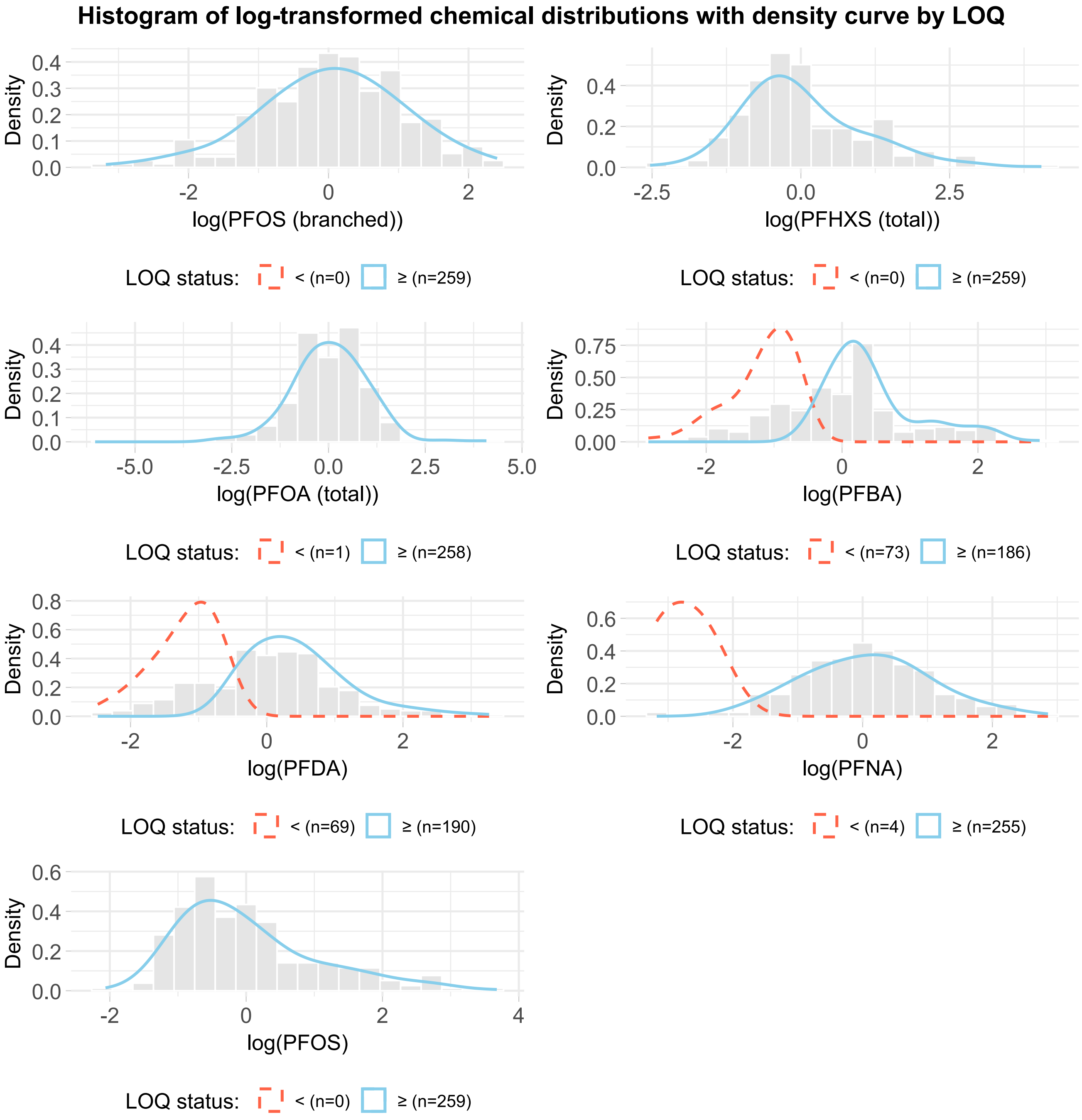


Figure E.1: *Log-transformed and standardised distributions of PFAS compounds were analysed after applying exclusion criteria and excluding adolescents with missing values for exposures or confounding variables. Histograms illustrate the standardised distributions of log-transformed concentrations for each PFAS compound. The overlaid density curves compare imputed observations that fall below the LOQ (represented by red dashed lines) with those that were observed above the LOQ (represented by blue solid lines).*

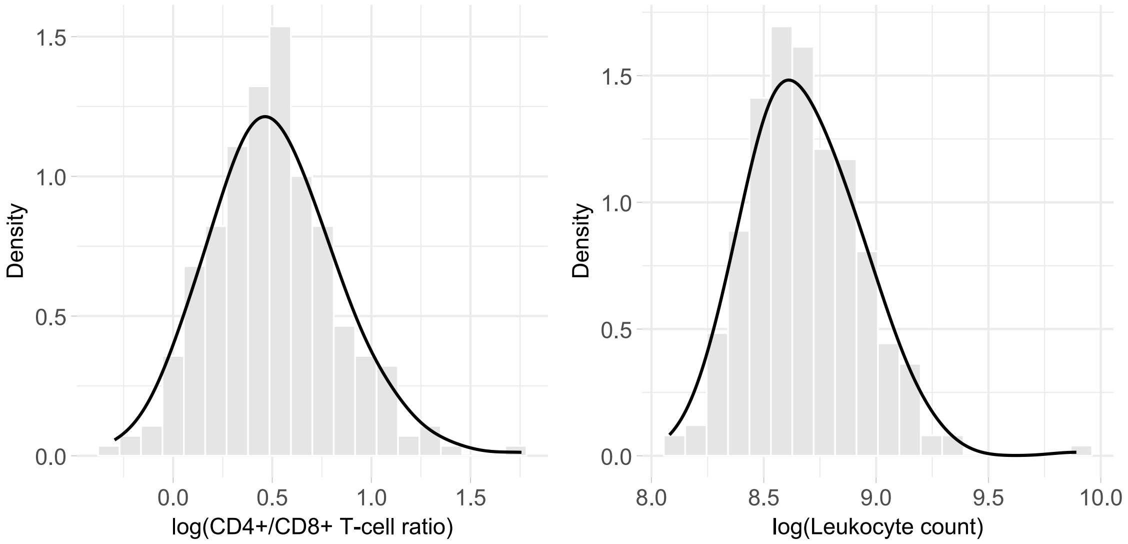


Figure E.2: *Histograms and density curve of (left) the log-transformed CD4+/CD8+ T-cell ratio (n=259) and (right) the log-transformed leukocyte count (n=260) based on all adolescents in the final complete case datasets.*

E.2 Results

Table E.2: *Bootstrap inclusion probabilities for pair-wise interactions, calculated as the proportion of 2000 bootstrap samples where each feature's coefficient is non-zero in a LASSO regression model with confounding covariates, with lambda selected via 5-fold cross-validation.*

	Outcome			Outcome	
Interactions	Leukocyte	CD4+/CD8+	Interactions	Leukocyte	CD4+/CD8+
PFOS (branched) * PFHXS (total)	0.4345	0.2050	PFOA (total) * PFBA	0.0825	0.1780
PFOS (branched) * PFOA (total)	0.1335	0.3195	PFOA (total) *PFDA	0.1190	0.1230
PFOS (branched) * PFBA	0.2470	0.3190	PFOA (total) * PFNA	0.1305	0.2285
PFOS (branched) * PFDA	0.2245	0.3730	PFOA (total) * PFOS	0.0920	0.2610
PFOS (branched) * PFNA	0.4135	0.0760	PFBA * PFDA	0.0965	0.2950
PFOS (branched) * PFOS	0.0380	0.1545	PFBA * PFNA	0.1340	0.0740
PFHXS (total) * PFOA (total)	0.1540	0.0780	PFBA * PFOS	0.1385	0.4980
PFHXS (total) * PFBA	0.2890	0.1250	PFDA * PFNA	0.1140	0.1735
PFHXS (total) * PFDA	0.0145	0.2965	PFDA * PFOS	0.1220	0.0785
PFHXS (total) * PFNA	0.0200	0.2560	PFNA * PFOS	0.0720	0.0985
PFHXS (total) * PFOS	0.0910	0.1880			

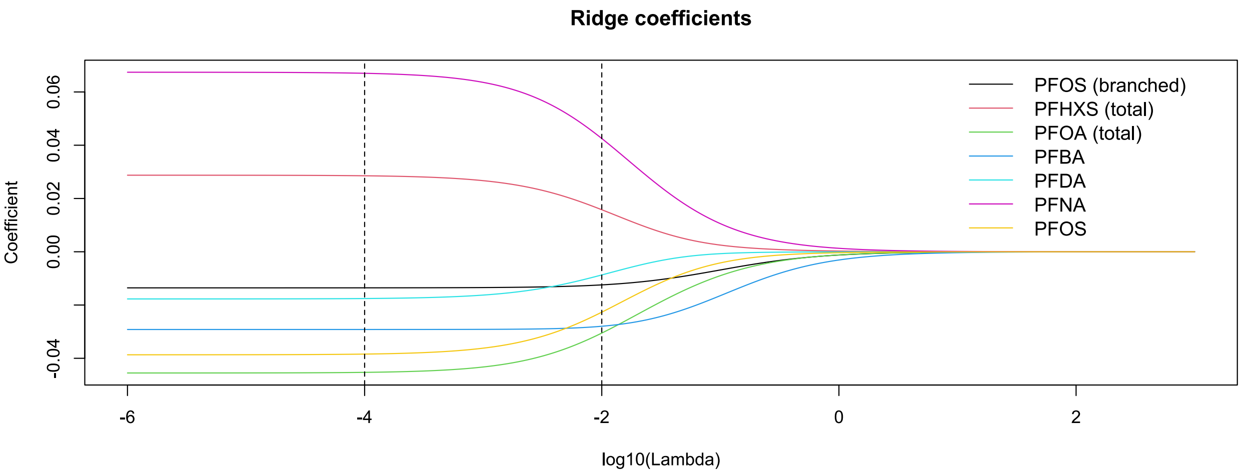


Figure E.3: *Each line presents the estimated coefficients from Ridge regression using the log-transformed leukocyte outcome, dependent on the regularisation parameter* $\log(\lambda)$. *As* λ *increases, the coefficients are progressively shrunk toward zero. The grid within the vertical dashed lines will be used to select the optimal lambda value through cross-validation.*

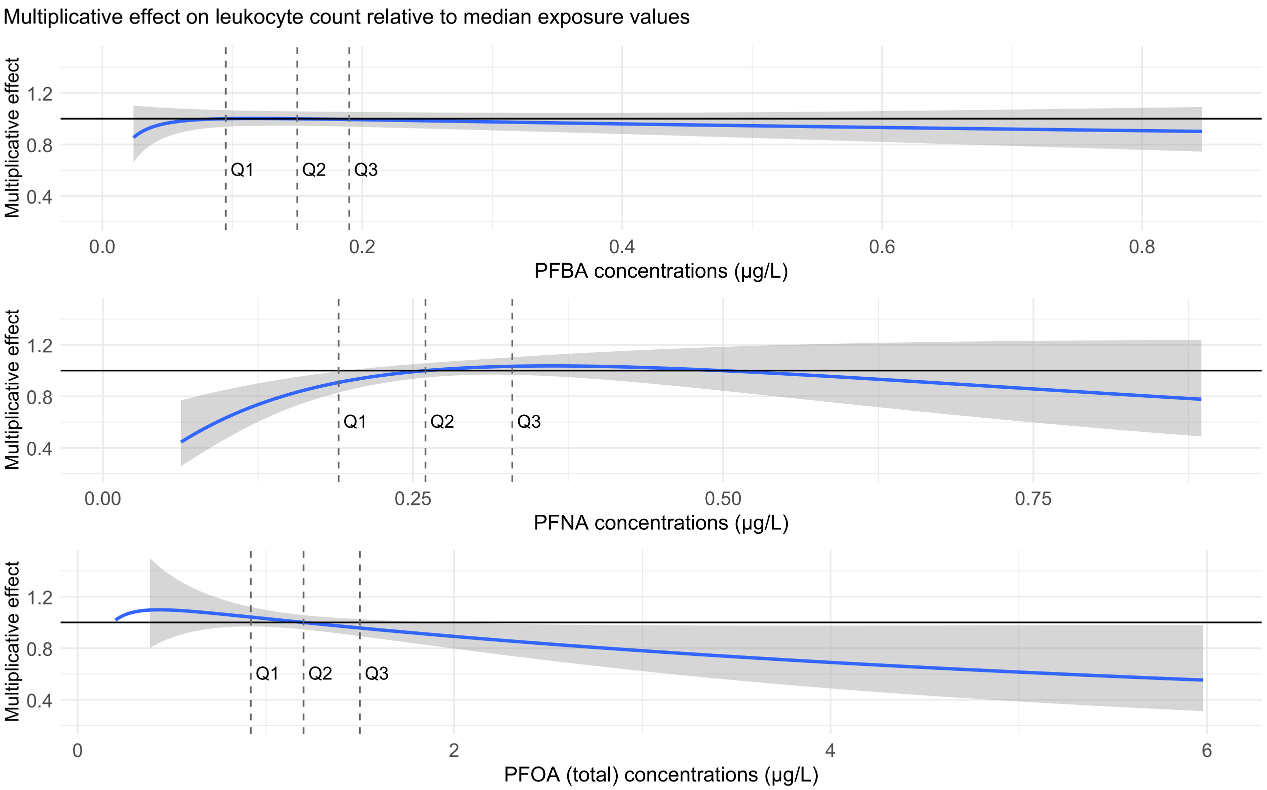


Figure E.4: *The estimated multiplicative effects on leukocyte counts using BKMR are presented for an increase/decrease in a single exposure from its median value, with all other exposures maintained at their median levels. The shaded areas indicate approximate 95% confidence intervals, assuming asymptotic normality. Vertical dashed lines mark the 25th (Q1), 50th (Q2), and 75th (Q3) percentiles of the exposure distribution.*

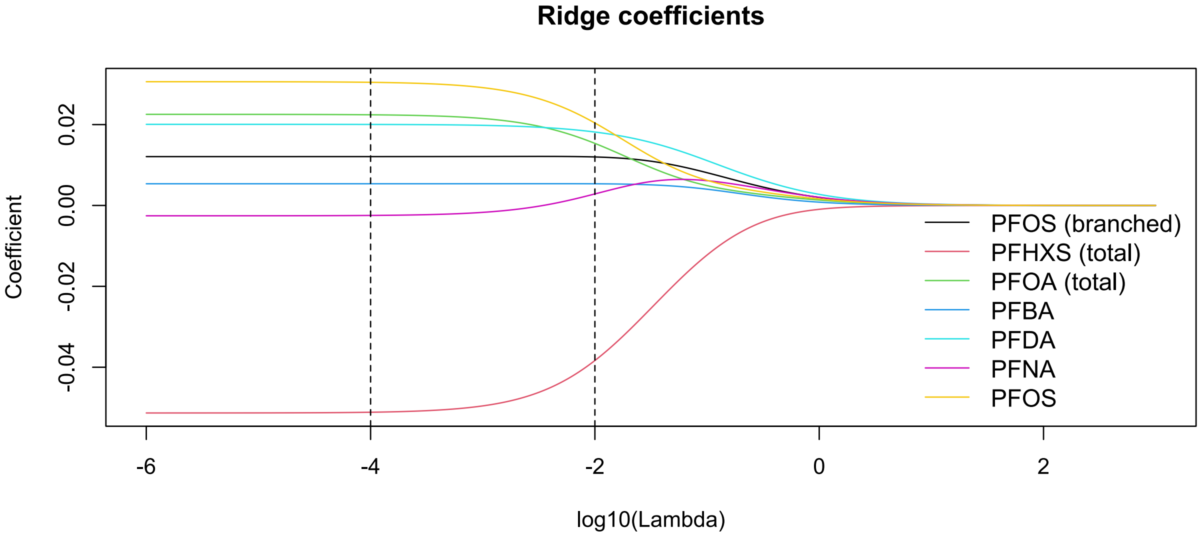


Figure E.5: *Each line presents the estimated coefficients from Ridge regression using the log-transformed CD4+/CD8+ ratio, dependent on the regularisation parameter* $\log(\lambda)$. *As* λ *increases (moving to the right), the coefficients are progressively shrunk toward zero. The grid within the vertical dashed lines will be used to select the optimal lambda value through cross-validation.*

E.3 Implementation details

Interquartile fold change in exposure concentration

The coefficients $\hat{\beta}_j$ from Ridge regression, multiple pollutant linear regression (OLS) and Bayesian (horseshoe) regression were transformed to represent the average interquartile (Q1 to Q3) fold effect on the untransformed outcome Y. Given the general model formulation in (6.1), without loss of generality, for a joint exposure, the multiplicative effect was defined as

$$\begin{aligned}
\frac{E[Y|\boldsymbol{Q3},\mathbf{z}]}{E[Y|\boldsymbol{Q1},\mathbf{z}]} &= \frac{\exp\left\{\hat{\beta}_0 + \sum_{j=1}^{7}\hat{\beta}_j\left(\frac{\log(Q3_j)-\mu_{\log(\boldsymbol{a}_j)}}{\sigma_{\log(\boldsymbol{a}_j)}}\right) + \mathbf{z}^\top\hat{\boldsymbol{\varphi}} + \frac{\hat{\sigma}^2}{2}\right\}}{\exp\left\{\hat{\beta}_0 + \sum_{j=1}^{7}\hat{\beta}_j\left(\frac{\log(Q1_j)-\mu_{\log(\boldsymbol{a}_j)}}{\sigma_{\log(\boldsymbol{a}_j)}}\right) + \mathbf{z}^\top\hat{\boldsymbol{\varphi}} + \frac{\hat{\sigma}^2}{2}\right\}} \\
&= exp\left\{\sum_{j=1}^{7}\hat{\beta}_j\left(\frac{\log(Q3_j)-\log(Q1_j)}{\sigma_{\log(\boldsymbol{a}_j)}}\right)\right\} \\
&= exp\left\{\sum_{j=1}^{7}\frac{\hat{\beta}_j}{\sigma_{\log(\boldsymbol{a}_j)}}log\left(\frac{Q3_j}{Q1_j}\right)\right\} \\
&= \prod_{j=1}^{7}\left(\frac{Q3_j}{Q1_j}\right)^{\hat{\beta}_j/\sigma_{\log(\boldsymbol{a}_j)}}
\end{aligned} \tag{E.1}$$

using the same notation as in (6.1) and $\frac{\hat{\sigma}^2}{2}$ is the adjustment for the log-normal distribution of the outcome Y. In the case of an individual's exposure effect, the remaining 6 terms in the product will cancel out.

Ridge regression

A ridge regression model including all seven standardised log-transformed chemical exposures, along with all the confounders, was fitted using the *glmnet* package in R. The log-transformed outcome was considered Gaussian. The functional form included additive linear exposures and covariates. A QQ-plot is used to assess normality. Residuals and squared residuals are plotted against each predictor to assess linearity and homoskedasticity. The analysis was conducted using 2000 non-parametric bootstrap samples. For each bootstrap replicate, a ridge regression model was applied with the regularisation parameter (k) selected via 5-fold cross-validation. Given our interest in accurately interpreting the size and direction of effects, a range of possible k was chosen based on Figure E.3. The grid considered was $k = 1E-4$ to 1E-2. Specifically, the k value chosen was the largest value for which the cross-validated MSE was within one standard error of the minimum MSE. The confounding covariates were not penalised. Across the bootstrap replicates, percentile-based 95% confidence intervals were constructed for the interquartile fold change effect using (E.1). The mean across the bootstrap samples was used as the final estimate.

Horseshoe regression

A Bayesian regression model was constructed to analyse all seven standardised log-transformed chemical exposures, along with all the confounders. The model was fitted using the *nimble* package. The likelihood assumed normally distributed outcomes with the mean specified by a linear predictor including all exposures and covariates. All variables were modelled as additive and linear. The model employed a Bayesian horseshoe prior on the exposure coefficients (see (3.8)). Additional covariates were assigned flat normal priors and were not subject to shrinkage. The residual standard deviation was assigned an improper Jeffreys prior via a normal prior on $log(\sigma)$.

Two MCMC chains were run for 100 000 iterations, with a burn-in of 5 000 and a thinning interval of 10, yielding 9 500 posterior samples per chain. Trace plots, Gelman-Rubin statistic ($\hat{R}$) and effective sample size will be used to confirm convergence. A QQ-plot is used to assess normality. Residuals and squared residuals are plotted against each predictor to assess linearity and homoskedasticity. Posterior predictive check was done using a density overlay plot comparing the observed outcome distribution with the posterior predictive distribution. The predictive mean bias was evaluated using a posterior predictive p-value, defined as

$$p = \frac{1}{N}\sum_{j=1}^{N} I(\bar{y}_{\text{posterior,j}} > \text{mean}(\log(Y))) \tag{E.2}$$

where $N = 19,000$ represents the total number of MCMC iterations, $\bar{y}_{\text{posterior,j}}$ is the mean of the 260 predictive replicates for the j-th iteration and $I(\cdot)$ is an indicator function. A p-value close to 0.5 suggests no bias. The posterior samples were used to calculate the posterior mean of the interquartile fold change, along with its 95% equal-tailed credible interval.

Repeated holdout WQS regression

Repeated holdout WQS regression was used to estimate the joint effect and weights, adjusting for additional covariates. The functional form is assumed to be additive and linear. In contrast to the literature, we utilised standardised log-transformed continuous exposures, which demonstrated enhanced power, as shown in Chapter 5. The model was implemented using the *gWQS* R package with a Gaussian outcome. The analysis used 100 repeated holdouts, where in each iteration the data were randomly split into training (40%) and validation (60%) sets. Within each training set, weights were estimated using 100 bootstrap samples and the joint effect was evaluated in the corresponding validation set. Linearity of the WQS index was assessed by inspecting scatter plots of the observed outcome versus the WQS index, as well as a plot of the residual versus fitted values. The residuals versus fitted values plot was also used to assess homoskedasticity. The final estimate of the joint effect and weights was defined as the average of coefficients across all holdout repetitions and 95% confidence intervals were computed based on the empirical distribution of these estimates. The analysis was conducted twice to assess the homogeneity assumption in both the positive and negative directions.

Bayesian kernel machine regression

Bayesian kernel machine regression (BKMR) was applied to estimate the joint effect of the exposure mixture and to assess the relative importance of individual exposures, while adjusting for additional covariates. Exposures were standardised after log-transformation. The model was implemented using the *bkmr* package in R, which provides a user-friendly interface using the Markov chain Monte Carlo (MCMC) algorithm with 10 000 iterations. It did not apply thinning and used a burn-in of 50%. The default settings from the *bkmr* packages are followed for the prior choices. A Gaussian likelihood for the outcome was assumed with normal, weakly informative priors for the confounding variables. For the residual variance modelled as precision σ^{-2}, both the shape and rate of the Gamma prior are set to 0.001. The notation used here refers to subsection 4.4.1. Component-wise variable selection (spike and slab prior) was used for the smoothness parameter r with a uniform prior on the inclusion probability δ, reflecting no strong prior belief about which variables are included. The slab $f_1(r)$ was defined as an inverse uniform prior with boundaries 0 and 100. For the kernel scale parameter $\lambda \equiv \tau\sigma^{-2}$, a Gamma prior is used with a mean of 10 and a standard deviation of 10. The posterior mean of δ will be interpreted to assess the relative importance. Trace plots will be used to confirm convergence. The posterior inclusion probability (PIP) for each variable is computed as the proportion of posterior samples in which the corresponding indicator δ equals one. For some of the selected exposures, univariate response curves will be shown. These are calculated using a grid of 500 points across the exposure range of interest while fixing all other exposures at their median value.

G-computation with random forest

A random forest regression model was used to estimate the joint effect of simultaneously increasing all chemical exposures from the 25th percentile (Q1) to the 75th percentile (Q3) on the outcome, while including additional covariates. Exposures were standardised after log-transformation. First, a repeated k-fold cross-validation was used to tune the random forest hyperparameters. Specifically, we evaluated combinations of the number of variables considered at each split (2–6) and the number of trees (ranging from 10 to 500 in increments of 10) using 20 repetitions of 5-fold cross-validation. For each combination, we calculated the mean squared error (MSE) on the held-out folds and summarised the average performance and its standard error across repetitions to select an optimal model. This model was fit using the *randomForest* package in R with 500 trees and 2 variables at each split. Using this model, G-computation was implemented by predicting the outcome under Q1 and Q3. The ratio of the average predictions was used as the final estimate. To quantify uncertainty, a non-parametric bootstrap with 2000 resamples was performed. For each bootstrap sample, the entire G-computation procedure was repeated to generate a distribution of joint effect estimates. The 95% confidence interval was derived from the empirical quantiles of this distribution.

Joint effect using OLS

We first fit a linear additive regression model including all PFAS exposures and covariates using ordinary least squares (OLS). To evaluate the appropriateness of the linearity assumption, we examined plots of residuals versus each exposure. Loess smoothers were overlaid to detect systematic patterns. Additionally, we assessed the distribution of residuals using histograms and QQ plots to evaluate the normality assumption. The residuals versus fitted values plot was used to check homoskedasticity. To quantify the joint effect of the PFAS exposures, we extracted the estimated regression coefficients and filled these in (E.1) to represent an interquartile range fold change in exposures. To quantify uncertainty, a non-parametric bootstrap with 2000 resamples was performed. For each bootstrap sample, the interquartile range fold change was calculated to generate a distribution of joint effect estimates. The 95% confidence interval was derived from the empirical quantiles of this distribution.

Bibliography

Bobb, J. F., Claus Henn, B., Valeri, L., and Coull, B. A. (2018). Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environmental Health*, 17(1):67.

Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J., and Coull, B. A. (2014). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, 16(3):493–508.

Braun, J. M., Gennings, C., Hauser, R., and Webster, T. F. (2016). What can epidemiological studies tell us about the impact of chemical mixtures on human health? *Environmental Health Perspectives*, 124(1):A6–A9.

Breiman, L., Cutler, A., Liaw, A., and Wiener, M. (2022). *Breiman and Cutler's Random Forests for Classification and Regression*. R package version 4.7-1.2.

Carrico, C., Gennings, C., Wheeler, D. C., and Factor-Litvak, P. (2015). Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *Journal of Agricultural, Biological, and Environmental Statistics*, 20(1):100–120.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 73–80. PMLR.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298.

Christensen, K. L. Y., Carrico, C. K., Sanyal, A. J., and Gennings, C. (2013). Multiple classes of environmental chemicals are associated with liver disease: Nhanes 2003–2004. *International Journal of Hygiene and Environmental Health*, 216(6):703–709.

Clyde, M. (2024). *BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling*. R package version 1.7.5.

Clyde, M. A., Joyee, G., and Michael L., L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics*, 20(1):80–101.

Consortium UAntwerpen, VITO, PIH, UHasselt and VUB (2023). Jongerenstudie hbm - omgeving 3m – resultatenrapport. Research report, Departement Omgeving, Vlaams

Planbureau voor Omgeving. In opdracht van het Departement Omgeving, Vlaams Planbureau voor Omgeving.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Curtin, P., Kellogg, J., Cech, N., and Gennings, C. (2019). A random subset implementation of weighted quantile sum (wqs$_{RS}$) regression for analysis of high-dimensional mixtures. *Communications in Statistics - Simulation and Computation*, 50(4):1119–1134.

de Valpine, P., Turek, D., Paciorek, C. J., Anderson-Bergman, C., Lang, D. T., and Bodik, R. (2017). Programming with models: Writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, 26(2):403–413.

Dominici, F., Peng, R., Barr, C., and Bell, M. (2010). Protecting human health from air pollution: Shifting from a single-pollutant to a multipollutant approach. *Epidemiology*, 21(2):187–194.

Ehrlich, V., Bil, W., Vandebriel, R., Granum, B., Luijten, M., Lindeman, B., Grandjean, P., Kaiser, A.-M., Hauzenberger, I., Hartmann, C., Gundacker, C., and Uhl, M. (2023). Consideration of pathways for immunotoxicity of per- and polyfluoroalkyl substances (pfas). *Environmental Health*, 22.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Garnier, Simon, Ross, Noam, Rudis, Robert, Camargo, Pedro, A., Sciaini, Marco, Scherer, and Cédric (2024). *viridis(Lite) - Colorblind-Friendly Color Maps for R*. viridis package version 0.6.5.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg.

George, E. and McCulloch, R. (1993). Variable selection via gibbs sampling. *Journal of The American Statistical Association*, 88:881–889.

Gilles, L., Govarts, E., Rodriguez Martin, L., Andersson, A.-M., Appenzeller, B. M. R., Barbone, F., Castaño, A., Coertjens, D., Den Hond, E., Dzhedzheia, V., Eržen, I., López, M. E., Fábelová, L., Fillol, C., Franken, C., Frederiksen, H., Gabriel, C., Haug, L. S., Horvat, M., Halldórsson, T. I., Janasik, B., Holcer, N. J., Kakucs, R., Karakitsios, S., Katsonouri, A., Klánová, J., Kold-Jensen, T., Kolossa-Gehring, M., Konstantinou, C., Koponen, J., Lignell, S., Lindroos, A. K., Makris, K. C., Mazej, D., Morrens, B., Murínová, P., Namorado, S., Pedraza-Diaz, S., Peisker, J., Probst-Hensch, N., Rambaud, L., Rosolen, V., Rucic, E., Rüther, M., Sarigiannis, D., Tratnik, J. S., Standaert, A., Stewart, L., Szigeti, T., Thomsen, C., Tolonen, H., Eiríksdóttir, , Van Nieuwenhuyse, A., Verheyen, V. J., Vlaanderen, J., Vogel, N., Wasowicz, W., Weber, T., Zock, J.-P., Sepai, O., and Schoeters, G. (2022). Harmonization of human biomonitoring studies in europe: Characteristics of the hbm4eu-aligned studies participants. *International Journal of Environmental Research and Public Health*, 19(11).

Grammarly Inc. (2025). Grammarly. https://www.grammarly.com. Accessed: 2025-08-11.

Hassen, H., Govarts, E., Portengen, L., Kalina, J., Komprdová, K., Tratnik, J. S., Kocman, D., Iszatt, N., Peeters, R., de Souza, C. M. T., Martin, L. R., Gilles, L., Santonen, T., Porras, S., Aimonen, K., Scheepers, P., Viegas, S., Bessonneau, H., Riou, M., Remy, S., Rodriguez-Carrillo, A., Vlaanderen, J., Gabriel, C., da Silva, S. d. N. P., Ogura, J. H., Bruckers, L., Engel, J., and Cano-Sancho, G. (2023). Statistical analysis plan (sap) for t4.1 projects: Human biomonitoring. Technical Report T4.1.4, VITO, Mol, Belgium.

HBM4EU (2021). PFAS – Factsheet. Accessed: 2025-06-01.

Hernan, M. and Robins, J. (2025). *Causal Inference: What If*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science*, 14(4):382–417.

Kassambara, A. (2025). *ggpubr: 'ggplot2' Based Publication Ready Plots*. R package version 0.6.0.

Keil, A. P. (2020). 2020 isee causal inference tutorial slides. https://github.com/alexpkeil1/2020_ISEE_causal/blob/master/slides/2020_ISEE_Keil_talk.pdf. Presentation slides from the International Society for Environmental Epidemiology (ISEE) 2020.

Keil, A. P., Buckley, J. P., O'Brien, K. M., Ferguson, K. K., Zhao, S., and White, A. J. (2020). A quantile-based g-computation approach to addressing the effects of exposure mixtures. *Environmental Health Perspectives*, 128(4):047004.

Kuhn and Max (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, 28(5):1–26.

Kyung, M., Gill, J., Ghosh, M., and Casella, G. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5:369–412.

Lempers, F. (1971). *Posterior Probabilities of Alternative Linear Models: Some Theoretical Considerations and Empirical Experiments*. Rotterdam University Press.

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics*, 63(4):1079–1088.

Marx-Stoelting, P., Rivière, G., Luijten, M., Aiello-Holden, K., Bandow, N., Baken, K., Cañas, A., Castano, A., Denys, S., Fillol, C., Herzler, M., Iavicoli, I., Karakitsios, S., Klanova, J., Kolossa-Gehring, M., Koutsodimou, A., Vicente, J. L., Lynch, I., Namorado, S., Norager, S., Pittman, A., Rotter, S., Sarigiannis, D., Silva, M. J., Theunis, J., Tralau, T., Uhl, M., van Klaveren, J., Wendt-Rasch, L., Westerholm, E., Rousselle, C., and Sanders, P. (2023). A walk in the parc: developing and implementing 21st century chemical risk assessment in europe. *Archives of Toxicology*, 97(3):893–908.

Microsoft and Weston, S. (2022a). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.17.

Microsoft and Weston, S. (2022b). *foreach: Provides Foreach Looping Construct.* R package version 1.5.2.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Morris, T. P., White, I. R., and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102.

OpenAI (2023). Chatgpt: GPT-4 language model. https://chat.openai.com/. Accessed: 2025-08-11.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pedersen, T. L. (2025). *patchwork: The Composer of Plots.* R package version 1.3.0.

Pelgrims, I., Devleesschauwer, B., Vandevijvere, S., De Clercq, E. M., Van der Heyden, J., and Vansteelandt, S. (2024). The potential impact fraction of population weight reduction scenarios on non-communicable diseases in belgium: application of the g-computation approach. *BMC Medical Research Methodology*, 24.

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2).

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.

R Core Team (2024). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Renzetti, S., Curtin, P., and Gennings, C. (2023). *gWQS: Generalized Weighted Quantile Sum Regression.* R package version 3.0.5.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512.

Rudis, B. (2024). *hrbrthemes: Additional Themes, Theme Components and Utilities for 'ggplot2'.* R package version 0.8.7.

Silva, E., Rajapakse, N., and Kortenkamp, A. (2002). Something from "nothing"-eight weak estrogenic chemicals combined at concentrations below noecs produce significant mixture effects. *Environmental science technology*, 36:1751–6.

Snowden, J. M., Rose, S., and Mortimer, K. M. (2011). Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7):731–738.

Tanner, E. M., Bornehag, C.-G., and Gennings, C. (2019). Repeated holdout validation for weighted quantile sum regression. *MethodsX*, 6:2855–2860.

Tay, J. K., Narasimhan, B., and Hastie, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software*, 106(1):1–31.

Thas, O. (2023). Linear models. https://othas.github.io/LIMO/. Accessed: 2025-05-24.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Vansteelandt, S. and Keiding, N. (2011). Invited commentary: G-computation–lost in translation? *American Journal of Epidemiology*, 173(7):739–742.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

VITO (2025). Vito - flemish institute for technological research. https://vito.be/en. Accessed: 2025-06-05.

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, 21(12):1–20.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

Wickham, H. and Bryan, J. (2023). *readxl: Read Excel Files*. R package version 1.4.3.

Wickham, H., François, R., Henry, L., Müller, K., and Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4.

Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. J. (1984). The collinearity problem in linear regression. the partial least squares (pls) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, 5(3):735–743.

Youngflesh, C. (2018). Mcmcvis: Tools to visualize, manipulate, and summarize mcmc output. *Journal of Open Source Software*, 3(24):640.

Yu, L., Liu, W., Wang, X., Ye, Z., Tan, Q., Qiu, W., Nie, X., Li, M., Wang, B., and Chen, W. (2022). A review of practical statistical methods used in epidemiological studies to estimate the health effects of multi-pollutant mixture. *Environmental Pollution*, 306:119356.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67:301 – 320.