Bayesian Interaction Estimation with High-Dimensional Dependent Predictors

by

Federico Ferrari

Department of Statistical Science
Duke University

Date: __________________
Approved:

_______________________
David B. Dunson, Advisor

_______________________
Amy H. Herring

_______________________
Mike West

_______________________
Jennifer Hoffman

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2021
Abstract
Bayesian Interaction Estimation with High-Dimensional Dependent Predictors
by
Federico Ferrari
Department of Statistical Science
Duke University

Date: __________________
Approved:

_______________________
David B. Dunson, Advisor

_______________________
Amy H. Herring

_______________________
Mike West

_______________________
Jennifer Hoffman

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2021
Abstract

Humans are constantly exposed to mixtures of different chemicals arising from environmental contamination. While certain compounds, such as heavy metals and mercury, are well known to be toxic, there are many complex mixtures whose health effects are still unknown. It is of fundamental public health importance to understand how these exposures interact to impact risk of disease and the health effects of cumulative exposure to multiple agents. The goal of this thesis is to build data-driven models to tackle major challenges in modern health applications, with a special interest in estimating statistical interactions among correlated exposures. In Chapter 1, we develop a flexible Gaussian process regression model (MixSelect) that allows to simultaneously estimate a complex nonparametric model and provide interpretability. A key component of this approach is the incorporation of a heredity constraint to only include interactions in the presence of main effects, effectively reducing dimensionality of the model search. Next, we focus our modelling effort on characterizing the joint variability of chemical exposures using factor models. In fact, chemicals usually co-occur in the environment or in synthetic mixtures; as a result, their exposure levels can be highly correlated. In Chapter 3, we build a Factor analysis for INteractions (FIN) framework that jointly provides dimensionality reduction in the chemical measurements and allows to estimate main effects and interactions. Through appropriate modifications of the factor modeling structure, FIN can accommodate higher order interactions and multivariate outcomes. Further, we
extend FIN to survival analysis and exponential families in Chapter 4, as medical studies often include collect high-dimensional data and time-to-event outcomes. We address these cases through a joint factor analysis modeling approach in which latent factors underlying the predictors are included in a quadratic proportional hazards regression model, and we provide expressions for the induced coefficients on the covariates. In Chapter 5, we combine factor models and nonparametric regression. We build a copula factor model for the chemical exposures and use Bayesian B-splines for flexible dose-response modeling. Finally, in Chapter 6 we propose a post-processing algorithm that allows for identification and interpretation of the factor loadings matrix and can be easily applied to the models described in the previous chapters.
To my Grandmas,
Adriana and Anna.
# Contents

Abstract iv  
List of Tables x  
List of Figures xi  
List of Abbreviations and Symbols xiii  
Acknowledgements xv  
1 Introduction 1  
2 Identifying main effects and interactions among exposures using Gaussian processes 5  
2.1 Introduction ......................................................... 5  
2.2 MixSelect Modeling Framework ................................... 7  
   2.2.1 Non-Identifiability and Projection ......................... 9  
   2.2.2 Variable Selection .......................................... 11  
2.3 Computational Challenges and Inference ...................... 13  
2.4 Simulations ....................................................... 17  
2.5 Environmental Epidemiology Application ........................ 22  
   2.5.1 Motivation .................................................. 22  
   2.5.2 Data Description ........................................... 23  
   2.5.3 Missing data and LOD ..................................... 25  
   2.5.4 Statistical Analysis ...................................... 27
List of Tables

2.1 MixSelect simulation results with $p = 25$, $n = 250$ .......................... 19
2.2 MixSelect simulation results with $p = 50$, $n = 250$ .......................... 20
2.3 MixSelect simulation results with $p = 25$, $n = 500$ .......................... 21
2.4 MixSelect simulation results with $p = 50$, $n = 500$ .......................... 22
2.5 Correlation matrix of chemical measurements ................................. 24
2.6 Coverage for in sample and out of sample predictive intervals .......... 28
2.7 Performance of MixSelect, BKMR, RAMP, hierNet, Family and PIE .... 30
2.8 Posterior inclusion probabilities when including interactions between chemical measurements with non-Hispanic Black ethnicity and Sex ... 32
3.1 FIN simulation results with $p = 25$ and dense $\Omega_0$ ......................... 61
3.2 FIN simulation results with $p = 25$ and sparse $\Omega_0$ ......................... 62
3.3 FIN simulation results with $p = 50$ and dense $\Omega_0$ ......................... 63
3.4 FIN simulation results with $p = 50$ and sparse $\Omega_0$ ......................... 64
4.1 Summary of Microarray data ............................................................. 88
# List of Figures

2.1 Realizations of projected GP ............................................. 10
2.2 Graphical representation of the MixSelect model ...................... 14
2.3 Estimated regression surface of model (a) with $n = 250$ and $p = 25$ . 18
2.13 Estimated main effects using MixSelect including interactions of chemical measurements with non-Hispanic Black ethnicity and Sex 32
2.4 Pattern of missingness in the chemical exposure, cholesterol and creatinine measurements ............................................. 35
2.5 Pattern of data below the limit of detection in the chemical measurements ........................................................................ 36
2.6 MixSelect residual plots ....................................................... 36
2.7 Histogram of the predictive mean at each MCMC iteration .......... 37
2.8 Estimated dose response curves for the chemicals Cadmium, Tungsten, Lead and Cobalt ......................................................... 38
2.9 Estimated interaction effects using MixSelect ........................... 39
2.10 Estimated covariate effects using MixSelect ............................ 40
2.11 Estimated main effects using MixSelect ............................... 40
2.12 Estimated regression surface for Cadmium interacted with Sex and non-Hispanic Black ethnicity ........................................ 41
3.1 Induced priors on main effects, pairwise interactions and 3rd order interactions ................................................................. 71
3.2 Histograms of the chemicals measurements ........................... 72
3.3 Pattern of Missing data of chemical measurements .................. 73
3.4 Correlation matrix and estimated factor loadings matrix . . . . . . . . 74
3.5 Estimated main effects in NHANES . . . . . . . . . . . . . . . . . . . 74
3.6 Estimated matrices of interactions in NHANES . . . . . . . . . . . 75
4.1 ROC curves fixed at different time points . . . . . . . . . . . . . . 87
4.2 Survival of high and low risk groups. . . . . . . . . . . . . . . . . . . 88
4.3 Factor loadings ovarian cancer dataset . . . . . . . . . . . . . . . . . 92
4.4 Out of sample accuracy for GLM factor, hiernet, RAMP and Family . 93
5.1 Histogram of chemical exposure and logged chemical exposure . . . 96
5.2 Histogram of three chemical exposures in the transformed space . . 97
5.3 Estimated regression surface of Selenium on BMI . . . . . . . . . . 101
5.4 Estimated regression surfaces of Cadmium, Copper and Mercury on
BMI . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 101
6.1 Non-identifiable trace plots . . . . . . . . . . . . . . . . . . . . . . . . 109
6.2 Example of postprocess Λ . . . . . . . . . . . . . . . . . . . . . . . . 110
6.3 Distribution of the condition numbers . . . . . . . . . . . . . . . . . 112
6.4 Comparison of running times in \( \log_{10} \) scale between MatchAlign and
rsp with full-SA. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 114
6.5 Matrix of factor loadings for the CNS cancer data . . . . . . . . . . 116
List of Abbreviations and Symbols

Symbols

\text{diag}(x_1, \ldots, x_p) \quad A \ p \times p \ diagonal \ matrix \ with \ entries \ x_1, \ldots, x_p \\
I_n \quad n \times n \ Identity \ matrix \\
\text{Ga}(a, b) \quad \text{Gamma} \ distribution \ with \ rate \ a \ and \ shape \ b \\
\text{Dirichlet}(a_1, \ldots, a_k) \quad \text{Dirichlet} \ distribution \ with \ parameters \ (a_1, \ldots, a_k) \\
\text{GP}(\mu(\cdot), c(\cdot)) \quad \text{Gaussian} \ process \ with \ mean \ \mu(\cdot) \ and \ covariance \ c(\cdot) \\
\text{N}(\mu, \sigma^2) \quad \text{Normal} \ distribution \ with \ mean \ \mu \ and \ variance \ \sigma^2 \\
\text{N}_p(\mu, \Sigma) \quad \text{Multivariate} \ Normal \ distribution \ with \ mean \ \mu \ and \ covariance \ \Sigma \\
\otimes \quad \text{Tensor} \ product \\
||A||_F \quad \text{Frobenius} \ norm \ of \ matrix \ A \\
\text{KL}(\Theta_0, \Theta) \quad \text{Kullback-Leibler} \ divergence \ between \ p(y|\Theta_0) \ and \ p(y|\Theta) \\
\text{DE}(\phi) \quad \text{zero} \ mean \ double-exponential \ with \ parameter \ \phi \\
\mathcal{O}(\cdot) \quad \text{Run} \ time \ or \ storage \ complexity \\
\mathbb{R} \quad \text{Real} \ numbers \\
\text{InvGauss}(a, b) \quad \text{Inverse-Gaussian} \ distribution \ with \ parameters \ (a, b) \\
\text{GInvGauss}(a, b, c) \quad \text{Generalized} \ Inverse-Gaussian \ distribution \ with \ parameters \ (a, b, c) \\

Abbreviations

\text{AUC} \quad \text{Area} \ Under \ the \ Curve \\
\text{BKMR} \quad \text{Bayesian} \ Kernel \ Machine \ Regression
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMI</td>
<td>Body-mass index</td>
</tr>
<tr>
<td>cdf</td>
<td>cumulative distribution function</td>
</tr>
<tr>
<td>DE</td>
<td>Double-exponential</td>
</tr>
<tr>
<td>DL</td>
<td>Dirichlet-Laplace</td>
</tr>
<tr>
<td>FIN</td>
<td>Factor INteractions</td>
</tr>
<tr>
<td>FR</td>
<td>Frobenius</td>
</tr>
<tr>
<td>GP</td>
<td>Gaussian Process</td>
</tr>
<tr>
<td>GP-LVM</td>
<td>GP Latent Variable Models</td>
</tr>
<tr>
<td>iid</td>
<td>independent and identically distributed</td>
</tr>
<tr>
<td>KL</td>
<td>Kullback-Leibler</td>
</tr>
<tr>
<td>MALA</td>
<td>Metropolis-Adjusted Langevin Algorithm</td>
</tr>
<tr>
<td>MCMC</td>
<td>Markov Chain Monte Carlo</td>
</tr>
<tr>
<td>MH</td>
<td>Metropolis-Hastings</td>
</tr>
<tr>
<td>MSE</td>
<td>Mean Squared Error</td>
</tr>
<tr>
<td>NHANES</td>
<td>National Health and Nutrition Examination Survey</td>
</tr>
<tr>
<td>pdf</td>
<td>probability density function</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>SEM</td>
<td>Structural Equation Models</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
</tbody>
</table>

xiv
Acknowledgements

First, I would like to thank David Dunson for being my advisor during the four years I spent at Duke. It was certainly an honor to work with him. As an advisor, David was always direct, frank and certainly did not hold back when I made a mistake, all qualities that I learnt to appreciate and value, and that made me grow as a researcher and a person. More importantly, I always admired his thinking process and his excitement towards new ideas has been really contagious.

I would also like to thank my defense committee members, Amy Herring, Mike West and Kate Hoffman for their support, helpful comments and guidance. Thank you to the DSS staff and in particular to Lori who has always been friendly and very helpful.

I would like to acknowledge the support received from the grant 1R01ES028804-01 of the National Institute of Environmental Health Sciences of the United States Institutes of Health. In particular, I would like to thank Bonnie Joubert for many fruitful discussions about the applied sides of the statistical methods I worked on.

I would like to thank those people I had the pleasure of sharing my time with in Durham for making my entire PhD experience incredibly enjoyable, Alessandro, Becky, Bernardo, Bora, Evan, Fan, Federica, Francesca, Ghassen, Heather, Jae Yeon, Jordan, Luca, Max, Maurizio, Meme, Michele, Roi, Sally, Sarah, Victor, Vittorio, the King’s kitchen mates and most importantly my girlfriend Zhuoli. Thanks again to Vitto for re-writing the previous sentence. I will also never forget the support received
from my close friends and family back home, especially in the most challenging times.
Humans are constantly exposed to mixtures of different chemicals arising from environmental contamination. While certain compounds, such as heavy metals and mercury, are well known to be toxic, there are many complex mixtures whose impacts on health are still unknown. The primary focus of epidemiology and toxicology studies has been on examining chemicals one at a time. However, chemicals usually co-occur in the environment or in synthetic mixtures, and hence assessing joint effects is of critical public health concern. Certainly, findings from one chemical at a time studies may be misleading (Dominici et al., 2010), (Mauderly and Samet, 2008). Hence, it is of fundamental public health importance to understand how these exposures interact to impact risk of disease and the health effects of cumulative exposure to multiple agents (Braun et al., 2016).

Building a flexible joint model for mixtures of chemicals is suggested by the National Research Council (Mauderly et al., 2010), (Vedal and Kaufman, 2011), (Council et al., 2004). Several attempts have been made to simultaneously detect the effect of different chemicals on health outcomes, using either parametric or nonparametric regression techniques. The former include regularization methods, like LASSO
(Roberts and Martin, 2005) or Ridge Regression, and deletion/substitution/addition algorithms (Sinisi and van der Laan, 2004), (Mortimer et al., 2008). Some of these techniques have also been applied to high dimensional spaces (Hao and Zhang, 2014). While providing interpretability in terms of linear effects and pairwise interactions, the resulting dose response surface is typically too restrictive, as chemicals often have nonlinear effects.

In Chapter 1, we propose a MixSelect framework to simultaneously estimate a flexible nonparametric surface and interpret the association between chemicals and the health outcome. To do so, we decompose the expected health outcome into a linear main effect, pairwise interactions, and a nonlinear deviation. Our interest is in model selection for these different components, accounting for uncertainty and addressing non-identifiability between the linear and nonparametric components of the semiparametric model. We propose a Bayesian approach to inference, placing variable selection priors on the different components, and developing a Markov chain Monte Carlo (MCMC) algorithm. A key component of our approach is the incorporation of a heredity constraint to only include interactions in the presence of main effects, effectively reducing dimensionality of the model search. We adapt a projection approach developed in the spatial statistics literature to enforce identifiability in modeling the nonparametric component using a Gaussian process. We also employ a dimension reduction strategy to sample the nonlinear random effects that aids the mixing of the MCMC algorithm.

Chemicals usually co-occur in the environment or in synthetic mixtures; as a result, their exposure levels can be highly correlated. High correlation usually occurs within blocks of chemicals; for example, this can arise when an individual is exposed to a product having a mixture of chemicals and when chemical measurements consist of metabolites or breakdown products of a parent compound. There is a large public health interest in studying \( E \times E \), \( E \times G \) and \( G \times G \) interactions, with \( E = \) environ-
mental exposures and $G =$ genetic factors. However, current methods for quadratic regression are not ideal in these applications due to the level of correlation in the predictors, the fact that strong sparsity assumptions are not appropriate, and the need for uncertainty quantification. Regarding the issue of sparsity, some exposures are breakdown products of the same compound, so it is unlikely that only one exposure has an effect on the outcome. Also, it is statistically challenging to tell apart highly correlated covariates with limited data. For this reason, it is appealing given the data structure to select blocks of correlated exposures together instead of arbitrarily selecting one chemical in a group.

In Chapter 3, we propose a latent factor joint model, which includes shared factors in both the predictor and response components while assuming conditional independence. By including a quadratic regression in the latent variables in the response component, we induce flexible dimension reduction in characterizing main effects and interactions. We propose a Bayesian approach to inference under this Factor analysis for INteractions (FIN) framework. Through appropriate modifications of the factor modeling structure, FIN can accommodate higher order interactions and multivariate outcomes.

It is also very common to collect high-dimensional data and time-to-event outcomes in medical studies. In such cases, Cox proportional hazards models are commonly used with variable selection or regularization to deal with the issue of high dimensionality. Such approaches are typically not specifically designed to accommodate interactions among predictors. Cases in which predictors can be moderately to highly correlated and heredity assumptions commonly used in high-dimensional interaction detection problems are unwarranted are especially of interest. In Chapter 4, we propose to address these cases through a joint factor analysis modeling approach in which latent factors underlying the predictors are included in a quadratic proportional hazards regression model. We provide expressions for the induced coefficients
on the covariates. We further show how our results are applicable to any generalized linear model with logarithmic link, such as Poisson and negative binomial regression.

In Chapter 5 we combine dimensionality reduction and flexible nonparametric regression. We provide a transformation for the chemical measurements that coherently allows for sampling of data points that are missing or are below the limit of detection (LOD), and flexibly and efficiently model dose-response curves. We propose to use a Gaussian copula model that allows to separate the dependence structure among exposures with the marginal distribution of the chemical measurements. The Gaussian copula model admits a latent representation of the measurements, where the transformed chemicals have a normal distribution. As exposures can be moderately high-dimensional with high correlations within blocks of variables, we model the dependence among the transformed chemicals using factor models (Murray et al., 2013). We then extend the model to regression settings by linking the latent variables and the outcome with a flexible specification using Bayesian splines (DiMatteo et al., 2001). Using a one-to-one transformation via the latent representation of the Gaussian copula model, we provide inference in the original scale of the ‘exposures’.

The Gaussian factor model is the crucial building block of the models introduced in Chapter 3, 4 and 5. Without identifiability constraints, it is well known that the factor loadings display rotational ambiguity. To perform inference and interpret the factor loadings matrix, we propose a computationally efficient post-processing algorithm in Chapter 6. We firstly orthogonalize the posterior samples using Varimax (Kaiser, 1958) and then tackle label and sign switching with a greedy matching algorithm. We compare the performance and computational complexity with other methods using a simulation study and high-dimensional gene expression data.

To the facilitate the reader, every chapter can be read independently. The code is available on Github and in the infinite R package on CRAN.
Identifying main effects and interactions among exposures using Gaussian processes

2.1 Introduction

Humans are exposed to mixtures of different chemicals arising due to environmental contamination. Certain compounds, such as heavy metals and mercury, are well known to be toxic to human health, whereas very little is known about how complex mixtures impact health outcomes. One of the key questions that epidemiology should address according to Braun et al. (2016) is: *What is the interaction among agents?* The primary focus of epidemiology and toxicology studies has been on examining chemicals one at a time. However, chemicals usually co-occur in the environment or in synthetic mixtures, and hence assessing joint effects is of critical public health concern. Certainly, findings from one chemical at a time studies may be misleading (Dominici et al., 2010), (Mauderly and Samet, 2008).

Building a flexible joint model for mixtures of chemicals is suggested by the National Research Council (Mauderly et al., 2010), (Vedal and Kaufman, 2011), (Council et al., 2004). Recently, several studies have shown relationships between complex
mixtures of chemicals and health or behavior outcomes. For example, Sanders et al. (2015) review findings on perinatal and childhood exposures to Cadmium (Cd), Manganese (Mn) and metal mixtures. Several attempts have been made to simultaneously detect the effect of different chemicals on health outcomes, using either parametric or nonparametric regression techniques. The former include regularization methods, like LASSO (Roberts and Martin, 2005) or Ridge Regression, and deletion/substitution/addition algorithms (Sinisi and van der Laan, 2004), (Mortimer et al., 2008). Some of these techniques have also been applied to high dimensional spaces (Hao and Zhang, 2014). While providing interpretability in terms of linear effects and pairwise interactions, the resulting dose response surface is typically too restrictive, as chemicals often have nonlinear effects.

Nonparametric models can also be used to estimate interactions among chemicals, ranging from tree based methods (Hu et al., 2008), (Lampa et al., 2014), to Bayesian Kernel Machine Regression (BKMR) (Bobb et al., 2014), (Valeri et al., 2017), (Liu et al., 2017) and Bayesian P-splines (Lang and Brezger, 2004). Although tree based methods, like Boosted Trees or Random Forests, are convenient computationally and often provide accurate predictions, interpretation of covariate effects is typically opaque. While providing good predictive performance, nonparametric regression surfaces like BKMR provide excessive flexibility when a simple parametric model provides an adequate approximation. On the other hand, the estimation of interactions with Bayesian P-splines becomes extremely challenging when $p$ is larger than $\sim 10$, which is common in environmental epidemiology.

Our goal is to simultaneously estimate a flexible nonparametric model and provide interpretability. To do so, we decompose the regression surface on the health outcome into a linear effect, pairwise interactions and a nonlinear deviation. This specification, which we describe in Section 2.2, allows one to interpret the parametric portion of the model while also providing flexibility via the nonparametric
component. We address identifiability between the parametric and nonparametric part of the model by adapting a projection approach developed in spatial statistics, see Section 2.2.1. We accurately take into account uncertainty in model selection on the different components of the model with a Bayesian approach to inference. We choose spike and slab priors for main effects and pairwise interactions (George and McCulloch, 1997) and allow for variable selection of nonlinear effects adapting the approach of Savitsky et al. (2011), which introduces spike and slab priors in the Gaussian process setting. We reduce computation imposing a heredity condition (Chipman, 1996), described in Section 2.2.2, and applying a dimension reduction approach to the Gaussian process surface (Guan and Haran, 2018), (Banerjee et al., 2012), which we describe in Section 2.3.

We describe our efficient Bayesian inference procedure in Section 2.3 and we propose a Markov chain Monte Carlo (MCMC) algorithm. We compare our method with the state of the art nonparametric models and with methods for interaction estimation in Section 2.4. Finally, in Section 2.5 we assess the association of metal concentrations on BMI using data from the National Health and Nutrition Examination Survey (NHANES). This application shows the practical advantages of our method and how it could be used as a building block for more complex analysis.

2.2 MixSelect Modeling Framework

Let $y_i$ denote a continuous health outcome for individual $i$, let $x_i = (x_{i1}, \ldots, x_{ip})^T$ denote a vector of ‘exposure’ measurements, and let $z_i = (z_{i1}, \ldots, z_{iq})^T$ denote covariates. For example, ‘exposure’ may consist of the levels of different chemicals in a blood or urine sample, while covariates correspond to demographic factors and potential confounders. For interpretability our focus is on decomposing the impact of the exposures into linear main effects, linear pairwise interactions, and a nonparametric deviation term, while including an adjustment for covariates. Each of the exposure
effect components will include a variable selection term so that some exposures may have no effect on the health response, while others only have linear main effects, and so on. This carefully structured semiparametric model differs from usual black-box nonparametric regression analyses that can characterize flexible joint effects of the exposures but lack interpretability and may be subject to overfitting and the curse of dimensionality. By including variable selection within our semiparametric model, we greatly enhance interpretability, while also favoring a more parsimonious representation of the regression function.

Our model structure can be described as follows:

\[
y_i = x_i^T \beta + \sum_{j=1}^{p} \sum_{k>j} \lambda_{jk} x_{ij} x_{ik} + g^*(x_i) + z_i^T \alpha + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),
\]

\[
g_n^* = Pg_n, \quad g \sim \text{GP}(0, c),
\]

where \( \beta = (\beta_1, \ldots, \beta_p)^T \) are linear main effects of exposures, \( \lambda = \{\lambda_{jk}\} \) are pairwise linear interactions, \( g_n = [g(x_1), \cdots, g(x_n)] \) is a nonparametric deviation, and \( \alpha = (\alpha_1, \ldots, \alpha_q)^T \) are coefficients for the covariates. We include variable selection in each of the three terms characterizing the exposure effects, as we will describe in detail in Section 2.2.2. In addition, a key aspect of our model is the inclusion of a constraint on the nonparametric deviation to enforce identifiability separately from the linear components. This is the reason for the \( P \) term multiplying \( g \) in the above expression, with \( P \) a projection matrix to be described in Section 2.2.1. The notation \( \text{GP}(0, c) \) denotes a Gaussian process (GP) centered at zero with covariance function \( c \) controlling the uncertainty and smoothness of the realizations.

In spatial statistics it is common to choose a Matern covariance function, but in our setting we instead use a squared exponential covariance to favor smooth depen-
tures from linearity; in particular, we let

\[ c(x, x') = \text{cov}(g(x), g(x')) = \tau^2 \exp \left\{ \sum_{j=1}^{p} \rho_j (x_j - x'_j)^2 \right\}, \tag{2.2} \]

where \( \rho_j \) is a smoothness parameter specific to the \( j^{th} \) exposure and \( \tau^2 \) is the signal variance. Similar covariance functions are common in the machine learning literature, and are often referred to as automatic relevance determination (ARD) kernels (Qi et al., 2004). They have also been employed by Bobb et al. (2014). However, to our knowledge previous work has not included linear main effects and interactions or a projection adjustment for identifiability. The proposed GP covariance structure allows variable selection (\( \rho_j = 0 \) eliminates the \( j^{th} \) exposure from the nonparametric deviation) and different smoothness of the deviations across the exposures that are included. For example, certain exposures may have very modest deviations while others may vary substantially from linearity.

The proposed model structure is quite convenient computationally, leading to an efficient Markov chain Monte Carlo (MCMC) algorithm, which mostly employs Gibbs sampling steps. We will describe the details of this algorithm in Section 2.3, but we note that the projection adjustment for identifiability greatly aids mixing of the MCMC, and our code can be run efficiently for the numbers of exposures typically encountered in environmental epidemiology studies (up to one hundred). Code for implementation is available at \texttt{https://github.com/fedfer/MixSelect}.

2.2.1 Non-Identifiability and Projection

Confounding between the Gaussian process prior and parametric functions is a known problem in spatial statistics and occurs when spatially dependent covariates are strongly correlated with spatial random effects, see Hanks et al. (2015) or Guan and Haran (2018). This problem is exacerbated when the same features are included in
both the linear term and in the nonparametric surface. For this reason we project the nonlinear random effects $g$ on the orthogonal column space of the matrix containing main effects.

The usual projection matrix on the column space of $X$ is equal to $P_X = X(X^TX)^{-1}X^T$. We define $P = P_X^\perp = I_n - P_X$ and set $g_n^* = Pg_n$. Firstly, notice that the projection has an effect on the variance of the generated nonlinear effects, in particular:

$$\sum_{i=1}^n (g_{i,n}^*)^2 \leq \sum_{i=1}^n (g_{i,n})^2$$

This follows from

$$(g_n^*)^Tg_n^* = [(I_n - P_X)g_n]^T[(I_n - P_X)g_n] = g_n^Tg_n - (P_Xg_n)^T(P_Xg_n) \leq g_n^Tg_n.$$ 

Figure 2.1 shows examples of realizations of $g_n$ and $g_n^*$. The curvature of the functions drawn from the projected GP is greater than the curvature in the non-projected case.

Another possibility would be to project the nonlinear random effects $g_n$ on the orthogonal column space of the matrix containing both main effects and interactions.
However, we noticed in our simulations that this would make the resulting nonparametric surface too restrictive, especially when the number of possible interactions $\frac{p(p-1)}{2}$ is greater than $n$, resulting in a worse performance of the model. We did not experience significant confounding between the interaction effects and the nonlinear regression surface. Finally, notice that rather than sampling $g$ and then projecting onto the orthogonal column space of $X$, we can equivalently sample $g^*$ from a Gaussian process with covariance matrix $PcP^T$. Another option that we explore in Section 2.3 consists in integrating out the nonlinear effects.

### 2.2.2 Variable Selection

In this section we describe the variable selection approach that we develop in order to provide uncertainty quantification and achieve parsimonious model specification. We assume that the chemical measurements and the covariates have been standardized prior to the analysis. We choose spike and slab priors for the main effects and nonlinear effects. Regarding main effects, we choose a mixture of a normal distribution with a discrete Dirac delta at zero. Let us define as $\gamma_k$ the indicator variable that is equal to 1 if the $k^{th}$ variable is active in the linear main effect component of the model and equal to 0 otherwise. We have that $\beta_k \sim \gamma_k N(0,1) + (1 - \gamma_k)\delta_0$. For the $\gamma_k$ we assume independent Bernoulli priors with success probability $\pi$. We endow $\pi$ with a Beta distribution with parameters $(a_\pi, b_\pi)$. The prior expected number of predictors included in the model is $p_{a_\pi, b_\pi}$, which can be used to elicitate the hyper-parameters $(a_\pi, b_\pi)$. As a default we choose $a_\pi = b_\pi = 1$, which corresponds to a Uniform distribution on $\pi$. We endow the main effects of covariate adjustments $\alpha_l$ with a Normal prior $N_q(0, I)$, for $l = 1, \ldots, q$.

We impose an heredity condition for the interactions. The heredity condition is commonly employed for datasets with $p \in [20, 100]$ by one-stage regularization methods like Bien et al. (2013) and Haris et al. (2016) or two stage-approaches as
Hao et al. (2018) when \( p > 100 \). Strong heredity means that an interaction between two variables is included in the model only if the main effects are. For weak heredity it suffice to have one main effect in the model in order to estimate the interaction of the corresponding variables. Formally:

S: \[
\lambda_{j,k} | \gamma_j = \gamma_k = 1 \sim N(0, 1), \quad \lambda_{j,k} | (\gamma_j = \gamma_k = 1)^C \sim \delta_0
\]

W: \[
\lambda_{j,k} | (\gamma_j = \gamma_k = 0)^C \sim N(0, 1), \quad \lambda_{j,k} | \gamma_j = \gamma_k = 0 \sim \delta_0
\]

where \( S \) and \( W \) stand for strong and weak heredity respectively, and \( \delta_0 \) is a Dirac distribution at 0. Models that satisfy the strong heredity condition are invariant to translation transformations in the covariates. Weak heredity provides greater flexibility with the cost of considering a larger number of interactions, leading to a potentially substantial statistical and computational cost. Consider the case when the \( j^{th} \) covariate has a low effect on the outcome but the interaction with the \( k^{th} \) feature is significantly different than zero. Strong heredity will sometimes prevent us from discovering this pairwise interaction. Heredity reduces the size of the model space from \( 2^p + \binom{p}{2} \) to \( \sum_{i=0}^{p} \binom{p}{i} 2^{\binom{i}{2}} \) or \( \sum_{i=0}^{p} \binom{p}{i} 2^p-i-i+1/2 \) for strong and weak heredity, respectively. The heredity condition can also be extended to higher order interactions.

As for the main effects and interactions, we apply a variable selection strategy for the nonlinear effects. We endow the signal standard deviation \( \tau \) with a spike and slab prior, i.e. \( \tau \sim \gamma^\tau F_\tau(\cdot) + (1 - \gamma^\tau) \delta_0 \), where \( F_\tau(\cdot) \) is a Gamma distribution with parameters \((1/2, 1/2)\) and \( \gamma^\tau \) has a Bernoulli(1/2) prior. We noticed that this spike and slab prior prevents overfitting of the nonlinear term in high dimensional settings, in particular when the variables are highly correlated and the true regression does not include nonlinear effects. This added benefit is highlighted in Section 2.4 when comparing with BKMR. Finally, when \( \gamma^\tau = 0 \), the regression does not include nonlinear effects, resulting in faster computations. In this case, the computational
complexity of the model equals the one of a Bayesian linear model with heredity constraints.

With respect to the covariate specific nonlinear effects, we follow the strategy of Savitsky et al. (2011), which is also employed by Bobb et al. (2014), and endow the smoothness parameters $\rho_1, \ldots, \rho_p$ with independent spike and slab priors. In particular $\rho_k \sim \gamma^\tau \gamma_k^p F_\rho(\cdot) + (1 - \gamma^\tau)(1 - \gamma_k^p)\delta_0$, where $F_\rho(\cdot)$ is a Gamma distribution with parameters $(1/2, 1/2)$. Only when $\gamma^\tau$ is different than zero, we allow the covariate specific nonlinear effects $\gamma_j^p$ to be different than zero. When $\gamma_k^p = 0$, the $k^{th}$ exposure is eliminated from the nonparametric term $g$ in (2.1). As before, we choose a Bernoulli prior for $\gamma_k^p$ with mean $\varphi$, and we endow $\varphi$ with a Beta prior with parameters $(a_\varphi, b_\varphi)$. As a default we choose $a_\varphi = b_\varphi = 1$, which corresponds to a Uniform distribution on $\varphi$. A graphical representation of the model can be found in Figure 2.2.

2.3 Computational Challenges and Inference

In this section we describe how we conduct inference for model (2.1). We also address the computational challenges associated with Gaussian process regression in the Bayesian framework and summarize the MCMC algorithm at the end of the section.

We defined a mixture of Normal priors for the main effects, interactions and the coefficients of the covariate adjustments, namely $\beta, \lambda$ and $\alpha$, in Section 2.2.2. Having a Gaussian likelihood, the full conditionals for these parameters are conjugate, hence we can directly sample from multivariate normal distributions within a Gibbs sampler. This operation could be quite expensive since the number of parameters is of order $p^2$. However, thanks to the strong heredity condition, we only need to sample the interactions between the variables with non-zero main effects and we set
to zero all the others. Given each of the elements of $\beta$, $\lambda$ and $\alpha$ we can update the labels $\gamma$ with a Bernoulli draw. We also re-parametrize the model setting $\tau = \tau^* \sigma$, so that we can directly update $\sigma^2$ from an inverse Gamma distribution.

Dealing with the nonlinear term $g$ can also be expensive since we need to sample $n$ parameters at each iteration. For this reason, we integrate out the GP term so that marginally the likelihood of model (2.1) is equivalent to:

$$y|\beta, \Lambda, c \sim N(X\beta + diag(X\Lambda X^T) + \alpha Z, \sigma^2 I_n + PcP^T),$$

(2.3)

where $\Lambda$ is a upper triangular matrix such that $\Lambda_{j,k} = \lambda_{j,k}$ when $k > j$ and zero otherwise.

The covariance matrix depends on the hyperparameters $\rho_j$, for $j = 1, \ldots, p$, that define the variable selection scheme for the nonlinear effects. The priors for the smoothness parameters $\rho_j$ and $\tau^2$ defined in Section 2.2.2 are not conjugate.
so that we need a Metropolis-Hastings step within the Gibbs sampler to sample these parameters. In order to compute the acceptance ratio, we need to evaluate the likelihood of (2.3) and invert the matrix \( \sigma^2 I_n + PcP^T \) of dimension \( n \): such operation is of complexity \( O(n^3) \) and needs to be done \( p \) times. For this reason we approximate the matrix \( PcP^T \) with the strategy described in Algorithm 1 of Guan and Haran (2018). This approach is a generalization of Banerjee et al. (2012) and uses random projections to find an approximation of the Eigen Decomposition of \( PcP^T \). In particular we approximate this matrix as \( U_mD_mD_m^T \), where \( m \) is related to the order of the approximation, with \( m \) usually being much smaller than \( n \). \( D_m \) is a diagonal matrix of dimension \( m \) and \( U_m \) is of dimension \( n \times m \). We can now apply the Sherman-Morrison-Woodbury formula to compute the inverse of \( \Sigma = \sigma^2 I_n + PcP^T \):

\[
\Sigma^{-1} = (\sigma^2 I_n + PcP^T)^{-1} \approx (\sigma^2 I_n + U_mD_mD_m^T)^{-1} = \\
\frac{1}{\sigma^2}(I_n + U_m(\sigma^2 D_m + U_m^T U_m)^{-1} U_m^T),
\]

which now involves the inversion of an \( m \times m \) matrix. Similarly we can simplify the computations for the determinant of \( \Sigma \) using the Determinant Lemma (Harville, 1998):

\[
|\Sigma| = |\sigma^2 I_n + PcP^T| \approx \sigma^{2n} \prod_{j=1}^{m} (D_{m;j,j}^{-1} + \sigma^{-2}) D_{m;j,j}. 
\]

It is challenging to design a sampler with satisfactory mixing for the smoothness parameters \( \{\rho_j\} \). However we obtained good performance for an add-delete sampler, which updates \( \rho_j \) at every iteration. When the previous \( \rho_j = 0 \), we perform add move: sample from a distribution with support on \( \mathbb{R}_+ \). When \( \rho_j \neq 0 \), we perform a delete move and propose \( \rho_j = 0 \). Then, for the \( \rho_j \neq 0 \), we also perform the Gibbs-type move and sample from the same proposal as in the add move. The MCMC sampler is summarized in Algorithm 1.
Algorithm 1 MCMC algorithm for sampling the parameters of model (2.1)

Step 1 Sample $\gamma_j$ for $j = 1, \ldots, p$ from

$$
\pi(\gamma_j | \cdot) \sim Bernoulli\left(\frac{1}{1 + \frac{1}{\pi} R_j}\right)
$$

where $R_j = \left|X_j^T \Sigma^{-1} X_0 + I\right|^{-1/2} \exp\left(\frac{1}{2} m_q V_{0,m0}\right)$, $\Sigma = \sigma^2 I_n + P c P^T$, $m_0 = X_0^T \Sigma^{-1} y$ and $V_0 = (X_0^T \Sigma^{-1} X_0 + I)^{-1}$. $X_0$ is the matrix of covariates such that $\gamma_k = 1$ for $k \neq j$. $X_1$ is the matrix of covariates such that $\gamma_k = 1$ for $k = 1, \ldots, p$, with $X_j$ included.

Step 2 Sample $\pi$ from $\pi(\pi | \cdot) \sim Beta(a_\pi + \sum_{j=1}^{p} \gamma_j, b_\pi + p - \sum_{j=1}^{p} \gamma_j)$

Step 3 Sample the main coefficients $\beta_j$ from $\pi(\beta_j | \cdot) \sim N(V X_j^T \Sigma^{-1} (y - \alpha Z - diag(X\Lambda X^T)), V)$, where $V = (X_j \Sigma^{-1} X_j^T + I)^{-1}$ and the subscript $\gamma$ indicates that we are including only the variables such that $\gamma_j = 1$

Step 4 Set $\lambda_{j,k}$ equal to zero according to the chosen heredity condition. Then update $\lambda_{j,k}$ following an appropriate modification of Step 2

Step 5 Sample $\alpha$ following an appropriate modification of Step 2

Step 6 If $\gamma_{\tau} = 0$, set $\rho_j = 0$ and $\gamma_j^* = 0$ and move to Step 7, else go to Step 6’.

Step 6’ If $\rho_j \neq 0$, perform delete move: propose $\rho_j^* = 0$ and $\gamma_j^* = 0$. If $\rho_j = 0$ perform add move: propose $\rho_j^* > 0$ and $\gamma_j^* = 1$, for $j = 1, \ldots, p$. Compute $U_* D_* U_m^T$ with the approximation of Section 2.3, $\Sigma^*$ with Sherman-Woodbury formula and $|\Sigma^*|$ with determinant lemma. Then compute:

$$
-2 \log(r) = \log|\Sigma^*| - \log|\Sigma| + \frac{1}{2} \mu^T (\Sigma^* - \Sigma) \mu,
$$

where $\mu = y - (Z \alpha + X \beta + diag(X \Lambda X^T))$. Sample $u$ from a Uniform distribution in the interval $(0, 1)$ and if $\log(r) > \log(u)$, set $\rho_j = \rho_j^*$, $\gamma_j = \gamma_j^*$, $\Sigma = \Sigma^*$, $|\Sigma^{-1}| = |\Sigma^{-1}|$

Step 7 For all $j = 1, \ldots, p$ such that $\rho_j \neq 0$, perform a Gibbs-type move: sample $\rho_j^*$ from a symmetric proposal distribution and then follow Step 5.

Step 8 Sample $\varphi$ following an appropriate modification of Step 2.

Step 9 Sample $\tau^2$ from a symmetric proposal distribution and update following an appropriate modification of Step 5. If $\tau^2 \neq 0$ perform a Gibbs-type move.

Step 10 sample $\sigma^2$ from $\pi(\sigma^2 | \cdot) \sim InvGamma\left(\frac{1+n}{2}, \frac{1+n+P c P^T \tau^2}{2}\right)$ where $c'(x, x^*) = (\tau^*)^2 \exp\left\{\sum_{j=1}^{p} \rho_j (x_j - x_j^*)^2\right\}$

\[16\]
2.4 Simulations

In this section we compare the performance of our model with respect to five other methods: BKMR (Bobb et al., 2014), Family (Haris et al., 2016), hierNet (Bien et al., 2013), PIE (Wang et al., 2019) and RAMP (Hao et al., 2018). BKMR is a nonparametric Bayesian method that employs Gaussian process regression with variable selection in a similar fashion as model (2.1). Family, hierNet, PIE and RAMP are designed for interaction selection in moderate to high dimensional settings. We generate the covariates independently $X_i \sim N_p(0, I_p)$ for $i = 1, \cdots, n$, for $n = 250, 500$ and $p = 25, 50$, so that the number of parameters that we estimate with model (2.1) is 353 and 1352, respectively. We generate the outcome as follows:

(a) $y_i = x_1 - x_2 + x_3 + 2x_1x_2 - x_1x_3 + \frac{1}{2}x_4^2 + \frac{4}{\exp(-2x_5) + 1} + \epsilon_i$

(b) $y_i = x_1 + x_2 - x_3 - x_4 + 2x_1x_2 - x_1x_3 - x_2x_3 - 2x_3x_4 + \epsilon_i$

(c) $y_i = \sin(x_1 + 3x_3) - \frac{1}{2}x_3^2 + \exp(-0.1 \cdot x_1) + \epsilon_i$

where $\epsilon_i \sim N(0,1)$. The first setting involves a model with strong heredity and nonlinear effects, whereas the second is an interaction model and the third a nonlinear model. We evaluate the performance on a test dataset of 100 units with predictive mean squared error for all the models. We compute the Frobenious norm for the matrix containing pairwise interactions for Family, hierNet, RAMP and PIE. The Frobenious norm between two square matrices $\Lambda$ and $\hat{\Lambda}$ of dimension $p$ is defined as

$$\sqrt{\text{trace}((\Lambda - \hat{\Lambda})^T(\Lambda - \hat{\Lambda}))}.$$

We also compute posterior inclusion probabilities of nonlinear effects, so that we can calculate the percentage of True positive and True negative nonlinear effects for our method and BKMR. We average the results across 50 simulations. The results for
n = 250 are summarized in Table 2.1 and Table 2.2 and n = 500 are summarized in Table 2.3 and Table 2.4.

Across all the simulation scenarios, our model consistently achieves nearly the best predictive performance in terms of prediction error and Frobenious norm, and is able to identify main effects, interactions and nonlinear effects. The experiments highlight the advantages of MixSelect in the context of the application, where the dose-response surfaces usually have roughly linear, hill-shaped or sigmoid shapes. Hence constraining the flexible nonparametric surface allows MixSelect to have a predictive and inference advantage over BKMR, which is the main nonparametric method used in environmental epidemiology applications. For model (a), we achieve a better performance because of the decomposition of the regression surface, and we correctly identify linear and nonlinear effects. With respect to model (b), our method is able to correctly estimate a regression surface without nonlinear effects, thanks to the spike and slab prior on the term $\tau$. We also achieve a similar if not better performance in the nonlinear scenario of method (c). Finally, Figure 2.3 shows the estimated regression surface versus the true surface for model (a), when $n = 250$ and $p = 25$.

**Figure 2.3:** Estimated regression surface of model (a) with $n = 250$ and $p = 25$. The red line indicates the true curve, the black line the estimated function, the grey bands the pointwise 99% posterior credible intervals for the mean function and the marks on the x-axis the data points in the training set.
Table 2.1: Results from the simulation study under the three scenarios with $p = 25$, $n = 250$. We computed test error, FR for interaction effects, percentage of true positives and true negatives for main effects and interactions for MixSelect, BKMR, hierNet, Family, PIE and RAMP. We divided each value of test error and FR by the best (lowest) result for that metric. This makes the metric of the best model equal to 1.

<table>
<thead>
<tr>
<th></th>
<th>MixSelect</th>
<th>BKMR</th>
<th>hierNet</th>
<th>Family</th>
<th>PIE</th>
<th>RAMP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>model (b)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test MSE</td>
<td>1</td>
<td>1.178</td>
<td>1.444</td>
<td>3.299</td>
<td>1.001</td>
<td>1.193</td>
</tr>
<tr>
<td>FR</td>
<td>1</td>
<td>2.962</td>
<td>7.234</td>
<td>4.027</td>
<td>2.652</td>
<td></td>
</tr>
<tr>
<td>TP main</td>
<td>1</td>
<td>1</td>
<td>0.974</td>
<td>0.974</td>
<td>0.962</td>
<td></td>
</tr>
<tr>
<td>TN main</td>
<td>0.941</td>
<td>0.865</td>
<td>0.869</td>
<td>0.692</td>
<td>0.921</td>
<td></td>
</tr>
<tr>
<td>TP int</td>
<td>1</td>
<td>1</td>
<td>0.962</td>
<td>1</td>
<td>0.923</td>
<td></td>
</tr>
<tr>
<td>TN int</td>
<td>1</td>
<td>0.993</td>
<td>0.946</td>
<td>0.996</td>
<td>0.997</td>
<td></td>
</tr>
<tr>
<td>TP nl</td>
<td>0.596</td>
<td>0.981</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TN nl</td>
<td>0.988</td>
<td>0.656</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>model (b)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test MSE</td>
<td>1</td>
<td>1.882</td>
<td>2.872</td>
<td>8.762</td>
<td>2.081</td>
<td>1.598</td>
</tr>
<tr>
<td>MSE beta</td>
<td>1</td>
<td>4.755</td>
<td>64.751</td>
<td>48.033</td>
<td>12.071</td>
<td></td>
</tr>
<tr>
<td>FR</td>
<td>1</td>
<td>9.526</td>
<td>16.658</td>
<td>4.196</td>
<td>2.121</td>
<td></td>
</tr>
<tr>
<td>TP main</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.974</td>
<td>0.954</td>
<td></td>
</tr>
<tr>
<td>TN main</td>
<td>0.999</td>
<td>0.878</td>
<td>0.847</td>
<td>0.680</td>
<td>0.990</td>
<td></td>
</tr>
<tr>
<td>TP int</td>
<td>1</td>
<td>0.901</td>
<td>0.967</td>
<td>0.921</td>
<td>0.947</td>
<td></td>
</tr>
<tr>
<td>TN int</td>
<td>1</td>
<td>0.990</td>
<td>0.937</td>
<td>0.987</td>
<td>0.999</td>
<td></td>
</tr>
<tr>
<td>TN nl</td>
<td>0.992</td>
<td>0.696</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>model (c)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test MSE</td>
<td>1.301</td>
<td>1.050</td>
<td>1.000</td>
<td>2.244</td>
<td>1</td>
<td>2.442</td>
</tr>
<tr>
<td>FR</td>
<td>1</td>
<td>10.400</td>
<td>3.047</td>
<td>13.246</td>
<td>4.715</td>
<td></td>
</tr>
<tr>
<td>TN main</td>
<td>0.881</td>
<td>0.839</td>
<td>0.834</td>
<td>0.890</td>
<td>0.920</td>
<td></td>
</tr>
<tr>
<td>TN int</td>
<td>1</td>
<td>0.991</td>
<td>0.970</td>
<td>0.993</td>
<td>0.995</td>
<td></td>
</tr>
<tr>
<td>TP nl</td>
<td>0.500</td>
<td>0.898</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TN nl</td>
<td>0.998</td>
<td>0.772</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.2: Results from the simulation study under the three scenarios with $p = 50$, $n = 250$. We computed test error, FR for interaction effects, percentage of true positives and true negatives for main effects and interactions for MixSelect, BKMR, hierNet, Family, PIE and RAMP. We divided each value of test error and FR by the best (lowest) result for that metric. This makes the metric of the best model equal to 1.

<table>
<thead>
<tr>
<th></th>
<th>MixSelect</th>
<th>BKMR</th>
<th>hierNet</th>
<th>Family</th>
<th>PIE</th>
<th>RAMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>model (a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test MSE</td>
<td>1.011</td>
<td>4.057</td>
<td>1.361</td>
<td>3.210</td>
<td>1.153</td>
<td>1</td>
</tr>
<tr>
<td>FR</td>
<td>1.099</td>
<td>2.284</td>
<td>5.113</td>
<td>2.879</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>TP main</td>
<td>0.973</td>
<td>1</td>
<td>0.973</td>
<td>0.953</td>
<td>0.993</td>
<td></td>
</tr>
<tr>
<td>TN main</td>
<td>0.961</td>
<td>0.927</td>
<td>0.946</td>
<td>0.794</td>
<td>0.964</td>
<td></td>
</tr>
<tr>
<td>TP int</td>
<td>0.950</td>
<td>1</td>
<td>0.970</td>
<td>1</td>
<td>0.980</td>
<td></td>
</tr>
<tr>
<td>TN int</td>
<td>1.000</td>
<td>0.998</td>
<td>0.991</td>
<td>0.999</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>TP nl</td>
<td>0.540</td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TN nl</td>
<td>0.995</td>
<td>0.012</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>model (b)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FR</td>
<td>1</td>
<td>9.055</td>
<td>16.709</td>
<td>4.637</td>
<td>1.466</td>
<td></td>
</tr>
<tr>
<td>TP main</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>0.970</td>
<td>0.985</td>
</tr>
<tr>
<td>TN main</td>
<td>0.998</td>
<td>0.903</td>
<td>0.897</td>
<td>0.810</td>
<td>0.997</td>
<td></td>
</tr>
<tr>
<td>TP int</td>
<td>1</td>
<td>0.930</td>
<td>0.970</td>
<td>0.905</td>
<td>0.970</td>
<td></td>
</tr>
<tr>
<td>TN int</td>
<td>1</td>
<td>0.997</td>
<td>0.974</td>
<td>0.996</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>TN nl</td>
<td>0.991</td>
<td>0.022</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>model (c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test MSE</td>
<td>1.468</td>
<td>3.901</td>
<td>1</td>
<td>2.322</td>
<td>1.007</td>
<td>2.427</td>
</tr>
<tr>
<td>FR</td>
<td>1</td>
<td>9.183</td>
<td>2.617</td>
<td>11.659</td>
<td>3.827</td>
<td></td>
</tr>
<tr>
<td>TP main</td>
<td>0.939</td>
<td>0.902</td>
<td>0.898</td>
<td>0.936</td>
<td>0.966</td>
<td></td>
</tr>
<tr>
<td>TN main</td>
<td>1.000</td>
<td>0.997</td>
<td>0.987</td>
<td>0.998</td>
<td>0.999</td>
<td></td>
</tr>
<tr>
<td>TP nl</td>
<td>0.500</td>
<td>0.980</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TN nl</td>
<td>0.999</td>
<td>0.020</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2.3: Results from the simulation study under the three scenarios with $p = 25$, $n = 500$. We computed test error, FR for interaction effects, percentage of true positives and true negatives for main effects and interactions for MixSelect, BKMR, hierNet, Family, PIE and RAMP. We divided each value of test error and FR by the best (lowest) result for that metric. This makes the metric of the best model equal to 1.

<table>
<thead>
<tr>
<th></th>
<th>MixSelect</th>
<th>BKMR</th>
<th>hierNet</th>
<th>Family</th>
<th>PIE</th>
<th>RAMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>test MSE</td>
<td>1.138</td>
<td>1</td>
<td>1.098</td>
<td>5.645</td>
<td>4.400</td>
<td>1.217</td>
</tr>
<tr>
<td>FR</td>
<td>1.033</td>
<td>5.659</td>
<td>5.820</td>
<td>2.465</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TP main</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TN main</td>
<td>0.758</td>
<td>0.798</td>
<td>0.947</td>
<td>0.679</td>
<td>0.919</td>
<td>0.997</td>
</tr>
<tr>
<td>TP int</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TN int</td>
<td>1.000</td>
<td>0.989</td>
<td>0.984</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td>TP nl</td>
<td>0.947</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TN nl</td>
<td>0.977</td>
<td>0.821</td>
<td>0.977</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>MixSelect</th>
<th>BKMR</th>
<th>hierNet</th>
<th>Family</th>
<th>PIE</th>
<th>RAMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>test MSE</td>
<td>1</td>
<td>1.902</td>
<td>1.430</td>
<td>8.928</td>
<td>1.363</td>
<td>1.061</td>
</tr>
<tr>
<td>FR</td>
<td>1</td>
<td>18.162</td>
<td>22.572</td>
<td>1.723</td>
<td>1.433</td>
<td>1</td>
</tr>
<tr>
<td>TP main</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TN main</td>
<td>0.998</td>
<td>0.863</td>
<td>0.907</td>
<td>0.688</td>
<td>0.992</td>
<td>0.992</td>
</tr>
<tr>
<td>TP int</td>
<td>1</td>
<td>1</td>
<td>0.978</td>
<td>1</td>
<td>1.089</td>
<td>0.999</td>
</tr>
<tr>
<td>TN int</td>
<td>1</td>
<td>0.988</td>
<td>0.958</td>
<td>0.993</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>TP nl</td>
<td>1</td>
<td>0.984</td>
<td>0.673</td>
<td>0.982</td>
<td>0.997</td>
<td>0.997</td>
</tr>
<tr>
<td>TN nl</td>
<td>0.808</td>
<td>0.719</td>
<td>0.868</td>
<td>0.834</td>
<td>0.851</td>
<td>0.851</td>
</tr>
<tr>
<td>TP nl</td>
<td>1.000</td>
<td>0.984</td>
<td>0.980</td>
<td>0.992</td>
<td>0.992</td>
<td>0.991</td>
</tr>
<tr>
<td>TN nl</td>
<td>0.645</td>
<td>0.985</td>
<td>0.989</td>
<td>0.893</td>
<td>0.893</td>
<td>0.893</td>
</tr>
</tbody>
</table>
Table 2.4: Results from the simulation study under the three scenarios with $p = 50$, $n = 500$. We computed test error, FR for interaction effects, percentage of true positives and true negatives for main effects and interactions for MixSelect, BKMR, hierNet, Family, PIE and RAMP. We divided each value of test error and FR by the best (lowest) result for that metric. This makes the metric of the best model equal to 1.

<table>
<thead>
<tr>
<th></th>
<th>MixSelect</th>
<th>BKMR</th>
<th>hierNet</th>
<th>Family</th>
<th>PIE</th>
<th>RAMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>test MSE</td>
<td>1.135</td>
<td>11.409</td>
<td>1</td>
<td>5.630</td>
<td>4.057</td>
<td>1.181</td>
</tr>
<tr>
<td>FR</td>
<td>1.808</td>
<td>8.718</td>
<td>9.642</td>
<td>3.949</td>
<td>3.849</td>
<td>1</td>
</tr>
<tr>
<td>TP main</td>
<td>1</td>
<td>1</td>
<td>0.993</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TN main</td>
<td>0.863</td>
<td>0.868</td>
<td>0.976</td>
<td>0.789</td>
<td>0.967</td>
<td>1</td>
</tr>
<tr>
<td>TP int</td>
<td>1</td>
<td>1</td>
<td>0.989</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TN int</td>
<td>1</td>
<td>0.996</td>
<td>0.996</td>
<td>0.999</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TP nl</td>
<td>0.826</td>
<td>1</td>
<td>0.999</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TN nl</td>
<td>0.999</td>
<td>0.037</td>
<td>0.046</td>
<td>0.094</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>model (b)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test MSE</td>
<td>1.000</td>
<td>12.987</td>
<td>1.420</td>
<td>9.485</td>
<td>1.364</td>
<td>1</td>
</tr>
<tr>
<td>FR</td>
<td>1.222</td>
<td>20.973</td>
<td>25.820</td>
<td>1.849</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TP main</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TN main</td>
<td>0.999</td>
<td>0.880</td>
<td>0.977</td>
<td>0.822</td>
<td>0.999</td>
<td>1</td>
</tr>
<tr>
<td>TP int</td>
<td>1</td>
<td>1</td>
<td>0.990</td>
<td>0.995</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TN int</td>
<td>1</td>
<td>0.996</td>
<td>0.993</td>
<td>0.999</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TN nl</td>
<td>1</td>
<td>0.046</td>
<td>0.046</td>
<td>0.094</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>model (c)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test MSE</td>
<td>1.360</td>
<td>4.139</td>
<td>1</td>
<td>2.589</td>
<td>1.070</td>
<td>2.519</td>
</tr>
<tr>
<td>FR</td>
<td>1</td>
<td>7.990</td>
<td>2.078</td>
<td>8.885</td>
<td>3.562</td>
<td>1</td>
</tr>
<tr>
<td>TN main</td>
<td>0.894</td>
<td>0.815</td>
<td>0.950</td>
<td>0.901</td>
<td>0.942</td>
<td>1</td>
</tr>
<tr>
<td>TP int</td>
<td>1.000</td>
<td>0.994</td>
<td>0.997</td>
<td>0.998</td>
<td>0.999</td>
<td>1</td>
</tr>
<tr>
<td>TN int</td>
<td>0.994</td>
<td>0.978</td>
<td>0.998</td>
<td>0.999</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>TP nl</td>
<td>0.523</td>
<td>0.983</td>
<td>0.999</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>TN nl</td>
<td>0.984</td>
<td>0.043</td>
<td>0.046</td>
<td>0.094</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

2.5 Environmental Epidemiology Application

2.5.1 Motivation

The goal of our analysis is to assess the association of fourteen metals (Barium, Cadmium, Cobalt, Caesium, Molybdenum, Manganese, Mercury, Lead, Antimony, Tin, Strontium, Thallium, Tungsten and Uranium) with body mass index (BMI). Recently, several studies showed the relation between complex mixtures of metals and health or behavioral outcomes. See Sanders et al. (2015) for example for a literature
review on perinatal and childhood exposures to Cadmium (Cd), Manganese (Mn) and metal mixtures. The authors state that there is suggestive evidence that Cadmium is associated with poorer cognition. Claus Henn et al. (2014) report associations between mixtures and pediatric health outcomes, cognition, reproductive hormone levels and neurodevelopment. With respect to obesity indices, metals have already been associated with an increase in waist circumference and BMI, see Padilla et al. (2010) and Shao et al. (2017), using data from the National Health and Nutrition Examination Survey (NHANES).

2.5.2 Data Description

We consider data from NHANES collected in 2015. We select a subsample of 2532 individuals for which at least one measurement of metals and BMI have been recorded. We also include in the analysis cholesterol, creatinine, sex, age and ethnicity, which has five categories (Hispanic, other Hispanic, non-Hispanic White, non-Hispanic Black and other Ethnicity). We choose Hispanic as a reference group for ethnicity. Table 2.5 shows the correlations among chemicals, Figure 2.4 and Figure 2.5 show the missingness pattern and the cases below the limit of detection (LOD). In NHANES, different groups of chemicals, such as metals or Phthalates, are only measured for a subsample of individuals. This subsampling only depends on demographic characteristics of the individuals and hence the missing at random assumption should be appropriate in our context.

We apply the base 10 logarithm transformation to the chemical exposure values, cholesterol and creatinine. We also apply the $\log_{10}$ transformation to BMI in order to make its distribution closer to normality, which is the assumed marginal distribution in our model. The log-transformation is commonly applied in environmental epidemiology in order to reduce the influence of outliers and has been employed in several studies using NHANES data (Nagelkerke et al., 2006), (Lynch et al., 2010),
Table 2.5: Correlation matrix between Barium (Ba), Cadmium (Cd), Cobalt (Co), Cesium (Cs), Molybdenum (Mo), Manganese (Mn), Lead (Pb), Antimony (Sb), Tin (Sn), Strontium (Sr), Thallium (TI), Tungsten (W), Uranium (U), Mercury (Hg) in the NHANES 2015 dataset.
(Buman et al., 2013). We leave these transformations implicit for the remainder of the section.

2.5.3 Missing data and LOD

In this subsection we describe how to explicitly model the covariates to allow imputation of observations that are missing or below the limit of detection. We are particularly motivated by studies of environmental health collecting data on mixtures of chemical exposures. These exposures can be moderately high-dimensional with high correlations within blocks of variables. For this reason we decide to endow the chemical measurements, cholesterol and creatinine with a latent factor model. Let $X$ be the $n \times p$ matrix containing the chemical measurements, $Z$ an $n \times q$ matrix containing the covariates and let $W_i = (X_i, z_{i1}, z_{i2})^T$ be a $d \times 1$ vector containing the 14 chemical measurements, cholesterol and creatinine. The factor model is as follows:

$$W_i = \Lambda \eta_i + \epsilon_i, \quad \epsilon_i \sim N_d(0, \Sigma),$$

$$\eta_i \sim N_k(0, I),$$

where we center the data $W_i$ to have zero mean prior to the analysis, $\Sigma = diag(\sigma_1^2, \ldots, \sigma_d^2)$ is as residual variance matrix, $\Lambda$ is a $d \times k$ factor loadings matrix, and $\eta_i$ are i.i.d. standard normal latent factors. We assume an element wise standard normal prior for $\Lambda$ and endow $\sigma_j^2$ with independent Inverse-Gamma priors with parameters $(1/2, 1/2)$, for $j = 1, \ldots, d$. From an eigendecomposition of the correlation matrix, the first 9 eigenvectors explain more than 85% of the total variability; hence we set the number of factors equal to 9. Algorithm 2 describes how to sample the parameters of (2.4) within an MCMC algorithm.

In addition to missingness due to chemicals that have not been assayed, 13.5% of chemicals have been recorded under the limit of detection (LOD). We can impute
**Algorithm 2** MCMC algorithm for imputing missing observations and those under the LOD while simultaneously sampling the parameters of model (2.4)

**Step 1** Sample $\eta_i, i = 1, \ldots, n$ from a multivariate normal distribution:

$$\pi(\omega | \eta) \sim N_k \left( (I_k + \Lambda^T \Sigma^{-1} \Lambda)^{-1} \Lambda^T \Sigma^{-1} W, (I_k + \Lambda^T \Sigma^{-1} \Lambda)^{-1} \right).$$

**Step 2** Denote $\lambda_j$ the rows of $\Lambda$, for $j = 1, \ldots, d$. Sample $d$ conditionally independent posteriors:

$$\pi(\lambda_j | \eta) \sim N \left( (I_k + \frac{\eta^T \eta}{\sigma_j^2})^{-1} \eta^T \sigma_j^{-2} W^{(j)}, (I_k + \frac{\eta^T \eta}{\sigma_j^2})^{-1} \right),$$

where $W^{(j)}$ is the $j^{th}$ column of the matrix $W$ and $\eta$ is the matrix with rows equal to $\eta_i$.

**Step 3** Sample $\sigma_j^{-2}$ for $j = 1, \ldots, d$ from conditionally independent gamma distributions

$$\pi(\sigma_j^{-2} | \eta) \sim Gamma \left( \frac{1 + n}{2}, \frac{1}{2} + \frac{1}{2} \sum_{i=1}^{n} (W_{ij} - \lambda_j^T \eta_i) \right).$$

**Step 4** Sample missing observations from conditionally independent distributions; if $W_{ij}$ is missing sample its value from

$$N(\eta_i^T \lambda_j, \sigma_j^2).$$

**Step 5** Sample observations below the LOD from conditionally independent truncated normal distributions:

$$X_{ij} | X_{ij} \in [-\infty, \log_{10}(\text{LOD}_j)] \sim TN(\eta_i^T \lambda_j, \sigma_j^2, -\infty, \log_{10}(\text{LOD}_j)),$$

where LOD$_j$ is the limit of detection for exposure $j$ and $TN(\mu, \sigma^2, a, b)$ is a truncated normal distribution with mean $\mu$, variance $\sigma^2$ and support in $[a, b]$. 

26
these observations as:

\[ X_{ij} \mid X_{ij} \in [-\infty, \log_{10}(\text{LOD}_j)] \sim TN(\eta_i^T \lambda_j, \sigma^2_j, -\infty, \log_{10}(\text{LOD}_j)), \]

where LOD\(_j\) is the limit of detection for exposure \(j\) and \(TN(\mu, \sigma^2, a, b)\) is a truncated normal distribution with mean \(\mu\), variance \(\sigma^2\) and support in \([a, b]\). A related approach was used in Ferrari and Dunson (2020) to impute chemicals below the LOD within an MCMC algorithm.

To simplify data imputation under the above model and improve robustness to model misspecification, we apply a common “cut of feedback” approach (Lunn et al., 2009). In particular, in imputing the missing values and those below the limit of detection, we use the conditional posterior given only the data in the \(W_i\) component of the model and not taking into account that \(W_i\) also appears in the outcome model.

2.5.4 Statistical Analysis

We estimate a quadratic regression with nonlinear effects for the transformed chemicals, which are included in the matrix \(X\), and we control for covariates, which are included in the matrix \(Z\), according to model (2.1). We use the specified priors in Section 2.2.2 and alternate between the steps of Algorithm 1 and Algorithm 2 at each MCMC iteration to obtain the posterior samples. In environmental epidemiology, the signal to noise ratio is usually low; hence we use the weak heredity specification in order to have greater flexibility in our model and to enhance power in discovery of linear interactions. We run the MCMC chain for a total of 5000 iterations, with a burn-in of 4000.

We observed good mixing for main effect and interaction coefficients. In particular, the average Effective Sample Size (ESS) for main effects and interactions was equal to 725. For the smoothness parameters, the effective sample size for each \(\rho_j\) was on average 3 times higher with respect to the corresponding parameters in BKMR.
Table 2.6: Coverage computed at different levels $\alpha$ for the *in sample* and *out of sample* predictive intervals of BMI.

<table>
<thead>
<tr>
<th>level $\alpha$</th>
<th>in sample</th>
<th>out of sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>0.025</td>
<td>0.979</td>
<td>0.968</td>
</tr>
<tr>
<td>0.05</td>
<td>0.96</td>
<td>0.942</td>
</tr>
<tr>
<td>0.1</td>
<td>0.91</td>
<td>0.88</td>
</tr>
</tbody>
</table>

We also computed the Geweke diagnostic for main and interaction effects, for a total of 105 parameters. The Geweke diagnostic tests for a difference of the mean in the first 25% of the MCMC samples and the last 25% of the samples. All computed p-values were not significant at the 0.01 level. Residual plots are included in Figure 2.6. The residual diagnostics suggest that the model assumptions are satisfied fairly well. Firstly, approximate normality holds, with only a mild deviation in the tails. Secondly, inspecting the scatter plot of predicted BMI vs standardized residual, we did not find any clear patterns, suggesting homoskedasticity and adequate fit of our regression model. Lastly, we conducted posterior predictive checks, comparing the mean of the in-sample predictions at each MCMC iteration to the data mean. Figure 2.7 shows that the two means align very well. We also observed good in sample and out of sample coverage of $100(1-\alpha)\%$ predictive intervals for different $\alpha$ values; refer to Table 2.6.

The complexity per iteration of Gibbs sampling is $O(n^2m)$ when $\tau \neq 0$, where $m$ is related to the approximation described in Section 2.3. When $\tau = 0$, the complexity per iteration of Gibbs sampling is $O(d^2)$, where $d$ is the number of active main effects.
In our analysis, we found significant nonlinear associations with BMI for Cadmium and Tungsten with posterior predictive probabilities of having an active nonlinear effect of 1 and 0.79, respectively. Figure 2.8 shows the estimated nonlinear surfaces for Cadmium and Tungsten, when all the other variables are set to their median. The nonlinear effect of Cadmium has a hill-shaped dose response, with a monotone increase at lower doses followed by a downturn leading to a reverse in the direction of association; presumably as toxic effects at high doses lead to weigh loss. We also found a significant negative linear association between BMI and Lead and Molybdenum, and the main effect estimates suggested a negative linear association with Cesium, Cobalt and Tin. A similar negative effect for higher doses of Cadmium, Cobalt and Lead was found in Shao et al. (2017) and Padilla et al. (2010), where both authors found an inverse linear association between these metals and BMI, suggesting that they can create a disturbance of metabolic processes.

We found positive linear interactions between Molybdenum × Strontium, Lead × Antimony, and negative interaction between Lead × Uranium. Figure 2.9 shows the estimated coefficients for interactions. With respect to covariate adjustments, we found a positive association between BMI and Age, Creatinine and Cholesterol, as expected, and also a negative association with ethnicities Other Hispanic, non-Hispanic White, non-Hispanic Black and Other Ethnicity with respect to the reference group Hispanic, refer to Figure 2.10. Finally, even if some of the chemicals were moderately correlated, see Molybdenum and Tungsten for example in Table 2.5, our model was able to distinguish the two effects, estimating a linear association for Molybdenum and no association for Tungsten.

We compared the performance of our model with the methods described in Section 2.4: BKMR (Bobb et al., 2014), Family (Haris et al., 2016), hierNet (Bien et al.,
Table 2.7: Performance of MixSelect, BKMR, RAMP, hierNet, Family and PIE for in sample mean squared error when training on the complete cases and out of sample mean squared error when holding out 500 data points.

<table>
<thead>
<tr>
<th></th>
<th>MixSelect</th>
<th>BKMR</th>
<th>hierNet</th>
<th>Family</th>
<th>PIE</th>
<th>RAMP</th>
</tr>
</thead>
<tbody>
<tr>
<td>in sample MSE</td>
<td>0.530</td>
<td>0.031</td>
<td>0.573</td>
<td>0.879</td>
<td>0.626</td>
<td>0.572</td>
</tr>
<tr>
<td>out of sample MSE</td>
<td>0.687</td>
<td>0.919</td>
<td>0.611</td>
<td>0.927</td>
<td>0.710</td>
<td>0.604</td>
</tr>
</tbody>
</table>

2013), PIE (Wang et al., 2019) and RAMP (Hao et al., 2018). For simplicity in making comparisons across methods that mostly lack an approach to accommodate missing exposures, we focus on complete case analyses, discarding all observations having any values that are missing. Table 2.7 shows the performance of the models for in sample MSE when training on the full dataset and out of sample MSE when holding out 500 data points. Notice that BKMR overfits the training data in the presence of highly correlated covariates and consequently has worse performance on the test set. In addition, BKMR estimates a posterior probability of a nonlinear effect greater than 0.87 for each chemical, which could be a result of overfitting. On the other hand, MixSelect is able to distinguish a simple regression surface from a more complex one thanks to the identifiability constraint, which prevents overfitting.

Figure 2.11 shows the estimated main effects of the chemicals, and 95% credible intervals for MixSelect. Notice that most of the main effect estimates of the other models are equal to 0, perhaps due to low power. The method PIE also estimates a negative association for Lead and Molybdenum; RAMP and hierNet estimate a negative association for Lead. Finally, there is suggestive evidence of a negative association between BMI with Cesium, Tin and Cobalt, which is also detected by PIE.
2.5.6 Application with Sex and Ethnicity Interaction

In this section, we assess whether the association between the metals and BMI changes with sex or non-Hispanic Black ethnicity. In epidemiology, it is common to conduct separate analyses for Blacks and non-Blacks as these groups can be very different with respect to certain exposures and outcomes. In NHANES studies, (Shim et al., 2017) show suggestive evidence of age and sex interactions as well as interactions between age and ethnicity for Lead, Cadmium, Mercury and Arsenic.

We run the analysis on the dataset with 2029 complete cases: 49% of observations are Male and 19% are non-Hispanic Black. We preprocess the data following Section 2.5.2. We estimate a quadratic regression with nonlinear effects for the transformed chemicals interacted separately with Sex and non-Hispanic Black ethnicity, which are included in the matrix $X$, and we control for covariates, which are included in the matrix $Z$, according to model 2.1. We estimate the model using the strong heredity specification and we compare the estimates of MixSelect with BKMR (Bobb et al., 2014), Family (Haris et al., 2016), hierNet (Bien et al., 2013), PIE (Wang et al., 2019) and RAMP (Hao et al., 2018).

Figure 2.12 shows the estimated nonlinear curves for Cadmium interacted with Sex and non-Hispanic Black ethnicity, when all the other variables are set to their median. The non linear effect of Cadmium in Females and non-Hispanic Blacks has a hill-shaped dose response as in Figure 2.8, whereas it is negatively associated with BMI in the Male subgroup. Table 2.8 contains the posterior inclusion probabilities of nonlinear effects. Figure 2.13 shows the estimated main effects of the chemicals, and 95% credible intervals for MixSelect. Notice that Lead and Molybdenum exposures have a stronger negative effect on Females than Males, and we observe the opposite behavior for Tin and Cobalt. These associations are also estimated by PIE, and they are partially supported by RAMP and hierNet. We found positive linear interactions
between Cadmium×Cobalt and Cobalt×Tin for Males.

**Figure 2.13:** Estimated main effects using MixSelect with 95% credible intervals and estimated coefficients using RAMP, hierNet, Family and PIE. We trained all the other models on the dataset with complete cases and included interactions of chemical measurements with non-Hispanic Black ethnicity and Sex.

<table>
<thead>
<tr>
<th>Chemical</th>
<th>Antimony</th>
<th>Barium</th>
<th>Cadmium</th>
<th>Cesium</th>
<th>Cobalt</th>
<th>Lead</th>
<th>Manganese</th>
<th>Mercury</th>
<th>Molybdenum</th>
<th>Strontium</th>
<th>Thallium</th>
<th>Tin</th>
<th>Tungsten</th>
<th>Uranium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>0.594</td>
<td>0.150</td>
<td>0.338</td>
<td>0.524</td>
<td>0.658</td>
<td>0.595</td>
<td>0.211</td>
<td>0.376</td>
<td>0.154</td>
<td>0.227</td>
<td>0.329</td>
<td>0.324</td>
<td>0.442</td>
<td>0.355</td>
</tr>
<tr>
<td>Female</td>
<td>0.989</td>
<td>0.093</td>
<td>1</td>
<td>0.138</td>
<td>0.833</td>
<td>0.171</td>
<td>0.435</td>
<td>0.239</td>
<td>1</td>
<td>0.135</td>
<td>0.380</td>
<td>0.448</td>
<td>0.444</td>
<td>0.122</td>
</tr>
<tr>
<td>Male</td>
<td>0.117</td>
<td>0.060</td>
<td>0.832</td>
<td>0.030</td>
<td>0.305</td>
<td>0.322</td>
<td>0.271</td>
<td>0.176</td>
<td>0.229</td>
<td>0.174</td>
<td>0.009</td>
<td>0.091</td>
<td>0.010</td>
<td>0.142</td>
</tr>
</tbody>
</table>

**Table 2.8:** Posterior inclusion probabilities of nonlinear effects when including interactions between chemical measurements with non-Hispanic Black ethnicity and Sex.
2.6 Discussion

We proposed a MixSelect framework that allows identification of main effects and interactions. We also allow flexible nonlinear deviations from the parametric specification relying on a Gaussian process prior. We showed that MixSelect improves on the state-of-the-art for assessing associations between chemical exposures and health outcomes. To our knowledge, this is the first flexible method that is designed to provide interpretable estimates for main effects and interactions of chemical exposures while not constraining the model to have a simple parametric form. We also included variable selection, uncertainty quantification, missingness in the predictors, and limit of detection. The proposed specification provides a nice building block for more complicated data structures; for example, there are straightforward extensions to allow censored outcomes, longitudinal data, spatial dependence, and other issues.

NHANES data are obtained using a complex sampling design, that includes over-sampling of certain population subgroups, and contains sampling weights for each observation that are inversely proportional to the probability of being sampled. We did not employ sampling weights in our analysis because our goal was to study the association between metals and BMI rather than providing population estimates. One possibility to include the sampling weights in our method is to jointly model the outcome and the survey weights (Si et al., 2015), without assuming that the population distribution of strata is known.

With correlated features, variable selection techniques can lead to multiple models having almost the same posterior probability of being the best one, and with few observations the interpretation of results becomes difficult. However, our method provided better inference under correlated predictors than BKMR (Bobb et al., 2014). We believe this is due to the projection approach, which protects against overfitting by adding a constraint to the highly flexible nonparametric surface. An alternative
solution is to cluster the predictors at each iteration of the MCMC algorithm using a nonparametric prior specification for the coefficients (MacLehose et al., 2007).

Instead of focusing on mean regression, we can easily modify MixSelect to accommodate quantile regression. In order to induce a regression on a specific quantile, one can use (2.1) but with the residual $\epsilon_i$ having an asymmetric Laplace distribution (Yu and Moyeed, 2001). The asymmetric Laplace can be represented as a scale mixture of Gaussians, facilitating a straightforward modification to our MCMC algorithm; refer to Chen et al. (2013a) for related work. Alternatively, it is possible to allow main effects and interactions to vary with quantiles of $y_i$, see for example Reich et al. (2011). We can also induce a quantile dependence on the nonlinear deviation $g^*(x_i)$. In particular, we can introduce uniformly distributed latent variables $\eta_i$ modifying the nonlinear deviation as $g^*(x_i, \eta_i)$, which is referred to as the Gaussian process transfer prior (Kundu and Dunson, 2014).

Chemical studies usually involve up to dozens of exposures, but recent developments employing novel data collection techniques are starting to produce interesting datasets in which the number of exposures is in the order of the number of data points, so that the estimation of statistical interactions becomes infeasible with standard techniques. In this paper we impose heredity constraints and an approximation to the Gaussian process surface in order to deal with this problem, but new developments for dimension reduction are needed to scale up to allow massive number of exposures.
Figure 2.4: Pattern of missingness in the chemical exposure, cholesterol and creatinine measurements.
Figure 2.5: Pattern of data below the limit of detection in the matrix $X$ including the chemical measurements.

Figure 2.6: Histogram and QQ-plot of residuals and scatter plot of predicted BMI values vs standardized residuals.
Figure 2.7: Histogram of the predictive mean at each MCMC iteration. The horizontal red line shows the mean of BMI and the blue line the mean of the predictions.
Figure 2.8: Estimated dose response curves for the chemicals Cadmium, Tungsten, Lead and Cobalt, when all the other quantities are equal to their median. The black line corresponds to the posterior median, the shaded bands indicate 95% posterior credible intervals, and the marks on the x-axis indicate the observed data points.
Figure 2.9: Estimated interaction effects using MixSelect with 95% credible intervals.
Figure 2.10: Estimated covariate effects using MixSelect with 95% credible intervals. Hispanic is the reference group for ethnicity.

Figure 2.11: Estimated main effects using MixSelect with 95% credible intervals and estimated coefficients using RAMP, hierNet, Family and PIE. We trained all the methods on the dataset with complete cases. Exposure measurements are on the log scale.
**Figure 2.12:** Estimated regression surface for Cadmium interacted with Sex and non-Hispanic Black ethnicity, when all the other quantities are equal to their median. The black line corresponds to the posterior median, the shaded bands indicate 95% posterior credible intervals, and the marks on the x-axis indicate the observed data points.
3

Bayesian Factor Analysis for Inference on Interactions

3.1 Introduction

There is broad interest in incorporating interactions in linear regression. Extensions of linear regression to accommodate pairwise interactions are commonly referred to as quadratic regression. In moderate to high-dimensional settings, it becomes very challenging to implement quadratic regression since the number of parameters to be estimated is $2p + \binom{p}{2}$. Hence, classical methods such as least squares cannot be used and even common penalization and Bayesian methods can encounter computational hurdles. Reliable inferences on main effects and interactions is even more challenging when certain predictors are moderately to highly correlated.

A lot of effort has been focused on estimating pairwise interactions in moderate high-dimensional and ultra high-dimensional problems. We refer to the former when the number of covariates is between 20 and 100 and to the latter when $p > 100$. When $p = 100$, the number of parameters to be estimated is greater than 5000. When $p \in [20, 100]$, one-stage regularization methods like Bien et al. (2013) and
Haris et al. (2016) can be successful. Some of these methods require a so-called heredity assumption (Chipman, 1996) to reduce dimensionality. Strong heredity means that the interaction between two variables is included in the model only if both main effects are. For weak heredity it suffices to have one main effect in the model. Heredity reduces the number of models from \(2^{p+(\binom{p}{2})}\) to \(\sum_{i=0}^{p} \binom{p}{i} 2^{\binom{i}{2}}\) or \(\sum_{i=0}^{p} \binom{p}{i} 2^{p-i(i+1)/2}\) for strong or weak heredity, respectively (Chipman, 1996). For ultra high-dimensional problems, two stage-approaches have been developed, see Hao et al. (2018) and Wang et al. (2019). However, these methods do not report uncertainties in model selection and parameter estimation, and rely on strong sparsity assumptions.

We are particularly motivated by studies of environmental health collecting data on mixtures of chemical exposures. These exposures can be moderately high-dimensional with high correlations within blocks of variables; for example, this can arise when an individual is exposed to a product having a mixture of chemicals and when chemical measurements consist of metabolites or breakdown products of a parent compound. There is a large public health interest in studying \(E \times E\), \(E \times G\) and \(G \times G\) interactions, with \(E = \) environmental exposures and \(G = \) genetic factors. However, current methods for quadratic regression are not ideal in these applications due to the level of correlation in the predictors, the fact that strong sparsity assumptions are not appropriate, and the need for uncertainty quantification. Regarding the issue of sparsity, some exposures are breakdown products of the same compound, so it is unlikely that only one exposure has an effect on the outcome. Also, it is statistically challenging to tell apart highly correlated covariates with limited data. For this reason, it is appealing given the data structure to select blocks of correlated exposures together instead of arbitrarily selecting one chemical in a group.

To address these problems, one possibility is to use a Bayesian approach to infer-
ence in order to include prior information to reduce dimensionality while characterizing uncertainty through the posterior distribution. There is an immense literature on Bayesian methods for high-dimensional linear regression, including recent algorithms that can scale up to thousands of predictors (Bondell and Reich, 2012), (Rossell and Telesca, 2017), (Johndrow et al., 2017), (Nishimura and Suchard, 2018). In addition some articles have explicitly focused on quadratic regression and interaction detection (Zhang and Liu, 2007), (Cordell, 2009), (Mackay, 2014). Bayes variable selection and shrinkage approaches will tend to have problems when predictors are highly correlated; this has motivated a literature on Bayesian latent factor regression (Lucas et al., 2006), (Carvalho et al., 2008).

Latent factor regression incorporates shared latent variables in the predictor and response components. This provides dimensionality reduction in modeling of the covariance structure in the predictors and characterizing the impact of correlated groups of predictors on the response. Such approaches are closely related to principal components regression, but it tends to be easier to simultaneously incorporate shrinkage and uncertainty quantification within the Bayesian framework. In addition, within the Bayes latent factor regression paradigm, typical identifiability constraints such as orthogonality are not needed (see, for example Bhattacharya and Dunson (2011)). The main contribution of this chapter is to generalize Bayesian latent factor regression to accommodate interactions using an approach inspired by Wang et al. (2019). This is accomplished by including pairwise interactions in the latent variables in the response component. We refer to the resulting framework as Factor analysis for INteractions (FIN). There is a rich literature on quadratic and nonlinear latent variable modeling, largely in psychometrics (refer, for example, to Arminger and Muthén (1998)). However, to our knowledge, such approaches have not been used for inferences on interactions in regression problems.

In Section 3.2 we describe the proposed FIN framework, including extensions for
higher order interactions. In Section 3.3 we provide theory on model misspecification and consistency. Section 3.4 contains a simulation study. Section 3.5 illustrates the methods on NHANES data. Code is available in R package infinitefactor.

3.2 Model

3.2.1 Model and Properties

Let $y_i$ denote a continuous health response for individual $i$, and $X_i = (x_{i1}, \cdots, x_{ip})^T$ denote a vector of exposure measurements. We propose a latent factor joint model, which includes shared factors in both the predictor and response components while assuming conditional independence. We include interactions among latent variables in the response component. We also assume that, given the latent variables, the explanatory variables and the response are continuous and normally distributed. We assume that the data have been normalized prior to the analysis so that we omit the intercept. The model is as follows:

\[
\begin{align*}
    y_i &= \eta_i^T \omega + \eta_i^T \Omega \eta_i + \epsilon_{y,i}, \quad \epsilon_{y,i} \sim N(0, \sigma_y^2), \\
    X_i &= \Lambda \eta_i + \epsilon_i, \quad \epsilon_i \sim N_p(0, \Psi), \\
    \eta_i &\sim N_k(0, I),
\end{align*}
\]

where $\Psi = diag(\sigma_1^2, \cdots, \sigma_p^2)$. In a Bayesian fashion, we assume a prior for the parameters $\Theta = (\omega, \Omega, \Lambda, \Psi, \sigma^2)$ that will be specified in 3.2.2. Model (1) is equivalent to classical latent factor regression models; refer, for example, to West (2003), except for the $\eta_i^T \Omega \eta_i$ term. Here, $\Omega$ is a $k \times k$ symmetric matrix inducing a quadratic latent variable regression that characterizes interactions among the latent variables.

The above formulation can be shown to induce a quadratic regression of $y$ on $X$. To build intuition consider the case in which $\sigma_j^2 = 0$ as done in West (2003) for the special case in which $\Omega = 0$. The many-to-one map $X_i = \Lambda \eta_i$ has multiple generalized inverses $\eta_i = \Lambda^T X_i + b$ such that $\Lambda b = 0$. If we substitute in the regression
equation, we obtain
\[ E(y_i|X_i) = (\Lambda^TX_i + b)^T\omega + (\Lambda^TX_i + b)^T\Omega(\Lambda^TX_i + b) = \]
\[ = X_i^T\Lambda\omega + X_i^T\Lambda\Omega^T\Lambda X_i + g(b) \]

The following proposition gives a similar result in the non deterministic case:

**Proposition 1.** Under model (1), the following are true:

(i) \[ E(y_i|X_i) = \text{tr}(\Omega V) + (\omega^T A)X_i + X_i^T(A^T\Omega A)X_i, \]

(ii) \[ \text{Cov}(y_i, X_i) = \Lambda\omega, \]

where \( V = (\Lambda^T\Psi^{-1}\Lambda + I)^{-1} \) and \( A = V\Lambda^T\Psi^{-1} = (\Lambda^T\Psi^{-1}\Lambda + I)^{-1}\Lambda^T\Psi^{-1}. \)

This shows that the induced regression of \( y \) on \( X \) from model (1) is indeed a quadratic regression. Let us define the induced main effects as \( \beta_X = A^T\omega \) and the matrix containing the first order interactions as \( \Omega_X = A^T\Omega A. \) Notice that we could define \( \Omega \) as a diagonal matrix and we would still estimate pairwise interactions between the regressors, further details are given in 3.2.4 and 3.2.4.

In epidemiology studies, it is of interest to include interactions between chemical exposures and demographic covariates. The covariates are often binary variables, like race or sex, or continuous variables that are non-normally distributed, like age. Hence, we do not want to assume a latent normal structure for the covariates. Letting \( Z_i = (z_{i1}, \cdots, z_{iq})^T \) be a vector of covariates, we modify model (1) to include a main effect for \( Z_i \) and an interaction term between \( Z_i \) and the latent factor \( \eta_i: \)

\[ y_i = \eta_i^T\omega + \eta_i^T\Omega\eta_i + Z_i^T\alpha + \eta_i^T\Delta Z_i + \epsilon_{y,i}, \quad \epsilon_{y,i} \sim N(0, \sigma^2), \]

\[ X_i = \Lambda\eta_i + \epsilon_i, \quad \epsilon_i \sim N_p(0, \Psi), \]

\[ \eta_i \sim N_k(0, I), \] (3.2)

where \( \Delta \) is a \( k \times q \) matrix of interaction coefficients between the latent variables and the covariates, and \( \alpha = (\alpha_1, \cdots, \alpha_q) \) are main effects for the covariates. Following Proposition 1 we have that

\[ \mathbb{E}(\eta_i^T\Delta Z_i|X_i, Z_i) = \mathbb{E}(\eta_i^T|X_i)\Delta Z_i = X_i^T(A^T\Delta)Z_i, \]
where \((A^T \Delta)\) is a \(p \times q\) matrix of pairwise interactions between exposures and covariates. In the sequel, we focus our development on model (1) for ease in exposition, but all of the details can be easily modified to pertain to model (2).

### 3.2.2 Priors and MCMC Algorithm

In this section we define the priors for \((\omega, \Omega, \Lambda, \Psi, \sigma^2)\), briefly describe the computational challenges given by model (1) and summarize our Markov Chain Monte Carlo sampler in Algorithm 3. We choose an Inverse-Gamma distribution with parameters \((\frac{1}{2}, \frac{1}{2})\) for \(\sigma^2\) and \(\sigma^2_j\) for \(j = 1, \cdots, p\). The elements of \(\omega\) and \(\Omega\) are given independent Gaussian priors. For \(\Lambda = \{\lambda_{i,j}\}\), a typical choice to attain identifiability requires \(\lambda_{i,j} = 0\) for \(j > i\) and \(\lambda_{j,j} > 0\) for \(j = 1, \cdots, k\) (Geweke and Zhou, 1996). However, some Bayesian applications, like covariance estimation (Bhattacharya and Dunson, 2011), do not require identifiability of \(\Lambda\). The same holds for inference on induced main effects and interactions for model (1). Notice that model (1) is invariant to rotations:

\[
y_{i} = \eta_{i}^T P P^T \omega + \eta_{i}^T P P^T \Omega P P^T \eta_{i} + \epsilon_{y,i}, \quad \epsilon_{y,i} \sim N(0, \sigma^2),
X_{i} = \Lambda P P^T \eta_{i} + \epsilon_{i}, \quad \epsilon_{i} \sim N_p(0, \Psi),
\]

where \(P\) is a \(k \times k\) orthogonal matrix \((PP^T = I)\). However, the induced main effects satisfy

\[
\beta_{X} = \Psi^{-1} \Lambda P (P^T \Lambda^T \Psi^{-1} \Lambda P + P^T P)^{-1} P^T \omega = \Psi^{-1} \Lambda (\Lambda^T \Psi^{-1} \Lambda + I)^{-1} \omega.
\]

The same holds for induced interactions, showing that we do not need to impose identifiability constraints on \(\Lambda\). We choose the Dirichlet-Laplace (DL) prior of Bhattacharya et al. (2015) row-wise, corresponding to

\[
\lambda_{j,k} | \phi_{jh}, \tau_j \sim DE(\phi_{jh} \tau_j) \quad h = 1, \cdots, k
\]

\[
\phi_j \sim Dir(a, \cdots, a) \quad \tau_j \sim Gamma(ka, 1/2),
\]

47
where \( j = 1, \ldots, p, \phi_j = (\phi_{j1}, \ldots, \phi_{jk}) \), DE refers to the zero mean double-exponential or Laplace distribution, and \( k \) is an upper bound on the number of factors, as the prior allows effective deletion of redundant factor loadings through row-wise shrinkage. The DL prior provides flexible shrinkage on the factor loadings matrix, generalizing the Bayesian Lasso (Park and Casella, 2008) to have a carefully chosen hierarchical structure on exposure-specific (\( \tau_j \)) and local (\( \phi_{jh} \)) scales. This induces a prior with concentration at zero, to strongly shrink small signals, and heavy-tails, to avoid over-shrinking large signals. The DL prior induces near sparsity row-wise in the matrix \( \Lambda \), as it is reasonable to assume that each variable loads on few factors.

In 3.2.4, we describe how the above prior specification induces an appealing shrinkage prior on the main effects and interactions, and discuss hyperparameter choice. In practice, we recommend the rule of thumb that chooses \( k \) such that
\[
\frac{\sum_{j=1}^{k} v_j}{\sum_{j=1}^{p} v_j} > 0.9,
\]
where \( v_j \) is the \( j \)th largest singular value of the correlation matrix of \( X \).

Proposition 2 in Section 3.3 provides theoretical justification for this criterion. As an alternative to row-wise shrinkage, we could have instead used column-wise shrinkage as advocated in Bhattacharya et al. (2015) and Legramanti et al. (2019). Although such approaches can be effective in choosing the number of factors, we found in our simulations that they can lead to over-shrinkage of the estimated main effects and interactions.

The inclusion of pairwise interactions among the factors in the regression of the outcome \( y_i \) rules out using a simple data augmentation Gibbs sampler, as in West (2003), Bhattacharya and Dunson (2011). The log full conditional distribution for \( \eta_i \) is:
\[
- \frac{1}{2} \left[ \eta_i^T \left( \frac{\omega_i^T}{\sigma_y^2} + \Lambda^T \Psi^{-1} \Lambda + I - 2 \frac{\Omega Y_i}{\sigma_y^2} \right) \eta_i - 2 \eta_i^T \left( \Lambda^T \Psi^{-1} X_i + \frac{\omega Y_i}{\sigma_y^2} \right) \right] - \\
- \frac{1}{2} \left[ 2 \frac{\eta_i^T \Omega \eta_i}{\sigma_y^2} + \frac{(\eta_i^T \Omega \eta_i)^2}{\sigma_y^2} \right] + C,
\]
where $C$ is a normalizing constant. We update the factors $\eta_i$ using the Metropolis-Adjusted Langevin Algorithm (MALA) (Grenander and Miller, 1994), (Roberts et al., 1996). Sampling the factors is the main computational bottleneck of our approach since we have to update $n$ vectors, each of dimension $k$. The overall MCMC algorithm and the MALA step are summarized in Algorithm 3.

### 3.2.3 Higher Order Interactions

FIN can be generalized to allow for higher order interactions. In particular, we can obtain estimates for the interaction coefficients up to the $Q^{th}$ order with the following model:

$$
E(y_i|\eta_i) = \sum_{h=1}^{k} \omega_h(1) \eta_{ih} + \sum_{h=1}^{k} \omega_h(2) \eta_{ih}^2 + \cdots + \sum_{h=1}^{k} \omega_h(Q) \eta_{ih}^Q,
$$

which is a polynomial regression in the latent variables. We do not include interactions between the factors, so that the number of parameters to be estimated is $Qk$. When $Q = 2$, this model is equivalent to $\Omega$ being a diagonal matrix. Recall that $\eta_i|X_i \sim N(AX,V)$, where $A$ and $V$ are defined in Proposition 1. Since we do not include interactions among the factors, let us just focus on the marginal distribution of the $j^{th}$ factor, i.e $\eta_{ih}|X_i \sim N(\mu_h,\sigma_h^2)$ where $\mu_h = \sum_{j=1}^{p} a_{hj} X_{ij}$ and $\sigma_h^2 = V_{hh}$. Below we provide an expression for $E(\sum_{q=1}^{Q} \omega_h^{(q)} \eta_{ih}^q|X)$, which can be calculated using non-central moments of a Normal distribution, see Winkelbauer (2012) for a reference.

$$
E(\sum_{q=1}^{Q} \omega_j^{(q)} \eta_j^q|X) = \sum_{f=0}^{[Q/2]} \sum_{q=f}^{[Q/2]} \omega_h^{(2q-1)} \sigma_h^{2q-2f} b_q^{q} \sum_{k_+=2f-1}^{p} \left( \frac{2f-1}{k_1 \cdots k_p} \right) \prod_{j=1}^{p} (a_{hj} X_{ij})^{k_{j+}} \left( \frac{2f}{k_1 \cdots k_p} \right) \prod_{j=1}^{p} (a_{hj} X_{ij})^{k_{j}},
$$

where $b_q^{q} = \frac{(2q-1)!}{(2f-1)(q-f)2^{q-f}}$, $b_q^{e} = \frac{(2q)!}{(2f)! (q-f)! 2^{q-f}}$ and $k_+ = \sum_{j=1}^{p} k_{h}$. We just need to sum up over the index $h$ in (3) and we can read out the expressions for the
Algorithm 3 MCMC algorithm for sampling the parameters of model (1)

Step 1 Sample $\eta_i$, $i=1,\ldots,n$ via Metropolis-Hastings using as a proposal distribution $a N(\eta_i + \frac{1}{2} \nabla \log(p(\eta|\cdot)),\epsilon I_k)$.

Step 2 Sample the main effects coefficients $\omega$ from a multivariate normal distribution:

$$\pi(\omega|\cdot) \sim N\left(\frac{\eta^T \eta}{\sigma^2} + I_n/100)^{-1}\eta(y - \text{diag}(\eta \Omega \eta^T)) \sigma^2, \left(\frac{\eta^T \eta}{\sigma^2} + I_n/100)^{-1}\right)$$

where $\eta$ is the matrix with rows equal to $\eta_i$.

Step 3 Sample upper triangular part of $\Omega$, namely $\Omega^U$, from a multivariate normal distribution:

$$\pi(\Omega^U|\cdot) \sim N\left(\frac{\eta^* \eta^*}{\sigma^2} + \frac{p(p+1)}{2} (y - \eta \omega - \text{diag}(\eta \Omega \eta^T))^T(y - \eta \omega - \text{diag}(\eta \Omega \eta^T)) \sigma^2, \left(\frac{\eta^* \eta^*}{\sigma^2} + I_n/100)^{-1}\right)$$

where $\eta^*$ is a matrix containing the pairwise interactions of among the columns of $\eta$. Then set $\Omega = \Omega + \Omega^U$.

Step 4 Sample $\sigma^{-2}$ from a Gamma distribution:

$$\pi(\sigma^{-2}|\cdot) \sim \text{Gamma}\left(1 + \frac{n}{2}, \frac{1}{2} \left(\frac{1}{2} y - \eta \omega - \text{diag}(\eta \Omega \eta^T)\right)^T \left(\frac{1}{2} y - \eta \omega - \text{diag}(\eta \Omega \eta^T)\right)\right)$$

Step 5 Denote $\lambda_j$ the rows of $\Lambda$, for $j = 1,\ldots,p$. Sample $p$ conditionally independent posteriors:

$$\pi(\lambda_j|\cdot) \sim N\left(D_j^{-1} + \frac{\eta^T \eta}{\sigma_j^2})^{-1}\eta^T \sigma_j^{-2} X^{(j)}, \left(D_j^{-1} + \frac{\eta^T \eta}{\sigma_j^2}\right)^{-1}\right)$$

where $X^{(j)}$ is the $j^{th}$ column of the matrix $X$, $D_j = \text{diag}(\tau_j^2 \psi_j \phi_{j1},\cdots,\tau_j^2 \psi_j \phi_{jk})$.

Step 6 Sample $\psi_{jh}$ for $j = 1,\ldots,p$ and $h = 1\cdots,k$ from independent Inverse Gaussian distributions:

$$\pi(\psi_{jh}|\cdot) \sim \text{InvGauss}\left(\tau_j \phi_{jh},1\right)$$

Step 7 Sample $\tau_j$ for $j = 1,\ldots,p$ from independent Generalized Inverse Gaussian distributions:

$$\pi(\tau_j|\cdot) \sim G\text{InvGauss}(1-k,1,2 \sum_{h=1}^{k} \frac{\lambda_j}{\phi_{jh}})$$

Step 8 In order to update $\phi_{jh}$ for $j = 1,\ldots,p$ and $h = 1\cdots,k$, sample $T_{jh}$ from independent Generalized Inverse Gaussian distributions:

$$\pi(T_{jh}|\cdot) \sim G\text{InvGauss}(a-1,1,2|\lambda_{jh})$$

Then set $\phi_{jh} = \frac{T_{jh}}{\sum_{h=1}^{k} T_{jh}}$.

Step 9 Sample $\sigma_j^{-2}$ for $j = 1,\ldots,p$ from conditionally independent gamma distributions:

$$\pi(\sigma_j^{-2}|\cdot) \sim \text{Gamma}\left(1 + \frac{n}{2}, \frac{1}{2} \left(\frac{1}{2} X_{ij} - \lambda_j^2 \eta_i\right)\right)$$
intercept, main effects and interactions up to the $Q^{th}$ order. In particular, we have
that the intercept is equal to $\sum_{h=1}^{k} \sum_{q=1}^{\frac{Q-1}{2}} \omega_h^{(2q+1)} V_{hh} b_h^{(q)}$. When $Q = 2$ this reduces
to $\sum_{h=1}^{k} \omega_h^{(2)} V_{hh} = tr(\Omega V)$, where $\Omega = diag(\omega_1^{(2)}, \ldots, \omega_k^{(2)})$. The expression for the
main effects coefficients on $X_j$ is $\sum_{h=1}^{k} \sum_{q=1}^{\frac{Q-1}{2}} \omega_h^{(2q+1)} V_{hh} b_h^{(q)} a_{hj}$. When $Q = 2$ this
becomes $\sum_{h=1}^{k} \omega_h^{(1)} a_{hj}$, hence $\beta_X = A^T \omega$. Similarly the expression for the interaction
between $X_j$ and $X_l$ is equal to $\sum_{h=1}^{k} \sum_{q=1}^{\frac{Q-1}{2}} 2 \omega_h^{(2q)} V_{hh} b_h^{(q)} a_{hj} a_{hl}$ and when $Q = 2$ we
have $\sum_{h=1}^{k} 2 \omega_h^{(2)} a_{hj} a_{hl}$ which is equal to $2[A^T \Omega A]_{(j,l)}$.

In general, if we are interested in the $q^{th}$ order interactions, we can find the
expression on the top summation for $f = \frac{q+1}{2}$ when $q$ is odd and on the bottom
summation for $f = \frac{q}{2}$ when $q$ is even. Finally notice that with $Qk$ parameters
we manage to estimate $\sum_{q=0}^{Q} \binom{p}{q}$ parameters thanks to the low dimensional factor
structure in the covariates.

### 3.2.4 Induced Priors

In this section, we show the behavior of the induced priors on the main effects and
pairwise interaction coefficients under model (1) using simulated examples, and we
show the induced grouping of coefficients when we have prior information on the
covariance structure of $X$. We endow $\omega$ with a normal prior having zero mean and
covariance equal to $\Xi$, where $\Xi$ is a diagonal matrix. Then, conditional on $\Lambda$ and
$\Psi$, the induced prior on $\beta_X$ is also Normal with mean 0 and covariance equal to
$A^T A$. Recall from Proposition 1 that the induced main effect coefficients are equal
to $\beta_X^T = \omega^T (A^T \Psi^{-1} A)^{-1} A^T \Psi^{-1}$. This expression is equivalent to West (2003) and we
can similarly characterize the limiting case of $\Psi \rightarrow 0$, i.e. when the factors explain
all of the variability in the matrix of regressors $X_j$. Let $\Psi = sI$ and $s \rightarrow 0$, together
with enforcing $\Lambda$ to be orthogonal, we have that $\beta_X = \Lambda \omega$. It follows that $\beta_X$ has the generalised singular $g$-prior (or $gsg$-prior) distribution defined by West (2003),
whose density is proportional to $\exp(-\frac{1}{2} \beta_X^T \Lambda^T \Xi^{-1} \Lambda \beta)$.

Now, consider the extension presented in the previous section, where we include powers of the factors in the regression of $y_i$. In Figure 3.1, we show the induced marginal priors for main effects, pairwise interactions and 3$^{rd}$ order interactions when $p = 20$ and $k = 5, 10$ when $\omega$ and $\Omega$ are given $N(0, 1)$ priors element-wise. Increasing (or decreasing) the variance of the priors on $\omega$ and $\Omega$ will directly increase (or decrease) the variance of the induced main effects and pairwise interactions, as $\beta_X$ and $\Omega_X$ are linear functions of $\omega$ and $\Omega$ respectively. For a fixed $k$, there is increasing shrinkage towards zero with higher orders of interaction. However, we avoid assuming exact sparsity corresponding to zero values of the coefficients, a standard assumption of other methods. Although most of the mass is concentrated around zero, the distributions have heavy tails. We can indeed notice that the form of the priors resembles a mixture of two normal distributions with different variances, and that we place a higher mixture weight on the normal distribution concentrated around zero as we increase the order of interactions. This is because higher order interactions contain products of the elements of $A$, previously defined in Proposition 1, and the elements of $A$ are affected by the DL prior shrinkage, since $A$ is a function of $\Lambda$. Also, notice that the priors have higher variance as we increase the number of latent factors $k$.

In environmental epidemiology, it is common to have prior knowledge of groups of exposures that are highly correlated and it is natural to include such information in the specification of $\Lambda$. One possibility is to impose a block sparsity structure in which each group of chemicals is restricted to load on the same factor. Then, cross group dependence is allowed including additional factors and endowing the factor loadings with a DL prior. Suppose that the variables in $X$ can be divided in $l$ groups: $S_1, S_2, \ldots, S_l$ of dimensions $p_1, p_2, \ldots, p_l$, where $l < k$ and $p = \sum_{r=1}^l p_r$. Letting $\Lambda = [\Lambda^B \Lambda']$, where $\Lambda^B$ is $p \times l$, we can assign a block sparsity structure to
\[ \Lambda^B: \]

\[ \lambda^B_{p_1+1,1} = \ldots = \lambda^B_{p_1,1} = 0 \]
\[ \lambda^B_{1,2} = \ldots = \lambda^B_{p_1,2} = \lambda^B_{p_1+p_2+1,2} = \ldots = \lambda^B_{p,2} = 0 \]
\[ \ldots \]
\[ \lambda^B_{1,l} = \ldots = \lambda^B_{p-p_1,l} = 0 \]

We show the effect of the block sparsity structure on the a priori induced groupings of main effects and interactions when \( l = k \), so that \( \Lambda = \Lambda^B \). Let us consider the scenario when the true number of factors is equal to 2, and the variables in \( X \) can be divided in two groups. This structure is reasonable when we have measurements of different chemicals that are breakdown products of the same exposure, for example PCB metabolites (Longnecker et al., 2001). Then the a priori covariance between main effects is equal to \( \text{Cov}(\beta_g, \beta_h | g, h \in S_r) = \frac{\lambda_h \lambda_g}{(\sum_{j \in S_r} \lambda_j^2 + \sigma^2)^2} \), for \( r = 1, 2 \), when the chemicals belong to the same group, and is zero otherwise. Hence, in this case, \( A^T A \) is block diagonal. On the other hand, assume that the number of factors is equal to the number of covariates, with \( \Lambda \) being diagonal. In this case the induced covariance on \( \beta \) is diagonal with elements \( \frac{\lambda^2}{(\lambda_j^2 + \sigma^2)^2} \). In general when there are \( l \) groups, the variance of \( \beta_h \), with \( h \in S_r \), is equal to \( \frac{\lambda^2}{(\sum_{j \in S_r} \lambda_j^2 + \sigma^2)^2} \). Hence, the variance of \( \beta \) is lower with respect to the independent case since we are borrowing strength and information from the other covariates within the same group.

Let us now focus on the symmetric matrix \( \Omega \) of dimension \( k \), letting \( \nu(\Omega) \) be the vector of lower triangular elements of \( \Omega \). Define the duplication matrix \( D_k \) as the \( k^2 \times \frac{k(k+1)}{2} \) matrix such that \( D_k \nu(\Omega) = \text{vec}(\Omega) \), see Magnus (1988) as a reference. The duplication matrix can be easily calculated for orders 2 and 3, whereas the R package matrixcalc provides the duplication matrix for higher orders. We are interested in
the distribution of $\Omega_X = A^T \Omega A$. Notice that
\[
\text{vec}(\Omega_X) = \text{vec}(A^T \Omega A) = (A^T \otimes A^T)\text{vec}(\Omega) = (A^T \otimes A^T)D_k\nu(\Omega).
\]

We choose a normal prior on pairwise interactions, i.e., $\nu(\Omega) \sim N_{\frac{k(k+1)}{2}}(0, I_{\frac{k(k+1)}{2}})$. so that $\text{vec}(\Omega_X) \sim N_{\nu^2}(0, (A^T \otimes A^T)D_kD_k^T(A \otimes A))$. Computing the covariance $(A^T \otimes A^T)D_kD_k^T(A \otimes A)$ of the induced Normal prior on the matrix containing the pairwise interactions $\text{vec}(\Omega_X) = (\omega_{1,1}, \omega_{1,2}, \cdots, \omega_{4,3}, \omega_{4})$, we find that the variables are divided in three groups: $(\omega_{1,1}, \omega_{1,2}, \omega_{2,2})$, $(\omega_{3,3}, \omega_{3,4}, \omega_{4,4})$ and $(\omega_{1,3}, \omega_{1,4}, \omega_{2,3}, \omega_{2,4})$.

The quadratic effect of the first two covariates are correlated with each other and with the interaction between them. The same holds for the variables loading on the second factor. Finally, the third group contains the interactions between one variable loading on the first factor and one variable loading on the second factor. In general, with $p$ variables and $k$ factors, we will have in total $k + \binom{k}{2}$ groups. In particular, the first $k$ groups will be made by the interactions between variables loading on the same factor. On the other hand, we have groups for the interactions between variables loading on different factors, as in the previous example with $S_1$ and $S_2$.

3.2.5 Complexity Gains

Inference under existing approaches for Bayesian linear modeling for pairwise interactions when $p$ is moderately high is typically computationally infeasible. In fact the complexity per iteration of Gibbs sampling is $O(np^4 + p^6)$ and the storage is of order $O(p^2)$. This is without considering any heredity structure. On the other hand, with model (1) we just need samples of $\Psi, \Lambda, \omega$ and $\Omega$ to compute main effects and interactions of $X$ on $y$ thanks to Proposition 1. The complexity per iteration of Gibbs sampling is $O(k^3p + npk)$, where $k$ is the number of factors. In our motivating applications, we have $n > p > k$. Further, the storage complexity is only $O(pk)$ since we only need to save the samples of $\Lambda, \Phi, \omega$ and $\Omega$. 

54
The computational complexity could be further reduced using the algorithm of Sabnis et al. (2016), which allows one to distribute the covariance matrix estimation to multiple cores, efficiently using a divide and conquer strategy. Let \( g \geq 1 \) denote the number of cores at our disposal and assume that \( p \) is a multiple of \( g \). Letting \( p_g = \frac{p}{g} \), the computational complexity becomes \( \mathcal{O}(k^3 p_g + np_g k) \). If we want to estimate the interactions up to the \( Q^{th} \) order, the computational complexity becomes \( \mathcal{O}(k^3(p + Q) + npk) \). Moreover, the storage complexity is \( \mathcal{O}(p(k + Q)) \).

3.3 Properties of the Model

In this section we prove that the posterior distribution of \( \Theta = (\omega, \Omega, \sigma^2, \Lambda, \Psi) \) is weakly consistent for a broad set of models. Let \( KL(\Theta_0, \Theta) \) denote the Kullback-Leibler divergence between \( p(X, y|\Theta_0) \) and \( p(X, y|\Theta) \), where

\[
p(X, y|\Theta_0) = \int p(X|\Lambda_0, \Psi_0, \eta)p(y|\omega_0, \Omega_0, \sigma^2_0, \eta)p(\eta)d\eta.
\]

We will assume that \( p(X, y|\Theta_0) \) represents the true data-generating model. This assumption is not as restrictive as it may initially seem. The model is flexible enough to always characterize and model quadratic regression in the response component, while accurately approximating any covariance structure in the predictor component. In fact it always holds that:

\[
\mathbb{E}(y_i|X_i) = \beta_0 X_i + X_i \Omega_0 X_i,
\]

\[
X_i \sim N(0, \Lambda_0 \Lambda_0^T + \Psi_0),
\]

where \( \beta_0 \) and \( \Omega_0 \) are functions of \( \Theta_0 \) as in Proposition 1, and the true number of factors is \( k_0 \). When \( k_0 = p \), we can write any covariance matrix as \( \Lambda_0 \Lambda_0^T + \Psi_0 \). We take an “over-fitted” factor modeling approach, related to Bhattacharya and Dunson (2011), Rousseau and Mengersen (2011), and choose \( k \) to correspond to an upper bound on the number of factors. In practice, we recommend the rule of thumb that
chooses $k$ such that $\frac{\sum_{j=1}^{k} v_j}{\sum_{j=1}^{k} v_j} > 0.9$, where $v_j$ is the $j^{th}$ largest singular value of the correlation matrix of $X$. We have found this choice to have good performance in a wide variety of simulation cases. However, there is nonetheless a potential concern that $k$ may be less than $k_0$ in some cases. Proposition 2 quantifies the distance in terms of Kullback-Leibler divergence between the true data generating model and the likelihood under model miss-specification as $n$ approaches infinity.

**Proposition 2.** Fix $\Lambda_0, \Psi_0 = s_0 I_p, k_0$, and assume that $k < k_0$. As $n$ increases the posterior distribution of $\Lambda$ and $\Psi = s I_p$ concentrates around $\Lambda^*$ and $\Psi^*$, satisfying:

$$KL((\Lambda_0, \Psi_0); (\Lambda^*, \Psi^*)) \leq \sum_{j=k+1}^{k_0} \frac{v_j}{s_0},$$

where $v_j$ is the $j^{th}$ largest singular value of $\Lambda_0 \Lambda_0^T$.

Unsurprisingly, the bound of Proposition 2 resembles the Eckart-Young theorem for low-rank approximation based on the Singular Value Decomposition of a matrix. The Eckart-Young theorem states that the rank $k$ approximation $\hat{\Omega}$ of a matrix $\Omega$ minimizing the Frobenius norm is such that $||\hat{\Omega} - \Omega||_F = \sqrt{\sum_{j=k+1}^{p} v_j^2}$. In a similar fashion as Principal Component Analysis and Factor Analysis, we can inspect the singular values of the correlation matrix of the regressors in order to choose the number of factors to include in the model, and thanks to Proposition 2 we know how far the posterior distribution will concentrate from the truth.

The next proposition provides a sufficient condition in order to achieve posterior consistency when $k \geq k_0$. Notice that we achieve posterior consistency on the induced main effects and pairwise interactions.

**Proposition 3.** Fix $\Theta_0 = (\omega_0, \Omega_0, \sigma_0^2, \Lambda_0, \Psi_0, k_0)$. Whenever $k \geq k_0$, for any $\delta > 0$ there exists an $\epsilon > 0$ such that:

$$\{\Theta : d_{x}(\Theta_0, \Theta) < \delta\} \subset \{\Theta : KL(\Theta_0, \Theta) < \epsilon\}$$

56
where $d_\infty$ is the sup-norm.

One can easily define a prior on $\Theta$ such that it places positive probability in any small neighborhood of $\Theta_0$, according to $d_\infty$. The prior defined in 3.2.2 satisfies this condition. Consequently, the posterior distribution of $\Theta$ is weakly consistent due to Schwartz (1965).

3.4 Simulation Experiments

In this section we compare the performance of our FIN method with four other approaches: PIE (Wang et al., 2019), RAMP (Hao et al., 2018), Family (Haris et al., 2016) and HierNet (Bien et al., 2013). These methods are designed for inference on interactions in moderate to high dimensional settings. We generate 25 and 50 covariates in three ways:

- $X_i \sim N_p(0, \Lambda \Lambda^T + I_p)$, $\lambda_{i,j} \sim N(0, 1)$, (factor)
- $X_i \sim N_p(0, W)$, $[W]_{i,j} = 0.8|j - j|$, (linear)
- $X_i \sim N_p(0, I_p)$, (independent)

In the factor scenario we set the true number of factors equal to 7 for $p = 25$ and equal to 17 when $p = 50$. FIN achieved similar performance when we chose a smaller number of latent factors. The average absolute correlation in the covariates is between 0.25 and 0.3 for the factor and linear scenarios when $p = 25$. These two simulation scenarios are the most similar to the environmental epidemiology data analysis in Section 3.5.

For each scenario, we generate the continuous outcome according to a linear regression with pairwise interactions:

$$y_i = X_i^T \beta_0 + X_i^T \Omega_0 X_i + \epsilon_i,$$

where half of the main effects are different from zero and $\epsilon_i \sim N(0, 1)$ for $i = 1, \cdots, 500$. We distinguish between a sparse matrix of pairwise interactions $\Omega_0$, with
only 5% non-zero interactions, or dense, where 20% of the elements are different from zero.

For each value of $p$ we have six simulation scenarios: factor, linear or independent combined with sparse or dense pairwise interactions. We generate the non-zero main effects and interaction coefficients from a Uniform distribution in the interval $(-0.5, 1)$ such that the regression equation follows the strong heredity constraint. Strong heredity allows an interaction between two variables to be included in the model only if the main effects are. This is done to favor RAMP, Family and HierNet, which assume the heredity condition. We repeat the simulations 50 times and evaluate the performance on a test dataset of 500 units computing predictive mean square error, mean square error for main effects, Frobenious norm (FR) for interaction effects, and percentage of true positives (TP) and true negatives (TN) for main effects and interactions. The percentage of TP and TN main effects is defined as follows:

$$\text{TP(main effects)} = \frac{1}{p} \sum_{j=1}^{p} \mathbb{1} (\hat{\beta}_j \neq 0, \beta_{0j} \neq 0, \text{sign}(\hat{\beta}_j) = \text{sign}(\beta_{0j}))$$

$$\text{TN(main effects)} = \frac{1}{p} \sum_{j=1}^{p} \mathbb{1} (\hat{\beta}_j = 0, \beta_{0j} = 0),$$

where $\hat{\beta}_j$ is the estimated main effect for feature $j$ and $\beta_{0j}$ is the true coefficient. FIN is the only method reporting uncertainty quantification and we set $\hat{\beta}_j = 0$ whenever zero is included in the 95% credible interval. We equivalently define the percentage of true positive and true negative interactions.

The MCMC algorithm was run for 5000 iterations with a burn-in of 4000. We observed good mixing. In particular, the Effective Sample Size (ESS) was always greater than 900 across our simulations, both for main effects and interactions. We set the hyperparameter $a$ of the Dirichlet-Laplace prior equal to $1/2$. We obtained
similar results for \( a \) in the interval \([1/p, p]\). The results are summarized in Table 3.1-3.2 for \( p = 25 \) and in Table 3.3-3.4 for \( p = 50 \). Across all the simulations, we chose \( k \) such that \( \sum_{j=1}^{k} v_j > 0.9 \).

In the factor scenario, FIN outperforms the other methods in predictive performance and estimation of main effects and interactions, whereas the rate recovery of true main effects and interactions is comparable to HierNet and PIE with sparse \( \Omega_0 \) and outperforms the other methods when \( \Omega_0 \) is dense. The latter scenario is the most challenging with respect to selection of main effects and interactions. Most of the other methods either select or shrink to zero all the effects. In the linear scenario, FIN also shows the best performance together with PIE and Hiernet. Despite the model misspecification with independent covariates, FIN has a comparable predictive performance with respect to the other methods, which do not take into account correlation structure in the covariates. The 95% predictive intervals provided by FIN contained the true value of \( y_i \) on average approximately 95% of the time in the factor scenario, 89% for the linear scenario, and 79% for the independent scenario. The average bias in the posterior predictive mean is negligible in each simulation scenario.

The optimization method performed by HierNet (Bien et al., 2013) tends to favor interactions only in presence of large component main effects, and in doing so overshrinks interactions estimates, especially in the dense scenario. Penalized regression techniques PIE (Wang et al., 2019) and RAMP (Hao et al., 2018) tend to over-shrink coefficient estimates and select too few predictors, particularly in the dense scenario. On the other hand, FAMILY (Haris et al., 2016) performs a relaxed version of the penalized algorithm by refitting an unpenalized least squares model, which results in a high false positive rate of main effects. We also considered different signal-to-noise ratios with \( \epsilon_i \sim N(0, \frac{1}{4}) \) and \( \epsilon_i \sim N(0, 4) \). The results are very similar
to the results we have presented; hence, we omit them.

3.5 Environmental Epidemiology Application

The goal of our analysis is to assess the effect of ten phthalate metabolites, four perfluoroalkyl (pfas) metabolites and fourteen metals on body mass index (BMI). Phthalates are mainly used as plasticizers and can be found in toys, detergents, food packaging, and soaps. They have previously been associated with increased BMI (Hatch et al., 2008) and waist circumference (WC) (Stahlhut et al., 2007). There is a growing health concern for the association of phthalates (Kim and Park, 2014), (Zhang et al., 2014) and pfas metabolites (Braun, 2017) with childhood obesity. Metals have already been associated with an increase in waist circumference and BMI, see Padilla et al. (2010) and Shao et al. (2017), using data from the National Health and Nutrition Examination Survey (NHANES).

We also consider data from NHANES, using data from the years 2015 and 2016. We select a subsample of 7602 individuals for which the measurement of BMI is not missing, though FIN can easily accommodate missing outcomes through adding an imputation step to the MCMC algorithm. Figure 3.4 contains a plot of the correlation between exposures. Several pairwise correlations are missing, as for example between pfas and most metals, because some chemicals are only measured within subsamples of the data. The average absolute correlation between the 28 exposures is around 0.28, similarly to the factor and linear simulation scenarios presented in Section 3.4. We also include in the analysis cholesterol, creatinine, race, sex, education and age. We apply the log\textsubscript{10} transformation to the chemicals, cholesterol and creatinine. Histograms of the chemical measurements can be found in Figure 3.2. We also apply the log\textsubscript{10} transformation to BMI in order to make its distribution closer to normality, which is the assumed marginal distribution in our model. The log-transformation is commonly applied in environmental epidemiology in order
Table 3.1: Results from simulation study with $p = 25$ and dense $\Omega_0$ in the three scenarios: factor, linear and independent for $n = 500$. We computed test error, Frobenious norm, MSE for main effects, percentage of true positives and true negatives for main effects and interactions for Hiernet, Family, PIE, RAMP and FIN model with $a = 0.5$ across 50 simulations. Test error, FR, and main MSE are presented as ratios compared to the best performing model.

<table>
<thead>
<tr>
<th></th>
<th>HierNet</th>
<th>FAMILY</th>
<th>PIE</th>
<th>RAMP</th>
<th>FIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test error</td>
<td>1.974</td>
<td>16.689</td>
<td>7.067</td>
<td>64.717</td>
<td>1</td>
</tr>
<tr>
<td>FR</td>
<td>1.361</td>
<td>1.013</td>
<td>1.418</td>
<td>1.620</td>
<td>1</td>
</tr>
<tr>
<td>main MSE</td>
<td>1.167</td>
<td>1.062</td>
<td>1.807</td>
<td>4.225</td>
<td>1</td>
</tr>
<tr>
<td>TP main</td>
<td>0.920</td>
<td>0.988</td>
<td>0.155</td>
<td>0.270</td>
<td>0.753</td>
</tr>
<tr>
<td>TN main</td>
<td>0.067</td>
<td>0.007</td>
<td>0.921</td>
<td>0.773</td>
<td>0.475</td>
</tr>
<tr>
<td>TP int</td>
<td>0.151</td>
<td>0.807</td>
<td>0.105</td>
<td>0.037</td>
<td>0.699</td>
</tr>
<tr>
<td>TN int</td>
<td>0.889</td>
<td>0.233</td>
<td>0.929</td>
<td>0.962</td>
<td>0.387</td>
</tr>
<tr>
<td>linear</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test error</td>
<td>1</td>
<td>2.662</td>
<td>1.688</td>
<td>6.309</td>
<td>1.565</td>
</tr>
<tr>
<td>FR</td>
<td>1</td>
<td>1.049</td>
<td>1.075</td>
<td>1.289</td>
<td>1.016</td>
</tr>
<tr>
<td>main MSE</td>
<td>2.421</td>
<td>1</td>
<td>1.766</td>
<td>4.259</td>
<td>1.263</td>
</tr>
<tr>
<td>TP main</td>
<td>1</td>
<td>0.996</td>
<td>0.177</td>
<td>0.301</td>
<td>0.572</td>
</tr>
<tr>
<td>TN main</td>
<td>0.002</td>
<td>0.005</td>
<td>0.904</td>
<td>0.805</td>
<td>0.718</td>
</tr>
<tr>
<td>TP int</td>
<td>0.532</td>
<td>0.818</td>
<td>0.280</td>
<td>0.028</td>
<td>0.635</td>
</tr>
<tr>
<td>TN int</td>
<td>0.849</td>
<td>0.278</td>
<td>0.887</td>
<td>0.968</td>
<td>0.570</td>
</tr>
<tr>
<td>independent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test error</td>
<td>1</td>
<td>6.150</td>
<td>2.759</td>
<td>10.729</td>
<td>7.128</td>
</tr>
<tr>
<td>FR</td>
<td>1.175</td>
<td>1.548</td>
<td>1</td>
<td>2.042</td>
<td>1.654</td>
</tr>
<tr>
<td>main MSE</td>
<td>1</td>
<td>1.529</td>
<td>1.756</td>
<td>2.446</td>
<td>2.031</td>
</tr>
<tr>
<td>TP main</td>
<td>1</td>
<td>1</td>
<td>0.241</td>
<td>0.074</td>
<td>0.302</td>
</tr>
<tr>
<td>TN main</td>
<td>0</td>
<td>0.002</td>
<td>0.930</td>
<td>0.985</td>
<td>0.888</td>
</tr>
<tr>
<td>TP int</td>
<td>0.989</td>
<td>0.952</td>
<td>0.641</td>
<td>0.005</td>
<td>0.412</td>
</tr>
<tr>
<td>TN int</td>
<td>0.937</td>
<td>0.414</td>
<td>0.908</td>
<td>1.000</td>
<td>0.914</td>
</tr>
</tbody>
</table>
Table 3.2: Results from simulation study with $p = 25$ and \textit{sparse} $\Omega_0$ in the three scenarios: factor, linear and independent for $n = 500$. We computed test error, Frobenious norm, MSE for main effects, percentage of true positives and true negatives for main effects and interactions for Hiernet, Family, PIE, RAMP and FIN model with $a = 0.5$ across 50 simulations. Test error, FR, and main MSE are presented as ratios compared to the best performing model.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>HierNet</th>
<th>FAMILY</th>
<th>PIE</th>
<th>RAMP</th>
<th>FIN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>test error</td>
<td>1.284</td>
<td>5.274</td>
<td>1.206</td>
<td>4.225</td>
</tr>
<tr>
<td>factor</td>
<td>FR</td>
<td>1.189</td>
<td>1.259</td>
<td>1</td>
<td>2.157</td>
</tr>
<tr>
<td></td>
<td>main MSE</td>
<td>3.430</td>
<td>1.560</td>
<td>1</td>
<td>1.590</td>
</tr>
<tr>
<td></td>
<td>TP main</td>
<td>0.667</td>
<td>0.823</td>
<td>0.698</td>
<td>0.583</td>
</tr>
<tr>
<td></td>
<td>TN main</td>
<td>0.445</td>
<td>0.259</td>
<td>0.863</td>
<td>0.834</td>
</tr>
<tr>
<td></td>
<td>TP int</td>
<td>0.514</td>
<td>0.839</td>
<td>0.562</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td>TN int</td>
<td>0.959</td>
<td>0.580</td>
<td>0.974</td>
<td>0.965</td>
</tr>
<tr>
<td></td>
<td>test error</td>
<td>1.199</td>
<td>5.271</td>
<td>1</td>
<td>5.060</td>
</tr>
<tr>
<td>linear</td>
<td>FR</td>
<td>3.889</td>
<td>6.859</td>
<td>1</td>
<td>7.916</td>
</tr>
<tr>
<td></td>
<td>main MSE</td>
<td>1</td>
<td>3.563</td>
<td>1.387</td>
<td>1.392</td>
</tr>
<tr>
<td></td>
<td>TP main</td>
<td>1</td>
<td>0.845</td>
<td>0.857</td>
<td>0.952</td>
</tr>
<tr>
<td></td>
<td>TN main</td>
<td>0.484</td>
<td>0.272</td>
<td>0.845</td>
<td>0.815</td>
</tr>
<tr>
<td></td>
<td>TP int</td>
<td>0.970</td>
<td>0.887</td>
<td>0.964</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td>TN int</td>
<td>0.970</td>
<td>0.645</td>
<td>0.987</td>
<td>0.975</td>
</tr>
<tr>
<td></td>
<td>test error</td>
<td>1.425</td>
<td>9.685</td>
<td>1</td>
<td>12.746</td>
</tr>
<tr>
<td></td>
<td>main MSE</td>
<td>1</td>
<td>6.082</td>
<td>3.056</td>
<td>4.326</td>
</tr>
<tr>
<td></td>
<td>TP main</td>
<td>1</td>
<td>0.830</td>
<td>0.860</td>
<td>0.630</td>
</tr>
<tr>
<td></td>
<td>TN main</td>
<td>0.418</td>
<td>0.585</td>
<td>0.847</td>
<td>0.915</td>
</tr>
<tr>
<td></td>
<td>TP int</td>
<td>1</td>
<td>0.852</td>
<td>1</td>
<td>0.071</td>
</tr>
<tr>
<td></td>
<td>TN int</td>
<td>0.993</td>
<td>0.868</td>
<td>0.990</td>
<td>0.995</td>
</tr>
</tbody>
</table>
Table 3.3: Results from simulation study with \( p = 50 \) and dense \( \Omega_0 \) in the three scenarios: factor, linear and independent for \( n = 500 \). We computed test error, Frobenious norm, MSE for main effects, percentage of true positives and true negatives for main effects and interactions for Hiernet, Family, PIE, RAMP and FIN model with \( a = 0.5 \) across 50 simulations. Test error, FR, and main MSE are presented as ratios compared to the best performing model.

<table>
<thead>
<tr>
<th></th>
<th>HierNet</th>
<th>FAMILY</th>
<th>PIE</th>
<th>RAMP</th>
<th>FIN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>factor</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test err</td>
<td>1</td>
<td>5.011</td>
<td>4.080</td>
<td>29.464</td>
<td>1.207</td>
</tr>
<tr>
<td>FR</td>
<td>1.317</td>
<td>1</td>
<td>1.369</td>
<td>1.098</td>
<td>1.021</td>
</tr>
<tr>
<td>main MSE</td>
<td>1.557</td>
<td>1</td>
<td>1.385</td>
<td>3.383</td>
<td>1.255</td>
</tr>
<tr>
<td>TP main</td>
<td>0.974</td>
<td>1</td>
<td>0.053</td>
<td>0.056</td>
<td>0.741</td>
</tr>
<tr>
<td>TN main</td>
<td>0.027</td>
<td>0.002</td>
<td>0.975</td>
<td>0.961</td>
<td>0.467</td>
</tr>
<tr>
<td>TP int</td>
<td>0.135</td>
<td>0.858</td>
<td>0.072</td>
<td>0.002</td>
<td>0.662</td>
</tr>
<tr>
<td>TN int</td>
<td>0.911</td>
<td>0.171</td>
<td>0.951</td>
<td>0.998</td>
<td>0.388</td>
</tr>
<tr>
<td><strong>linear</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test err</td>
<td>1</td>
<td>1.547</td>
<td>2.100</td>
<td>4.717</td>
<td>1.781</td>
</tr>
<tr>
<td>FR</td>
<td>1.167</td>
<td>1</td>
<td>1.165</td>
<td>1.102</td>
<td>1.045</td>
</tr>
<tr>
<td>main MSE</td>
<td>14.006</td>
<td>1</td>
<td>1.873</td>
<td>3.267</td>
<td>1.827</td>
</tr>
<tr>
<td>TP main</td>
<td>1</td>
<td>1</td>
<td>0.071</td>
<td>0.034</td>
<td>0.114</td>
</tr>
<tr>
<td>TN main</td>
<td>0</td>
<td>0</td>
<td>0.945</td>
<td>0.968</td>
<td>0.913</td>
</tr>
<tr>
<td>TP int</td>
<td>0.332</td>
<td>0.798</td>
<td>0.068</td>
<td>0.002</td>
<td>0.147</td>
</tr>
<tr>
<td>TN int</td>
<td>0.866</td>
<td>0.280</td>
<td>0.966</td>
<td>0.998</td>
<td>0.902</td>
</tr>
<tr>
<td><strong>independent</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test err</td>
<td>1</td>
<td>1.072</td>
<td>1.251</td>
<td>1.322</td>
<td>1.239</td>
</tr>
<tr>
<td>FR</td>
<td>1.211</td>
<td>1</td>
<td>1.112</td>
<td>1.140</td>
<td>1.105</td>
</tr>
<tr>
<td>main MSE</td>
<td>9.340</td>
<td>1.565</td>
<td>1.295</td>
<td>1.378</td>
<td>1</td>
</tr>
<tr>
<td>TP main</td>
<td>1</td>
<td>1</td>
<td>0.091</td>
<td>0.013</td>
<td>0</td>
</tr>
<tr>
<td>TN main</td>
<td>0</td>
<td>0</td>
<td>0.947</td>
<td>0.996</td>
<td>1</td>
</tr>
<tr>
<td>TP int</td>
<td>0.516</td>
<td>0.861</td>
<td>0.082</td>
<td>0</td>
<td>0.0002</td>
</tr>
<tr>
<td>TN int</td>
<td>0.810</td>
<td>0.308</td>
<td>0.991</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 3.4: Results from simulation study with $p = 50$ and sparse $\Omega_0$ in the three scenarios: factor, linear and independent for $n = 500$. We computed test error, Frobenious norm, MSE for main effects, percentage of true positives and true negatives for main effects and interactions for Hiernet, Family, PIE, RAMP and FIN model with $a = 0.5$ across 50 simulations. Test error, FR, and main MSE are presented as ratios compared to the best performing model.

<table>
<thead>
<tr>
<th></th>
<th>HierNet</th>
<th>FAMILY</th>
<th>PIE</th>
<th>RAMP</th>
<th>FIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>factor</td>
<td>test error</td>
<td>2.755</td>
<td>16.096</td>
<td>3.381</td>
<td>24.331</td>
</tr>
<tr>
<td></td>
<td>FR</td>
<td>1</td>
<td>1.184</td>
<td>1.103</td>
<td>1.358</td>
</tr>
<tr>
<td></td>
<td>main MSE</td>
<td>1.367</td>
<td>1.209</td>
<td>1.183</td>
<td>1.853</td>
</tr>
<tr>
<td></td>
<td>TP main</td>
<td>0.825</td>
<td>0.958</td>
<td>0.415</td>
<td>0.264</td>
</tr>
<tr>
<td></td>
<td>TN main</td>
<td>0.225</td>
<td>0.067</td>
<td>0.890</td>
<td>0.934</td>
</tr>
<tr>
<td></td>
<td>TP int</td>
<td>0.631</td>
<td>0.940</td>
<td>0.485</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>TN int</td>
<td>0.966</td>
<td>0.264</td>
<td>0.967</td>
<td>0.995</td>
</tr>
<tr>
<td>linear</td>
<td>test error</td>
<td>1</td>
<td>5.894</td>
<td>1.168</td>
<td>8.743</td>
</tr>
<tr>
<td></td>
<td>FR</td>
<td>1.122</td>
<td>1.990</td>
<td>1</td>
<td>2.154</td>
</tr>
<tr>
<td></td>
<td>main MSE</td>
<td>1</td>
<td>1.751</td>
<td>1.614</td>
<td>2.148</td>
</tr>
<tr>
<td></td>
<td>TP main</td>
<td>1</td>
<td>0.900</td>
<td>0.496</td>
<td>0.342</td>
</tr>
<tr>
<td></td>
<td>TN main</td>
<td>0.208</td>
<td>0.141</td>
<td>0.856</td>
<td>0.954</td>
</tr>
<tr>
<td></td>
<td>TP int</td>
<td>0.952</td>
<td>0.875</td>
<td>0.731</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>TN int</td>
<td>0.972</td>
<td>0.560</td>
<td>0.972</td>
<td>0.997</td>
</tr>
<tr>
<td>independent</td>
<td>test error</td>
<td>1</td>
<td>7.881</td>
<td>1.738</td>
<td>10.984</td>
</tr>
<tr>
<td></td>
<td>FR</td>
<td>1.595</td>
<td>2.588</td>
<td>1</td>
<td>2.884</td>
</tr>
<tr>
<td></td>
<td>main MSE</td>
<td>1</td>
<td>2.687</td>
<td>2.859</td>
<td>3.573</td>
</tr>
<tr>
<td></td>
<td>TP main</td>
<td>1</td>
<td>1</td>
<td>0.350</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td>TN main</td>
<td>0.054</td>
<td>0.012</td>
<td>0.919</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>TP int</td>
<td>1</td>
<td>0.985</td>
<td>0.753</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>TN int</td>
<td>0.979</td>
<td>0.355</td>
<td>0.979</td>
<td>1</td>
</tr>
</tbody>
</table>
to reduce the influence of outliers and has been employed in several studies using NHANES data (Nagelkerke et al., 2006), (Lynch et al., 2010), (Buman et al., 2013). We leave these transformations implicit for the remainder of the section.

We assume a latent normal structure for the chemicals, which are included in the matrix $X$, and use the other variables as covariates, which are included in the matrix $Z$. We estimate a quadratic regression according to model (2). We specify independent Gaussian priors for elements of $\alpha$ and $\Delta$. Algorithm 3 can be easily adapted for model (2). The matrix $X$ has 60% missing data and Figure 3.3 contains a plot of the missingness pattern. Since we are modeling the chemical measurements, we can simply add a sampling step to the MCMC algorithm to sample the missing data according to (2). Similarly, 0.4% of chemicals have been recorded under the limit of detection (LOD). In order to be coherent with our model we can sample these observations as:

$$X_{ij} \mid X_{ij} \in [-\infty, \log_{10}(\text{LOD}_j)] \sim TN(\mu_i \lambda_j, \sigma_j^2, -\infty, \log_{10}(\text{LOD}_j))$$

where $\text{LOD}_j$ is the limit of detection for exposure $j$ and $TN(\mu, \sigma^2, a, b)$ is a truncated normal distribution with mean $\mu$, variance $\sigma^2$ and support in $[a, b]$. We imputed the missing data using MICE (White et al., 2011) to compute the correlation matrix of chemicals. We noticed from the Eigendecomposition of the correlation matrix that the first 13 eigenvectors explain more than 90% of the total variability; hence we set the number of factors equal to 13.

Figure 3.4 on the right shows the posterior mean of the matrix of factor loadings $\Lambda$, before and after applying the MatchAlign algorithm of Chapter 6, which resolves rotational ambiguity and column label switching for the posterior samples of $\Lambda$. The matrix of factor loadings reflects the correlation structure of the chemicals. We can distinguish three families of chemicals: metals collected from urine, PFAS and phthalates. The PFAS chemicals load mostly on the 1st factor, the metals from urine.
on the 8th factor together with the phthalates, which is expected since there is high
correlation between the two groups of chemicals. Finally, a group of highly correlated
phthalates loads on the 13th factor.

We also estimated a regression with pairwise interactions using the methods PIE,
RAMP, Family and HierNet introduced in Section 3.4. These methods do not directly
deal with missing data, so we imputed the missing data using MICE (White et al.,
2011). Figure 3.5 shows the estimated main effects of the chemicals. The signs of
the coefficients are generally consistent across different methods.

Figure 3.6 shows the posterior mean of the matrix of chemical interactions and
of the matrix $A^T\Delta$ of pairwise interactions between exposures and covariates. As
expected, we estimate a “dense” matrix of interactions. This is due to exposures
being breakdown products of the same compound and high correlation between
chemicals belonging to the same family. For example the correlation between the
pfas metabolites is equal to 0.7, with only 1977 observations containing complete
measurements. Interactions between highly correlated pfas metabolites have been
observed in animal studies (Wolf et al., 2014), (Ding et al., 2013). Linear (Henn
et al., 2011), (Lin et al., 2013) and nonlinear interactions (Valeri et al., 2017) be-
tween metals have been associated with neurodevelopment. Interactions between
phathalates and other chemicals have been related to human semen quality (Hauser
et al., 2005). Finally, we estimate several interactions between chemicals and age,
cholesterol and creatinine, which are usually expected in environmental epidemiology
applications (Barr et al., 2004). The code for reproducing the analysis is available

3.6 Structural Equation Modeling of Interactions

In applied fields such as epidemiology, multivariate outcomes are routinely collected
in many studies and it can be misleading to consider outcomes one at a time in
analyses (Schwartz et al. (2010), Samoilenko et al. (2018)). Deleforge et al. (2015) use inverse multivariate regression and addresses non-linearity using a kernel method. Using an interaction screening approach, Kong et al. (2017) suggest a two-stage method for identifying interactions for multivariate outcomes without relying on the heredity assumption, and Pan et al. (2019b) propose a more flexible model building on Kong et al. (2017). However, none these methods report parameter uncertainty.

To address the problem of quadratic regression in high-dimensional settings with correlated covariates and outcomes, we propose to use a Bayesian approach to inference with Structural Equation Models (SEM). SEMs are a natural generalization of factor analysis to multivariate outcomes and have been successfully employed for multilevel data (Ansari and Jedidi, 2001) and nonlinear structures (Lee et al., 2004). Dimension reduction in SEM is realized by the estimation of latent factors for outcomes and covariates. Moreover, similar to principal component regression, SEM directly characterizes correlated predictors and responses using latent factors.

Let $Y_i = (y_{i1}, \cdots, y_{iq})$ denote a vector of continuous health responses for individual $i$, and $X_i = (x_{i1}, \cdots, x_{ip})^T$ denote a vector of exposure measurements, for $i = 1, \ldots, n$. Our goal is to jointly model the impact of the exposure measurements on the outcomes and estimate interaction effects. We propose to use a structural equation model, and we model the dependence between the outcome and the predictors via the latent variables. Expanding on the Quasi-ML model proposed by Klein and Muthén (2007), which constrains the response vector to load on a single latent factor, we consider a SEM with a quadratic form. Normalizing data prior to analysis
so we can omit the intercepts, we let:

\[ Y_i = \Lambda_y \xi_i + \epsilon_{y,i}, \quad \epsilon_{y,i} \sim N_q(0, \Phi) \]
\[ X_i = \Lambda_x \eta_i + \epsilon_{x,i}, \quad \epsilon_{x,i} \sim N_p(0, \Psi) \]
\[ \xi_i = \Gamma \eta_i + \Omega(\eta_i) + \epsilon_{\xi,i}, \quad \epsilon_{\xi,i} \sim N_m(0, \Sigma_{\xi}) \]
\[ \eta_i \sim N_k(0, \Sigma_{\eta}), \]

where \( \Lambda_y \) and \( \Lambda_x \) are factor loading matrices of dimensions \( q \times m \) and \( p \times k \), respectively, \( \Phi, \Psi, \Sigma_{\eta} \) and \( \Sigma_{\xi} \) are all diagonal matrices, and \( \Gamma \) is a matrix of dimension \( m \times k \). \( \Omega(\eta_i) \) is an \( m \times 1 \) vector such that the \( j^{th} \) element is equal to \( \eta_i^T \Omega_j \eta_i \), with \( \Omega_j \) being a \( k \times k \) matrix containing the coefficients of the pairwise interactions of \( \eta_i \) on \( \xi_{ij} \).

**Proposition 4.** Under Model (3.4), the following is true:

\[ \mathbb{E}(Y_i|X_i) = \Lambda_y \left[ \Gamma A X_i + \mathbb{E}(\Omega(\eta_i)|X_i) \right], \]

where \( A = V \Lambda_x^T \Psi^{-1} \) and \( V = (\Lambda_x^T \Psi^{-1} \Lambda_x + \Sigma_{\eta}^{-1})^{-1} \). \( \mathbb{E}(\Omega(\eta_i)|X_i) \) is an \( m \times 1 \) vector such that the \( j^{th} \) element is equal to \( \mathbb{E}(\eta_i \Omega_j \eta_i|X_i) \), for \( j = 1, \ldots, m \). In particular we have that:

\[ \mathbb{E}(\eta_i \Omega_j \eta_i|X_i) = tr(\Omega_j V) + X_i^T (A^T \Omega_j A) X_i. \]  

**Corollary 5.** The induced main effects and interaction effects of Proposition (4) are identifiable.

In epidemiology studies, it is of interest to include interactions between chemical exposures and demographic covariates. The covariates are often binary variables, like *race* or *sex*, or continuous variables that are non-normally distributed, like *age*. Hence, we do not want to assume a latent normal structure for the covariates. Denoting \( Z_i = (z_{i1}, \cdots, z_{il})^T \) as a vector of covariates, we can modify Model (3.4) to include main effects for \( Z_i \):

\[ Y_i = \alpha Z_i + \Lambda_y \xi_i + \epsilon_{y,i}, \]  

(3.6)
where $\alpha$ is a $q \times \ell$ matrix, representing main effects for the covariates. If we are also interested in interactions between the effect modifiers and the chemical exposures, we can further modify (3.4) as follows:

$$
\xi_i = \Gamma \eta_i + \Omega(\eta_i) + \Delta(\eta_i) + \epsilon_{\xi,i},
$$

(3.7)

where $\Delta(\eta_i)$ is an $m \times 1$ vector whose $j^{th}$ element is $\eta_i \Delta_j Z_i$, with $\Delta_j$ being a $k \times \ell$ matrix of interaction coefficients between the latent variables and the covariates. Following Proposition 4, we can show that this specification induces interaction effects between covariates and chemical measurements.

**Corollary 6.** Adding (3.7) to Model (3.4), we have that:

$$
E(Y_i|X_i) = \Lambda_q[\Gamma A X_i + E(\Omega(\eta_i)|X_i) + E(\Delta(\eta_i)|X_i)] + \alpha Z_i
$$

where $A$ and $E(\Omega(\eta_i)|X_i)$ are defined in Proposition 4. $E(\Delta(\eta_i)|X_i)$ is an $m \times 1$ vector such that the $j^{th}$ element is equal to $E(\eta_i^T \Delta_j Z_i|X_i)$, for $j = 1, ..., m$. In particular,

$$
E(\eta_i^T \Delta_j Z_i|X_i) = X_i^T (A^T \Delta_j) Z_i.
$$

3.7 Discussion

We proposed a novel method that exploits the correlation structure of the predictors and allows us to estimate interaction effects in high dimensional settings, assuming a latent factor model. Using simulated examples, we showed that our method has a similar performance to state-of-the-art methods for interaction estimation when dealing with independent covariates and outperforms the competitors when there is moderate to high correlation among the predictors. We provided a characterization of uncertainty with a Bayesian approach to inference. Our FIN approach is particularly motivated by epidemiology studies with correlated exposures, as illustrated using data from NHANES.
NHANES data are obtained using a complex sampling design, that includes oversampling of certain population subgroups, and contains sampling weights for each observation that are inversely proportional to the probability of being sampled. We did not employ sampling weights in our analysis because our goal was to study the association between exposures and BMI rather than providing population estimates. One possibility to include the sampling weights in our method is to jointly model the outcome and the survey weights (Si et al., 2015), without assuming that the population distribution of strata is known.

Our MCMC algorithm can be efficiently employed for $n$ and $p$ in the order of thousands and hundreds respectively, which allows us to estimate around 5000 interactions when $p = 100$. However, it is necessary to speed up the computations in order to apply our method to bigger $p$, which is common with genomics data. The computational bottleneck is the Metropolis Hastings step described in 3.2.2. One possibility is to include the heredity constraint (Chipman, 1996) while estimating the factors.

In order to allow departures from linearity and Gaussianity, it is of interest to model the regression on the health outcome as a non-linear function of latent factors. Non parametric latent models have desirable properties in term of convergence rates (Zhou et al., 2017) and large support for density estimation (Kundu and Dunson, 2014). Verma and Engelhardt (2018) developed a dimension reduction approach with latent variables for single cell RNA-seq data building on Gaussian process latent variable models (GP-LVM). Although attractive from a modeling perspective, a major challenge is efficient posterior computation. Another promising direction to decrease modeling assumptions is to rely on a copula factor model related to Murray et al. (2013).
Figure 3.1: Induced priors on main effects, pairwise interactions and 3rd order interactions for $p = 20$ and $k = 5, 10$. The green lines corresponds to 0.25 and 0.75 quartiles and the red lines to the 0.05 and 0.95.
Figure 3.2: Histograms of the chemicals measurements included in the matrix $X$. 
Figure 3.3: Pattern of Missing data in the matrix $X$ including the chemical measurements.
Figure 3.4: On the left, correlation between the exposures, the colour grey indicates missing pairwise correlation. On the right, posterior mean of the matrix $\Lambda$ of factor loadings before and after applying the MatchAlign algorithm.

Figure 3.5: Estimated main effects using FIN with 95% credible intervals and estimated coefficients using RAMP, hierNet, Family and PIE.
Figure 3.6: On the left, posterior mean of the matrix of chemicals interactions. On the right, posterior mean of the matrix $A^T \Delta$ of pairwise interactions between exposures and covariates. The white boxes indicates that the 99% credible interval contains zero.
Quadratic Proportional Hazards Model with High-Dimensional Correlated Predictors

4.1 Introduction

In medical studies, it is common to collect high dimensional data and time-to-event outcomes. Cox proportional hazards (PH) models are routinely used with variable selection or regularization when the number of predictors $p$ is high. In many settings, it is important to assess covariate interactions. For example, there is often interest in interactions between exposures and between treatments and other factors. In clinical studies, identifying treatment-biomarker interactions can be critical to the development of personalized treatment regimes. While standard Cox PH methods can be used after including all quadratic terms as candidate predictors, unstructured methods that treat the interaction coefficients the same as main effects can have poor performance. This is particularly true when the number of covariates $p$ is large and/or certain covariates are moderately to highly correlated.

Several methods have been developed to tackle high dimensionality in the Cox model, including LASSO (Witten and Tibshirani, 2010), adaptive LASSO, and smoothly
clipped absolute deviation (SCAD) (Fan and Li, 2002). By shrinking some regression coefficients to zero, these methods performance variable selection and coefficient estimation simultaneously (Wang et al., 2014). Motivated by the lack of consideration of interactions, Wang et al. (2019) proposed a modified adaptive Lasso method to select main effects and interactions in Cox regression. This approach enforces the strong heredity constraint, meaning that interactions between two covariates are included in the model only if both main effects are. In contrast, the weak heredity constraint requires only one main effect to be in the model. Li and Sun (2020) propose a regularization procedure, RAMP-PPL, for selection of main and interaction effects under weak or strong heredity. A drawback of these methods is that they do not report uncertainties in model selection and parameter estimation and rely on strong sparsity assumptions. Moreover, none of the methods mentioned above are explicitly designed for correlated predictors. Penalization methods often have high type I and II error rates when predictors are highly correlated.

Focused on the problem of correlated predictors but not on variable selection or interaction modeling, McCurdy et al. (2017) propose a latent variable model for survival time prediction. Latent factor models provide dimensionality reduction by representing predictors as linear combinations of low-dimensional factors. Latent factor models are useful in characterizing the impact of correlated groups of predictors on the response. While being closely related to Principal Component Analysis (PCA), this approach has the advantages of simultaneously incorporating shrinkage and providing uncertainty quantification within a Bayesian framework. In addition, within the Bayes latent factor regression setting, typical identifiability constraints such as orthogonality are not needed (see, for example Bhattacharyya and Dunson (2011)). Latent factor survival models have previously been used in the Bayesian setting by Pan et al. (2019a), who introduce latent variables to model the hazard of diabetic complications with highly correlated predictors. Their approach has issues
in terms of interpretability, as they lack expressions for regression coefficients on the observed predictors. Also, they impose identifiability constraints that reduce model flexibility and impose an order dependence on the covariates (Carvalho et al., 2008).

The main contribution of this chapter is to propose a framework for inference on main effects and interactions in Cox models via an interpretable and flexible Bayesian approach. Generalizing the latent variable modeling method of Ferrari and Dunson (2020), which focused on Gaussian linear models, we provide expressions for the induced main effects and pairwise interactions. We do not require identifiability constraints on the factor loadings. We further show how our approach can be extended to estimate pairwise and higher order interactions in log linear models for count data.

In Section 4.2 we describe the proposed survival factor model with interactions and provide expressions for induced main and quadratic effects. Section 4.3 provides an overview of our MCMC algorithm. In Section 4.4, we discuss extensions of our results to generalized linear models with logarithmic link functions, such as Poisson and negative binomial regression. In Section 4.5 we illustrate our method using gene expression data from an ovarian cancer study to predict time until death.

### 4.2 Survival Factor Model

Let $X_i = (x_{i1}, \ldots, x_{ip})^T$ denote a vector of correlated covariates, $t_i$ a time-to-event and $c_i$ a fixed censoring time. The value of $t_i$ is only observed when $t_i \leq c_i$, and we let $\nu_i = 1(t_i \leq c_i)$ denote the censoring indicator. As a building block for our survival factor model, we propose to use a Cox proportional hazards (PH) model (Cox, 1972), with hazard function

$$\lambda(t|X_i) = \lambda_0(t)\exp(X_i^T\beta),$$

78
baseline hazard $\lambda_0(t)$ and linear predictor $X_i^T \beta$. The hazard $\lambda(t|X_i)$ provides the rate of event occurrence for individual $i$ conditional on their covariates. The Cox PH model is routinely used in time-to-event analyses, providing simple interpretations for covariate effects in terms of hazard ratios. We want to maintain this interpretation in developing a Bayesian method for inference on interactions accommodating high-dimensional correlated covariates.

With this goal in mind, we start by adapting the approach of Ferrari and Dunson (2020), and defining a latent factor joint model for the predictors $X_i$ and time-to-event response, including main effects and pairwise interactions in the survival function. We also assume that, given the latent variables, covariates are continuous and normally distributed. For simplicity, we omit the intercept in the $X_i$ component model. The model is as follows:

$$
\lambda(t|\eta_i) = \lambda_0(t)\exp(\eta_i^T \omega + \eta_i^T \Omega \eta_i),
$$

$$
X_i = \Lambda \eta_i + \epsilon_i, \quad \eta_i \sim N_k(0, I), \quad \epsilon_i \sim N_p(0, \Psi),
$$

where $\eta_i$ are latent factors, $\Lambda$ is a $p \times k$ tall and skinny factor loadings matrix, $\Psi = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$ is a diagonal residual covariance matrix, and $\omega, \Omega$ are coefficients controlling how the latent factors relate to survival, with $\omega$ controlling main effects and $\Omega$ interactions.

We assume a prior for the parameters $\Theta = (\omega, \Omega, \Lambda, \Psi)$ that will be specified in Section 4.3. Model (4.1) is equivalent to Pan et al. (2019a), except for the $\eta_i^T \Omega \eta_i$ term. Here, $\Omega$ is a $k \times k$ symmetric matrix inducing a quadratic latent variable regression that characterizes interactions among the latent variables. The above formulation can be shown to induce a quadratic effect of $X_i$ on the log hazard function after integrating out the latent variables $\eta_i$. In doing so, we also compute expressions for induced main effects, allowing for straightforward interpretation of coefficients, without the need of identifiability constraints as in Pan et al. (2019a).
Proposition 7. Under model (4.1), the following is true:

$$E(\lambda(t|X_i)) = \lambda_0(t)\sqrt{\frac{|V'|}{|V|}} \exp(H_i),$$

$$H_i = \frac{1}{2}\omega^TV'\omega + \beta_X^TX_i + X_i^T\Omega_XX_i,$$

where $\beta_X^T = \omega^T(I-2\Omega V)^{-1}A$, $\Omega_X = \frac{1}{2}A^T[(I-2\Omega V)^{-1}-I]V^{-1}A$, $V' = (V^{-1}-2\Omega)^{-1}$, $V = (\Lambda^T\Psi^{-1}\Lambda + I)^{-1}$, $A = VA^T\Psi^{-1}$.

This shows that model (4.1) induces main effects and interactions in the hazard function. Notice that the induced main effects without interactions are equal to $\omega^TA$, similarly to factor regression (West, 2003), (Ferrari and Dunson, 2020).

In epidemiology, it is of interest to include interactions between chemical exposures and demographic covariates. The covariates are often binary variables, like *ethnicity* or *sex*, or continuous variables that are non-normally distributed, like *age*. Here, we do not want to assume a latent normal structure for these covariates. Letting $Z_i = (z_{i1}, \cdots, z_{iq})^T$ be a vector of covariates, we modify model (4.2) to include an interaction term between $Z_i$ and the latent factor $\eta_i$:

$$\lambda(t|\eta_i) = \lambda_0(t)\exp(\eta_i^T\omega + \eta_i^T\Omega\eta_i + Z_i^T\alpha + \eta_i^T\Delta Z_i),$$

where $\alpha = (\alpha_1, \cdots, \alpha_q)$ are main effects for the covariates and $\Delta$ is a $k \times q$ matrix of interaction coefficients between the latent variables and the covariates. Following Proposition 7, we have an expression for the marginal hazard function integrating out the latent factors,

$$E(\lambda(t|X_i)) = \sqrt{\frac{|V'|}{|V|}} \exp(H_i),$$

$$H_i = \frac{1}{2}\omega^TV'\omega + \beta_X^TX_i + X_i^T\Omega_XX_i + \alpha_Z^TZ_i + Z_i^T\Delta_ZX_i + \frac{1}{2}Z_i^T\Omega_ZZ_i,$$

where $\Delta_Z = \Delta^T(I - 2\Omega V)^{-1}A$ is a $q \times p$ matrix of pairwise interactions between exposures and covariates, $\alpha_Z = \alpha + \Delta^TV'\omega$ is a vector of main effects for $Z_i$ and
\( \Omega_z = \Delta^T V \Delta \) is a \( q \times q \) matrix of pairwise interactions between covariates. In the sequel, we focus our development on model (7) for ease in exposition, but all of the details can be easily modified to pertain to model (4.2).

4.3 Priors and MCMC Algorithm

In this section we define the priors for \( \omega, \Omega, \Lambda \), briefly describe the computational challenges given by model (4.1) and summarize our Markov chain Monte Carlo sampler in Algorithm 4.

We assume independent normal priors element-wise for \( \omega \) and \( \Omega \). For \( \Lambda = \{\lambda_{i,j}\} \), a typical choice to attain identifiability requires \( \lambda_{i,j} = 0 \) for \( j > i \) and \( \lambda_{j,j} > 0 \) for \( j = 1, \ldots, k \) (Geweke and Zhou, 1996). However, some Bayesian applications, such as covariance estimation (Bhattacharya and Dunson, 2011), do not require the identifiability of \( \Lambda \). The same holds for inference on induced main effects and interactions for model (4.1). As it is reasonable to assume that each covariate loads only on a few factors, the Dirichlet-Laplace (DL) prior (Bhattacharya et al., 2015) is used to induce near sparsity row-wise in the matrix \( \Lambda \). The DL prior provides flexible shrinkage on the factor loadings matrix, \( \Lambda \), and induces a prior concentrated heavily at zero. In practice, we recommend the rule of thumb that chooses \( k \) such that

\[
\frac{\sum_{j=1}^{k} v_j}{\sum_{j=1}^{p} v_j} > 0.9, \quad \text{where} \quad v_j \text{ is the } j^{th} \text{ largest singular value of the correlation matrix of } X.
\]

Proposition 2 provides theoretical justification for this criterion. As an alternative to row-wise shrinkage, we could have instead used column-wise shrinkage as advocated in Bhattacharya and Dunson (2011) and Legramanti et al. (2019).

We use the following partial likelihood of Cox (1972) in place of the exact data
likelihood in defining the posterior distribution of the parameters.

\[ L(t|\eta_t) = \frac{\lambda(t|\eta_t)}{\sum_{j:t_j \geq t_i} \lambda(t|\eta_j)}, \]

\[ L(t_1, \ldots, t_n|\ldots) = \prod_{i, q=1} L(t|\eta_t). \]  

(4.3)

This is a common approach in the Bayesian survival analysis literature. A summary of our MCMC procedure for the parameter of model 4.2 can be found in Algorithm 4. This is a hybrid Gibbs-Metropolis algorithm. The Metropolis-Hastings updates are carried using the Metropolis-Adjusted Langevin Algorithm (MALA) (Grenander and Miller, 1994). For most of the steps, the full conditional posterior distributions have a simple form.

4.4 Extensions to GLMs

In this section we show how to extend the results of Proposition (7) to generalized linear models with logarithmic link and exponential families.

4.4.1 Regression with Count Data

Several methodologies exist to tackle high dimensionality in the setting of generalized linear models with count data outcomes. Lee et al. (2013) propose an efficient algorithm, Poisson Singular Value Decomposition with Offset (PSVDOS), to estimate the Poisson factor model. Using this methodology, the authors identify low-dimensional factors in gene-expression data. Similarly, Zhou et al. (2012) propose a flexible non-parametric Bayesian prior, the beta-negative binomial (BNB) process, for an infinite Poisson factor analysis model. Zhou et al. (2018) posit an alternative to Poisson factor analysis, Negative Binomial factor analysis (NBFA), to address the issue of overdispersion in the Poisson model. Despite the wide range of existing methods for generalized linear models with logarithmic links, there does not exist a strong
body of literature that addresses dimensionality reduction while also accounting for quadratic effects. Thus, the results in Proposition (7) for model (4.1) can be particularly useful for various generalized linear models with logarithmic links, such as the Poisson and Negative-Binomial.

Let $y_i$ denote the count response variable. We consider generalized linear model with a logarithmic link

$$\log(\mathbb{E}(y_i|X_i)) = \beta^T X_i$$

where the logarithmic function is used to relate the linear term, $\beta^T X_i$, to the conditional expectation of $y_i$ given $X_i$. Two notable generalized linear models with logarithmic link are the Poisson and the Negative-Binomial models. In the Poisson model,

$$\lambda_i = \mathbb{E}(y_i|X_i) = e^{\beta^T X_i}$$

$$Y_i|X_i \sim \text{Poisson}(\lambda_i)$$

where $\lambda_i$ is the mean parameter for the $i^{th}$ observation. A main limitation of the Poisson distribution is the fact that the mean and variance are equal. In order to deal with overdispersed count data, a possible alternative is to use Negative-Binomial regression, namely:

$$\mu_i = \mathbb{E}(y_i|X_i) = e^{\beta^T X_i}$$

$$Y_i|X_i \sim \text{NB}(\mu_i, \frac{\mu_i}{\mu_i + r_i})$$

where NB indicates the Negative Binomial distribution with expectation $\mu_i$. In both scenarios, we can add a latent factor model for the covariates and induce a dependence with the outcome by including latent factors in the regression equation. As before, we include main effects and interactions between the latent variables:

$$\log(\mathbb{E}(y_i|\eta_i)) = \omega^T \eta_i + \eta_i^T \Omega \eta_i$$

$$X_i = \Lambda \eta_i + \epsilon_i; \quad \epsilon_i \sim N_p(0, \Psi);$$

$$\eta_i \sim N(0, I)$$

(4.4)
Then, we can use Proposition (7) to compute \( \mathbb{E}(y_i|\eta_i) \), and show that model (4.4) induces a quadratic regression in the space of the covariates. In particular we have that:

\[
\mathbb{E}(y_i|X_i) = \sqrt{|V|} \exp(H),
\]

\[
H_i = \frac{1}{2} \omega^T V' \omega + \beta_X^T X_i + X_i^T \Omega_X X_i,
\]

which allows to estimate quadratic effects with high-dimensional correlated predictors in regression settings with count data.

### 4.4.2 Exponential Families

Now we consider the more general scenario when the outcome distribution is part of the exponential family. Chen et al. (2013b) propose a dimensionality reduction method, Generalized Linear Principal Component Analysis (GLPCA), that can be broadly applied to exponential family distributions and allows for a large range of link functions, including log, negative log-log, complementary log-log, negative binomial, and probit. Collins et al. (2002) and Mohamed et al. (2008) similarly discuss extensions of PCA to the exponential family, with the latter exploiting a Bayesian approach in order to minimize overfitting and improve generalizability. In addition to PCA, factor models have previously been used to deal with high dimensionality in the exponential family. Wedel and Kamakura (2001) discusses a general approach to factor analysis in which different observed and latent variables can come from different distributions, including those in the exponential family. Li and Tao (2013) similarly posit a probabilistic approach to factor analysis in which factors belong to various exponential family distributions.

Let us assume that the distribution of the outcome \( y_i \) is part of the exponential
family, i.e.:

\[ p(y_i|\xi_i) = e^{\xi_i T(y_i) - A(\xi_i)} \]

where \( \xi_i \) is the univariate natural parameter and \( T(y_i) \) is a sufficient statistic. We generalize the Gaussian linear factor model and set \( \xi_i = \eta_i^T \omega \), without considering interactions for simplicity. We know that \( E(y_i|\eta_i) = \mu_i = g^{-1}(\eta_i^T \omega) \), where \( g(\cdot) \) is a link function. Our goal is to compute the expectation of \( y_i \) given \( X_i \) after integrating out the latent variables:

\[
E(y_i|X_i) = E(E(y_i|\eta_i)|X_i) = E(g^{-1}(\eta_i^T \omega)|X_i) = \int g^{-1}(\eta_i^T \omega)p(\eta_i|X_i) d\eta_i.
\]

Endowing the covariates with a Gaussian factor model, i.e. \( X_i = \Lambda \eta_i + \epsilon_i \), we have that \( p(\eta_i|X_i) \) is pdf of a normal distribution with mean \( AX_i \) and variance \( V \) (see Proposition (7)). In this case, the above integral can be solved when \( g^{-1}(\cdot) \) is the identity function, as in linear regression, or the exponential function, as in regression for count data or survival analysis.

On the contrary, when we are dealing with a binary regression and \( g^{-1}(\cdot) \) is equal to the logit, the above integral does not have an analytical solution. However, letting \( \mu_i = p_i \), the probability of success, we can integrate out the latent variables and compute the expectation of the log-odds conditional on \( X_i \):

\[
E \left( \log \left( \frac{p_i}{1 - p_i} \right) | X_i \right) = E(\omega^T \eta_i|X_i) = \omega^T AX_i
\]

This expectation can be easily generalized to pairwise and higher-order interactions following Ferrari and Dunson (2020). More generally, we can compute the expectation of the natural parameter conditional on \( X_i \) for any distribution in the exponential family.
4.5 Genomics application

In this section we analyze gene expression data to predict time until death. In order to conduct our analysis, we use data from the curatedOvarianData R package, a comprehensive resource for gene expression and clinical data from ovarian cancer patients (Ganzfried et al., 2013). Data from this package has been extensively used for various survival applications. For example, Zhang et al. (2018) use patient data in order to identify single-gene prognostic signatures for advanced ovarian cancer. Similarly, Yang et al. (2017) utilize LASSO penalized regression to determine non-coding RNAs (IncRNAs) associated with the recurrence of ovarian cancer. Similar to our approach, Avalos-Pacheco et al. (2021) utilize clinical and genomic ovarian cancer data within a sparse latent factor regression model for survival outcome prediction. While their method performs dimensionality reduction and supervised regression separately, our approach does so jointly.

Our analysis specifically uses the Illumina Human microRNA expression dataset E.MTAB.386, which is provided by the curatedOvarianData package and has Angiogenic mRNA and microRNA gene expression data for 129 patients. To facilitate our analysis, we kept only the 10% of covariates that exhibited the highest variance. Hence, the final dimensionality of matrix $X$ is $129 \times 1036$. No gene expressions or times-to-event were missing or censored. We also standardized the covariates in $X$ prior to the analysis. We included in the analysis age at initial pathological diagnosis and tumor stage. Given the high-dimensionality of the data, we included only main effects in the analysis, as done in Avalos-Pacheco et al. (2021). We use the priors specified in Section 4.3 and set $k = 5$, we run the Gibbs sampler for 11,000 iterations with a burn-in of 1,000.

We observed a concordance index equal to 0.58, a result is similar to that of Avalos-Pacheco et al. (2021). The key difference in our approach was that we em-
Figure 4.1: ROC curves fixed at different time points during the study.

ployed only a portion of the observations and jointly estimated the factor and survival models. Figure (4.1) shows various ROC curves fixed at different points in time during the study, as well as the corresponding areas under curve (AUC). An alternative to the standard ROC curve, this time-dependent approach calculates the ROC curve for cumulative death incidence by time \( t \). Our model exhibits the best performance at 1080 days (approximately 3 years into the study) with an AUC of 0.6, while becoming less reliable past this threshold. We also estimated the hazard function for each unit, and divided the observations into high-risk and low-risk, based on whether their estimated hazard was higher or lower than the median one. Figure (4.2) shows the survival functions for the two groups, demonstrating that patients in the low-risk category have higher survival.

Figure (4.3) shows the posterior mean of the matrix of factor loadings \( \Lambda \) after applying the MatchAlign algorithm of Chapter 6, which resolves rotational ambiguity and column label switching for the posterior samples of \( \Lambda \). We used the function
jointRot contained in the R package \texttt{infinitefactor}. We rearranged the rows of \( \Lambda \) to highlight the five estimated clusters.

4.6 Application with Binary Data

We analyze four microarray cancer datasets from the R package \texttt{datamicroarray}. The specific cancers analyzed are breast (Gravier et al., 2010), CNS (Pomeroy et al., 2002), colon (Alon et al., 1999), and (Singh et al., 2002). Table 4.1 provides an overview of the four microarray data sources. We compare our method with hierarchical lasso (Bien et al., 2013), RAMP (Hao et al., 2018) and FAMILY (Haris et al., 2016).

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|c|}
\hline
Dataset & n & p & Labels & Positive/Negative \\
\hline
Breast & 168 & 2905 & Good/Poor & 111/57 \\
CNS & 60 & 7129 & Survival/Failure & 21/39 \\
Colon & 62 & 2000 & Normal/Tumor & 22/40 \\
Prostate & 102 & 12,600 & Tumor/Not Tumor & 52/50 \\
\hline
\end{tabular}
\caption{Summary of Microarray data}
\end{table}

No missingness was observed in the data. After standardizing \( X \), we run a uni-
variate logistic regression with every predictor singularly to identify those with significance levels below a certain threshold, $\alpha = 0.01$. We proceed with only these predictors to facilitate our analysis and include interactions among the predictors in every model. We then run every procedure 50 times, each time using a different randomized split of training data (75%) and testing data (25%). We use the priors specified in Section 4.3 and set $k = 5$, we run the Gibbs sampler for 11’000 iterations with a burn-in of 1’000. Figure 4.4 shows boxplot distributions of the accuracies obtained on each of the microarray datasets. Our model shows a good performance across the different datasets, with median accuracies of 0.8, 0.8, 0.87, and 0.93 respectively.

4.7 Discussion

We proposed a factor analysis modeling approach in which latent factors underlying predictors are used in a quadratic proportional hazards regression model. This novel approach is particularly useful for survival outcome prediction with high-dimensional and highly correlated predictors. It further allows for the estimation of interaction effects in high-dimensional settings, a task that can be critical in understanding and developing effective treatments for disease. We test our approach on mRNA and microRNA gene expression data from 129 ovarian cancer patients and observe a concordance index of 0.58, a value comparable to that of Avalos-Pacheco et al. (2021), despite using a single batch of observations. In addition to the main survival application, we showed how our results can be easily extended to regressions with count data and exponential families. We also provide code for the count data case and logistic regression.

As the partial likelihood given by the Cox proportional hazards model includes a summation over the observations, the update of latent factors with partial likelihood is computationally inefficient. Despite this limitation, our fast C++ implementation
was able to considerably speed up the MCMC computations. A possible improvement is related to the implementation of EM algorithm for factor models as in Avalos-Pacheco et al. (2021) or Variational Inference approximations (Yoshida and West, 2010).

In order to allow departures from linearity and Gaussianity, it is of interest to model the regression as a non-linear function of latent factors. A promising direction to decrease modeling assumptions is to rely on a copula model related to Murray et al. (2013). This would allow to separately model the marginal distribution of covariates and the dependence structure using a Gaussian copula factor model and then regress the latent factors on the survival outcome.

To improve our analysis, it would certainly help to include more data points. In order to do so, we would have to account for batch effects that are common to medical data. In such cases, data coming from different laboratories or projects, or generated under different experimental conditions cannot simply be aggregated together. Failing to account for batch effects can lead to unreliable conclusions distortions in mean and variance. By adopting the approach of Avalos-Pacheco et al. (2021), we can integrate other useful gene expression data provided by the curatedOvarianData package, thereby increasing the number of observations.
Algorithm 4 MCMC algorithm for sampling the parameters of model (4.1)

Step 1 Sample \( \omega, \Omega \) and \( \eta_i \), for \( i = 1, \ldots, n \), using MALA with appropriate proposal distributions.

Step 2 Denote \( \lambda_j \) the rows of \( \Lambda \), for \( j = 1, \ldots, p \). Sample \( p \) conditionally independent posteriors:

\[
\pi(\lambda_j | \cdot) \sim N \left( (D_j^{-1} + \frac{\eta^T \eta}{\sigma_j^2})^{-1} \eta^T \sigma_j^{-2} X^{(j)}, (D_j^{-1} + \frac{\eta^T \eta}{\sigma_j^2})^{-1} \right)
\]

where \( X^{(j)} \) is the \( j \)th column of the matrix \( X \), \( D_j = \text{diag}(\tau_j^2 \psi_{j1}, \phi_{j1}, \ldots, \tau_j^2 \psi_{jk}, \phi_{jk}) \).

Step 3 Sample \( \psi_{jh} \) for \( j = 1, \ldots, p \) and \( h = 1 \ldots, k \) from independent Inverse Gaussian distributions:

\[
\pi(\psi_{jh} | \cdot) \sim \text{InvGauss} \left( \tau_j \phi_{jh}, 1 \right)
\]

Step 4 Sample \( \tau_j \) for \( j = 1, \ldots, p \) from independent Generalized Inverse Gaussian distributions:

\[
\pi(\tau_j | \cdot) \sim \text{GInvGauss} \left( 1 - k, 1, 2 \sum_{h=1}^{k} \frac{|\lambda_j h|}{\phi_{jh}} \right)
\]

Step 5 In order to update \( \phi_{jh} \) for \( j = 1, \ldots, p \) and \( h = 1 \ldots, k \), sample \( T_{jh} \) from independent Generalized Inverse Gaussian distributions:

\[
\pi(T_{jh} | \cdot) \sim \text{GInvGauss} \left( a - 1, 1, 2 |\lambda_{jh}| \right)
\]

Then set \( \phi_{jh} = \frac{T_{jh}}{\sum_{k=1}^{k} T_{jh}} \)

Step 6 Sample \( \sigma_j^{-2} \) for \( j = 1, \ldots, p \) from conditionally independent gamma distributions

\[
\pi(\sigma_j^{-2} | \cdot) \sim \text{Gamma} \left( \frac{1 + n}{2}, \frac{1}{2} + \frac{1}{2} \sum_{i=1}^{n} (X_{ij} - \lambda_j^T \eta_i) \right)
\]
Figure 4.3: Matrix of factor loadings for the ovarian cancer dataset after postprocessing with the MatchAlign algorithm.
Figure 4.4: Out of sample accuracy in 4 datasets (breast, cns, colon, prostate for GLM factor, hiernet (Strong and Weak heredity), RAMP (Strong and Weak heredity) and Family (Weak heredity).
5.1 Introduction

Recent developments in the collection of untargeted assays enable the concurrent
detection of thousands of exposures and have become a powerful tool for character-
izing exogenous exposures and their downstream molecular responses (Niedzwiecki
et al., 2019). These exposures are often non-normally distributed, right-skewed, and
have a significant percentage of measurements below the limit of detection (LOD)
and of missing data, as shown by Figure 5.1. Characterizing the distribution of the
chemicals measurements and coherently allowing for sampling of missing data and
observations below the LOD is of crucial important when building a new methodol-
ogy, as models that do not take into account these attributes can lead to misleading
results.

Within the Bayesian paradigm, several methods can be adapted to sample the
observations below the LOD and missing data by adding an imputation step to the
MCMC algorithm. This can done by endowing chemical measurements with a model,
which usually results in making the posterior of the outcome model parameters non-conjugate. A solution to the non-conjugacy of the posterior distribution is the so-called “cut of feedback” (Lunn et al., 2009), which forces the outcome model parameters to be independent of the parameters related to missing data imputation. Although it has computational advantages, this procedure does not assure that the sampling distribution converges to the posterior distribution of the model.

The goal of this chapter is to provide a transformation of the chemical measurements that coherently allows to sample data points that are missing or measured below the limit of detection (LOD), while flexibly and efficiently estimating dose-response curves. We propose to use a Gaussian copula model that separates the dependence structure among exposures and the marginal distribution of the chemical measurements. The Gaussian copula model admits a latent representation of the measurements, where the transformed chemicals have a normal distribution. As exposures can be moderately high-dimensional with high correlations within blocks of variables, we model the dependence among the transformed chemicals using factor models (Murray et al., 2013). Then, we extend the method to regression settings by linking the latent variables and the outcome with a flexible model using Bayesian splines (DiMatteo et al., 2001). Using a one-to-one transformation via the latent representation of the Gaussian copula model, we are able to perform inference in the original scale of the ‘exposures’.

5.2 Model

Let \( y_i \) denote a continuous health response and let \( X_i = (X_{i1}, \ldots, X_{ip}) \) be a \( p \times 1 \) vector of chemical measurements for individual \( i \), for \( i = 1, \ldots, n \). We propose to separately model the marginal distribution of the chemical measurements and their dependence structure using a Gaussian copula factor model (Murray et al., 2013) and then regress the latent factors on the health outcome using Bayesian B-splines
(DiMatteo et al., 2001). We model the joint distribution of $X_i$ as

$$F(X_{i1}, \ldots, X_{ip}) = C(F_1(X_{i1}), \ldots, F_p(X_{ip})), \quad (5.1)$$

where $F_j$ is the univariate marginal distribution of chemical $j$ and $C(\cdot)$ is a distribution function on $[0,1]^p$ that encapsulates the dependence between covariates. Any joint distribution $F$ can be completely specified by its marginal distributions and a copula $C$ (Sklar, 1959), with the copula being uniquely determined when the covariates are continuous.

For our method we first employ the Gaussian copula, although we extend our method to the t-copula in Section 5.2.1. The Gaussian copula is defined as:

$$C(u_1, \ldots, u_p) = \Phi_p(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p) | \Sigma), \quad (5.2)$$

where $\Phi_p(\cdot | \Sigma)$ is the p-dimensional Gaussian cdf with correlation matrix $\Sigma$, $\Phi(\cdot)$ is the univariate standard Gaussian cdf and $(u_1, \ldots, u_p) \in [0,1]^p$. Combining equations (5.1) and (5.2), the implied joint distribution of $X_i$ is

$$F(X_{i1}, \ldots, X_{ip}) = \Phi_p(\Phi^{-1}(F_1(X_{i1})), \ldots, \Phi^{-1}(F_p(X_{ip}))). \quad (5.3)$$
The Gaussian distribution is used to model the dependence structure, whereas the chemicals have univariate marginal distributions $F_j(\cdot)$, for $j = 1, \ldots, p$. Model (5.3) admits the following latent variable representation:

$$X_{ij} = F_j^{-1}\left(\Phi\left(\frac{Z_{ij}}{c}\right)\right) \quad z_i \sim N_p(0, \Sigma),$$

where $F_j^{-1}(\cdot)$ is the pseudo-inverse of the univariate marginal of chemical $j$, $Z_{ij}$ is the latent variable related to chemical $j$ and observation $i$, and $c$ is a positive normalizing constant that depends on $\Sigma$. The high correlation between chemicals will be reflected in the latent space with the correlation matrix $\Sigma$. For this reason, we follow the approach of (Murray et al., 2013) and endow $Z_i = (Z_{i1}, \ldots, Z_{ip})$ with a latent factor model:

$$Z_i | \eta_i \sim N(\Lambda \eta_i, \Psi) \quad \eta_i \sim N(0, I),$$

where $\Lambda$ is a $p \times k$ factor loadings matrix, $\eta_i$ are $k \times 1$ latent factors and $\Psi$ is a $p \times p$ diagonal matrix of residual variances. An example of the chemicals in the transformed space can be seen in Figure 5.2.

We now consider a regression on a continuous health outcome $y_i$ via the latent...
variables:

\[ y_i = f(\eta_i) + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2). \]

When \( f(\cdot) \) is linear, i.e. \( f(\eta_i) = \omega^T \eta_i \), it is not difficult to show that the induced regression is linear in the space of \( Z_i \):

\[
\mathbb{E}(y_i|X_i) = \mathbb{E}(\omega^T \eta_i|X_i) = \omega^T \mathbb{E}(\eta_i|X_i) = \omega^T \mathbb{E}(\eta_i|Z_i) = \omega^T ((\Lambda^T \Lambda + I_k)^{-1} \Lambda^T Z_i | X_i) = \omega^T ((\Lambda^T \Lambda + I_k)^{-1} \Lambda^T \mathbb{E}(Z_i|X_i),
\]

where \( \mathbb{E}(Z_i|X_i) \) is a \( p \times 1 \) vector such that the \( j^{th} \) element is equal to \( c\Phi^{-1}(F_j(X_{ij})) \). This follows from the fact that the distribution of \( \eta_i|Z_i \) is Normal with covariance \( V = (\Lambda^T \Lambda + I)^{-1} \) and mean \( AZ_i \), where \( A = V \Lambda^T = (\Lambda^T \Lambda + I)^{-1} \Lambda^T \).

The linearity assumption is often too restrictive, so one possibility is to model the continuous outcome with a nonparametric function of the latent variables. This would however create several computational challenges, as many flexible models make the posterior distribution of the latent variables non-conjugate and challenging to sample. For this reason, we model \( f(\cdot) \) using Bayesian B-splines (DiMatteo et al., 2001) of degree \( d \):

\[
 f(\eta_i) = \sum_{i=1}^{k} \sum_{h=1}^{H+2} \alpha_i b_h(\eta_i),
\]

where \( b_h(\cdot) \) denotes the \( h^{th} \) function in a B-spline basis of degree \( d \) with natural boundary constraints. Let \( \xi = (\xi_1, \ldots, \xi_H) \) be the boundary knots, we set \( b_1(\cdot) \) and \( b_{H+2}(\cdot) \) to be linear functions in the intervals \([-\infty, \xi_1]\) and \([\xi_H, +\infty]\), respectively. With this specification we can estimate the induced regression in the space of the chemicals measurements by using the results of Proposition (1) and the extension to higher order interactions of Section 3.2.3. Without loss of generality, in the remainder we assume cubic splines \((d = 3)\).
5.2.1 \textit{t-copula}

Due to the exponential behaviour of the tails, the Gaussian copula does not allow for extreme values (or tail) dependence. In this section we provide an extension of (5.3) to tackle tail dependence. In particular, we employ the t-copula (Demarta and McNeil, 2005), which allows to model dependent extreme values. This is of particular relevance in our motivating application since whenever a chemical has a high measurement value, we also expect other correlated chemicals to have high values. We can modify (5.2) using the t-distribution:

\begin{equation}
F(X_{i1}, \ldots, X_{ip}) = T_{p,\nu,C}(T_{\nu}^{-1}(F_1(X_{i1})), \ldots, T_{\nu}^{-1}(F_p(X_{ip})))
\end{equation}

Where $T_{p,\nu,C}$ is the multivariate cdf of a t-distribution with $\nu$ degrees of freedom and correlation $C$ and $T_{\nu}$ is the univariate cdf of a standard t-distribution with $\nu$ degrees of freedom. Following the same steps as before, we have that the random vector $(X_1, \ldots, X_p)$ has copula t-distribution:

\begin{equation}
\eta_i \sim N(0, I) \quad Z_i | \eta_i, \phi \sim N(\Lambda \eta_i, \phi^{-1} I_p) \quad \phi \sim \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu}{2}\right)
\end{equation}

\begin{equation}
X_{ij} = F_j^{-1}\left(T_{\nu}\left(\frac{Z_{ij}}{c}\right)\right).
\end{equation}

5.3 Environmental Epidemiology Application

The goal of our analysis is to assess the association of 15 metals (Barium, Cadmium, Cobalt, Caesium, Molybdenum, Manganese, Mercury, Lead, Antimony, Tin, Selenium, Strontium, Thallium, Tungsten and Uranium) with body mass index (BMI). Similarly as in Section 2.5 we use data from the National Health and Nutrition Examination Survey (NHANES). Refer to Section 2.5 for a general description of the dataset and literature review on the association between BMI and metals. Differently from Section 2.5, we use data from the year 2016.
We set the parameters of the Bayesian B-splines to $d = 3$ and $H = 5$. Regarding the factor model, we endow the parameters with the priors specified in Section 3.2.2, using the Dirichlet-Laplace (Bhattacharya and Dunson, 2011) prior for the factor loadings row-wise. We set $k = 9$ as in Section 2.5. The computational burden of model (5.1) is sampling the parameters of $F_j$ whenever choosing a parametric model for it. However, we replace $F_j(\cdot)$ by the scaled empirical marginal cdf $\hat{F}_j(t) = \frac{n}{n+1} \sum_{i=1}^n \frac{1}{n} \mathbb{1}(X_{ij} \leq t)$, which is a common solution in practice. Refer to Klaassen et al. (1997) for a theoretical analysis of using the empirical marginal cdfs with copula models. For this analysis, we use the t-copula (5.6). We endow $\alpha_{lh}$ with standard Normal priors, for $l = 1, \ldots, 9$ and $h = 1, \ldots, 5$. To estimate the parameters, we carefully modify Algorithm 3 and run an MCMC procedure for 11,000 iterations with a burn-in of 1,000 iterations.

Figure 5.3 and 5.4 show the estimated regression surfaces of Selenium, Cadmium, Copper and Mercury on BMI when all the other chemical measurements are fixed at the median. Notice that the regression surfaces are shown in the original scale of the data. The bars in the x-axis shows the values of the data points and as expected the credible intervals are wider where less data points are observed. The estimated regression surfaces follow the common patterns in chemical exposure studies. In particular, the estimated surface for Selenium is a typical dose-response curve and Cadmium has a hill-shaped dose response. Finally, Cadmium has a similar association to BMI as that shown in Figure 2.6.
Figure 5.3: Estimated regression surface of Selenium on BMI using data from NHANES 2016 when all the other chemical measurements are fixed at the median.

Figure 5.4: Estimated regression surfaces of Cadmium, Copper and Mercury on BMI using data from NHANES 2016 when all the other chemical measurements are fixed at the median.
6

MatchAlign Algorithm

6.1 Introduction

Factor models are commonly used to characterize the dependence structure in correlated variables while providing dimensionality reduction. With continuous observations, this is usually achieved by assuming that the observed variables are linear combinations of a set of lower dimensional latent variables. Let \( X_i = (x_{i1}, \ldots, x_{ip})^T \) be a \( p \times 1 \) vector of correlated variables. Without loss of generality, we assume that the data have been centered prior to analysis so that we can omit the intercept. The typical Gaussian factor model has the following representation:

\[
X_i = \Lambda \eta_i + \epsilon_i, \quad \epsilon_i \sim N(0, \Sigma), \quad \eta_i \sim N_k(0, I),
\]

(6.1)

where \( \Lambda \) is a \( p \times k \) tall and skinny matrix of factor loadings, \( \eta_i \) is a \( k \times 1 \) vector of latent factors and \( \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2) \) is a diagonal matrix containing the residual variances. A typical choice for the latent dimension is to set \( k \ll p \) in order to provide dimensionality reduction. The factor model specification induces a decomposition of
the $p \times p$ covariance matrix of $X_i$:

$$\text{cov}(X_i) = \Lambda \Lambda^T + \Sigma. \quad (6.2)$$

This decomposition implicitly assumes that the correlation among variables is fully explained by the latent variables via the term $\Lambda \Lambda^T$.

It is well known that the covariance decomposition in (6.2) is not unique, and as a result $\Lambda$ is non-identifiable. For example, consider a $k \times k$ semi-orthogonal matrix $P \ (PP^T = I)$, then $\Lambda' = \Lambda P$ also satisfies the above equation. We refer to this as rotational ambiguity throughout the paper. Although the estimation of some parameters, such as the covariance, is not affected by this (Bhattacharya and Dunson, 2011), rotational ambiguity is still of crucial importance as it prevents inference on $\Lambda$ and interpretation of the induced groupings of variables from (6.1).

Non-identifiability of the factor loadings matrix is a well studied problem in the frequentist literature, where non-identifiability causes multimodality in the objective function. However, from a statistical perspective, each mode is equivalently optimal (Lawley and Maxwell, 1962) and it suffices to choose a single mode to perform inference. A possibility to select one mode is by applying an orthogonalization procedure, such as Varimax (Kaiser, 1958), Quartimax (Neuhaus and Wrigley, 1954), Promin (Lorenzo-Seva, 1999), or the Orthogonal Procrustes algorithm in Aßmann et al. (2016). However, in the Bayesian paradigm, multimodality in the posterior distribution of the factor loadings matrix allows the sampling of $\Lambda$ to switch between modes during the MCMC procedure, effectively preventing convergence to a single mode.

One possibility to attain identifiability of $\Lambda$ in Bayesian models is by imposing significant constraints (Ghosh and Dunson, 2009), (Erosheva and Curtis, 2011). A typical choice to enforce identifiability requires setting the upper diagonal elements of $\Lambda$ equal to zero and requiring the diagonal elements to be positive (Geweke and Zhou,
A more general solution was provided by Frühwirth-Schnatter and Lopes (2018) using generalized lower triangular (GLT) matrices. This structure has been used routinely in numerous settings; see, for example, Lucas et al. (2006), in order to perform inference on $\Lambda$, at the cost of introducing order dependence among the factors (Carvalho et al., 2008). Chen et al. (2020) showed how structural information on the matrix of factor loadings affects the identifiability and the estimation of $\Lambda$. However, Millsap (2001) showed that fixing different elements of $\Lambda$ to zero in different runs of the algorithm can still lead to a convergence failure, or even to a factor solution that is no longer in the same equivalence class. Similarly, Erosheva and Curtis (2011) showed that requiring loadings to be positive may result in nontrivial multimodality of the likelihood, and as a result chains with different starting values may produce solutions that are substantially different in fit. Finally, this structure on $\Lambda$ also constrains the class of estimable covariance matrices by forcing some entries of the matrix $\Lambda$ to be equal to zero.

To address rotational ambiguity, we propose to use a post-processing algorithm that allows for identification and interpretation of $\Lambda$. Crucially, this algorithm does not affect the choice of priors and structure for the matrix of factor loadings, which is a modeling choice left to the practitioner. Several post-processing algorithms have been proposed to deal with rotational ambiguity in specific settings. McAlinn et al. (2018) adapts an optimization procedure for posterior mode finding for Bayesian dynamic factor analysis in macroeconomic applications. Kaufmann and Schumacher (2017) propose a post-processing clustering procedure in order to obtain an identified posterior sample and Kaufmann and Schumacher (2019) use an empirical procedure in order to identify a mode posterior by exploiting correlation among factors. Erosheva and Curtis (2011) address sign switching across the samples of $\Lambda$ with an approach based on Stephens (2000).

We address the post-processing task with a more general approach related to Pa-
pastamoulis and Ntzoufras (2020) and Marin et al. (2005). We divide the algorithm into an orthogonalization step and a sign permutation step. We first solve orthogonal ambiguity by performing Varimax (Kaiser, 1958), as in Papastamoulis and Ntzoufras (2020), though our algorithm can be adapted to any orthogonalization procedure. The rotational ambiguity between samples of $\Lambda$ is then limited to switching in the column labels and column signs, which we will refer to as label switching and sign switching, respectively. We propose to solve both these problems by matching each posterior sample to a reference matrix, and using the matches to align the samples. The aligned samples can then be directly used for inference. By simplifying the alignment step and developing a greedy maximization procedure, we significantly improve the computational efficiency with respect to Papastamoulis and Ntzoufras (2020) and Marin et al. (2005), while maintaining good estimation performance. Also, our method is not constrained to factor models with a latent dimension less than 50, but can be used with a latent dimension in the order of hundreds. We focus our analysis on the latent factor model, but our algorithm can be applied to a much larger class of models involving rotational ambiguity in matrix valued parameters.

We define the identifiability setting in Section 6.2 and the MatchAlign algorithm in Section 6.3. We provide extensive simulations in Section 6.4 and analyze the performance of our post-processing algorithm compared to that of several alternative methods. In Section 6.5, we apply our algorithm to genomics. Our algorithm is implemented in the \texttt{infinitefactor R} package on CRAN.

6.2 Identifiability Setting

Let us consider the Gaussian Bayesian factor model introduced in (6.1). This model allows for dimensionality reduction in characterizing the covariance of $X_i$ through
the decomposition
\[ \text{cov}(X_i) = \Lambda \Lambda^T + \Sigma. \]

In order to tackle identifiability of \( \Lambda \), we firstly need identification between \( \Lambda \Lambda^T \) and the residual variance \( \Sigma \). The identification of the residual variance is often overlooked in the factor models literature. A one-factor model is identifiable only if at least 3 factor loadings are nonzero (Anderson et al., 1956), (Frühwirth-Schnatter and Lopes, 2018). For the remainder of the paper we will assume that \( k \leq \frac{p-1}{2} \), which ensures that \( \Sigma \) is identifiable when all rows of \( \Lambda \) are nonzero. As a result, this condition guarantees identification of \( \Lambda \Lambda^T \). See Frühwirth-Schnatter and Lopes (2018) and Papastamoulis and Ntzoufras (2020) for a comprehensive analysis of the topic and for detailed conditions to ensure identification of \( \Sigma \).

We now focus on uniqueness of the factor loading matrix \( \Lambda \). As noted in the Introduction, the covariance decomposition is not unique, and as a result \( \Lambda \) is non-identifiable. Consider a \( k \times k \) semi-orthogonal matrix \( P \) \((PP^T = I)\), then \( \Lambda' = \Lambda P \) also satisfies (6.2). We refer to this as rotational ambiguity. Rotational ambiguity is a well studied problem in the frequentist literature and can be solved with orthogonalization procedures such as Varimax (Kaiser, 1958). In the Bayesian paradigm, there will typically be rotational drift as posterior samples are collected via an MCMC algorithms. Such changes in rotation need to be adjusted for before calculating posterior summaries of \( \Lambda \); this can be accomplished through a post processing approach to rotationally align the samples or via imposing restrictions on \( \Lambda \) \textit{a priori} to avoid the rotational drift entirely.

Under the latter approach, a typical choice is to restrict \( \Lambda \) so that the upper triangular elements are zero and the diagonal elements are positive (Geweke and Zhou, 1996). This restriction has been used routinely in numerous settings; see, for example, Lucas et al. (2006), in order to perform inference on \( \Lambda \), at the cost of intro-
ducing order dependence among the factors (Carvalho et al., 2008). A more general solution was provided by Frühwirth-Schnatter and Lopes (2018) using generalized lower triangular (GLT) matrices. To address rotational ambiguity, we instead follow the former approach and propose a post-processing algorithm.

6.3 Algorithm

In this section we describe our MatchAlign algorithm that solves rotational ambiguity in the posterior samples of non identifiable matrix value parameters. The MatchAlign procedure is summarized in Algorithm 5. The three key steps are: 1) apply an orthogonalization procedure to each posterior sample of Λ, 2) choose a reference matrix (the pivot) and 3) match the columns of each posterior sample to the pivot’s columns. Figure 6.1 shows an example of the output of our algorithm.

Notably the computation in the algorithm is split into two for loops that can each be massively parallelized or distributed. The orthogonalization step can be completed for each sample completely independently with no shared memory access in a distributed environment. It is currently implemented with optional parallel computing using the parallel package in R. The for loop to align the samples can be easily parallelized by assigning groups of samples to several machines, as long as each process has access to the pivot. Next, we describe the three main steps in the algorithm.

Let \{Λ^{(t)}, \ t = 1, \ldots, T\} be the posterior samples of the factor loading matrix Λ. Our first goal is to tackle generic rotational invariance across Λ^{(t)}. In order to achieve that, we apply an orthogonalization procedure to each sample. Orthogonalization of factor loadings highlights the grouping of covariates, often inducing sparsity row-wise in Λ, which allows to represent the samples in an interpretable form. Figure 6.2 shows an example of the application of Varimax to a posterior sample of Λ. We define as \tilde{Λ} the factor loading matrix after applying Varimax. In our approach we focus
Algorithm 5 MatchAlign algorithm to solve rotational ambiguity in matrix valued parameters.

**Input** \{\Lambda^{(t)} : t = 1, \ldots, T\}

1. **for** \( t \in 1 : T \) **do**
   1. Orthogonalize \( \Lambda^{(t)} \) using Varimax and output \( \tilde{\Lambda}^{(t)} \)
   **end**
2. Choose a pivot \( \Lambda^P \) from \{\tilde{\Lambda}^{(t)} : t = 1, \ldots, T\}
3. **for** \( t \in 1 : T \) **do**
   1. **for** \( j \in 1 : k \) **do**
      1. Compute normed differences between \( c_j^{(t)} \) and \( \Lambda^P \) and \( -\Lambda^P \) columns
      2. Retain the \( j \)th column having minimum norm value
      3. Drop the matched column and its negative from the pivot
   **end**
   1. Reorder and re-sign
**end**

on orthogonal rotations, and apply Varimax (Kaiser, 1958) to each \( \Lambda^{(t)} \). Oblique rotations (Thurstone, 1947) can also provide accurate representations of correlated factors, or more generally linear association in the columns of the unidentifiable matrix. However, the rotated matrices would not necessarily be comparable across samples.

After the orthogonalization step, the samples of \( \Lambda \) still exhibit symptoms of non-identifiability due to sign and permutation ambiguity in the columns (Conti et al., 2014). Following Papastamoulis and Ntzoufras (2020), we define \( Q \) as a \( k \times k \) permutation matrix. Each row and column of \( Q \) has a single non-zero element, which is equal to 1. We also define \( S = \text{diag}(s_1, \ldots, s_k) \), with \( s_j \in \{-1, 1\} \), and \( P = SQ \), a signed permutation matrix. Our goal is to find the signed permutation matrices in order to align the samples, and to do that we minimize the following loss function:

\[
\min_{Q^{(t)}_\cdot, S^{(t)}_\cdot, t = 1, \ldots, T} \sum_{t=1}^{T} ||\tilde{\Lambda}^{(t)}Q^{(t)}S^{(t)} - \Lambda^P||_F
\]

where \( || \cdot ||_F \) is the Frobenious norm, \( c_j^{(t)} \) is the \( j \)th column of \( \tilde{\Lambda}^{(t)} \), and \( \Lambda^P = [c_1^P \cdots c_k^P] \) is an exemplar factor loading matrix, which we call the pivot. We use the pivot to
align the posterior samples, as in Marin et al. (2005). The pivot matrix is used in the algorithm as a proxy of a ‘true’ $\Lambda$, and we will provide the details of its choice later in this section. The loss function penalizes differences between the pivot and the posterior samples, and to minimize it we must find the signed rearrangement of $c^{(t)}_1, \ldots, c^{(t)}_k$ that best matches the pivot’s columns.

One possibility to minimize (6.3) is to compute the loss for all the possible signed permutations, which is computationally infeasible. Instead, we implement a greedy procedure and try to minimize the loss iteratively, i.e. column by column, as in Marin et al. (2005). We compute the $L_2$ normed differences between the columns of $\tilde{\Lambda}^{(t)}$ and the ones of $\Lambda^P$ and $-\Lambda^P$. We use the $L_2$ norm as it is also employed in the Varimax algorithm. Other rotation schemes such as quartimax (Neuhaus and Wrigley, 1954) work better with the absolute norm or the sup norm. Our empirical simulations showed that MatchAlign is not sensitive to the choice of norm. The computations are carried in a greedy fashion: we compute the normed differences between the
column of $\tilde{\Lambda}^{(t)}$ with largest norm and $c^P_1, -c^P_1, \ldots, c^P_k, -c^P_k$. After matching with $c^P_j$ or $-c^P_j$, for some $j = 1, \ldots, k$ that minimizes the $L_2$ norm, we match the next column of $\tilde{\Lambda}^{(t)}$ with $c^P_1, \ldots, c^P_{j-1}, -c^P_{j-1}, c^P_{j+1}, \ldots, -c^P_k$ and proceed iteratively. In this way, we only need to compute $k(2k+1)$ normed differences for each posterior sample.

Whenever using a prior for $\Lambda$ that introduces increasing shrinkage as the column order increases, as in Bhattacharya and Dunson (2011) or Legramanti et al. (2019), we can naturally start matching the columns from $c^P_1$ and proceed in column order.

Due to sampling noise one $c^{(t)}_j$ may be minimally distant from more than one $c^P_h$. Because of that, we need to avoid matching multiple $c^{(t)}_j$s to one $c^P_h$, which we refer to as duplication. Alignment that involves duplicating some columns over others destroys information and prevents the accurate reconstruction of identifiable parameters such as the covariance after post-processing. Allowing duplication of columns would also introduce numerical instability when performing operations for posterior samples with duplicate columns, as well as biasing ergodic summaries of
the parameters.

Several methods for factor models take an “over-fitted” approach (Bhattacharya and Dunson, 2011), (Rousseau and Mengersen, 2011), and choose $k$ to correspond to an upper bound on the number of factors. In this scenario, we might have multiple columns centered near $0_k$. For this reason, our algorithm may not detect when these orthogonalized columns switch labels or signs between samples. However, this does not significant bias the ergodic summaries of the factor loadings. In fact, columns might be mislabelled only if their normed difference is small. This is an important consideration because the matching method does not globally minimize differences between the reference and sample matrices. Instead, it operates on each column iteratively.

In our algorithm, every column of $\Lambda^{(t)}$ is matched to either $c_j^P$ or $-c_j^P$, for some $j = 1, \ldots, k$. In order to correctly and efficiently match columns, we must choose a pivot that is central in the distribution of a column statistic. As a proxy of unique information contained in each column, we consider the condition number:

$$\kappa^{(t)} = \kappa(\Lambda^{(t)}) = \frac{\sigma_{\text{max}}(\Lambda^{(t)})}{\sigma_{\text{min}}(\Lambda^{(t)})}$$

where $\sigma_{\text{max}}(\Lambda^{(t)})$ and $\sigma_{\text{min}}(\Lambda^{(t)})$ are the largest and smallest singular values of $\Lambda^{(t)}$, respectively. An example of the distribution of $\kappa^{(1)}, \ldots, \kappa^{(T)}$ can be seen in Figure 6.3. We choose as a pivot the matrix with the median condition number. Note that this number may approach infinity with an overspecified number of columns. In this case, we can use the largest singular value in place of the condition number. We find in practice that these two choices provide a similar performance of MatchAlign. One possibility to run the algorithm without selecting a pivot is by employing the solution of Papastamoulis and Ntzoufras (2020) based on the work of Stephens (2000) in the context of mixture models. However, this approach involves a significant increase in the computational cost.
Notice that identifiable parameters such as $ΛΛ^T$ are not affected by applying MatchAlign, since we process the samples of $Λ$ by post-multiplying with a semi-orthogonal matrix. Finally, even after post-processing, it is infeasible to recover the ‘true’ $Λ$ that generated the data as $Λ$ matrices with columns that underwent sign permutations are equivalent from a statistical point of view. However, if the pivot is the ‘true’ $Λ$ used in the data generating process, then unsurprisingly the posterior mean of aligned lambda corresponds to the true $Λ$.

6.4 Simulations

In this section we compare the performance of the MatchAlign algorithm with the Varimax-RSP algorithm of Papastamoulis and Ntzoufras (2020). In particular, we run their faster implementation with full simulated Annealing (rsp full-SA). We generate the data according to 6.1 by sampling each element of $Λ$ independently from a standard normal distribution and we sample the diagonal elements of $Σ$ from independent Inverse-Gammas distributions with parameters $(1/2, 1/2)$. We fit a Gaussian factor model with the Dirichlet-Laplace prior (Bhattacharya et al., 2015) with parameters $(1/2, 1/2)$ for each row of $Λ$ and an inverse-Gamma
prior on the diagonal elements of the matrix $\Sigma$. We do not enforce any identifiability constraints on the matrix $\Lambda$. We use the function `linearDL` contained in `infiniteR` package to estimate the model. We run the MCMC procedure for 11’000 iterations and a burn-in of 1’000, keeping 10’000 samples in total. We run the simulations for different values of $k$ and $p$ with $n = 500$ data points. For each set of parameters we average the running times of the two algorithms over 25 simulations.

The results are shown in Figure 6.4 with the running times being on the log$_{10}$ scale. Notice how the MatchAlign algorithm is often at least an order of magnitude faster (approximately 10 times faster) than rsp full-SA when $k = 5, 10$, and two orders of magnitude faster (approximately 100 times faster) when $k = 25, 50$.

6.5 Genomics Application

Dimensionality Reduction is of crucial importance in genomics, where data often present linearly dependent features in high dimensional spaces. Here, we analyze gene expressions data of CNS cancer patients (Pomeroy et al., 2002). We downloaded the dataset from the R package `datamicroarray`. In total, 7129 gene expressions are measured for 60 patients. No missingness is observed in the data. We estimate model (6.1) with the Dirichlet-Laplace prior (Bhattacharya and Dunson, 2011) for each row of $\Lambda$ and an inverse-Gamma prior with parameters $(1/2, 1/2)$ on the diagonal elements of the matrix $\Sigma$. We noticed from the eigendecomposition of the correlation matrix that the first 5 eigenvectors explain more than 50% of the total variability; hence, we set the number of factors equal to 5.

We run the MCMC procedure for 11’000 iterations with a burn-in of 1’000 and apply the MatchAlign algorithm to the posterior samples of $\Lambda$. Figure 4.3 shows the aligned posterior mean for the factor loadings matrix for the 7129 gene expressions. We re-ordered the rows of $\Lambda$ to highlight the covariate groupings. The time to align the samples using MatchAlign was around 65 seconds on a single thread of a 4.5GHz
Figure 6.4: Comparison of running times in log_{10} scale between MatchAlign and rsp with full-SA.

processor. The scaling with $p$ is limited to the increased computation of taking the 2-norm of the differences, but this can be efficiently parallelized using openBLAS or other multi-threaded linear algebra libraries. Comparison was not performed with Papastamoulis and Ntzoufras (2020) due to the excessive computational time required for the alignment.

6.6 Discussion

We proposed a computationally efficient post-processing algorithm that allows to solve label and sign switching in matrix valued parameters that are subject to rotational ambiguity. Our MatchAlign algorithm reduces the solution space by it-
eratively comparing columns, saving massive computational time in comparison to permutation searches. In section 6.4 we compared the computational performance of MatchAlign with respect to the algorithm of Papastamoulis and Ntzoufras (2020) and showed how MatchAlign can be applied for high-dimensional applications in section 6.5.

Crucially, this algorithm does not alter the estimation of identifiable parameters, such as $\Lambda\Lambda^T$ and allows for inference on the factor loadings matrix. The code is available in the `infinitefactor R` package on the CRAN repository and can be easily applied in several cases.
Figure 6.5: Matrix of factor loadings for the CNS cancer dataset after postprocessing with MatchAlign algorithm.
In Chapter 1, we built a flexible joint model for mixtures of chemicals with Gaussian processes, as done by Bobb et al. (2014). We modeled the joint effect of chemical exposures on human health outcomes with a nonparametric regression, with interest being focused on selection of main effects and interactions while allowing for flexible nonparametric estimation. For interpretability, we decomposed the expected health outcome into a linear main effect, pairwise interactions, and a nonlinear deviation term. We carefully used a projection approach to enforce identifiability in modeling the parametric and nonparametric components. The proposed method allows to perform variable selection of main effects, interactions, and nonlinear effects via spike and slabs priors (George and McCulloch, 1997), (Savitsky et al., 2011).

While we imposed heredity constraints and used an approximation to the Gaussian process surface in order to increase computation performance, new methods for dimensionality reduction are needed to scale up to a massive number of exposures. Recent developments in Gaussian process computations (Moran and Wheeler, 2020), (Peruzzi et al., 2020) use high-performance computing and Gibbs sampler parallel operations to scale up to millions of observations. The next challenge is to adapt
these scalable methods to environmental epidemiology, where many observations are missing or measured below the limit of detection. Adapting these techniques to the methods that we developed will allow to flexibly model massive datasets.

In Chapter 3, we used latent factor models to induce pairwise and higher order interactions among chemical exposures. Factor models are particularly attractive in this setting because they characterize the block correlation structure among correlated exposures, while providing dimensionality reduction. The proposed FIN method allows for the estimation of interaction effects in high dimensional settings, scaling up to hundreds of exposures. FIN demonstrated similar performance to state of-the-art methods for interaction estimation with independent covariates and outperformed competitors in settings with moderately to highly correlated predictors.

FIN has primarily been used as a predictive model which characterizes the correlation structure among chemicals. It is of critical importance to provide an extension that estimates the causal effect while accurately providing uncertainty quantification and dimensionality reduction. One possibility is to carefully modify the structure of the factor loadings matrix in order to estimate weighted averages of exposures, which are then regressed on the health outcome, while allowing other factors to explain the remaining variability in the chemical measurements. This will make possible the estimation of a causal effect while preserving the flexibility of the model.

Factor models are useful tools in dealing with high dimensional continuous data, but some work was needed to deal with different types of outcomes. For example, multivariate outcomes are routinely collected in applied fields such as epidemiology (Schwartz et al., 2010), (Samoilenko et al., 2018). In Chapter 3, we also proposed a Bayesian approach to inference with Structural Equation Models (SEM) to estimate a multivariate quadratic regression in high-dimensional settings with correlated covariates and outcomes. SEMs are a natural generalization of factor analysis to multivariate outcomes and have been successfully employed for multilevel data (Ansari
and Jedidi, 2001) and nonlinear structures (Lee et al., 2004). Dimension reduction in SEMs is realized by estimation of latent factors for outcomes and covariates, and dependence is induced in the latent space. This method allows practitioners to fully analyze popular datasets in environmental epidemiology, such as NHANES, rather than considering only a subpopulation of individuals with complete outcome measurements.

It is likewise very common to collect high-dimensional data and time-to-event outcomes in fields such as medicine. Cox proportional hazards models are routinely used with variable selection or regularization to deal high dimensional covariates. Such approaches do not usually account for interactions among correlated predictors, which would greatly increase dimensionality. Latent factor models have been used in this setting by Pan et al. (2019a), where the authors impose identifiability constraints for latent factors to identify main effects, which reduces the flexibility of the model. In Chapter 4, we found expressions for induced main effects and interactions assuming a latent structure similar to that of Chapter 3, and we allow for identification of these parameters in the space of chemical exposures for any proportional hazard model.

In Chapter 5, we built a copula factor model that provides interpretability via grouping of variables, flexibility for dose-response curves, scalability for new data challenges and uncertainty quantification. Interpretable grouping of variables is attained using a factor model (Murray et al., 2013) for the chemical exposures. This approach treats the marginal densities of the different exposures as nuisance parameters (Hoff et al., 2007) and models the covariance among them, effectively accommodating for exposures’ skewed distributions. Flexible dose-responses can be modeled using Bayesian B-splines in the latent space (DiMatteo et al., 2001), with the added benefit of allowing a simple modification of the efficient and adaptive Metropolis-Hastings algorithm developed in Chapter 3. Preliminary results show that this approach can run locally for tens of thousands observations and hundreds of correlated
covariates, correctly estimating complex dose-response surfaces and characterizing the correlation structure.

Finally, in Chapter 6 we developed a post-processing algorithm that allows for inference on latent factors and factor loadings matrix without imposing identifiability constraints. Crucially, this algorithm does not affect the choice of priors and structure for the matrix of factor loadings, which is a modeling choice left to the practitioner. By developing a greedy maximization procedure, we significantly improve the computational efficiency with respect to Papastamoulis and Ntzoufras (2020) and Marin et al. (2005), while maintaining a good estimation performance.

The statistical methods presented in this thesis are tailored at practitioners and focused on the motivating applications. They coherently provide solutions to the major challenges in environmental epidemiology, such as missing data or observations measured below the limit of detection, non-normality of chemical measurements distributions and high correlation. The code has been made publicly available on several GitHub repositories and included in the R package \texttt{infinitefactor} with an efficient C++ implementation.
Appendix A

Proofs

Proof of Proposition 1. Let us drop the index i for notation simplicity and always assume that we are conditioning on all the parameters. The posterior distribution of $\eta$ is Normal with covariance $V = (\Lambda^T \Psi^{-1} \Lambda + I)^{-1}$ and mean $AX$ where $A = V \Lambda^T \Psi^{-1} = (\Lambda^T \Psi^{-1} \Lambda + I)^{-1} \Lambda^T \Psi^{-1}$. This follows from a simple application of Bayes Theorem. Now:

$$
\mathbb{E}(y|X) = \mathbb{E}(\mathbb{E}(y|\eta)|X) = \mathbb{E}(\eta^T \omega + \eta^T \Omega \eta|X) = \\
= \omega^T \mathbb{E}(\eta|X) + \mathbb{E}(\eta^T \Omega \eta|X)
$$

Recall that the expectation of a quadratic form $\eta^T \Omega \eta$ of a random vector $\eta$ with mean $\mu$ and covariance matrix $\Sigma$ is equal to $tr(\Omega \Sigma) + \mu^T \Omega \mu^T$.

$$
\mathbb{E}(y|X) = \omega^T AX + tr(\Omega V) + (AX)^T \Omega (AX) = \\
= tr(\Omega V) + (\omega^T A)X + X^T (A^T \Omega A)X
$$

(ii) Recall that $\eta \sim N(0, I)$, $y = \eta^T \omega + \eta^T \Omega \eta_i + \epsilon_y$ and $X = \Lambda \eta + \epsilon$, from simple algebra it follows that

$$
\text{Cov}(y, X) = \omega^T \text{Cov}(\eta, \eta) \Lambda^T + \text{Cov}(\eta^T \Omega \eta, \Lambda \eta)
$$
From the prior specification $\text{Cov}(\eta, \eta) = I$, hence let us focus on the term $\text{Cov}(\eta^T \Omega \eta, \Lambda \eta)$ and show that it is equal to $0_p$:

$$\text{Cov}(\eta^T \Omega \eta, \Lambda \eta) = \text{Cov}(\sum_{j=1}^{p} \sum_{l=1}^{p} \omega_{j,l} \eta_j \eta_l, \Lambda \eta) =$$

$$= \sum_{j=1}^{p} \sum_{l=1}^{p} \omega_{j,l} \text{Cov}(\eta_j \eta_l, \begin{pmatrix} \lambda_{1,j} \eta_j + \ldots + \lambda_{1,k} \eta_k \\ \ldots \\ \lambda_{p,j} \eta_j + \ldots + \lambda_{p,k} \eta_k \end{pmatrix}) =$$

$$= \sum_{j=1}^{p} \sum_{l=1}^{p} \omega_{j,l} \text{Cov}(\eta_j \eta_l, \begin{pmatrix} \lambda_{1,j} \eta_j + \lambda_{1,l} \eta_l \\ \ldots \\ \lambda_{p,j} \eta_j + \lambda_{p,l} \eta_l \end{pmatrix}) =$$

$$= \sum_{j=1}^{p} \sum_{l=1}^{p} \omega_{j,l} \left[ \text{Cov}(\eta_j \eta_l, \begin{pmatrix} \lambda_{1,j} \eta_j \\ \ldots \\ \lambda_{p,j} \eta_j \end{pmatrix}) + \text{Cov}(\eta_j \eta_l, \begin{pmatrix} \lambda_{1,l} \eta_l \\ \ldots \\ \lambda_{p,l} \eta_l \end{pmatrix}) \right]$$

Now $\text{Cov}(\eta_j \eta_l, \eta_j) = E(\eta_j^2 \eta_l) = 0$. In fact when $j \neq l$, we have that $E(\eta_j^2 \eta_l) = E(\eta_j^2)E(\eta_l) = 0$ and when $j = l$, $E(\eta_j^3) = 0$ since $\eta_j \sim N(0, 1)$. 

Proof of Proposition 2. Let $p_0(X, y) = p(X, y|\Theta_0) = \int p(X, y|\Theta_0, \eta)p(\eta)d\eta$ where $\Theta_0 = (\omega_0, \Omega_0, \sigma^2_0, \Phi_0)$ and let $p'(X, y) = p(X, y|\eta')$ for a given vector $\eta'$. Also, we have that $p_0(X, y) = C_0k_0(X, y)$, where $k_0(X, y)$ is the kernel of a Multivariate normal distribution with parameters $\Theta_0$. We are interested in computing:

$$KL(p_0; p') = \int p_0(X, y) \log \left( \frac{p_0(X, y)}{p'(X, y)} \right) dX dy$$

Let us focus on the $p_0(X, y)$:

$$p_0(X, y) = \int p_0(X, y|\eta)p(\eta)d\eta =$$

$$= \int_{B_{1-\epsilon}} p_0(X, y|\eta)p(\eta)d\eta + \int_{B_{1-\epsilon}^C} p_0(X, y|\eta)p(\eta)d\eta$$

Where $B_{1-\epsilon}$ is a closed ball such that $p(\eta \in B_{1-\epsilon}) = 1 - \epsilon$ according to the prior $p(\eta)$. Now, on the closed ball $B_{1-\epsilon}$ the function $p_0(X, y|\eta)$ has a supremum which
we denote \( \eta^* = \eta(y, X, \Theta_0) = \arg\sup_{\eta \in B_{1-\epsilon}} p_0(X, y|\eta) \). Also recall that \( p_0(X, y|\eta) = C_0k_0(X, y|\eta) \) where \( k_0(X, y|\eta) \ll 1 \).

\[
p_0(X, y) = \int_{B_{1-\epsilon}} p_0(X, y|\eta)p(\eta)d\eta + \int_{B_{1-\epsilon}^C} p_0(X, y|\eta)p(\eta)d\eta \leq \int_{B_{1-\epsilon}} p_0(X, y|\eta^*)p(\eta)d\eta + \int_{B_{1-\epsilon}^C} C_0p(\eta)d\eta = \]

\[
= p_0(X, y|\eta^*)(1-\epsilon) + C_0\epsilon
\]

We now need to take the logarithm of the expression above, recall \textit{log sum inequality}:

\[
\log\left(\frac{a_1 + a_2}{b_1 + b_2}\right) \leq \frac{a_1}{a_1 + a_2}\log\left(\frac{a_1}{b_1}\right) + \frac{a_2}{a_1 + a_2}\log\left(\frac{a_2}{b_2}\right) \leq \log\left(\frac{a_1}{b_1}\right) + \log\left(\frac{a_2}{b_2}\right)
\]

We apply the \textit{log sum inequality} with \( a_1 = p_0(X, y|\eta^*)(1-\epsilon) \), \( a_2 = C_0\epsilon \), \( b_1 = 1 - \epsilon^* \) and \( b_2 = \epsilon^* \), so we have that:

\[
\log\left(\frac{p_0(X, y|\eta^*)(1-\epsilon) + C_0\epsilon}{1 - \epsilon^* + \epsilon^*}\right) \leq \log(p_0(X, y|\eta^*)) \frac{1 - \epsilon}{1 - \epsilon^*} + \log\left(\frac{C_0\epsilon}{\epsilon^*}\right) = \]

\[
= \log(p_0(X, y|\eta^*)) + \log\left(\frac{1 - \epsilon}{1 - \epsilon^*}\right) + \log\left(\frac{C_0\epsilon}{\epsilon^*}\right) \leq \]

\[
\leq \log(p_0(X, y|\eta^*)) - \log(1 - \epsilon^*) + \log\left(\frac{C_0\epsilon}{\epsilon^*}\right)
\]

We can choose \( \epsilon^* \) s.t. \(-\log(1 - \epsilon^*) \leq \epsilon_1 \) and \( \epsilon = \frac{\epsilon^*}{C_0} \) so that the last term in the above expression is equal to zero. We can also choose \( \epsilon \leq \frac{\epsilon^*}{C_0} \) and we would have \( \log\left(\frac{C_0\epsilon}{\epsilon^*}\right) \leq 0 \). Finally we have that:

\[
KL(p_0; p') \leq \int p_0(X, y)\log\left(\frac{p_0(X, y|\eta^*)}{p(X, y|\eta')}\right)dXdy + \epsilon_1
\]
We can now compute the above integral, find it below multiplied by 2:

\[
\log \left( \frac{\prod_{j=1}^{p} \psi_j^0}{\prod_{j=1}^{p} \psi_j} \right) + \left[ (\Lambda \eta')^T \Psi^{-1}(\Lambda \eta') - (\Lambda_0 \eta^*)^T \Psi^{-1}_0(\Lambda_0 \eta^*) \right] + tr \left( (\Psi^{-1} - \Psi^{-1}_0) Cov_0(X) \right) + \\
+ \log \left( \frac{\sigma^2}{\sigma_0} \right) + \mathbb{E}_0(y^2) \left( \frac{1}{\sigma^2} - \frac{1}{\sigma_0^2} \right) + 2\mathbb{E}_0(y) \left[ \frac{\eta^* \omega_0 + \eta^* \Omega_0 \eta^*}{\sigma_0^2} - \frac{\eta' \omega + \eta' \Omega \eta'}{\sigma^2} \right] + \\
+ \left[ \frac{(\eta' \omega + \eta' \Omega \eta')^2}{\sigma^2} - \frac{(\eta^* \omega_0 + \eta^* \Omega_0 \eta^*)^2}{\sigma_0^2} \right]
\]

We have that \( 2\mathbb{E}_0(X^T) [\Psi^{-1}_0 \Lambda_0 \eta^* - \Psi^{-1} \Lambda \eta'] = 0 \). These are all continuous functions of \( \Theta, \Theta, \eta^* \) and \( \eta' \), so we can choose \( \Theta \) and \( \eta' \) such that the above expression is \( \leq \epsilon_1 \) so that \( KL(p_0; p') \leq 2\epsilon_1 \). In particular there exist \( \delta \) such that this holds for any \( \eta' \in D_\delta = \{ \eta : ||\eta - \eta^*||_2 \leq \delta \} \). Hence we have that \( \Phi(D_\delta) > 0 \), where \( \Phi \) is the multivariate normal distribution and we can apply Proposition 6.28 of Ghosal and Van der Vaart (2017) to get the result. \( \square \)

**Proof of Proposition 3.** As \( n \) grows, we have that the posterior distribution of \( (\Lambda, \Psi) \) concentrates on the model that is closest to the true data-generating model in Kullback-Leibler divergence, see (Berk et al., 1966).

\[
KL((\Lambda_0, \Psi_0); (\Lambda, \Psi)) = tr \left( (\Lambda \Lambda^T + \Psi)^{-1}(\Lambda_0 \Lambda_0^T + \Psi_0) \right) - p + \log \left( \frac{|\Lambda \Lambda^T + \Psi|}{|\Lambda_0 \Lambda_0^T + \Psi_0|} \right)
\]

where \( KL((\Lambda_0, \Psi_0), (\Lambda, \Psi)) \) denotes the Kullback-Leibler divergence between \( p(X, y|\Lambda_0, \Psi_0) \) and \( p(X, y|\Lambda, \Psi_0) \). Let \( (\Lambda^*, \Psi^*) = \arg\inf d(\Lambda, \Psi) \). Now, \( \Lambda_0 \Lambda_0^T \) is symmetric and positive definite, so it admits an EigenDecomposition, i.e.:

\[
\Lambda_0 \Lambda_0^T + s_0 I_p = Q_0 \Sigma_0 Q_0^{-1} + s_0 I_p = Q_0(\Sigma_0 + s_0 I_p)Q_0^{-1},
\]

where \( Q_0 \) is the \( p \times p \) matrix containing the eigenvectors of \( \Lambda_0 \Lambda_0^T \) and \( \Sigma_0 \) is a diagonal matrix containing the eigenvalues, i.e. \( \text{diag}(\Sigma_0) = (v_1, \ldots, v_{k_0}, 0, \ldots, 0) \). Define \( \Lambda_1 \Lambda_1^T \) as the best \( k \)th rank approximation to \( \Lambda_0 \Lambda_0^T \), when the approximation is based
on the Frobenius norm. From the Eckart-Young theorem, we know that $\Lambda_1 \Lambda_1^T = Q_0 \Sigma_1 Q_0^{-1}$ where $\text{diag}(\Sigma_1) = (v_1, \ldots, v_k, 0, \ldots, 0)$. By definition of $(\Lambda^*, \Psi^*)$:

$$KL((\Lambda_0, \Psi_0); (\Lambda^*, \Psi^*)) \leq KL((\Lambda_0, \Psi_0); (\Lambda_1, \Psi_0)) =$$

$$= \text{tr} \left( (Q_0(\Sigma_1 + s_0 I_p)Q_0^{-1})(Q_0(\Sigma_0 + s_0 I_p)Q_0^{-1}) \right) - p + \log \left( \frac{|Q_0(\Sigma_1 + s_0 I_p)Q_0^{-1}|}{|Q_0(\Sigma_0 + s_0 I_p)Q_0^{-1}|} \right) =$$

$$= \text{tr} \left( (\Sigma_1 + s_0 I_p)^{-1}(\Sigma_0 + s_0 I_p) \right) - p + \sum_{j=k+1}^{k_0} \left( \log(s_0) - \log(v_j + s_0) \right) \leq$$

$$\leq k + \sum_{j=k+1}^{k_0} \frac{s_0}{s_0 + v_j} + (p - k_0) - p =$$

$$= \sum_{j=k+1}^{k_0} \left( \frac{s_0 + v_j}{s_0} - 1 \right) = \sum_{j=k+1}^{k_0} \frac{v_j}{s_0}$$

\[ \square \]

**Proof of Proposition 4.** The conditional distribution of $\eta_i$ given $X_i$ is Normal with covariance $V = (\Lambda_x^T \Psi^{-1} \Lambda_x + \Sigma_{\eta}^{-1})^{-1}$ and mean $A X_i$ where $A = V \Lambda_x^T \Psi^{-1} = (\Lambda_x^T \Psi^{-1} \Lambda_x + \Sigma_{\eta}^{-1})^{-1} \Lambda_x^T \Psi^{-1}$. Now:

$$E(Y_i | X_i) = E(\mathbb{E}(Y_i | \xi_i) | X_i) = E(\Lambda_y \mathbb{E}(\xi_i | \eta_i) | X_i) =$$

$$= \Lambda_y E(\Gamma \eta_i + \Omega(\eta_i) | X_i) =$$

$$= \Lambda_y [\Gamma E(\eta_i | X_i) + E(\Omega(\eta_i) | X_i)],$$

where $E(\Omega(\eta_i) | X_i)$ is a vector such that the $j^{th}$ element is equal to $E(\eta_i \Omega_{j} \eta_i | X_i)$, and it holds:

$$E(\eta_i \Omega_{j} \eta_i | X_i) = \text{tr}(\Omega_j V) + X_i^T (A^T \Omega_j A) X_i,$$

\[ \square \]

**Proof of Corollary 5.** Model (3.4) is invariant to rotations. For simplicity we consider
the model without interactions, but everything holds for that as well.

\[ Y_i = (\Lambda_y P_y)(P_y^T \xi_i) + \epsilon_{y,i}, \quad \epsilon_{y,i} \sim N_q(0, \Phi) \]

\[ X_i = (\Lambda_x P_x)(P_x^T \eta_i) + \epsilon_{x,i}, \quad \epsilon_{x,i} \sim N_p(0, \Psi) \]

\[ P_y^T \xi_i = (P_y^T P_x)(P_x^T \eta_i) + P_y^T \epsilon_{\xi,i}, \quad \epsilon_{\xi,i} \sim N_m(0, \Sigma_{\xi}) \]

\[ \eta_i \sim N_k(0, \Sigma_{\eta}), \]

where \( P_y \) and \( P_x \) are \( m \times m \) and \( k \times k \) orthogonal matrices respectively, i.e. \( P_y P_y^T = I \) and \( P_x P_x^T = I \). Let us consider the induced main effects:

\[
E(Y_i | X_i) = \left( \Lambda_y P_y (P_y^T \Gamma P_x) \right) \left( (\Lambda_x P_x)^T \Psi^{-1}(\Lambda_x P_x) + \Sigma_{\eta}^{-1} \right)^{-1} \left( \Lambda_x P_x \right)^T \Psi^{-1} X_i = \\
= \Lambda_y \Gamma P_x \left[ P_x^T \Lambda_x \Psi^{-1} \Lambda_x P_x + P_x^T \Sigma_{\eta}^{-1} P_x \right]^{-1} P_x^T \Lambda_x \Psi^{-1} X_i = \\
= \Lambda_y \Gamma \left[ \Lambda_x^T \Psi^{-1} \Lambda_x + \Sigma_{\eta}^{-1} \right]^{-1} \Lambda_x^T \Psi^{-1} X_i
\]

which corresponds to the expression in Proposition (4). We used the fact that \( \Sigma_{\eta} \) is a diagonal matrix. The same can be shown with interaction effects. \( \square \)

**Proof of Proposition 7.** Our goal is to compute the hazard function conditional on \( X_i \) by integrating out the latent variable, i.e.:

\[
\mathbb{E}(\exp(\omega^T \eta_i + \eta_i^T \Omega \eta_i) | X_i) = \int \exp(\omega^T \eta_i + \eta_i^T \Omega \eta_i) p(\eta_i | X_i)
\]

where \( \eta_i | X_i \sim N(\mu_i = AX_i, V) \), with \( V = (\Lambda^T \Psi^{-1} \Lambda + I)^{-1} \), \( A = VA^T \Psi^{-1} = (\Lambda^T \Psi^{-1} \Lambda + I)^{-1} \Lambda^T \Psi^{-1} \). After completing the squares, the result of the above integral is:

\[
\mathbb{E}(\exp(\omega^T \eta_i + \eta_i^T \Omega \eta_i) | X_i) = \sqrt{\frac{|V'|}{|V|}} \exp(H)
\]

where \( H = \frac{1}{2} (\Delta^T V'^{-1} \Delta - \mu_i^T V^{-1} \mu_i) \), \( \Delta = V'(\omega + V^{-1} \mu_i) \) and \( V' = (V^{-1} - 2\Omega)^{-1} = \)
\[(I - 2V\Omega)^{-1}V = V(I - 2\Omega V)^{-1}\]. Then we rearrange:

\[
2H = [\Delta^TV'^{-1}\Delta - \mu_i^TV^{-1}\mu_i] = \\
= [(V'(\omega + V^{-1}\mu_i))^TV'^{-1}(V'(\omega + V^{-1}\mu_i)) - \mu_i^TV^{-1}\mu_i] = \\
= [(\omega + V^{-1}\mu_i)^TV'(\omega + V^{-1}\mu_i) - \mu_i^TV^{-1}\mu_i] = \\
= [\omega^TV'\omega + 2\omega^TV^{-1}V'\mu_i + \mu_i^TV^{-1}V'V^{-1} - \mu_i^TV^{-1}\mu_i] = \\
= [\omega^TV'\omega + 2\omega^T(I - 2\Omega V)^{-1}\mu_i + \mu_i^T(I - 2\Omega V)^{-1}V^{-1}\mu_i - \mu_i^TV^{-1}\mu_i] = \\
= [\omega^TV'\omega + 2\omega^T(I - 2\Omega V)^{-1}\mu_i + \mu_i^T[(I - 2\Omega V)^{-1} - I]V^{-1}\mu_i].
\]

We now have expressions for the induced main effects, \(\omega^T(I - 2\Omega V)^{-1}A\), and the interaction effects, \(\frac{1}{2}A^T[(I - 2\Omega V)^{-1} - I]V^{-1}A\). \(\square\)

**Corollary 8.** The induced main effects and interaction effects of Proposition (7) are identifiable.

**Proof.** Notice that model (4.2) is invariant to rotations:

\[
\lambda(t|\eta_i) = \lambda_0(t)\exp((P^T\eta_i)^T(P^T\omega) + (P^T\eta_i)^T(P^T\Omega P)(P^T\eta_i)), \\
X_i = (\Lambda P)(P^T\eta_i) + \epsilon_i, \quad \epsilon_i \sim N_p(0, \Psi), \\
\eta_i \sim N_k(0, I).
\]

where \(P_y\) and \(P_x\) are \(m \times m\) and \(k \times k\) orthogonal matrices respectively, i.e. \(P_yP_y^T = I\) and \(P_xP_x^T = I\). Let us consider the induced main effects:

\[
E(Y_i|X_i) = (\Lambda_yP_y)(P_y^T\Gamma P_x)[(\Lambda_xP_x)^T\Psi^{-1}(\Lambda_xP_x) + \Sigma^{-1}_\eta]^{-1}(\Lambda_xP_x)^T\Psi^{-1}X_i = \\
= \Lambda_y\Gamma P_x\big[P_x^T\Lambda_x^T\Psi^{-1}\Lambda_xP_x + P_x^T\Sigma^{-1}_\eta P_x\big]^{-1}P_x^T\Lambda_x^T\Psi^{-1}X_i = \\
= \Lambda_y\Gamma[\Lambda^T\Psi^{-1}\Lambda + \Sigma^{-1}_\eta]^{-1}\Lambda^T\Psi^{-1}X_i
\]

which corresponds to the expression in Proposition (7). \(\square\)


Thurstone, L. L. (1947), *Multiple-factor analysis*, Univ. of Chicago Press.


Biography

Federico Ferrari obtained a Bachelor degree in “Economics and Finance” at Università degli studi di Bologna, with distinction. During his time in Bologna, he also did an exchange semester at Lund University in Sweden. He continued his studies in Turin where he received a master’s degree in “Stochastics and Data Science” from the department of Mathematics at the University of Turin and in parallel a master in “Statistics and Applied Mathematics” at Collegio Carlo Alberto, both with distinction. In Turin, he wrote a thesis in Nonparametric Bayesian multi-armed bandits supervised by Professor Stefano Favaro, which he later published in the Annals of Applied Statistics with an application to single cell experiment design (Camerlenghi et al., 2020).

In August 2017, Federico started his graduate studies in the Department of Statistical Science at Duke University and has been advised by David Dunson. He won the Data Expeditions award with the project “Poverty Indexes from World Bank Data”. Next, he will be moving to Dana Farber cancer institute and Harvard School of Public Health as a Postdoctoral Researcher.