Bayesian variable selection for multi-dimensional semiparametric regression models

Joseph Antonelli, Maitreyi Mazumdar, David Bellinger,
David Christiani, Robert Wright, Brent A. Coull

Abstract

Humans are routinely exposed to mixtures of chemical and other environmental factors, making the quantification of health effects associated with environmental mixtures a critical goal for establishing environmental policy sufficiently protective of human health. The quantification of the effects of exposure to an environmental mixture poses several statistical challenges. It is often the case that exposure to multiple pollutants interact with each other to affect an outcome. Further, the exposure-response relationship between an outcome and some exposures, such as some metals, can exhibit complex, nonlinear forms, since some exposures can be beneficial and detrimental at different ranges of exposure. To estimate the health effects of complex mixtures we propose sparse tensor regression, which uses tensor products of marginal basis functions to approximate complex functions. We induce sparsity using multivariate spike and slab priors on the number of exposures that make up the tensor factorization. We allow the number of components required to estimate the health effects of multiple pollutants to be unknown and estimate it from the data. The proposed approach is interpretable, as we can use the posterior probabilities of inclusion to identify pollutants that interact with each other. We illustrate our approach's ability to estimate complex functions using simulated data, and apply our method to a study of the association between exposure to metal mixtures and neurodevelopment.

1 Introduction

The study of the health impacts of environmental mixtures, including both chemical and non-chemical stressors, on human health is a research priority in the environmental health sciences. (Carlin *et al.* , 2013; Taylor *et al.* , 2016; Braun *et al.* , 2016). In the past, the majority of studies estimate the independent associations between an outcome and a given exposure, either marginally or while controlling for co-exposures. Other approaches allow for interactions between exposures, perhaps with some variable selection technique applied, but these approaches are often based on models assuming simple, linear associations between exposures or other strong parametric assumptions. It has been recently shown that in some settings, such as in studies on the impacts of exposure to metal mixtures on neurodevelopment in children, the exposure-response relationship between the multivariate exposure of interest and a health outcome can be a complex, non-linear and non-additive function (Henn *et al.* , 2010, 2014; Bobb *et al.* , 2014). Therefore there is a need for statistical approaches that can handle complex interactions among a large number of exposures. Even in settings with a relatively small number of exposures, the number of potential interactions can become large, making some form of variable selection or dimension reduction necessary.

There exist a vast statistical literature on high-dimensional regression models and approaches to reducing the dimension of the covariate space. The LASSO (Tibshirani, 1996) utilized an l1 penalty on the absolute value of regression coefficients in a regression model, which forces many of them to zero, effectively removing them from the model. Many approaches have built upon this same methodology (Zou & Hastie, 2005; Zou, 2006); however these all typically rely on relatively simple, parametric models. More recently, a number of papers (Zhao *et al.* , 2009; Bien *et al.* , 2013; Hao *et al.* , 2014; Lim & Hastie, 2015) have adopted these techniques to identify interactions, although these approaches are only applicable to linear models with pairwise interactions. While similar ideas were used in nonlinear models in Radchenko & James (2010), again the analysis was restricted to pairwise interactions. Another issue is that inferring any measure of uncertainty around these estimates can be quite challenging in high-dimensions.

A number of approaches have been introduced that embed variable selection within nonlinear regression models. Shively *et al.* (1999) utilized variable selection in Bayesian nonparametric regression where they placed integrated Wiener processes for each covariate in the model, and then extended these ideas in Wood *et al.* (2002) to non additive models that allowed for interactions between covariates. They utilized the BIC approximation to the marginal likelihood in order to calculate posterior model probabilities, which requires enumerating all potential models. Even if the number of covariates in the model is not prohibitively large, considering even just two-way interactions can lead to a massive number of models to consider. Reich *et al.* (2009) utilized Gaussian process priors for all main effects and two-way interactions within a regression model and perform variable selection to see which of these are important for predicting the outcome. They do not

allow for explicit identification of higher-order interactions as all remaining interactions are absorbed into a function of all the covariates that accounts for higher-order interactions. Yau *et al.* (2003) and Scheipl *et al.* (2012) utilized variable selection in a additive regression model framework with interaction terms within a Bayesian framework, though one must specify the terms that enter the model *a priori* and there can be a large number of potential terms to consider. Qamar & Tokdar (2014) use Gaussian processes to separately model subsets of the covariate space and used stochastic search variable selection to identify which covariates should enter into the same Gaussian process function, effectively implying that they interact with each other. While this can be used to capture and identify complex interactions, it relies on Gaussian processes and therefore does not scale to larger data sets.

Complex models that allow for nonlinearities and interactions are becoming popular in the machine learning literature, with Gaussian process (or kernel machine) regression (O'Hagan & Kingman, 1978), support vector machines (Cristianini & Shawe-Taylor, 2000), deep neural networks (Bengio *et al.* , 2003) and random forests (Breiman, 2001) being just a few examples. Bobb *et al.* (2014) used Bayesian kernel machine regression to quantify the health effects of an environmental mixtures. This approach imposed sparsity within the Gaussian process to flexibly model multiple exposures. While these approaches and their many extensions can provide accurate predictions in the presence of complex relationships among exposures, they typically only provide a variable importance measure for each feature included in the model, and do not allow for formal inference on higher-order effects. An important question remains how to estimate the joint effect of a set of exposures that directly provides inference on interactions, while scaling to larger data sets without needing to search over such a large space of potential models that is considered when looking at higher-order interactions. Further, full Markov Chain Monte Carlo (MCMC) inference in kernel machine models does not scale with the sample size, limiting its applicability to large datasets. While one can potentially use approximate Bayesian methods, such as variational inference to simplify computation, these methods are only approximate and can yield biased estimators for some parameters (Guo *et al.* , 2016; Miller *et al.* , 2016).

In this paper we introduce a flexible, tensor regression approach that addresses several of the aforementioned issues as it allows for nonlinear interactions between variables, identifies explicitly which exposures are interacting with each other, and scales to much larger data sets than those allowed by traditional Gaussian process regression. We will introduce sparsity via spike and slab priors George & McCulloch (1993) within a tensor regression framework that reduces the dimension of the parameter space, while yielding inference on interactions among individual exposures in the mixture.

2 Model formulation

Throughout, we will assume that we observe $\boldsymbol{D_i} = (Y_i, \boldsymbol{X_i}^\mathsf{T}, \boldsymbol{C_i}^\mathsf{T})^\mathsf{T}$ for $i, \ldots, n$, where n is the sample size of the observed data, Y_i is the outcome, $\boldsymbol{X_i}$ is a p-dimensional vector of exposure variables, and $\boldsymbol{C_i}$ is an m-dimensional vector of covariates for subject i. Building on the construction of Wood (2006), we will assume that we have low rank bases to represent smooth functions for each of the exposures. For instance, the main effect of X_1, and an interaction effect between X_1 and X_2 could be modeled as follows:

$$f(X_1) = \sum_{l=1}^{L} \eta_l g_l(X_1) \qquad f(X_1, X_2) = \sum_{l=1}^{L} \sum_{m=1}^{M} \eta_{lm} g_l(X_1) h_m(X_2), \tag{1}$$

where $g()$ and $h()$ represent smooth basis functions of X_1 and X_2 respectively. For simplicity we will let the degrees of freedom for each exposure all be set to d, though this can easily be extended to allow differing degrees of freedom. Throughout the paper we will let $g_1() = 1$ and $h_1() = 1$ representing an intercept, while the remaining basis functions are defined using natural splines. Importantly, the inclusion of the intercept into the basis expansion implies that main effect terms are a subset of the terms included when we have an interaction. As we include higher order interactions, the intercept implies that all lower level interactions between the exposures being modeled are included in the expansion. To ease the presentation of our model, we will use subscripts to denote what type of parameter we are referring to. $\boldsymbol{\beta}_{12}$ refers to parameters associated with two-way interaction terms between X_1 and X_2 (excluding intercepts and main effect terms), while $\boldsymbol{\beta}_1$ refers to the main effect parameters for X_1 (excluding the intercept). Using this formulation we can re-write the above functions as

$$f(X_1) = \beta_0 + \widetilde{\boldsymbol{X}}_1 \boldsymbol{\beta}_1 \qquad f(X_1, X_2) = \beta_0 + \widetilde{\boldsymbol{X}}_1 \boldsymbol{\beta}_1 + \widetilde{\boldsymbol{X}}_2 \boldsymbol{\beta}_2 + \widetilde{\boldsymbol{X}}_{12} \boldsymbol{\beta}_{12}, \tag{2}$$

where the matrices $\widetilde{\boldsymbol{X}}_1, \widetilde{\boldsymbol{X}}_2, \widetilde{\boldsymbol{X}}_{12}$ contain the appropriate basis functions. This process could be trivially extended to any order interaction up to a $p-$way interaction, which is the maximum possible interaction in our model. We will use the following fully Bayesian formulation:

$$Y_i \sim Normal\left(f(\boldsymbol{X}_i) + \boldsymbol{C}_i \boldsymbol{\beta}_c, \sigma^2\right)$$

$$
\begin{aligned}
f(\boldsymbol{X}_i) &= \sum_{h=1}^{k} f_h(\boldsymbol{X}_i) \\
f_h(\boldsymbol{X}_i) &= \beta_0^{(h)} + \sum_{j=1}^{p} \widetilde{\boldsymbol{X}}_j \boldsymbol{\beta}_j^{(h)} + \sum_{j=1}^{p}\sum_{k=1}^{p} \widetilde{\boldsymbol{X}}_{jk} \boldsymbol{\beta}_{jk}^{(h)} + \dots \\
P(\boldsymbol{\beta}_S^{(h)}) &= \left(1 - \prod_{j \in S} \zeta_{jh}\right) \delta_{\mathbf{0}} + \left(\prod_{j \in S} \zeta_{jh}\right) \psi_1(\boldsymbol{\beta}_S^{(h)}) \\
&\quad \text{where } S \text{ is some subset of } \{1, 2, \dots, p\} \\
P(\zeta_{jh}) &= \tau_h^{\zeta_{jh}} (1 - \tau_h)^{1 - \zeta_{jh}} \mathbf{1}(A_h \not\subset A_m \forall m \text{ or } A_h = \{\}) \\
&\quad \text{where } A_h = \{j : \zeta_{jh} = 1\} \\
\tau_h &\sim B(\alpha\gamma/k, \gamma) \\
\sigma^2 &\sim IG(c, d),
\end{aligned}
$$

where any probability statements are conditional on parameters introduced in later rows. The set $\{\zeta_{jh}\}$ is a collection of binary indicators of whether the exposure j is included in the h^{th} component of the model. We use spike and slab priors to effectively eliminate exposures from the model, and in this paper we consider the spike to be a point mass at $\mathbf{0}$, while the slab, $\psi_1()$, is a multivariate normal distribution centered at $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}$ with covariance $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$. The dimension of the spike and slab prior depends on how many exposures are currently in the model, since their are more parameters for higher-order interactions. Due to this, we will let $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ be a diagonal matrix with $\sigma^2_{\boldsymbol{\beta}}$ on the diagonals. We adopt a finite-dimensional approximation to the two-dimensional Indian Buffet Process (Griffiths & Ghahramani, 2005; Ghahramani *et al.* , 2007) as a prior distribution over the matrix of binary indicators, $\boldsymbol{\zeta}$. As $k \to \infty$, this approximation approaches an Indian Buffet Process; in practice we set k to a large number that flexibly captures all features in the data, with a subset of the k column vectors in $\boldsymbol{\zeta}$ equaling zero.

Notice that each of the $h = 1, \dots, k$ functions $f_h(\boldsymbol{X}_i)$ take the same form. The difference will come due to the spike and slab formulation and the structure of the $\boldsymbol{\zeta}$ matrix, which says that each of the functions will be comprised of a different subset of the exposures. Importantly, the exposures included in one of the functions can not be a subset of the exposures included in another function as this would lead to redundant information in the two functions.

2.1 Identifiability of $\boldsymbol{\zeta}$

Two scientific goals in the study of environmental mixtures are to identify which exposures are associated with the outcome and to identify interactions among exposures. To this end, our prior imposes that for any two sets of variables included in a given feature, A_h and A_m, either $A_h \not\subset A_m$ and $A_m \not\subset A_h$ or one of the two is an empty set. These conditions state that each column vector of the matrix, $\boldsymbol{\zeta}$ must either be unique and not a subset of another column vector, or be a vector of zeros. An illustrative example of why this condition is necessary is a scenario in which all elements of $\boldsymbol{\zeta}$ are zero with the exception of ζ_{11} and ζ_{12}. In this scenario, the first covariate has a nonzero effect in both the first and second features of the model. Regardless of the values of $\boldsymbol{\beta}_1^{(1)}$ and $\boldsymbol{\beta}_1^{(2)}$, this model could be equivalently expressed as a function of one feature, rather than two. For instance, while $\boldsymbol{\beta}_1^{(1)} = -\boldsymbol{\beta}_1^{(2)}$ represents the same model as $\boldsymbol{\beta}_1^{(1)} = \boldsymbol{\beta}_1^{(2)} = \mathbf{0}$, inference regarding $\boldsymbol{\zeta}$ would differ across these two models. In the first scenario it would appear that the first covariate is associated with the outcome, even though it is not. Our conditions preclude this from happening by restricting the columns of $\boldsymbol{\zeta}$ to be unique, and therefore $\boldsymbol{\zeta}$ will provide a meaningful and interpretable measure of the importance of each exposure and potential interaction.

Another issue affecting our model selection indicators is that low-order interactions are subsets of the models defined by higher-order interactions. For example, if the true model is such that there exist a main effect of both X_1 and X_2, but no interaction between them, the model could mistakenly estimate that there is an interaction between exposure 1 and 2 simply due to the main effects. This issue has been addressed in similar contexts (Wood, 2006; Reich *et al.* , 2009) by utilizing constraints such that the two-way interaction between X_1 and X_2 only includes information that is orthogonal to the main effects of the individual exposures. In our context, this is not so much an issue of identifiability, but rather an issue with MCMC chains getting stuck in local modes. When exploring the model space via MCMC we can accept a move to including the two-way interaction, which is a local mode, and never be able to leave this model to find the true model of only including main effects.

To avoid this issue we simply need to update particular parameters jointly when searching the model space. When proposing the inclusion of an exposure j in feature h that leads to higher-order interactions, we also propose excluding exposure j in feature h, but including exposure j in a different feature with a subset of the exposures already included in feature h. This can be seen intuitively in the two exposure setting as

illustrated below.

$$\boldsymbol{\zeta}_{p_1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \boldsymbol{\zeta}_{p_2} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \quad \boldsymbol{\zeta}_{p_3} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Imagine that $\zeta_{11} = 1$ and we are updating ζ_{21}, which indicates whether exposure 2 enters into feature 1. If we did this naively, we would simply compare models $\boldsymbol{\zeta}_{p_1}$ and $\boldsymbol{\zeta}_{p_2}$, which would lead to us choosing $\boldsymbol{\zeta}_{p_2}$ even if there were only main effects of X_1 and X_2. If, however, we also update ζ_{22} at the same time, we will also evaluate model $\boldsymbol{\zeta}_{p_3}$. If the true relationship is a main effect of the two exposures, then model $\boldsymbol{\zeta}_{p_3}$ will be favored by the data and we will estimate $\boldsymbol{\zeta}$ accurately. This setting generalizes to higher-order interactions: When proposing a move to a higher-order interaction we must also evaluate all models that could lead to the data favoring a move to the higher-order interaction when in truth there is no need for a higher-order interaction.

2.2 Selection of spline degrees of freedom

An important tuning parameter in the proposed tensor model is the number of degrees of freedom, d, used to flexibly model the effects of the p exposures. Throughout we will make the simplifying assumption that the degrees of freedom is the same for all exposures, though this can trivially be relaxed to allow for differing amounts of flexibility across the marginal functions. We propose the use of the WAIC model selection criterion (Watanabe, 2010; Gelman *et al.* , 2014) to select d. If we let $s = 1, \ldots, S$ represent posterior draws from our model and let $\boldsymbol{\theta}$ represent all parameters in the model, the WAIC is defined as

$$WAIC = -2 \left(\sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(Y_i | \boldsymbol{\theta}^s) \right) - \sum_{i=1}^{n} \text{var}(\log p(Y_i | \boldsymbol{\theta})) \right). \tag{3}$$

The WAIC approximates leave one out cross validation, and therefore provides an assessment of model fit based on out of sample prediction performance. We first draw samples from the posterior distributions of the model parameters under a range of values for d, and then select the model that minimizes WAIC.

2.3 Prior elicitation

There are a number of hyperparameters that need to be chosen to complete the model specification. First, one must choose an appropriate value for $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$, which controls the amount of variability and shrinkage of $\boldsymbol{\beta}_S^{(h)}$ for those coefficients that come from the slab component of the prior. We will make the simplifying assumption that $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is a diagonal matrix with $\sigma^2_{\boldsymbol{\beta}}$ on the diagonals. It is well known that variable selection is very sensitive to the choice of this parameter (Reich *et al.* , 2009), so it is crucially important to select an appropriate value. We will utilize an empirical Bayes strategy where we estimate $\sigma^2_{\boldsymbol{\beta}}$ and then obtain posterior samples conditional on this value. A description of the algorithm to find the empirical Bayes estimate is available in The Appendix.

In addition, one must also select values for α and γ, which dictate the prior probability that an exposure has a nonzero association in a given feature. One can choose values to induce a desired level of sparsity in the model. In practice, however, this is typically not known and therefore we choose values that scale with the number of exposures in order to accommodate higher-dimensional data. We set $\alpha = Ck$ and $\gamma = p$, which allows the prior sparsity level to increase as the number of covariates increases. This strategy closely resembles other strategies in high-dimensional Bayesian models, such as those proposed by Zhou *et al.* (2014) and Ročková & George (2016). We will discuss the implications of scaling the prior with p in Section 3.

2.4 Lower bound on slab variance

One issue with the aforementioned empirical Bayes estimation of $\sigma^2_{\boldsymbol{\beta}}$ occurs if there is little to no association between the exposures and the outcome. In this case, $\sigma^2_{\boldsymbol{\beta}}$ will be estimated to be very small, which can lead to problems in the estimation of $\boldsymbol{\zeta}$. If the prior variance for the slab is small enough, then it is barely distinguishable from the spike as it begins to approximate a point mass at zero. In this situation, updating ζ_{jh} becomes essentially a coinflip for all j and h, which will lead to high posterior inclusion probabilities even when there is no true association. To avoid this we will estimate a lower bound, above which we are comfortable believing that any high posterior inclusion probabilities that we see are caused by true signals in the data.

To estimate this lower bound, we will first permute the rows of Y, thereby breaking any associations between the exposures and the outcome. We can run our MCMC algorithm on the permuted data for a number of iterations and keep track of the probability that $\zeta_{jh} = 1$. We can do this for a decreasing grid of potential values for $\sigma^2_{\boldsymbol{\beta}}$ and we can take the smallest value of $\sigma^2_{\boldsymbol{\beta}}$ such that a pre-chosen threshold is met. Examples of thresholds would be the smallest value such that the probability of including a main effect for any of the exposures is less than 0.25, or the smallest value such that the probability of any two-way interaction is

less than 0.05. This strategy could also be used to control for Bayesian counterparts to traditional frequentist error rates such as the false discovery rate, since we can calculate the average false discovery rate for each value of $\sigma^2_{\boldsymbol{\beta}}$.

To illustrate how this approach works within a given data set, we simulated a data set with no associations between the exposure and outcome. In this case, the empirical Bayes estimate of $\sigma^2_{\boldsymbol{\beta}}$ will be very small as it is roughly estimated to be the average squared value of the nonzero coefficients in $\boldsymbol{\beta}$. Figure 1 shows the posterior inclusion probabilities averaged over all exposures in this example as a function of $\sigma^2_{\boldsymbol{\beta}}$. The gray region represents the area where false positives start to arise because $\sigma^2_{\boldsymbol{\beta}}$ is too small, and we see that the empirical Bayes estimate is far into this region of false positives. If we were to proceed with the empirical Bayes estimate, then we would conclude that all of the exposures had posterior inclusion probabilities near 0.5, incorrectly indicating they are important for predicting the outcome. Our permutation strategy to identifying a lower bound is able to find a value of $\sigma^2_{\boldsymbol{\beta}}$ that avoids high posterior inclusion probabilities for false positives. In practice, we will estimate both the lower bound and empirical Bayes estimate and we will proceed with the maximum of these two values, which in this case would be the lower bound.

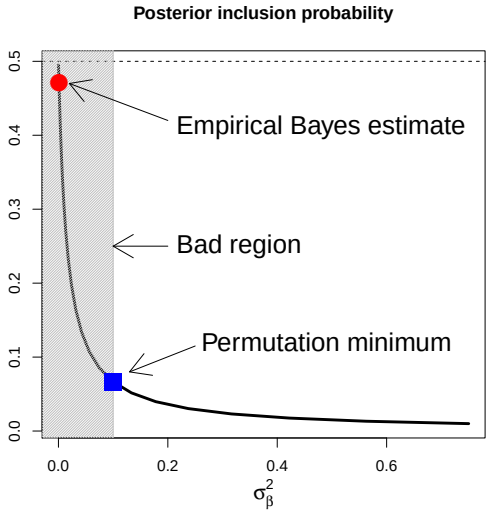

Figure 1: Illustration of empirical Bayes estimate and lower bound estimate of $\sigma^2_{\boldsymbol{\beta}}$

3 Properties of Prior Distributions

A primary goal in many analyses of the health effects of an environmental mixture is to identify interactions among exposures on risk. Therefore, it is important to understand the extent to which the proposed prior specification induces shrinkage on the interaction terms. The amount of shrinkage is dictated by both τ_h and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$. Prior elicitation of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is accomplished via empirical Bayes. Therefore, we focus on the role of τ_h, which directly controls the probability that a j^{th} order interaction is nonzero in the model and can be controlled through selection of values for α and γ. The covariates are exchangeable a priori, so we examine the probability that a j^{th} order interaction is zero by deriving the probability that the first j covariates have an interaction term between them.

Lemma 3.1: Assuming Model 3, the probability that there does not exist a feature, k' in the model such that $\zeta_{1k'} = \zeta_{2k'} = \cdots = \zeta_{jk'} = 1$ and $\zeta_{j+1k'} = \cdots = \zeta_{pk'} = 0$ is

$$\left(1 - \frac{\Gamma(j + \alpha\gamma/k)\Gamma(p + \gamma - j)}{\Gamma(p + \gamma + \alpha\gamma/k)}\right)^k. \tag{4}$$

The Appendix contains a proof of this result, which is useful because it expresses the probability of any order interaction being zero as a function of α and γ. This can guide the user in specifying values of α and β that best represent prior beliefs. It has been noted in the literature that a desirable feature of the prior is to place more shrinkage on higher-order interactions (Gelman *et al.* , 2008; Zhou *et al.* , 2014). Therefore We examine (4) to derive conditions that accomplish this type of shrinkage. Specifically, Theorem 3.2 shows that a standard choice of α and γ lead to this increasing amount of shrinkage as a function of the interaction order.

Theorem 3.2: Assuming Model 3, setting $\alpha = k/\gamma$ so that $\tau_h \sim B(1, \gamma)$, then the probability in Lemma 3.1 increases as a function of j if $\gamma \geq p$

The Appendix contains a proof of this result, which shows that higher-order interactions are more aggressively forced out of the model under either the standard $B(1,p)$ prior, which is often used in high-dimensional Bayesian models (Zhou *et al.* , 2014; Ročková & George, 2016), or if $\gamma > p$, which increasingly shrinks higher-order interactions even more aggressively than this Beta prior. In practice, a common prior choice for τ_h is $B(C,p)$, with C being some constant that doesn't change with the number of exposures. Nonetheless, this result shows that one can achieve increasing penalization of higher-order interaction terms as long as we scale the prior with p.

4 Posterior computation

Posterior samples from all unknown parameters in Model 3 can easily be obtained via Gibbs sampling, as all of the full conditional distributions have closed form. The algorithm is as follows:

1. Sample σ^2 from the following full conditional:

$$\sigma^2|\bullet \sim IG\left(c+\frac{n}{2}, d+\frac{\sum_{i=1}^{n}\left(Y_i-\sum_{h=1}^{k} f_h(\boldsymbol{X_i})-\boldsymbol{C_i}\boldsymbol{\beta_c}\right)^2}{2}\right)$$

2. Sample τ_h for $h=1,\ldots,k$ from the following full conditional:

$$\tau_h|\bullet \sim Beta\left(\alpha\gamma/k+\sum_{j=1}^{p}\zeta_{jh}, \gamma+\sum_{j=1}^{p}(1-\zeta_{jh})\right)$$

3. To update $\boldsymbol{\zeta}$ let us first introduce some additional notation. Let $\boldsymbol{\zeta_B}$ be the subset of $\boldsymbol{\zeta}$ we are updating and $\boldsymbol{\beta_B}$ be the corresponding regression coefficients. Further, let $\boldsymbol{\Lambda}$ be the set of all parameters in our model excluding $\boldsymbol{\zeta_B}$ and $\boldsymbol{\beta_B}$. To update $\boldsymbol{\zeta_B}$ we need to calculate $p(\boldsymbol{\zeta_B}|\boldsymbol{D},\boldsymbol{\Lambda})$, which we can do via the following:

$$\begin{aligned}
p(\boldsymbol{\zeta_B}=\boldsymbol{z_B}|\boldsymbol{D},\boldsymbol{\Lambda}) &= \frac{p(\boldsymbol{\beta_B}=\boldsymbol{0},\boldsymbol{\zeta_B}=\boldsymbol{z_B}|\boldsymbol{D},\boldsymbol{\Lambda})}{p(\boldsymbol{\beta_B}=\boldsymbol{0}|\boldsymbol{\zeta_B}=\boldsymbol{z_B},\boldsymbol{D},\boldsymbol{\Lambda})} && (5)\\
&= \frac{p(\boldsymbol{D},\boldsymbol{\Lambda}|\boldsymbol{\beta_B}=\boldsymbol{0},\boldsymbol{\zeta_B}=\boldsymbol{z_B})p(\boldsymbol{\beta_B}=\boldsymbol{0},\boldsymbol{\zeta_B}=\boldsymbol{z_B})}{p(\boldsymbol{D},\boldsymbol{\Lambda})p(\boldsymbol{\beta_B}=\boldsymbol{0}|\boldsymbol{\zeta_B}=\boldsymbol{z_B},\boldsymbol{D},\boldsymbol{\Lambda})} && (6)\\
&= \frac{p(\boldsymbol{D},\boldsymbol{\Lambda}|\boldsymbol{\beta_B}=\boldsymbol{0})p(\boldsymbol{\beta_B}=\boldsymbol{0},\boldsymbol{\zeta_B}=\boldsymbol{z_B})}{p(\boldsymbol{D},\boldsymbol{\Lambda})p(\boldsymbol{\beta_B}=\boldsymbol{0}|\boldsymbol{\zeta_B}=\boldsymbol{z_B},\boldsymbol{D},\boldsymbol{\Lambda})} && (7)\\
&\propto \frac{p(\boldsymbol{\beta_B}=\boldsymbol{0},\boldsymbol{\zeta_B}=\boldsymbol{z_B})}{p(\boldsymbol{\beta_B}=\boldsymbol{0}|\boldsymbol{\zeta_B}=\boldsymbol{z_B},\boldsymbol{D},\boldsymbol{\Lambda})} && (8)\\
&= \frac{\Phi(\boldsymbol{0};\boldsymbol{\mu_\beta},\boldsymbol{\Sigma_\beta})}{\Phi(\boldsymbol{0};\boldsymbol{M},\boldsymbol{V})}\prod_{h,j:\zeta_{jh}\in\boldsymbol{\zeta_B}}\tau_h^{\zeta_{jh}}(1-\tau_h)^{1-\zeta_{jh}} && (9)
\end{aligned}$$

   where $\Phi()$ represents the multivariate normal density function. $\boldsymbol{M}$ and $\boldsymbol{V}$ represent the conditional posterior mean and variance for the elements of $\boldsymbol{\beta}$ corresponding to the elements of $\boldsymbol{\zeta_B}$ that are equal to one and are defined as

$$\boldsymbol{M}=\left(\frac{\boldsymbol{X^*}^T\boldsymbol{X^*}}{\sigma^2}+\boldsymbol{\Sigma_\beta^{-1}}\right)^{-1}(\boldsymbol{X^*}^T Y^*+\boldsymbol{\Sigma_\beta^{-1}}\boldsymbol{\mu_\beta}), \qquad \boldsymbol{V}=\left(\frac{\boldsymbol{X^*}^T\boldsymbol{X^*}}{\sigma^2}+\boldsymbol{\Sigma_\beta^{-1}}\right)^{-1}, \tag{10}$$

   where $\boldsymbol{X^*}$ is a design matrix with the appropriate main effects and interactions induced by the model $\boldsymbol{z_B}$, and $Y_i^*=Y_i-\sum_{h:h\notin\boldsymbol{\zeta_B}} f_h(\boldsymbol{X})-\boldsymbol{C_i}\boldsymbol{\beta_c}$. Intuitively, the numerator of expression 9 is the prior probability that all elements of $\boldsymbol{\beta}$ corresponding to features in $\boldsymbol{\zeta_B}$ are zero and $\boldsymbol{\zeta_B}=\boldsymbol{z_B}$. The denominator of the expression represents the conditional posterior probability that all elements of $\boldsymbol{\beta}$ corresponding to features in $\boldsymbol{\zeta_B}$ are zero.

4. Update $\boldsymbol{\beta_S^{(h)}}$ from a normal distribution with mean $\boldsymbol{M}$ and variance $\boldsymbol{V}$ as defined above.

5. Update $\boldsymbol{\beta_c}$ from the following distribution:

$$\text{MVN}\left(\left(\frac{\boldsymbol{C}^T\boldsymbol{C}}{\sigma^2}+\boldsymbol{\Sigma_c^{-1}}\right)^{-1}(\boldsymbol{C}^T\widetilde{Y}+\boldsymbol{\Sigma_c^{-1}}\boldsymbol{\mu_c}), \left(\frac{\boldsymbol{C}^T\boldsymbol{C}}{\sigma^2}+\boldsymbol{\Sigma_c^{-1}}\right)^{-1}\right), \tag{11}$$

   where $\widetilde{Y}_i=Y_i-\sum_{h=1}^{k} f_h(\boldsymbol{X_i})$

5 Simulation study

Here we present results from a simulation study designed to assess the operating characteristics of the proposed approach in a variety of settings. In all scenarios we simulated 500 data sets with a sample size of $n = 200$ and $p = 10$ exposures. We generate the exposures, $\boldsymbol{X}$, from a multivariate normal distribution with all pairwise correlations set to 0.3 and marginal variances set to 1. We generate one additional covariate, C, from a standard normal distribution. We set the residual variance to $\sigma^2 = 1$ in all scenarios. In all cases we take the maximum of the empirical Bayes and lower bound estimates of $\sigma^2_{\boldsymbol{\beta}}$. For each simulated data set, we run a Gibbs sampler for 40,000 iterations with two chains, and fit the model for values of $d \in \{1, 2, 3, 4, 5, 6\}$. We present results for each value of d, the value that is selected based on the WAIC, and for the Bayesian kernel machine regression (BKMR) approach of Bobb *et al.* (2014). We assess the performance of the various models by calculating the mean squared error (MSE) of the estimates $f(X_1, ..., X_p)$, representing the true relationship between the exposures and the outcome, at all observed values of the exposure. We also assess the ability of the proposed model to identify nonzero exposure effects and interaction terms.

5.1 Polynomial interactions

We investigate two simulation scenarios. We first generate outcome data from a polynomial model that contains both a linear and a nonlinear interaction term as follows:

$$Y = 0.5X_2X_3 + 0.6X_4^2X_5 + \epsilon \tag{12}$$

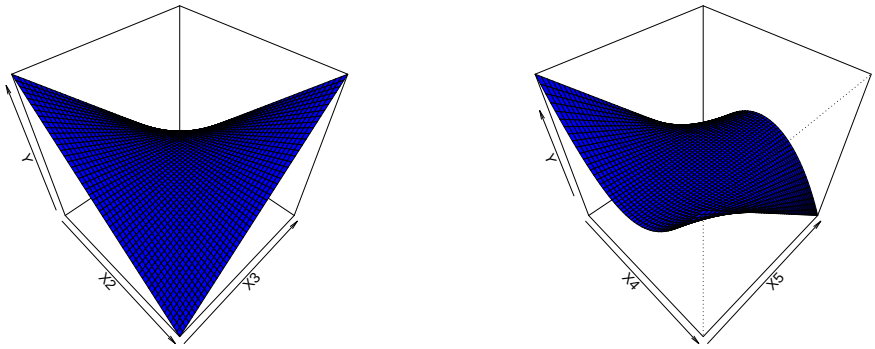

Figure 2: Data generating mechanism for the polynomial simulation. The left panel shows the relationship between X_2, X_3, and Y. The right panel shows the relationship between X_4, X_5, and Y

Figure 2 illustrates both of the regression terms in the model, while Figure 3 presents the results of the simulation study. The left panel of Figure 3 shows the matrix of posterior probabilities that exposures are interacting with each other. Cell (i, j) of this heatmap shows the posterior probability that exposures i and j are interacting with each other in the model averaged over all simulations. Results show that the model correctly identifies the pairs (X_2, X_3) and (X_4, X_5) as interacting with each other, as the average posterior interaction probabilities are 0.82 and 1.00, respectively. The middle panel highlights the ability of our procedure to estimate the marginal effect of X_4, which has a quadratic effect on the outcome, while holding constant the values of other exposures. The grey lines, which represent individual simulation estimates, approxiate the true, red line, indicating that the model adequately captures the nonlinear relationship. Finally, the right panel of Figure 3 shows the mean squared error of the global effect estimates of $f(\boldsymbol{X_i})$. Results show that well chosen values of d, including the automated choice using WAIC, lead to improved performance.

5.2 Exponential interactions

A second simulation scenario corresponds to a model that includes a three-way, nonlinear interaction as follows:

$$Y = 0.45X_2X_3^2\exp\left(0.75X_1\right) + \epsilon \tag{13}$$

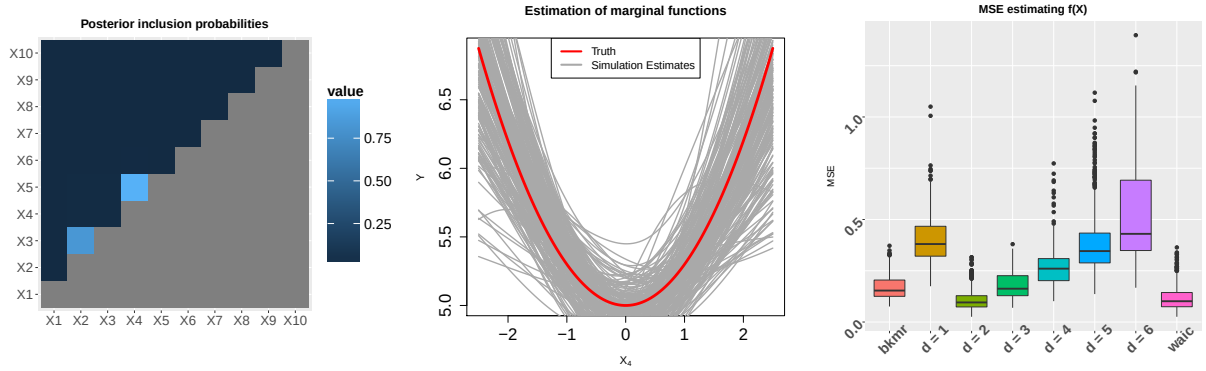

Figure 3: Results of the quadratic simulation. The left panel shows the posterior probabilities of exposures interacting with each other for the model selected by the WAIC. The middle panel shows the estimated marginal functions of X_4 on Y for the model selected by the WAIC, where the red line represents the truth and the grey lines represent individual estimates from separate simulations. The right panel shows the MSE in estimating $f(\boldsymbol{X})$ for all models considered.

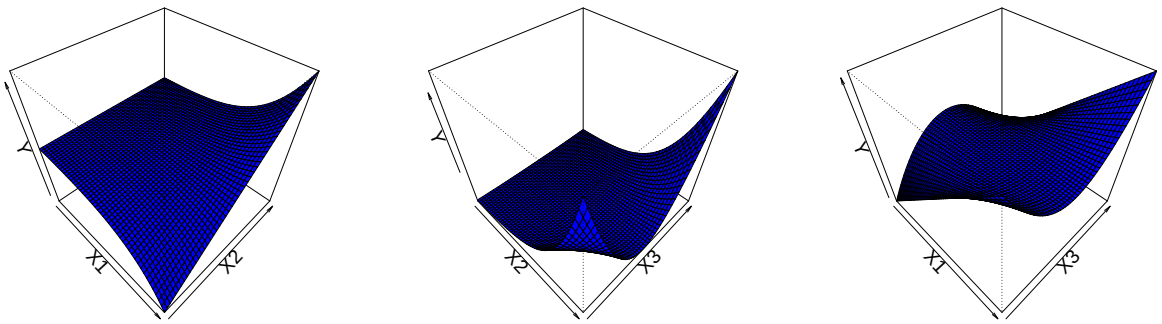

Figure 4: Data generating mechanism for the exponential simulation. Each panel corresponds to the relationship between two of the three covariates involved in the three-way interaction while fixing the third covariate at a given value.

Figure 4 highlights the bivariate associations between the three exposures that are active in the model, while holding the third exposure constant. It is clear from the figure that the relationship between the exposures is nonlinear and non-additive. Figure 5 shows results from this scenario. The posterior interaction probabilities are high for any combination of X_1, X_2, and X_3, due to the fact that the model is picking up on the three-way interaction between them. The average posterior probability that all three exposures interact in a three-way interaction is 0.75. The middle panel of Figure 5 shows the estimated marginal effect of X_1 on the outcome, while fixing the other exposures constant. We see that in some cases the estimated effect is flat, which corresponds to simulations in which X_1 was excluded. In other scenarios, the curve looks quadratic instead of exponential, which corresponds to scenarios in which the WAIC selects 2 degrees of freedom splines. The right panel of Figure 5 shows that the MSE of the model chosen by WAIC is almost as good as any individual choice of d, highlighting that WAIC is an effective method for choosing the spline degrees of freedom.

6 Analysis of Data on Metal Mixtures and Neurodevelopment

Here we present the results of an analysis of data from a study of the impact of metal mixtures on neurodevelopment of young Bangladeshi children using the proposed tensor model. The data consists of 375 children where the main outcome is the z-scored motor composite score (MCS). The MCS is a summary measure of psychomotor development derived from the Bayley Scales of Infant and Toddler Development (Bayley, 1993). Log-transformed values of prenatal exposure to Arsenic (As), Manganese (Mn), and Lead (Pb) were measured from umbilical cord blood. As potential confounders of the neurodevelopment - exposure associations, we control for the following covariates: gender, age at the time of neurodevelopment assessment, mother's education, mother's IQ, an indicator for which clinic the child visited, and the HOME score, which is a proxy for socioeconomic status.

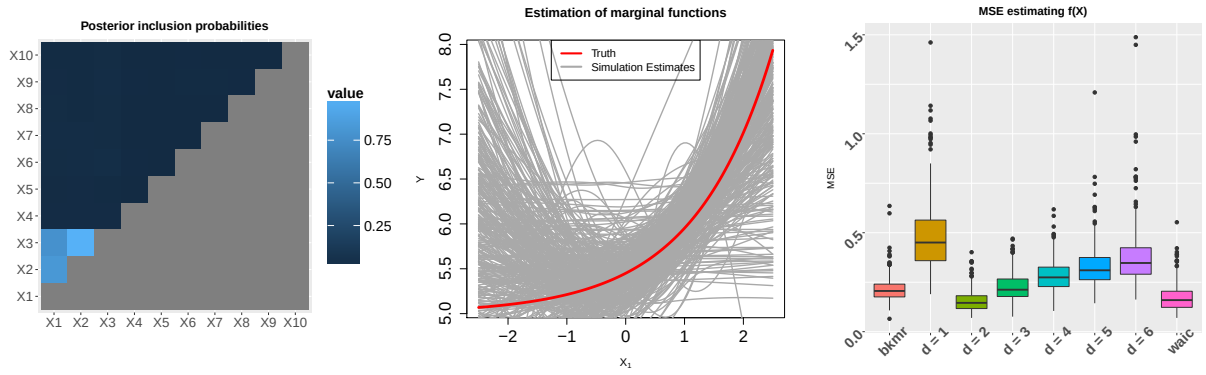

Figure 5: Results of the exponential simulation. The left panel shows the posterior probabilities of exposures interacting with each other for the model selected by the WAIC. The middle panel shows the estimated marginal functions of X_1 on Y for the model selected by the WAIC, where the red line represents the truth and the grey lines represent individual estimates from separate simulations. The right panel shows the MSE in estimating $f(\boldsymbol{X})$ for all models considered.

We fit our proposed approach to identify the complex relationship between the three metals and neurodevelopment. We ran the MCMC for 40,000 iterations with $\alpha = 3/p, \gamma = p, c = 0.001$, $d = 0.001$, and a non-informative, normal prior on $\boldsymbol{\beta_c}$. We ran the model for values of $d \in \{1, 2, 3, 4, 5, 6, 7\}$ and present results for $d = 4$, as it provided the smallest WAIC value.

6.1 Posterior inclusion probabilities

One feature of our approach is that we can examine the posterior distribution of $\boldsymbol{\zeta}$ to identify whether individual exposures are important for predicting the outcome, whether there is evidence of an interaction between two metals in the data, or if any higher-order interactions exist. Figure 6 shows the posterior probability of two-way interactions between all metals in the model, and the marginal posterior inclusion probability for each exposure. Figure 6 provides strong evidence that each exposure is important for predicting a child's MCS score. It further shows that there is strong evidence of an interaction between Mn and As, as well as moderate evidence of an interaction between Pb and As. Not shown in the figure is the posterior probability of all three metals interacting, which was estimated to be 0.0005.

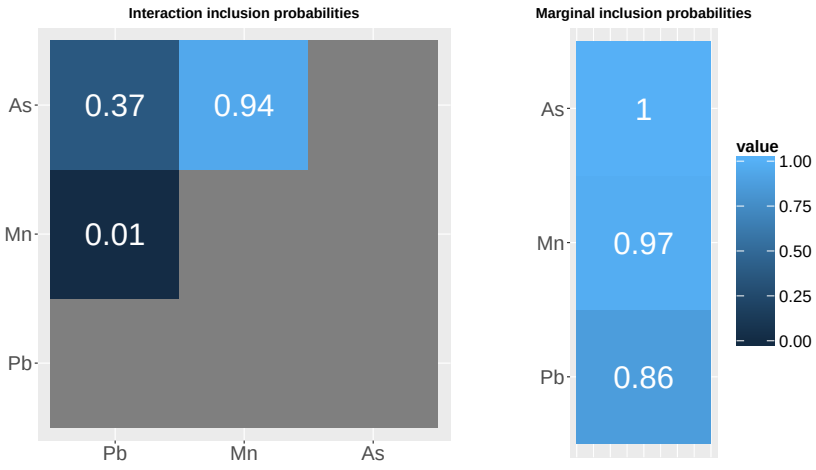

Figure 6: The left panel contains posterior inclusion probabilities for two-way interactions, and the right panel shows the marginal inclusion probabilities for each exposure

6.2 Visualization of identified interactions

Now that we have identified interactions from the model, we can plot posterior predictive distributions to assess the nature of these associations. Figure 7 plots the relationship between Mn and MCS for different values of As, while fixing Pb at its median. We see that at low and high levels of As, there appears to be a slightly positive impact of Mn on neurodevelopment. At the median level of As, however, there appears

to be an inverted U-shaped effect of Mn on MCS scores. This highlights results seen previously in a kernel machine regression analysis of these data Bobb *et al.* (2014), which indicated a strong interaction between Mn and As on the MCS score but didn't provide a measure of the strength of this interaction.

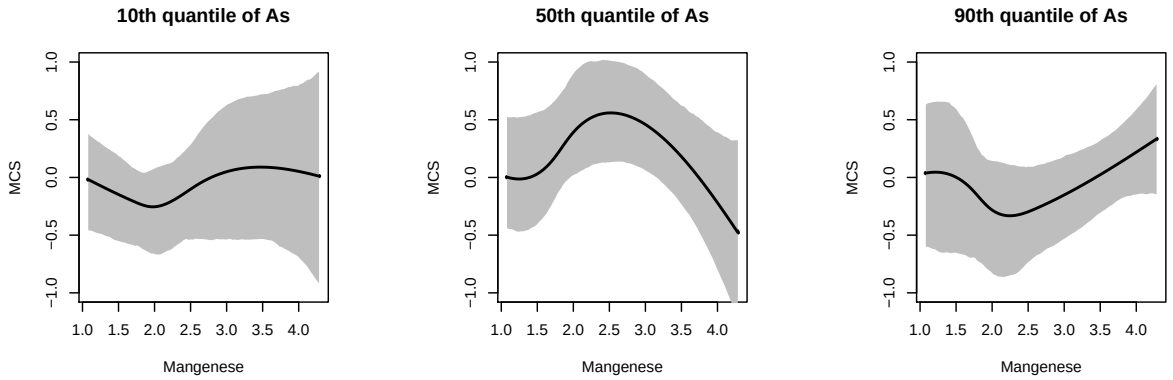

Figure 7: Posterior means and pointwise credible intervals of the effect of Mangenese on MCS at different quantiles of Arsenic while fixing Lead at its median.

We can also examine surface plots that show the relationship between As, Mn, and MCS at all levels of As, not just specific quantiles. Figure 8 shows both the posterior mean and pointwise standard deviations of this relationship. We see that the median levels of Mn and As are the most beneficial towards children in terms of their MCS score, which corresponds to the hump in the middle panel of Figure 7. Elevated levels of As tend to lower the MCS score, which is expected given that As is very toxic. In terms of standard deviations, we see that, as expected, the uncertainty around this surface is the smallest in the center, where most of the data is observed, and greatly increases on the fringes of the observed data, where we are effectively extrapolating effects.

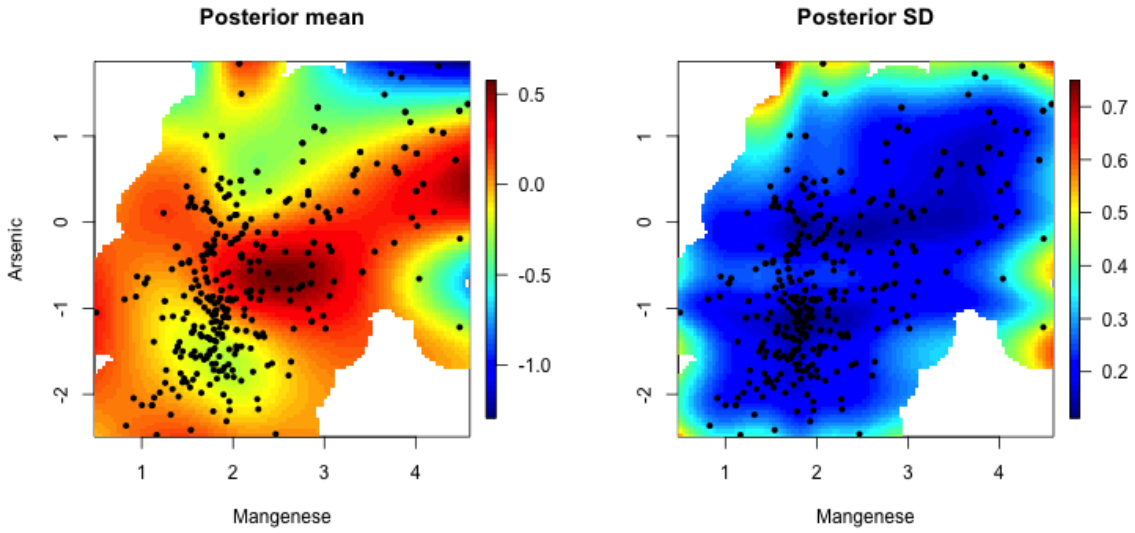

Figure 8: Posterior means and pointwise credible intervals of the relationship between Mangenese and Arsenic on MCS. The black dots represent the location of the observed data points.

7 Discussion

We have proposed a flexible method for analyzing complex, nonlinear associations between a large number of exposures on an outcome. Our method improves upon on existing methodology for environmental mixtures in several ways. By not relying on Gaussian processes to introduce flexibility in the exposure-response relationship, our approach can handle vastly larger data sets on the order of tens of thousands of subjects. The spike and slab formulation of our tensor factorization model also allows for formal inference on the presence of interactions among exposures, an important question in environmental epidemiology. Previous

studies have used graphical tools to identify interactions between pollutants. While useful, this approach does not provide formal inference on these features. Moreover, it can be difficult to visualize higher order interactions. We have illustrated the strengths of this approach through both simulated data and a study of the health effects of metal mixtures on child neurodevelopment.

While our formulation allows us to analyze much larger data sets, there are some disadvantages relative to fully nonparametric techniques such as Gaussian processes. Our model requires us to specify the functional form of the relationships between the predictors and the outcome a priori. In the current implementation, we employed natural splines to flexibly model these relationships. While other formulations are possible, any formulation would require a subjective choice. Further, one must select the number of degrees of freedom of the splines, and we have restricted this to be the same for all exposures across all features. This can be a limitation if some of the functions to be estimated are quite smooth while others are more complicated. Future extensions could consider a data-adaptive approach to selection of the degrees of freedom for each function, or the utilization of other flexible structures that are both computationally fast and do not require pre-specification of the functional forms of the exposure-response functions.

Acknowledgements

This publication was made possible by USEPA grant RD-835872-01. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication. This work was also supported by grants from the National Institutes of Health (ES000002, ES024332, ES007142).

A Derivation of empirical Bayes variance

As seen in Casella (2001) the Monte Carlo EM algorithm treats the parameters as missing data and iterates between finding the "complete data" log likelihood where the parameter values are filled in with their expectations from the gibbs sampler, and maximizing this expression as a function of the hyperparameter values. After removing the terms not involving $\sigma_{\boldsymbol{\beta}}$ the "complete data" log likelihood can be written as

$$-\frac{|\boldsymbol{\beta}|_0}{2}\log(\sigma^2_{\boldsymbol{\beta}}) - \sum_{S:\beta_S\neq 0}\frac{(\beta_S-\mu_{\beta_S})^2}{2\sigma^2_{\boldsymbol{\beta}}}, \tag{14}$$

where μ_{β_S} is the prior mean for coefficient β_S, and $|\boldsymbol{\beta}|_0$ is the number of nonzero regression parameters. The estimate of $\sigma^2_{\boldsymbol{\beta}}$ at iteration m uses the posterior samples from iteration $m-1$, which were sampled with $\sigma^2_{\boldsymbol{\beta}} = \sigma^{2\,(m-1)}_{\boldsymbol{\beta}}$. Adopting the same notation that is typically used with the EM algorithm we can take the expectation of this expression, where expectations are taken as the averages over the previous iterations posterior samples

$$Q(\sigma^2_{\boldsymbol{\beta}}|\sigma^{2\,(m-1)}_{\boldsymbol{\beta}}) = -\frac{E_{\sigma^{2\,(m-1)}_{\boldsymbol{\beta}}}[|\boldsymbol{\beta}|_0]}{2}\log(\sigma^2_{\boldsymbol{\beta}}) - \frac{E_{\sigma^{2\,(m-1)}_{\boldsymbol{\beta}}}\left[\sum_{S:\beta_S\neq 0}(\beta_S-\mu_{\beta_S})^2\right]}{2\sigma^2_{\boldsymbol{\beta}}} \tag{15}$$

And now we can take the derivative of this expression with respect to $\sigma^2_{\boldsymbol{\beta}}$ to find that the maximum occurs at

$$\sigma^{2\,(m)}_{\boldsymbol{\beta}} = \frac{E_{\sigma^{2\,(m-1)}_{\boldsymbol{\beta}}}\left[\sum_{S:\beta_S\neq 0}(\beta_S-\mu_{\beta_S})^2\right]}{E_{\sigma^{2\,(m-1)}_{\boldsymbol{\beta}}}[|\boldsymbol{\beta}|_0]} \tag{16}$$

Intuitively, to obtain the empirical Bayes estimate, one must run an MCMC chain and update $\sigma^2_{\boldsymbol{\beta}}$, every T MCMC iterations. This process is continued until the estimate of $\sigma^2_{\boldsymbol{\beta}}$ converges, and then the MCMC can be run conditioning on this estimated value.

B Results about prior shrinkage

Here we will investigate the prior probability that a particular interaction term is equal to zero. Our goal is to penalize interactions of higher order, and the probability that we force them to zero is one aspect of this. Since the covariates are exchangeable a priori, we will just investigate the probability that $\zeta_{1k'} = \zeta_{2k'} = \cdots = \zeta_{jk'} = 1$ and $\zeta_{j+1k'} = \cdots = \zeta_{pk'} = 0$.

$$p(\zeta_{1k'} = \zeta_{2k'} = \cdots = \zeta_{jk'} = 1, \zeta_{j+1k'} = \cdots = \zeta_{pk'} = 0|\tau_1,\ldots,\tau_k) =$$

$$\prod_{h=1}^{k} \left(p(\text{any}(\zeta_{1h}, \ldots, \zeta_{jh}) = 0) + p(\text{all}(\zeta_{1h}, \ldots, \zeta_{jh}) = 1 \text{ and any } (\zeta_{(j+1)h}, \ldots, \zeta_{ph}) = 1)\right)$$

$$= \prod_{h=1}^{k} \left((1 - \tau_h^j) + \tau_h^j (1 - (1 - \tau_h)^{(p-j)})\right)$$

$$= \prod_{h=1}^{k} \left(1 - \tau_h^j (1 - \tau_h)^{(p-j)}\right)$$

So now we can take the expectation over the τ_h to get the unconditional probability

$$= E\left[\prod_{h=1}^{k} \left[1 - \tau_h^j (1 - \tau_h)^{(p-j)}\right]\right]$$

$$= \prod_{h=1}^{k} \left(E_{\tau_h}\left[1 - \tau_h^j (1 - \tau_h)^{(p-j)}\right]\right)$$

$$= \prod_{h=1}^{k} \left(1 - \int_0^1 \tau_h^j (1 - \tau_h)^{(p+\gamma-j-1)} d\tau_h\right)$$

$$= \prod_{h=1}^{k} \left(1 - \frac{\Gamma(j+1)\Gamma(p+\gamma-j)}{\Gamma(p+\gamma+1)}\right)$$

$$= \left(1 - \frac{\Gamma(j+1)\Gamma(p+\gamma-j)}{\Gamma(p+\gamma+1)}\right)^k$$

B.1 Probability increases as j increases if $\gamma \geq p$

Now it is of interest to see how this changes as a function of j, the order of interaction. Ideally we would want to show that the probability increases as a function of j. We don't expect this to happen in general, but we can find out conditions at which it does hold. To simplify expressions with the gamma function we will restrict γ to be an integer, though we expect the results to extend to all real values of γ.

The expression increasing as a function of j is equivalent to $\Gamma(j+1)\Gamma(p+\gamma-j)$ decreasing with j, so we will show that $\Gamma(j+1)\Gamma(p+\gamma-j)$ is a decreasing function of j under certain conditions.

Let $f(j) = \Gamma(j+1)\Gamma(p+\gamma-j) = j!(p+\gamma-j-1)!$ and let us examine $f(j)/f(j+1)$.

$$\begin{aligned}\frac{f(j)}{f(j+1)} &= \frac{j!(p+\gamma-j-1)!}{(j+1)!(p+\gamma-j-2)!} \\ &= \frac{p+\gamma-j-1}{j+1}\end{aligned}$$

We want this ratio to be greater than 1 for all $j \in \{1, \ldots, p-1\}$ for the function to be increasing. Clearly this ratio is getting smaller as j gets larger so we can look at the largest value of j, which is $p-1$.

$$\frac{f(p-1)}{f(p)} = \frac{\gamma}{p}$$

So clearly if we set $\gamma \geq p$, we have the desired result.

References

Bayley, Nancy. 1993. *Bayley scales of infant development: Manual*. Psychological Corporation.

Bengio, Yoshua, Ducharme, Réjean, Vincent, Pascal, & Jauvin, Christian. 2003. A neural probabilistic language model. *Journal of machine learning research*, **3**(Feb), 1137–1155.

Bien, Jacob, Taylor, Jonathan, & Tibshirani, Robert. 2013. A lasso for hierarchical interactions. *Annals of statistics*, **41**(3), 1111.

Bobb, Jennifer F., Valeri, Linda, Claus Henn, Birgit, Christiani, David C., Wright, Robert O., Mazumdar, Maitreyi, Godleski, John J., & Coull, Brent A. 2014. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, **16**(3), 493–508.

Braun, Joseph M., Gennings, Chris, Hauser, Russ, & Webster, Thomas F. 2016. What can epidemiological studies tell us about the impact of chemical mixtures on human health? *Environmental Health Perspectives*, **124**(1), A6–A9.

Breiman, Leo. 2001. Random forests. *Machine learning*, **45**(1), 5–32.

Carlin, Danielle J, Rider, Cynthia V, Woychik, Rick, & Birnbaum, Linda S. 2013. Unraveling the health effects of environmental mixtures: an NIEHS priority. *Environmental health perspectives*, **121**(1), a6.

Casella, George. 2001. Empirical bayes gibbs sampling. *Biostatistics*, **2**(4), 485–500.

Cristianini, Nello, & Shawe-Taylor, John. 2000. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press.

Gelman, Andrew, Jakulin, Aleks, Pittau, Maria Grazia, & Su, Yu-Sung. 2008. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 1360–1383.

Gelman, Andrew, Carlin, John B, Stern, Hal S, & Rubin, Donald B. 2014. *Bayesian data analysis*. Vol. 2. Chapman & Hall/CRC Boca Raton, FL, USA.

George, Edward I, & McCulloch, Robert E. 1993. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**(423), 881–889.

Ghahramani, Zoubin, Griffiths, Thomas L, & Sollich, Peter. 2007. Bayesian nonparametric latent feature models.

Griffiths, Thomas L, & Ghahramani, Zoubin. 2005. Infinite latent feature models and the Indian buffet process. *Pages 475–482 of: NIPS*, vol. 18.

Guo, Fangjian, Wang, Xiangyu, Fan, Kai, Broderick, Tamara, & Dunson, David B. 2016. Boosting Variational Inference. *arXiv preprint arXiv:1611.05559*.

Hao, Ning, Feng, Yang, & Zhang, Hao Helen. 2014. Model Selection for High Dimensional Quadratic Regression via Regularization. **85721**, 26.

Henn, Birgit Claus, Ettinger, Adrienne S, Schwartz, Joel, Téllez-Rojo, Martha María, Lamadrid-Figueroa, Héctor, Hernández-Avila, Mauricio, Schnaas, Lourdes, Amarasiriwardena, Chitra, Bellinger, David C, Hu, Howard, *et al.* . 2010. Early postnatal blood manganese levels and childrens neurodevelopment. *Epidemiology (Cambridge, Mass.)*, **21**(4), 433.

Henn, Birgit Claus, Coull, Brent A, & Wright, Robert O. 2014. Chemical mixtures and childrens health. *Current opinion in pediatrics*, **26**(2), 223.

Lim, Michael, & Hastie, Trevor. 2015. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, **24**(3), 627–654.

Miller, Andrew C, Foti, Nicholas, & Adams, Ryan P. 2016. Variational Boosting: Iteratively Refining Posterior Approximations. *arXiv preprint arXiv:1611.06585*.

O'Hagan, Anthony, & Kingman, JFC. 1978. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–42.

Qamar, Shaan, & Tokdar, ST. 2014. Additive Gaussian Process Regression. *arXiv preprint arXiv:1411.7009*, 28.

Radchenko, Peter, & James, Gareth M. 2010. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, **105**(492), 1541–1553.

Reich, Brian J, Storlie, Curtis B, & Bondell, Howard D. 2009. Variable selection in Bayesian smoothing spline ANOVA models: Application to deterministic computer codes. *Technometrics*, **51**(2), 110–120.

Ročková, Veronika, & George, Edward I. 2016. The spike-and-slab lasso. *Journal of the American Statistical Association*.

Scheipl, Fabian, Fahrmeir, Ludwig, & Kneib, Thomas. 2012. Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association*, **107**(500), 1518–1532.

Shively, Thomas S, Kohn, Robert, & Wood, Sally. 1999. Variable selection and function estimation in additive nonparametric regression using a data-based prior. *Journal of the American Statistical Association*, **94**(447), 777–794.

Taylor, Kyla W, Joubert, Bonnie R, Braun, Joe M, Dilworth, Caroline, Gennings, Chris, Hauser, Russ, Heindel, Jerry J, Rider, Cynthia V, Webster, Thomas F, & Carlin, Danielle J. 2016. Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: lessons from an innovative workshop. *Environmental health perspectives*, **124**(12), A227.

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

Watanabe, Sumio. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**(Dec), 3571–3594.

Wood, Sally, Kohn, Robert, Shively, Tom, & Jiang, Wenxin. 2002. Model selection in spline nonparametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(1), 119–139.

Wood, Simon N. 2006. Low-Rank Scale-Invariant Tensor Product Smooths for Generalized Additive Mixed Models. *Biometrics*, **62**(4), 1025–1036.

Yau, Paul, Kohn, Robert, & Wood, Sally. 2003. Bayesian variable selection and model averaging in high-dimensional multinomial nonparametric regression. *Journal of Computational and Graphical Statistics*, **12**(1), 23–54.

Zhao, Peng, Rocha, Guilherme, & Yu, Bin. 2009. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 3468–3497.

Zhou, Jing, Bhattacharya, Anirban, Herring, Amy H., & Dunson, David B. 2014. Bayesian factorizations of big sparse tensors. *Journal of the American Statistical Association*, **1459**(1994), 00–00.

Zou, Hui. 2006. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, **101**(476), 1418–1429.

Zou, Hui, & Hastie, Trevor. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.