Published in final edited form as:
Ann Appl Stat. 2020 December ; 14(4): 1743–1758. doi:10.1214/20-aoas1363.

IDENTIFYING MAIN EFFECTS AND INTERACTIONS AMONG EXPOSURES USING GAUSSIAN PROCESSES

FEDERICO FERRARI[*], **DAVID B. DUNSON**[†]
Department of Statistical Science, Duke University

Abstract

This article is motivated by the problem of studying the joint effect of different chemical exposures on human health outcomes. This is essentially a nonparametric regression problem, with interest being focused not on a black box for prediction but instead on selection of main effects and interactions. For interpretability 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 nonidentifiability 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. The proposed MixSelect framework is evaluated using a simulation study, and is illustrated using data from the National Health and Nutrition Examination Survey (NHANES). Code is available on GitHub.

Keywords

Bayesian modeling; chemical mixtures; Gaussian process; interaction selection; semiparametric; strong heredity; variable selection

[*] ff31@duke.edu;. [†] dunson@duke.edu .

SUPPLEMENTARY MATERIAL

Supplementary Material to “Nonparametric Bayesian multi-armed bandits for single cell experiment design” (DOI: 10.1214/20-AOAS1363SUPPA;.pdf). The supplementary material contains a description of the predictive distribution, a comparison with P-splines, Figures and Tables relative to Sections 4 and 5, the algorithm for sampling the parameters of (5.1) and an extension of the analysis in Section 5 with sex or non-Hispanic Black ethnicity interactions.

Source code “Nonparametric Bayesian multi-armed bandits for single cell experiment design” (DOI: 10.1214/20-AOAS1363SUPPB;.zip). R source code for Algorithm 1 and Algorithm 2 of the Supplementary Material (Ferrari and Dunson (2020a)).

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, Gennings and Hauser (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 cooccur 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 (2009)).

Building a flexible joint model for mixtures of chemicals is suggested by the National Research Council (Mauderly et al. (2010), National Research Council et al. (2004), Vedal and Kaufman (2011)). Recently, several studies have shown relationships between complex mixtures of chemicals and health or behavior outcomes. For example, Sanders, Claus Henn and Wright (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 (Mortimer et al. (2008), Sinisi and van der Laan (2004)). 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-methods (Hu et al. (2008), Lampa et al. (2014)), to Bayesian Kernel Machine Regression (BKMR) (Bobb et al. (2015), Liu et al. (2018), Valeri 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 ~10, which is common in environmental epidemiology; refer to Section 2 of the Supplementary Material (Ferrari and Dunson (2020a)) for additional details.

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, 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 2.1. 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, Vannucci and Sha (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, and applying a dimension reduction approach to the Gaussian process surface (Banerjee, Dunson and Tokdar (2013), Guan and Haran (2018)), which we describe in Section 3.

We describe our efficient Bayesian inference procedure in Section 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 4. Finally, in Section 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. MixSelect modeling framework.

Let y_i denote a continuous health outcome for individual i, let $x_i = (x_{i1}, \dots, x_{ip})^T$ denote a vector of “exposure” measurements, and let $z_i = (z_{i1}, \dots, z_{iq})^T$ 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:

$$\begin{aligned} y_i &= x_i^T\beta + \sum_{j=1}^{p}\sum_{k>j}^{p}\lambda_{jk}x_{ij}x_{ik} + g^*(x_i) + z_i^T\alpha + \epsilon_i, \quad \epsilon_i \sim N\left(0, \sigma^2\right), \\ g_n^* &= Pg_n, \quad g \sim \mathrm{GP}(0, c), \end{aligned} \tag{2.1}$$

where $\beta = (\beta_1, \dots, \beta_p)^T$ are linear main effects of exposures, $\lambda = \{\lambda_{jk}\}$ are pairwise linear interactions, $g_n = [g(x_1), \dots, g(x_n)]$ is a nonparametric deviation and $\alpha = (\alpha_1, \dots, \alpha_q)^T$ 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. 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 2.1. The notation 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 departures from linearity; in particular, we let

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

where ρ_j is a smoothness parameter specific to the jth exposure and τ^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. (2015). 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 jth 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 3, but we note that the projection adjustment for identifiability greatly aids mixing of the MCMC; our code can be run efficiently for the numbers of exposures typically encountered in environmental epidemiology studies (up to 100). Code for implementation is available at https://github.com/fedfer/MixSelect and in the Supplementary Material (Ferrari and Dunson (2020a)).

2.1. Nonidentifiability 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 $P_X = X(X^T X)^{-1} X^T$. We define $P = P\frac{\perp}{X} = I_n - P_X$ and set $g_n^* = P g_n$. First, notice that the projection has an effect on the variance of the generated nonlinear effects; in particular,

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

This follows from

$$\begin{aligned}\left(g_n^*\right)^T g_n^* &= \left[(I_n - P_X)g_n\right]^T\left[(I_n - P_X)g_n\right] \\ &= g_n^T g_n - \left(P_X g_n\right)^T\left(P_X g_n\right) \leq g_n^T g_n.\end{aligned}$$

Figure 1 in the Supplementary Material (Ferrari and Dunson (2020a)) 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 nonprojected 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 3 consists in integrating out the nonlinear effects.

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 γ_k the indicator variable that is equal to 1 if the kth 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 γ_k we assume independent Bernoulli priors with success probability π. We endow π with a Beta distribution with parameters (a_π, b_π). The prior expected number of predictors included in the model is $p\frac{a_\pi}{a_\pi + b_\pi}$ which can be used to elicitate the hyperparameters (a_π, b_π). As a default we choose $a_\pi = b_\pi = 1$ which corresponds to a uniform distribution on π. We endow the main effects of covariate adjustments α_l with a normal prior $N_q(0, I)$, for $l = 1, \ldots, q$.

We impose a 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, Taylor and Tibshirani (2013) and Haris, Witten and Simon (2016) or two-stage approaches as Hao, Feng and Zhang (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 suffices to have one main effect in the model in order to estimate the interaction of the corresponding variables. Formally:

$$\begin{aligned} &S: \quad \lambda_{j,k} \Big| \gamma_j = \gamma_k = 1 \sim N(0,1), \qquad \lambda_{j,k} \Big| \left(\gamma_j = \gamma_k = 1\right)^C \sim \delta_0, \\ &W: \quad \lambda_{j,k} \Big| \left(\gamma_j = \gamma_k = 0\right)^C \sim N(0,1), \qquad \lambda_{j,k} \Big| \gamma_j = \gamma_k = 0 \sim \delta_0, \end{aligned}$$

where S and W stand for strong and weak heredity, respectively, and δ_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 jth covariate has a low effect on the outcome, but the interaction with the kth 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 $\sum_{i=0}^{p} \binom{p}{i} 2^{pi - i(i+1)/2}$ 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 τ with a spike and slab prior, that is, $\tau \sim \gamma^{\tau} F_{\tau}(\cdot) + (1 - \gamma^{\tau})\delta_0$, where $F_{\tau}(\cdot)$ is a gamma distribution with parameters (1⁄2, 1⁄2) and γ^{τ} 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 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, Vannucci and Sha (2011), which is also employed by Bobb et al. (2015), and endow the smoothness parameters $\rho_1, \ldots, \rho_p$ with independent spike and slab priors. In particular, $\rho_k \sim \gamma^{\tau} \gamma_k^{\rho} F_{\rho}(\cdot) + (1 - \gamma^{\tau})\left(1 - \gamma_k^{\rho}\right)\delta_0$, where $F_{\rho}(\cdot)$ is a gamma distribution with parameters (1⁄2, 1⁄2). Only when γ^{τ} is different than zero, we allow the covariate specific nonlinear effects γ_j^{ρ} to be different than zero. When $\gamma_k^{\rho} = 0$, the kth exposure is eliminated from the nonparametric term g in (2.1). As before, we choose a Bernoulli prior for γ_k^{ρ} with mean φ, and we endow φ with a Beta prior with parameters $(a_{\varphi}, b_{\varphi})$. As a default we choose a_{φ} b_{φ} 1 which corresponds to a Uniform distribution on φ. A graphical representation of the model can be found in Figure 1.

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, β, λ and α, in Section 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 nonzero main effects, and we set to zero all the others. Given each of the elements of β, λ and α, we can update the labels γ with a Bernoulli draw. We also reparametrize the model setting $\tau = \tau * \sigma$, so that we can directly update σ^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

$$y \Big| \beta, \Lambda, c \sim N\left(X\beta + \mathrm{diag}\left(X \Lambda X^T\right) + \alpha Z, \sigma^2 I_n + PcP^T\right), \tag{3.1}$$

where Λ 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 ρ_j, for $j = 1, \ldots, p$, that define the variable selection scheme for the nonlinear effects. The priors for the smoothness parameters ρ_j and τ^2 defined in Section 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 (3.1) 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, Dunson and Tokdar (2013) and uses random projections to find an approximation of the Eigen Decomposition of PcP^T. In particular, we approximate this matrix as $U_m D_m U_m^T$, 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 $\Sigma = \sigma^2 I_n + PcP^T$,

$$\begin{aligned}\Sigma^{-1} &= \left(\sigma^2 I_n + PcP^T\right)^{-1} \approx \left(\sigma^2 I_n + U_m D_m U_m^T\right)^{-1} = \\ &= \frac{1}{\sigma^2}\left(I_n + U_m\left(\sigma^2 D_m + U_m^T U_m\right)^{-1} U_m^T\right)\end{aligned}$$

which now involves the inversion of an $m \times m$ matrix. Similarly, we can simplify the computations for the determinant of Σ using the determinant lemma (Harville (1997)),

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

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 ρ_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.

4. Simulations.

In this section we compare the performance of our model with respect to five other methods: BKMR (Bobb et al. (2015)), Family (Haris, Witten and Simon (2016)), hierNet (Bien, Taylor and Tibshirani (2013)), PIE (Wang and Jiang (2019)) and RAMP (Hao, Feng and Zhang (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

```
Algorithm 1 MCMC algorithm for sampling the parameters of model (2.1)
  Step 1 Sample $\gamma_j$ for $j = 1, \ldots, p$ from
      $\pi(\gamma_j \mid \cdot) \sim \text{Bernoulli}\left(\frac{1}{1 + \frac{1-\pi}{\pi} R_j}\right),$
  where $R_j = \frac{\left|X_{0j}^T \Sigma^{-1} X_{0j} + I\right|^{-1/2} \exp\left(\frac{1}{2} m_0^T V_0 m_0\right)}{\left|X_{1j}^T \Sigma^{-1} X_{1j} + I\right|^{-1/2} \exp\left(\frac{1}{2} m_1^T V_1 m_1\right)}$, $\Sigma = \sigma^2 I_n + PcP^T$, $m_0 = X_{0j}^T \Sigma^{-1} y$ and $V_0 =$
  $\left(X_{0j}^T \Sigma^{-1} X_{0j}^T + I\right)^{-1}$. $X_{0j}$ is the matrix of covariates such that $\gamma_k = 1$ for $k \neq j$. $X_{1j}$ is the
  matrix of covariates such that $\gamma_k = 1$ for $k = 1, \ldots, p$, with $X_i$ included.
  Step 2 Sample $\pi$ from $\pi(\pi \mid \cdot) \sim \text{Beta}\left(a_\pi + \sum_{j=1}^{T} \gamma_j, b_\pi + p - \sum_{j=1}^{p} \gamma_j\right)$
  Step 3 Sample the main coefficients $\beta_\gamma$ from the distribution:
      $\pi(\beta_\gamma \mid \cdot) \sim N\left(V X_\gamma^T \Sigma^{-1} \left(y - \alpha Z - \text{diag}\left(X \Lambda X^T\right)\right), V\right),$
  where $V = \left(X_\gamma \Sigma^{-1} X_\gamma + I\right)^{-1}$ 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 $\gamma_j^\rho = 0$ 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 $U_m^* D^* U_m^{*T}$ with the
  approximation of Section 3, $\Sigma^{*-1}$ with Sherman – Woodbury formula and $\left|\Sigma^{*-1}\right|$ with de-
  terminant lemma. Then compute
      $-2\log(r) = \log\left|\Sigma^{*-1}\right| - \log\left|\Sigma^{-1}\right| + \frac{1}{2}\mu^T\left(\Sigma^{*-1} - \Sigma^{-1}\right)\mu,$
  where $\mu = y - \left(Z\alpha + X\beta + \text{diag}\left(X \Lambda X^T\right)\right)$. Sample $u$ from a Uniform distribution in the
  interval $(0, 1)$ and if $\log(r) > \log(u)$, set $\rho_j = \rho_j^*, \gamma_j$, $\Sigma = \Sigma^*$, $\left|\Sigma^{-1}\right| = \left|\Sigma^{*-1}\right|$
  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 appro-
  priate modification of Step 5. If $\tau^{*2} \neq 0$ perform a Gibbs – type move.
  Step 10 Sample $\sigma^2$ from $\pi\left(\sigma^2 \mid \cdot\right) \sim \text{InvGamma}\left(\frac{1+n}{2}, \frac{1 + \mu^T\left(I_n + Pc' P^T\right)^{-1}\mu}{2}\right)$ where
  $c'(x, x^*) = (\tau^*)^2 \exp\left\{\sum_{j=1}^{p} \rho_j \left(x_j - x_j^*\right)^2\right\}$
```

settings. We generate the covariates independently $X_i \sim N_p(0, I_p)$ for $i = 1, \ldots, n$, 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. $$y_i = \sin(x_1 + 3x_3) - \frac{1}{2}x_3^2 + \exp(0.1 * x_1) + \epsilon_i,$$

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 Λ and $\widehat{\Lambda}$ of dimension p is defined as

$$\sqrt{\text{trace}\left(\left(\Lambda - \widehat{\Lambda}\right)^T\left(\Lambda - \widehat{\Lambda}\right)\right)}.$$

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 $n = 500$ are summarized in Table 1 and Table 2 and are summarized for $n = 250$ in Table 1 and Table 2 of the Supplementary Material (Ferrari and Dunson (2020a)).

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 τ. We also achieve a similar, if not better performance, in the nonlinear scenario of method (c). Finally, Figure 2 of the Supplementary Material (Ferrari and Dunson (2020a)) shows the estimated regression surface vs. the true surface for model (a), when $n = 250$ and $p = 25$.

5. Environmental epidemiology application.

5.1. Motivation.

The goal of our analysis is to assess the association of 14 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, Claus Henn and Wright (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, Coull and Wright (2014) report associations between mixtures and pediatric health outcomes, cognition, reproductive hormone levels and neurodevelopment. With respect to obesity indices and using data from the National Health and Nutrition Examination Survey (NHANES), metals have already been associated with an increase in waist circumference and BMI; see Padilla et al. (2010) and Shao et al. (2017).

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 Etnicity). We choose Hispanic as a reference group for ethnicity. Table 3 in the Supplementary Material (Ferrari and Dunson (2020a)) shows the correlations among chemicals; Figure 3 and Figure 4 in the Supplementary Material (Ferrari and Dunson (2020a)) 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 (Buman et al. (2013), Lynch et al. (2010), Nagelkerke et al. (2006)). We leave these transformations implicit for the remainder of the section.

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:

$$\begin{aligned} &W_i = \Lambda \eta_i + \epsilon_i, \quad \epsilon_i \sim N_d(0, \Sigma), \\ &\eta_i \sim N_k(0, I), \end{aligned} \tag{5.1}$$

where we center the data W_i to have zero mean prior to the analysis, $\Sigma = \mathrm{diag}\left(\sigma_1^2, \ldots, \sigma_d^2\right)$ is as residual variance matrix, Λ is a $d \times k$ factor loadings matrix and η_i are i.i.d. standard normal latent factors. We assume an elementwise standard normal prior for Λ and endow σ^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 *nine* eigenvectors explain more than 85% of the total variability; hence, we set the number of factors equal to 9. Algorithm 2 in the Supplementary Material (Ferrari and Dunson (2020a)) describes how to sample the parameters of (5.1) 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 these observations as

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

where LOD_j is the limit of detection for exposure j and $TN\left(\mu, \sigma^2, a, b\right)$ is a truncated normal distribution with mean μ, variance σ^2 and support in $[a, b]$. A related approach was used in Ferrari and Dunson (2020b) 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.

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 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 ρ_j was on average *three* times higher with respect to the corresponding parameters in BKMR. 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 5 of the Supplementary Material (Ferrari and Dunson (2020a)). The residual diagnostics suggest that the model assumptions are satisfied fairly well. First, approximate normality holds with only a mild deviation in the tails. Second, 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 6 in the Supplementary Material (Ferrari and Dunson (2020a)) 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 α values; refer to Table 4 in Supplementary Material (Ferrari and Dunson (2020a)).

The complexity per iteration of Gibbs sampling is $\mathcal{O}\left(n^2 m\right)$ when $\tau \neq 0$, where m is related to the approximation described in Section 3. When $\tau = 0$, the complexity per iteration of Gibbs sampling is $\mathcal{O}\left(d^2\right)$, where d is the number of active main effects.

5.5. Results.

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 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 among 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 7 in the Supplementary Material (Ferrari and Dunson (2020a)) 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 8 of the Supplementary Material (Ferrari and Dunson (2020a)). Finally, even if some of the chemicals were moderately correlated (see molybdenum and tungsten, e.g., in Table 3 in the Supplementary Material, Ferrari and Dunson (2020a)), 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 4: BKMR (Bobb et al. (2015)), Family (Haris, Witten and Simon (2016)), hierNet (Bien, Taylor and Tibshirani (2013)), PIE (Wang and Jiang (2019)) and RAMP (Hao, Feng and Zhang (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 3 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 3 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. In the Supplementary Material (Ferrari and Dunson (2020a)) we consider possible chemical interactions with Sex and non-Hispanic Black ethnicity. The nonlinear effect of cadmium in Females and non-Hispanic Blacks has a hill-shaped dose response as in Figure 2, whereas it is negatively associated with BMI in the male subgroup. Moreover, we found that lead and molybdenum exposures have a stronger negative effect on females than males, and we observe the opposite behavior for tin and cobalt.

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, which 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 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, Pillai and Gelman (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. (2015)). 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 ϵ_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 Yu et al. (2013) for related work. Alternatively, it is possible to allow main effects and interactions to vary with quantiles of y_i; see, for example, Reich, Fuentes and Dunson (2011). We can also induce a quantile dependence on the nonlinear deviation $g^*(x_i)$. In particular, we can introduce uniformly distributed latent variables η_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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments.

This research was supported by grant 1R01ES028804–01 of the National Institute of Environmental Health Sciences of the United States Institutes of Health. The authors would like to thank Amy Herring, Phuc Nguyen and Alessandro Zito for helpful comments.

REFERENCES

BANERJEE A, DUNSON DB and TOKDAR ST (2013). Efficient Gaussian process regression for large datasets. Biometrika 100 75–89. MR3034325 10.1093/biomet/ass068 [PubMed: 23869109]

BIEN J, TAYLOR J and TIBSHIRANI R (2013). A LASSO for hierarchical interactions. Ann. Statist 41 1111–1141. MR3113805 10.1214/13-AOS1096

BOBB JF, VALERI L, CLAUS HENN B, CHRISTIANI DC, WRIGHT RO, MAZUMDAR M, GODLESKI JJ and COULL BA (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics 16 493–508. MR3365442 10.1093/biostatistics/kxu058 [PubMed: 25532525]

BRAUN JM, GENNINGS C and HAUSER R (2016). What can epidemiological studies tell us about the impact of chemical mixtures on human health? Environ. Health Perspect 124 A6–A9. [PubMed: 26720830]

BUMAN MP, WINKLER EAH, KURKA JM, HEKLER EB, BALDWIN CM, OWEN N, AINSWORTH BE, HEALY GN and GARDINER PA (2013). Reallocating time to sleep, sedentary behaviors, or active behaviors: Associations with cardiovascular disease risk biomarkers, NHANES 2005–2006. Am. J. Epidemiol 179 323–334. 10.1093/aje/kwt292 [PubMed: 24318278]

CHIPMAN H (1996). Bayesian variable selection with related predictors. Canad. J. Statist 24 17–36. MR1394738 10.2307/3315687

CLAUS HENN B, COULL BA and WRIGHT RO (2014). Chemical mixtures and children's health. Curr. Opin. Pediatr 26 223–229. [PubMed: 24535499]

DOMINICI F, PENG ROGER, BARR D, CHRISTOPHER D and BELL ML (2010). Protecting human health from air pollution: Shifting from a single-pollutant to a multi-pollutant approach. Epidemiology 21 187. [PubMed: 20160561]

FERRARI F and DUNSON DB (2020a). Supplement to “Identifying main effects and interactions among exposures using Gaussian processes.” 10.1214/20-AOAS1363SUPPA, 10.1214/20-AOAS1363SUPPB

FERRARI F and DUNSON DB (2020b). Bayesian factor analysis for inference on interactions. J. Amer. Statist. Assoc. To appear 10.1080/01621459.2020.1745813

GEORGE EI and MCCULLOCH RE (1997). Approaches for Bayesian variable selection. Statist. Sinica 7 339–373.

GUAN Y and HARAN M (2018). A computationally efficient projection-based approach for spatial generalized linear mixed models. J. Comput. Graph. Statist 27 701–714. MR3890863 10.1080/10618600.2018.1425625

HANKS EM, SCHLIEP EM, HOOTEN MB and HOETING JA (2015). Restricted spatial regression in practice: Geostatistical models, confounding, and robustness under model misspecification. Environmetrics 26 243–254. MR3340961 10.1002/env.2331

HAO N, FENG Y and ZHANG HH (2018). Model selection for high-dimensional quadratic regression via regularization. J. Amer. Statist. Assoc 113 615–625. MR3832213 10.1080/01621459.2016.1264956

HAO N and ZHANG HH (2014). Interaction screening for ultrahigh-dimensional data. J. Amer. Statist. Assoc 109 1285–1301. MR3265697 10.1080/01621459.2014.881741

HARIS A, WITTEN D and SIMON N (2016). Convex modeling of interactions with strong heredity. J. Comput. Graph. Statist 25 981–1004. MR3572025 10.1080/10618600.2015.1067217

HARVILLE DA (1997). Matrix Algebra from a Statistician’s Perspective Springer, New York. MR1467237 10.1007/b98818

HU W, MENGERSEN K, MCMICHAEL A and TONG S (2008). Temperature, air pollution and total mortality during summers in Sydney, 1994–2004. Int. J. Biometeorol 52 689–696. [PubMed: 18506490]

KUNDU S and DUNSON DB (2014). Latent factor models for density estimation. Biometrika 101 641–654. MR3254906 10.1093/biomet/asu019

LAMPA E, LIND L, LIND PM and BORNEFALK-HERMANSSON A (2014). The identification of complex interactions in epidemiology and toxicology: A simulation study of boosted regression trees. Environ. Health 13 57. 10.1186/1476-069X-13-57 [PubMed: 24993424]

LANG S and BREZGER A (2004). Bayesian P-splines. J. Comput. Graph. Statist 13 183–212. MR2044877 10.1198/1061860043010

LIU SH, BOBB JF, LEE KH et al. (2018). Lagged kernel machine regression for identifying time windows of susceptibility to exposures of complex mixtures. Biostatistics 19 325–341. MR3815175 10.1093/biostatistics/kxx036 [PubMed: 28968676]

LUNN D, BEST N, SPIEGELHALTER D, GRAHAM G and NEUENSCHWANDER B (2009). Combining MCMC with ‘sequential’ PKPD modelling. J. Pharmacokinet. Pharmacodyn 36 19. [PubMed: 19132515]

LYNCH BM, DUNSTAN DW, HEALY GN, WINKLER E, EAKIN E and OWEN N (2010). Objectively measured physical activity and sedentary time of breast cancer survivors, and associations with adiposity: Findings from NHANES (2003–2006). Cancer Causes Control 21 283–288. [PubMed: 19882359]

MACLEHOSE RF, DUNSON DB, HERRING AH and HOPPIN JA (2007). Bayesian methods for highly correlated exposure data. Epidemiology 18 199–207. 10.1097/01.ede.0000256320.30737.c0 [PubMed: 17272963]

MAUDERLY JL and SAMET JM (2009). Is there evidence for synergy among air pollutants in causing health effects? Environ. Health Perspect 117 1–6. 10.1289/ehp.11654 [PubMed: 19165380]

MAUDERLY JL, BURNETT RT, CASTILLEJOS M, OZKAYNAK H, SAMET JM, STIEB DM, VEDAL S and WYZGA RE (2010). Is the air pollution health research community prepared to support a multipollutant air quality management framework? Inhal. Toxicol 22 1–19. 10.3109/08958371003793846

MORTIMER K, NEUGEBAUER R, LURMANN F and ALCORN S (2008). Air pollution and pulmonary function in asthmatic children: Effects of prenatal and lifetime exposures. Epidemiology 19 550–557. [PubMed: 18520616]

NAGELKERKE NJD, BERNSEN RMD, SGAIER SK and JHA P (2006). Body mass index, sexual behaviour, and sexually transmitted infections: An analysis using the NHANES 1999–2000 data. BMC Public Health 6 199. [PubMed: 16884541]

NATIONAL RESEARCH COUNCIL et al.. (2004). Research Priorities for Airborne Particulate Matter: IV. Continuing Research Progress 4. National Academies Press.

PADILLA MA, ELOBEID M, RUDEN DM and ALLISON DB (2010). An examination of the association of selected toxic metals with total and central obesity indices: NHANES 99–02. Int. J. Environ. Res. Public Health 7 3332–3347. [PubMed: 20948927]

QI YA, MINKA TP, PICARD RW and GHAHRAMANI Z (2004). Predictive automatic relevance determination by expectation propagation. In Proceedings of the Twenty-First International Conference on Machine Learning 85 ACM, New York.

REICH BJ, FUENTES M and DUNSON DB (2011). Bayesian spatial quantile regression. J. Amer. Statist. Assoc 106 6–20. MR2816698 10.1198/jasa.2010.ap09237

ROBERTS S and MARTIN M (2005). A critical assessment of shrinkage-based regression approaches for estimating the adverse health effects of multiple air pollutants. Atmos. Environ 39 6223–6230.

SANDERS AP, CLAUS HENN B and WRIGHT RO (2015). Perinatal and childhood exposure to cadmium, manganese, and metal mixtures and effects on cognition and behavior: A review of recent literature. Curr. Environ. Health Rep 2 284–294. [PubMed: 26231505]

SAVITSKY T, VANNUCCI M and SHA N (2011). Variable selection for nonparametric Gaussian process priors: Models and computational strategies. Statist. Sci 26 130–149. MR2849913 10.1214/11-STS354

SHAO W, LIU Q, HE X, LIU H and GU A (2017). Association between level of urinary trace heavy metals and obesity among children aged 6–19 years: NHANES 1999–2011. Environ. Sci. Pollut. Res. Int 24 11573–11581. [PubMed: 28321702]

SI Y, PILLAI NS and GELMAN A (2015). Bayesian nonparametric weighted sampling inference. Bayesian Anal 10 605–625. MR3420817 10.1214/14-BA924

SINISI SE and VAN DER LAAN MJ (2004). Deletion/substitution/addition algorithm in learning with applications in genomics. Stat. Appl. Genet. Mol. Biol 3 Art. 18. MR2101467 10.2202/1544-6115.1069

VALERI L, MAZUMDAR MM, BOBB JF, CLAUS HENN B, RODRIGUES E, SHARIF OIA, KILE ML, QUAMRUZZAMAN Q, AFROZ S et al. (2017). The joint effect of prenatal exposure to metal mixtures on neurodevelopmental outcomes at 20–40 months of age: Evidence from rural Bangladesh. Environ. Health Perspect 125 067015. 10.1289/EHP614 [PubMed: 28669934]

VEDAL S and KAUFMAN JD (2011). What does multi-pollutant air pollution research mean? Am. J. Respir. Crit. Care Med 183 4–6. 10.1164/rccm.201009-1520ED [PubMed: 21193783]

WANG C and JIANG B (2019). Penalized interaction estimation for ultrahigh dimensional quadratic regression. Preprint Available at arXiv:1901.07147.

YU K and MOYEED RA (2001). Bayesian quantile regression. Statist. Probab. Lett 54 437–447. MR1861390 10.1016/S0167-7152(01)00124-9

YU K, CHEN CWS, REED C and DUNSON DB (2013). Bayesian variable selection in quantile regression. Stat. Interface 6 261–274. MR3066690 10.4310/SII.2013.v6.n2.a9

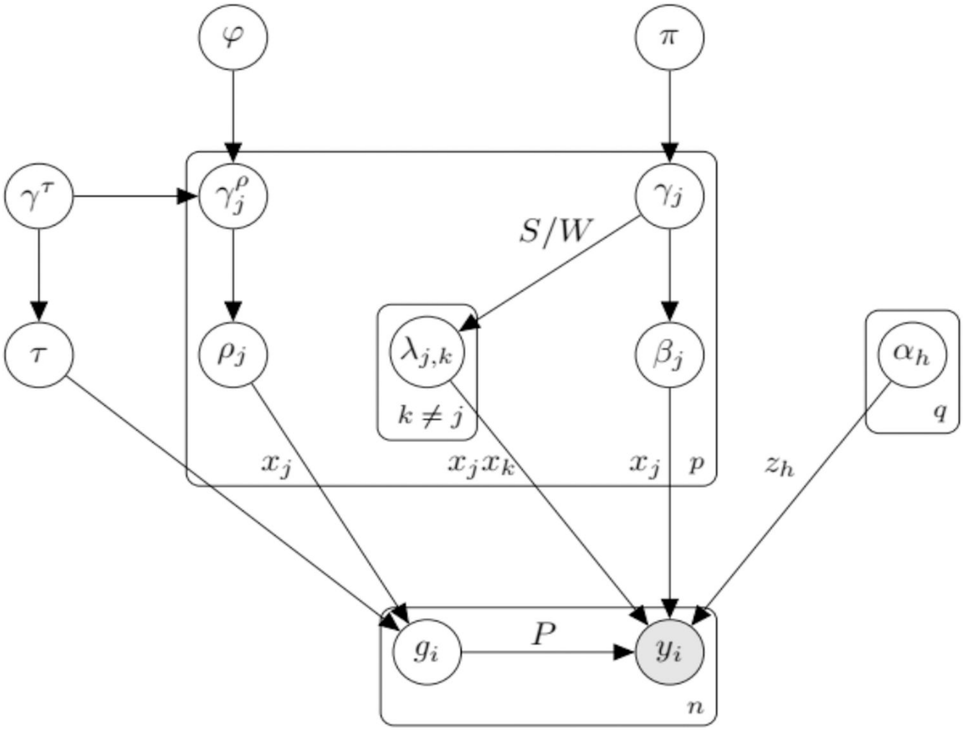

Fig. 1.

Graphical representation of the model. The arrows between two nodes indicate conditional dependence. Variables that are in the same plate share the same indices. S/W refers to strong or weak heredity.

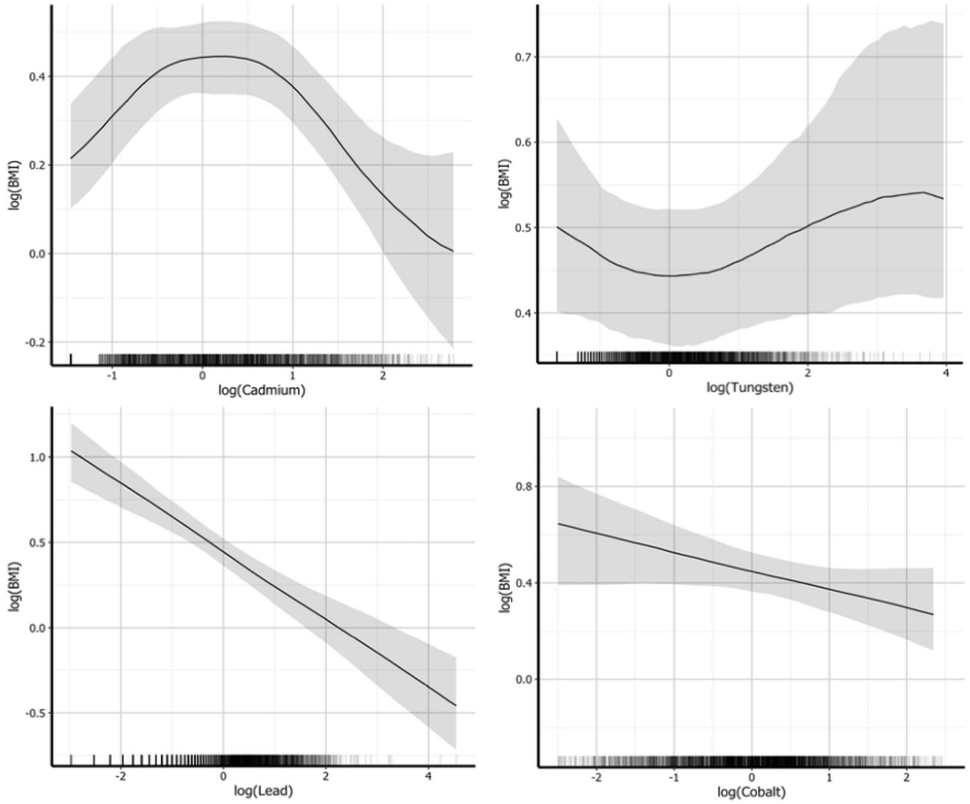

Fig. 2.

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.*

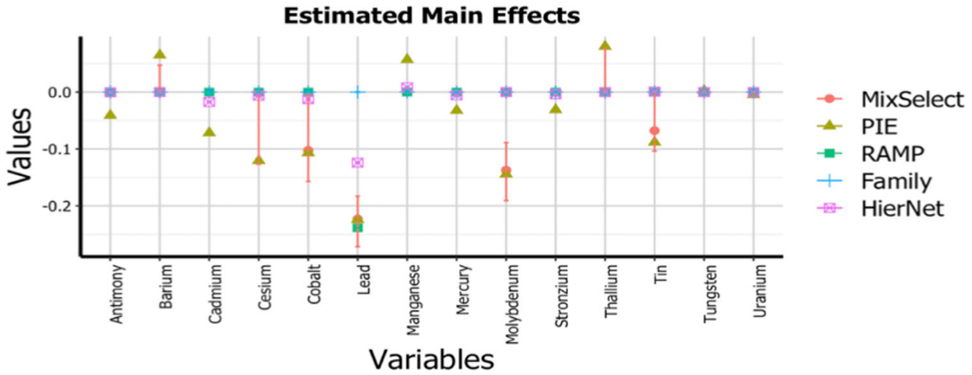

Fig. 3.

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.*

Table 1

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

		MixSelect	BKMR	hierNet	Family	PIE	RAMP
Model (a)	test MSE	1.138	1	1.098	5.645	4.400	1.217
	FR	1.033		5.659	5.820	2.465	1
	TP main	1		1	1	1	1
	TN main	0.758		0.798	0.947	0.679	0.919
	TP int	1		1	1	1	1
	TN int	1.000		0.989	0.984	0.997	0.997
	TP nl	0.947	1				
	TN nl	0.977	0.821				
Model (b)	test MSE	1	1.902	1.430	8.928	1.363	1.061
	FR	1		18.162	22.572	1.723	1.433
	TP main	1		1	1	1	1
	TN main	0.998		0.863	0.907	0.688	0.992
	TP int	1		1	0.978	1	0.989
	TN int	1		0.988	0.958	0.993	0.999
	TP nl	0.984	0.673				
	TN nl						
Model (c)	test MSE	1.359	1	1.203	2.927	1.285	2.641
	FR	1		8.759	2.508	9.600	5.542
	TP main						
	TN main	0.808		0.719	0.868	0.834	0.851
	TN int	1.000		0.984	0.980	0.992	0.991
	TP nl	0.645	0.985				
	TN nl	0.989	0.893				

Table 2

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

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

Table 3

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*

	MixSelect	BKMR	hierNet	Family	PIE	RAMP
In sample MSE	0.530	0.031	0.573	0.879	0.626	0.572
Out of sample MSE	0.687	0.919	0.611	0.927	0.710	0.604