Submitted to the Annals of Applied Statistics

SUPPLEMENTARY MATERIALS TO: IDENTIFYING MAIN EFFECTS AND INTERACTIONS AMONG EXPOSURES USING GAUSSIAN PROCESSES

By Federico Ferrari and David B. Dunson

Duke University

1. Predictive Distribution. Suppose we observe new data (X^*, Z^*) and our goal is to compute the predictive distribution. The predictive mean of y^* given X, y, Θ, where Θ is the vector containing all the parameters, is:

$$\mu^* + P^* c^* P^T (\sigma^2 I_n + PcP^T)^{-1}(y - \mu),$$

where $\mu = X\beta + diag(X\Lambda X) + \alpha Z$ and μ^* is equivalently defined for X^*, c^* is the covariance matrix such that element (i, j) is equal to $c(x_i^*, x_j)$ and P^* is the projection matrix on the column space of the matrix containing the new main effects and pairwise interactions.

2. Comparison with P-splines. In Lang and Brezger [2004], the authors propose to model interactions between x_j and x_s adding to the regression equation the term $f_{js}(x_j, x_s)$, which is an unknown surface approximated by the tensor product of B-splines:

$$f(x_{ij}, x_{ih}) = \sum_{\rho=1}^{m} \sum_{\nu=1}^{m} \beta_{js\rho\nu} B_{j\rho}(x_j) B_{s\nu}(x_s),$$

where m is the number of knots and $B_{..}(\cdot)$ are B-spline basis functions.

Bayesian P-splines require $\frac{p(p-1)}{2} m^2$ parameters to model nonlinear pairwise interactions, and Lang and Brezger [2004] propose to use $m = 20, 40, 80$. In the context of our application, p can be in the order of $[15, 100]$ and n is usually in the order of thousands, so that the estimation of interactions with Bayesian P-splines is extremely challenging. Conversely, our approach jointly models flexible nonlinear interactions between $x_{i1}, \ldots, x_{i1}$ with a Gaussian process and the selection of nonlinear effects is carried out with only $p+2$ parameters. Moreover, our model allows for interpretable linear interactions. To avoid estimating $\frac{p(p-1)}{2}$ parameters, we carefully employ the heredity structure to have a parsimonious interaction specification and estimation procedure.

3. Figures and Tables.

3.1. *Simulation.*

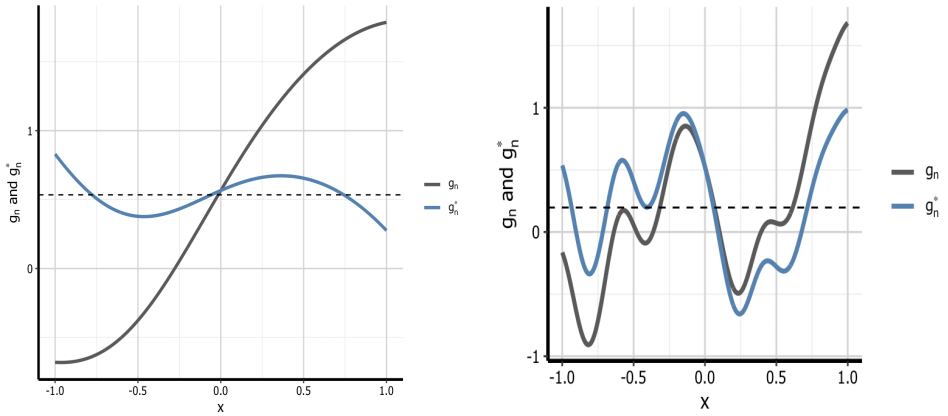

FIG 1. *Realizations of g_n in dark gray and $g_n^* = Pg_n$ in blue when $\rho = 1$ on the left and $\rho = 4$ on the right, the horizontal dashed line indicates the mean.*

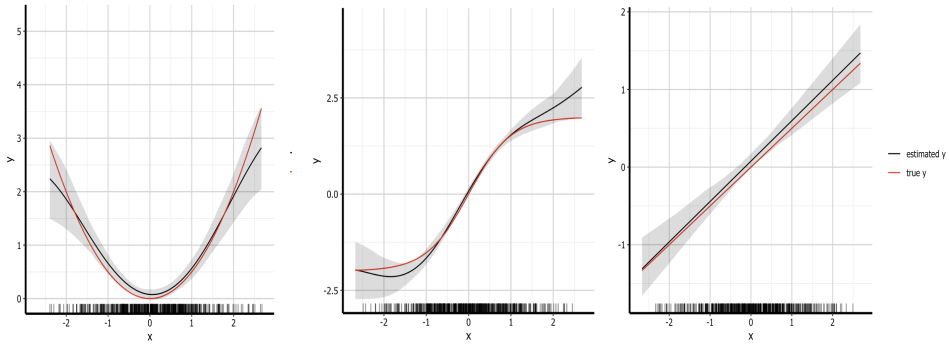

FIG 2. *Estimated regression surface of model (a) with $n = 250$ and $p = 25$. The red line indicates the true curve, the black line the estimated function, the grey bands the pointwise 99% posterior credible intervals for the mean function and the marks on the x-axis the data points in the training set.*

		MixSelect	BKMR	hierNet	Family	PIE	RAMP
model (b)	test MSE	1	1.178	1.444	3.299	1.001	1.193
	FR	1		2.962	7.234	4.027	2.652
	TP main	1		1	0.974	0.974	0.962
	TN main	0.941		0.865	0.869	0.692	0.921
	TP int	1		1	0.962	1	0.923
	TN int	1		0.993	0.946	0.996	0.997
	TP nl	0.596	0.981				
	TN nl	0.988	0.656				
model (b)	test MSE	1	1.882	2.872	8.762	2.081	1.598
	MSE beta	1		4.755	64.751	48.033	12.071
	FR	1		9.526	16.658	4.196	2.121
	TP main	1		1	1	0.974	0.954
	TN main	0.999		0.878	0.847	0.680	0.990
	TP int	1		0.901	0.967	0.921	0.947
	TN int	1		0.990	0.937	0.987	0.999
	TN nl	0.992	0.696				
model (c)	test MSE	1.301	1.050	1.000	2.244	1	2.442
	FR	1		10.400	3.047	13.246	4.715
	TN main	0.881		0.839	0.834	0.890	0.920
	TN int	1		0.991	0.970	0.993	0.995
	TP nl	0.500	0.898				
	TN nl	0.998	0.772				

TABLE 1

Results from the simulation study under the three scenarios with $p = 25$, $n = 250$. We computed test error, FR for interaction effects, percentage of true positives and true negatives for main effects and interactions for MixSelect, BKMR, hierNet, Family, PIE and RAMP. We divided each value of test error and FR by the best (lowest) result for that metric. This makes the metric of the best model equal to 1.

		MixSelect	BKMR	hierNet	Family	PIE	RAMP
model (a)	test MSE	1.011	4.057	1.361	3.210	1	1.153
	FR	1.099		2.284	5.113	2.879	1
	TP main	0.973		1	0.973	0.953	0.993
	TN main	0.961		0.927	0.946	0.794	0.964
	TP int	0.950		1	0.970	1	0.980
	TN int	1.000		0.998	0.991	0.999	1.000
	TP nl	0.540	1				
	TN nl	0.995	0.012				
model (b)	test MSE	1	13.501	2.826	9.174	2.383	1.314
	FR	1		9.055	16.709	4.637	1.466
	TP main	1		1	1	0.970	0.985
	TN main	0.998		0.903	0.897	0.810	0.997
	TP int	1		0.930	0.970	0.905	0.970
	TN int	1		0.997	0.974	0.996	1.000
	TN nl	0.991	0.022				
model (c)	test MSE	1.468	3.901	1	2.322	1.007	2.427
	FR	1		9.183	2.617	11.659	3.827
	TN main	0.939		0.902	0.898	0.936	0.966
	TN int	1.000		0.997	0.987	0.998	0.999
	TP nl	0.500	0.980				
	TN nl	0.999	0.020				

TABLE 2

Results from the simulation study under the three scenarios with $p = 50$, $n = 250$. We computed test error, FR for interaction effects, percentage of true positives and true negatives for main effects and interactions for MixSelect, BKMR, hierNet, Family, PIE and RAMP. We divided each value of test error and FR by the best (lowest) result for that metric. This makes the metric of the best model equal to 1.

3.2. *Environmental Epidemiology Application.*

	Ba	Cd	Co	Cs	Mo	Mn	Pb	Sb	Sn	Sr	TI	W	U
Ba													
Cd	0.16												
Co	0.56	0.29											
Cs	0.46	0.45	0.64										
Mo	0.4	0.26	0.62	0.64									
Mn	0.27	0.1	0.27	0.18	0.19								
Pb	0.41	0.53	0.51	0.61	0.5	0.22							
Sb	0.32	0.21	0.48	0.44	0.53	0.25	0.44						
Sn	0.27	0.23	0.42	0.43	0.46	0.18	0.43	0.46					
Sr	0.78	0.33	0.61	0.58	0.48	0.23	0.54	0.36	0.30				
TI	0.4	0.33	0.55	0.77	0.6	0.14	0.47	0.43	0.38	0.47			
W	0.35	0.06	0.49	0.47	0.66	0.2	0.35	0.49	0.39	0.35	0.42		
U	0.33	0.29	0.37	0.33	0.4	0.22	0.39	0.47	0.35	0.37	0.29	0.43	
Hg	0.11	0.38	0.12	0.3	0.18	0.04	0.26	0.12	0.13	0.18	0.28	0.08	0.15

TABLE 3

Correlation matrix between Barium (Ba), Cadmium (Cd), Cobalt (Co), Cesium (Cs), Molybdenum (Mo), Manganese (Mn), Lead (Pb), Antimony (Sb), Tin (Sn), Strontium (Sr), Thallium (TI), Tungsten (W), Uranium (U), Mercury (Hg) in the NHANES 2015 dataset.

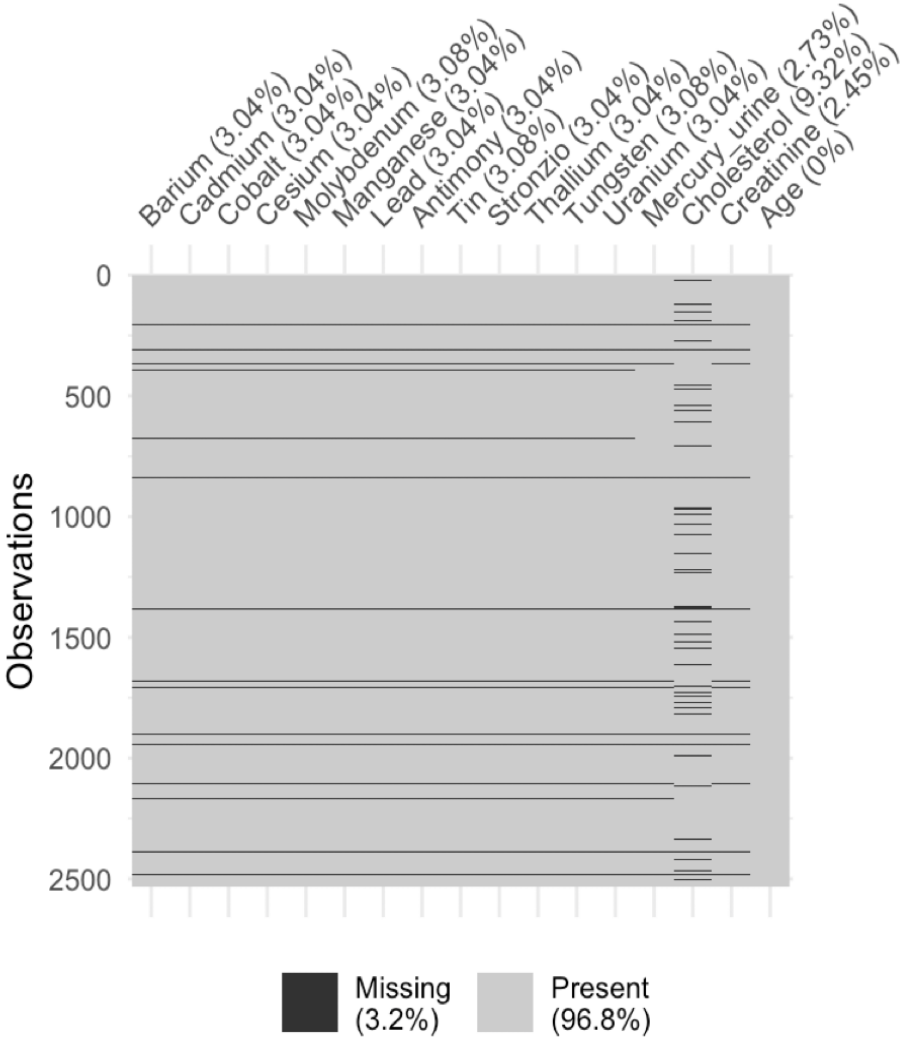

FIG 3. *Pattern of missingness in the chemical exposure, cholesterol and creatinine measurements.*

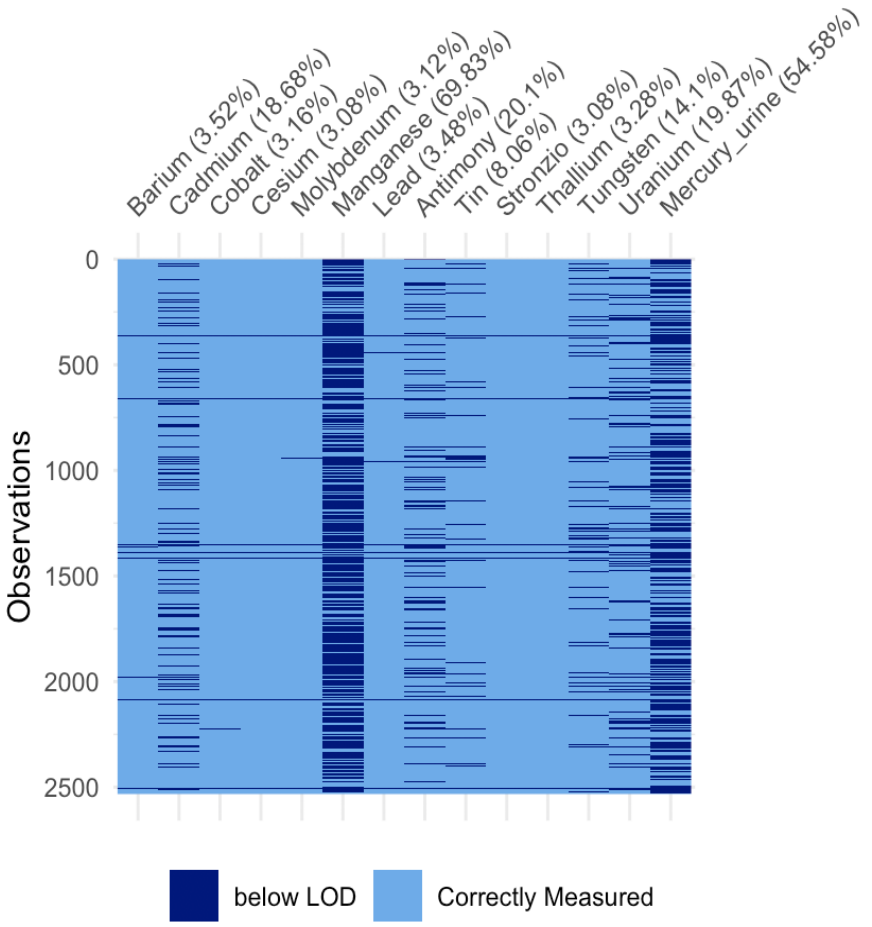

FIG 4. *Pattern of data below the limit of detection in the matrix X including the chemical measurements.*

Algorithm 2 MCMC algorithm for imputing missing observations and those under the LOD while simultaneously sampling the parameters of model (5.1)

Step 1 Sample η_i, $i = 1, \cdots, n$ from a multivariate normal distribution:

$$\pi(\omega| -) \sim N_k\bigg((I_k + \Lambda^T\Sigma^{-1}\Lambda)^{-1}\Lambda^T\Sigma^{-1}W_i, (I_k + \Lambda^T\Sigma^{-1}\Lambda)^{-1}\bigg).$$

Step 2 Denote λ_j the rows of Λ, for $j = 1, \cdots, d$. Sample d conditionally independent posteriors:

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

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

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

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

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

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

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

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

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

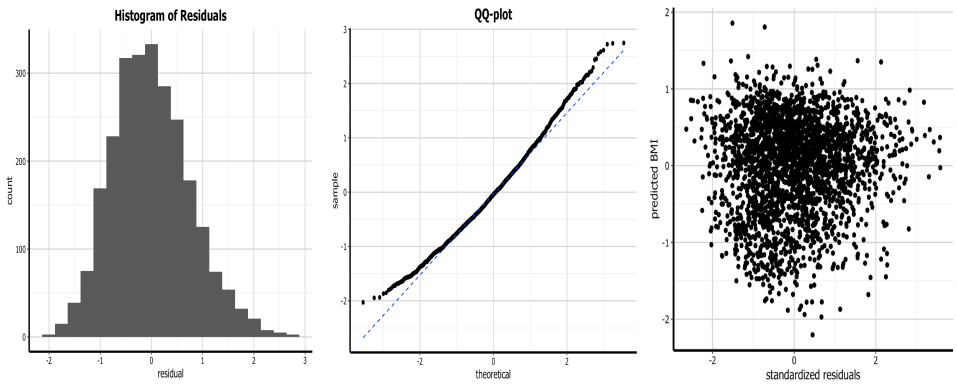

FIG 5. *Histogram and QQ-plot of residuals and scatter plot of predicted BMI values vs standardized residuals from the analysis described in Section 5.*

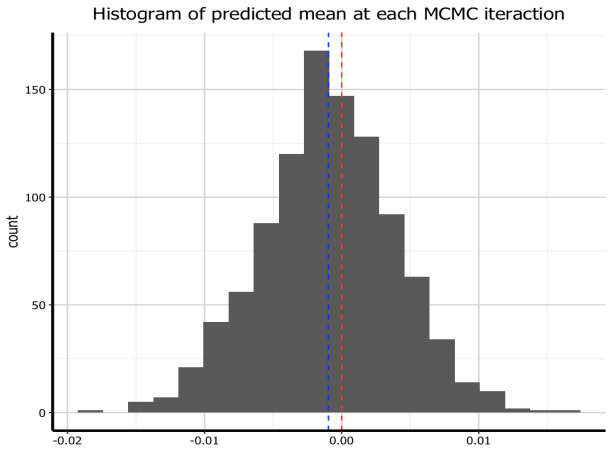

FIG 6. *Histogram of the predictive mean at each MCMC iteration. The horizotal red line shows the mean of BMI and the blue line the mean of the predictions.*

level α	in sample	out of sample
0.01	0.99	0.98
0.025	0.979	0.968
0.05	0.96	0.942
0.1	0.91	0.88

TABLE 4
Coverage computed at different levels α for the in sample and out of sample predictive intervals of BMI.

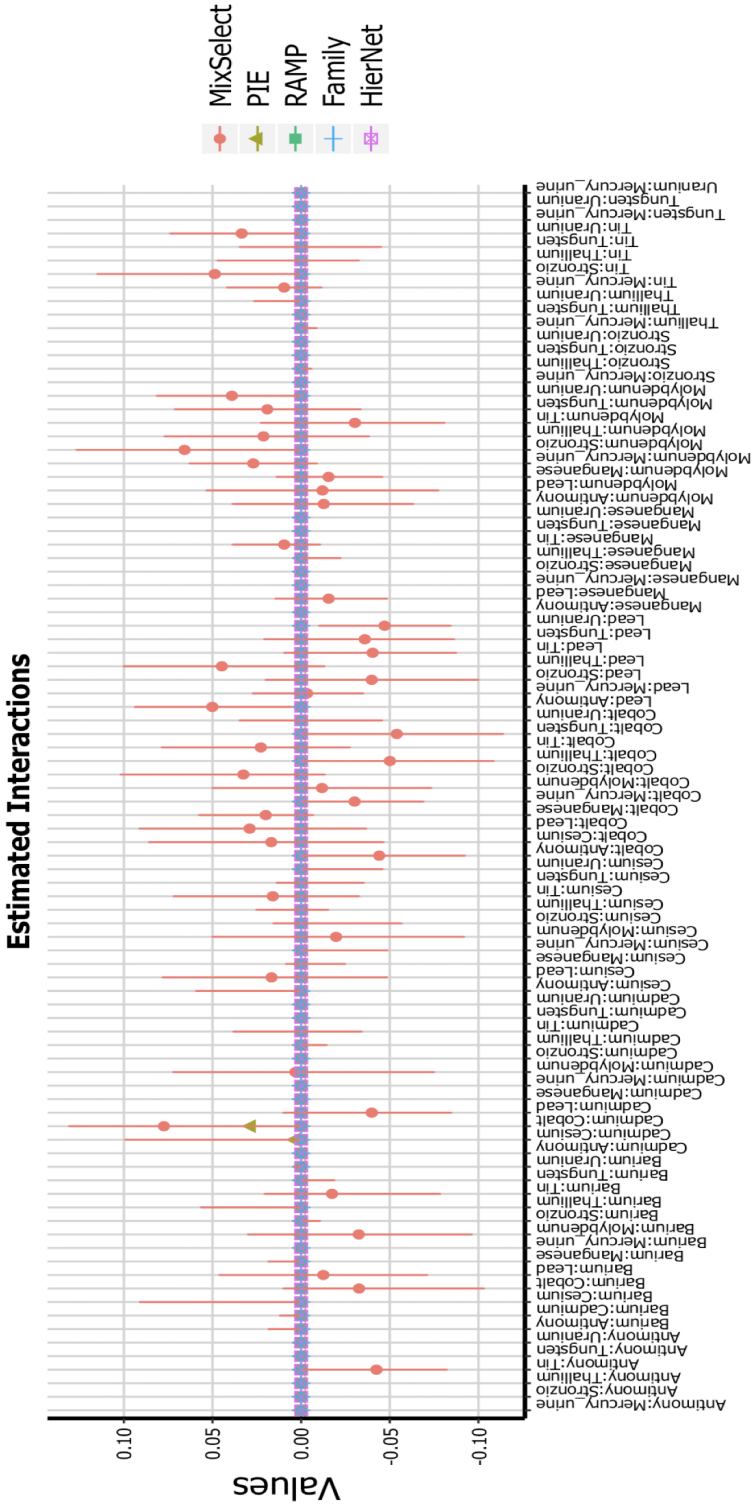

FIG 7. *Estimated interaction effects using MixSelect with* 95% *credible intervals.*

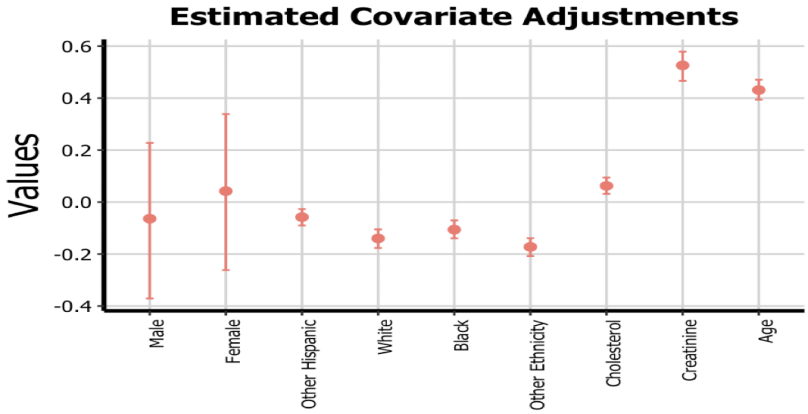

FIG 8. *Estimated covariate effects using MixSelect with* 95% *credible intervals. Hispanic is the reference group for ethnicity.*

	MixSelect	BKMR
Barium	0.12	1
Cadmium	1	1
Cobalt	0.20	0.87
Cesium	0.52	1
Molybdenum	0.59	1
Manganese	0.18	0.96
Lead	0.39	1
Antimony	0.24	1
Tin	0.22	1
Strontium	0.18	1
Thallium	0.18	1
Tungsten	0.79	1
Uranium	0.16	0.94
Mercury_urine	0.16	1

TABLE 5

Posterior inclusion probability of chemical measurements nonlinear effects for MixSelect and BKMR. The estimates of MixSelect have used Algorithm 2 to impute missing values and those under the LOD, while estimates of BKMR have been computed with complete cases as current BKMR code does not allow missingness or LOD.

4. Application with Sex and Ethnicity Interaction. In this section, we assess whether the association between the metals analyzed in *Section 5* and BMI changes with sex or non-Hispanic Black ethnicity. In epidemiology, it is common to conduct separate analyses for Blacks and non-Blacks as these groups can be very different with respect to certain exposures and outcomes. In NHANES studies, [Shim et al., 2017] show sug-

gestive evidence of age and sex interactions as well as interactions between age and ethnicity for Lead, Cadmium, Mercury and Arsenic.

We run the analysis on the dataset with 2029 complete cases: 49% of observations are Male and 19% are non-Hispanic Black. We preprocess the data following *Section 5.2*. We estimate a quadratic regression with nonlinear effects for the transformed chemicals interacted separately with Sex and non-Hispanic Black ethnicity, which are included in the matrix X, and we control for covariates, which are included in the matrix Z, according to model (2.1). We estimate the model using the strong heredity specification and we compare the estimates of MixSelect with the methods described in *Section 4*: BKMR [Bobb et al., 2014], Family [Haris et al., 2016], hierNet [Bien et al., 2013], PIE [Wang et al., 2019] and RAMP [Hao et al., 2018].

Figure 9 shows the estimated nonlinear curves for Cadmium interacted with Sex and non-Hispanic Black ethnicity, when all the other variables are set to their median. The non linear effect of Cadmium in Females and non-Hispanic Blacks has a hill-shaped dose response as in *Figure 2*, whereas it is negatively associated with BMI in the Male subgroup. *Table 6* contains the posterior inclusion probabilities of nonlinear effects. *Figure 10* shows the estimated main effects of the chemicals, and 95% credible intervals for MixSelect. Notice that Lead and Molybdenum exposures have a stronger negative effect on Females than Males, and we observe the opposite behavior for Tin and Cobalt. These associations are also estimated by PIE, and they are partially supported by RAMP and hierNet. We found positive linear interactions between Cadmium×Cobalt and Cobalt×Tin for Males.

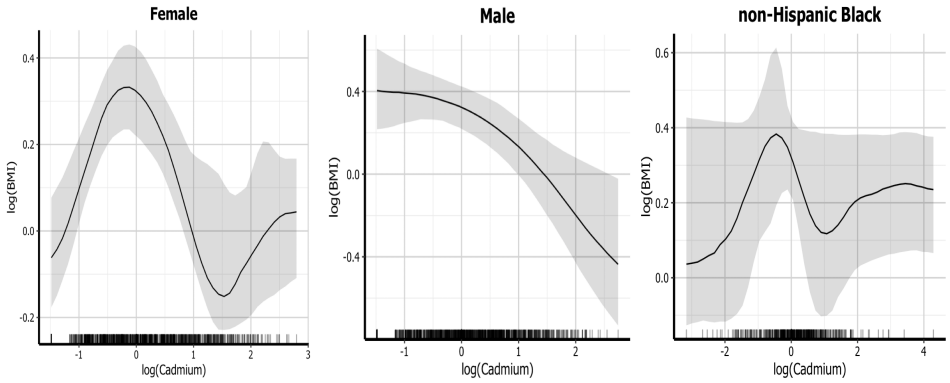

FIG 9. *Estimated regression surface for Cadmium interacted with Sex and non-Hispanic Black ethnicity, when all the other quantities are equal to their median. The black line corresponds to the posterior median, the shaded bands indicate 95% posterior credible intervals, and the marks on the x-axis indicate the observed data points.*

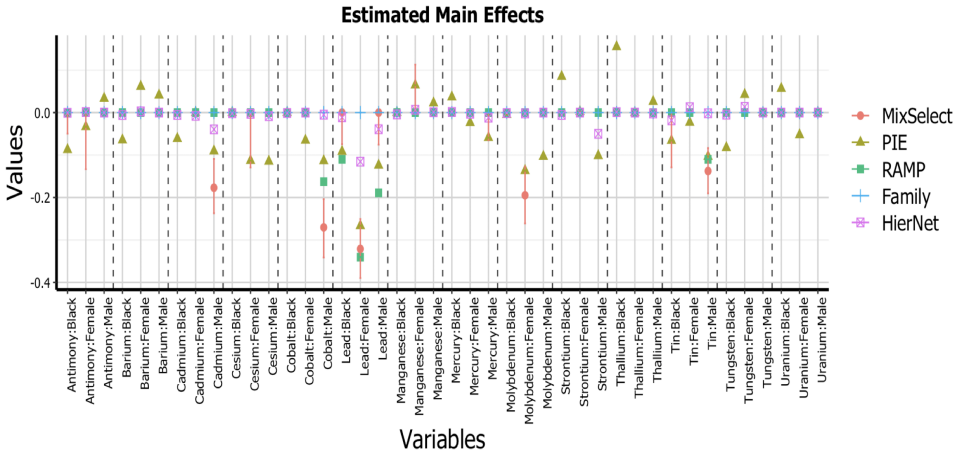

FIG 10. *Estimated main effects using MixSelect with* 95% *credible intervals and estimated coefficients using RAMP, hierNet, Family and PIE. We trained all the other models on the dataset with complete cases and included interactions of chemical measurements with non-Hispanic Black ethnicity and Sex. Exposure measurements are on the log scale.*

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

TABLE 6

Posterior inclusion probabilities of nonlinear effects when including interactions between chemical measurements with non-Hispanic Black ethnicity and Sex.

References.

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

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

Ning Hao, Yang Feng, and Hao Helen Zhang. Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association*, 113(522):615–625, 2018.

Asad Haris, Daniela Witten, and Noah Simon. Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics*, 25(4):981–1004, 2016.

Stefan Lang and Andreas Brezger. Bayesian p-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212, 2004.

Youn K Shim, Michael D Lewin, Patricia Ruiz, June E Eichner, and Moiz M Mumtaz. Prevalence and associated demographic characteristics of exposure to multiple metals and their species in human populations: The united states nhanes, 2007–2012. *Journal of Toxicology and Environmental Health, Part A*, 80(9):502–512, 2017.

Cheng Wang, Binyan Jiang, and Liping Zhu. Penalized interaction estimation for ultrahigh dimensional quadratic regression. *arXiv Preprint arXiv:1901.07147*, 2019.