Biostatistics (2015), **16**, 3, *pp.* 493–508
doi:10.1093/biostatistics/kxu058
Advance Access publication on December 22, 2014

Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures

JENNIFER F. BOBB*, LINDA VALERI

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA
jbobb@hsph.harvard.edu

BIRGIT CLAUS HENN

Department of Environmental Health, Harvard School of Public Health, Landmark Center, 401 Park Drive, Boston, MA 02215, USA

DAVID C. CHRISTIANI

Department of Environmental Health, Harvard School of Public Health, 665 Huntington Avenue, Boston, MA 02115, USA

ROBERT O. WRIGHT

Mount Sinai Hospital, 17 East 102 Street Floor 3, West Room D3-110, New York, NY 10029, USA

MAITREYI MAZUMDAR

Department of Environmental Health, Harvard School of Public Health, 665 Huntington Avenue, Boston, MA 02115, USA

JOHN J. GODLESKI

Department of Environmental Health, Harvard School of Public Health, Landmark Center, 401 Park Drive, Boston, MA 02215, USA

BRENT A. COULL

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

Summary

Because humans are invariably exposed to complex chemical mixtures, estimating the health effects of multi-pollutant exposures is of critical concern in environmental epidemiology, and to regulatory agencies such as the U.S. Environmental Protection Agency. However, most health effects studies focus on single agents or consider simple two-way interaction models, in part because we lack the statistical methodology

*To whom correspondence should be addressed.

to more realistically capture the complexity of mixed exposures. We introduce Bayesian kernel machine regression (BKMR) as a new approach to study mixtures, in which the health outcome is regressed on a flexible function of the mixture (e.g. air pollution or toxic waste) components that is specified using a kernel function. In high-dimensional settings, a novel hierarchical variable selection approach is incorporated to identify important mixture components and account for the correlated structure of the mixture. Simulation studies demonstrate the success of BKMR in estimating the exposure-response function and in identifying the individual components of the mixture responsible for health effects. We demonstrate the features of the method through epidemiology and toxicology applications.

Keywords: Air pollution; Bayesian variable selection; Environmental health; Gaussian process regression; Metal mixtures.

1. INTRODUCTION

Recognizing that populations are exposed to a mix of chemicals and other pollutants, the desire to quantify the health effects of complex multi-pollutant mixtures has grown in recent years. Mixtures of concern include air pollution (Kioumourtzoglou *and others*, 2013), mixtures of toxic waste (a focus of the U.S. Superfund Research Program; Hu *and others*, 2007), mixtures of persistent organic chemicals (Gennings *and others*, 2010), and the interplay between environmental exposures and psychosocial factors (Carlin *and others*, 2013).

To estimate the health effects of these multi-pollutant mixtures, several challenges must be addressed. First, the mixture components may have complex non-linear and non-additive relationships with health. Secondly, allowing for a flexible exposure-response function of multiple components and their interactions quickly leads to a high-dimensional problem with a large number of parameters relative to the number of observations, yielding unstable estimates. Thirdly, statistical methods must account for the complex structure of the mixture, which often consists of multiple highly correlated exposures. Current approaches to studying mixtures (Billionnet *and others*, 2012), while addressing some of these complexities, also have distinct disadvantages. For example, clustering methods result in a loss of information due to categorizing the continuous exposure concentrations. Statistical learning algorithms like random forests (Breiman, 2001) can provide a measure of variable importance for the mixture components, but this measure does not succinctly summarize the magnitude or direction of the association. Variable selection techniques within the regression framework (e.g. lasso methods; Tibshirani, 1994) shrink individual regression coefficients toward zero, but these are typically based on a relatively simple parametric model of the mixture components. Hierarchical model formulations address highly correlated pollutants by shrinking individual effect estimates toward group means (Thomas *and others*, 2007), but this approach also typically assumes linear and additive associations between each component and health.

In this paper, we introduce Bayesian kernel machine regression (BKMR) as a new approach for estimating the health effects of mixtures. For this approach, we model the health outcome as a smooth function h, represented using a kernel function, of the exposure variables, adjusted for possible confounding factors. Because the health outcome may depend on only a subset of the mixture components, we conduct variable selection in order to identify which of these components are responsible for the health effects of the mixture. Finally, to address collinearity of the mixture components, we develop a hierarchical variable selection extension to BKMR that can incorporate prior knowledge on the structure of the mixture.

Previous work on kernel machine regression (KMR) has focused on testing, variable selection, and risk prediction. In statistical genomics, KMR methods have been applied primarily to test for the overall effect of a genetic pathway (Liu *and others*, 2007) or for the effect of a gene in the presence of possible gene–gene or gene–environment interaction (Maity and Lin, 2011; Zou *and others*, 2010). In the

context of computer experiments, Linkletter *and others* (2006) applied Gaussian process models with variable selection to identify a subset of inputs with the largest impacts on the system being studied. Savitsky *and others* (2011) considered a general framework for Gaussian process models with variable selection and evaluated their performance in terms of their predictive power and ability to correctly select relevant variables.

This work provides several contributions. First, to our knowledge this is the first time KMR methods have been considered for estimating the health effects of multi-pollutant mixtures. Previous work focused mainly on variable selection and prediction, but in this setting estimating the exposure-response function is the major goal. Secondly, we develop a novel hierarchical variable selection approach within BKMR (Section 2) that is able to account for the structure of the mixture and systematically handle highly correlated exposures. We conduct simulation studies (Section 3) based on real multi-pollutant datasets, in which we compare our method to a two-stage frequentist KMR approach, which tests each mixture component sequentially and then estimates the exposure-response function. Finally, we apply BKMR to two environmental health datasets: (1) an epidemiology study of metal mixtures and psychomotor development (Section 4) highlights the ability of BKMR to estimate complex exposure-response functions in a setting where both non-linearity and interaction have been reported (Claus Henn *and others*, 2012), and (2) a toxicology study of air pollution mixtures and hemodynamics (Section 5) highlights the ability of hierarchical variable selection to identify important mixture components in a setting with several highly correlated pollutants.

2. Bayesian kernel machine regression

For each subject $i = 1, \ldots, n$, we assume

$$Y_i = h(\mathbf{z}_i) + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i, \tag{2.1}$$

where Y_i is a health endpoint, $\mathbf{z}_i = (z_{i1}, \ldots, z_{iM})^{\mathrm{T}}$ is a vector of M exposure variables (e.g. air pollution constituents), $\mathbf{x}_i$ contains a set of potential confounders, and $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. In the context of environmental mixtures $h(\cdot)$ typically characterizes a high-dimensional exposure-response function that may incorporate non-linearity and/or interaction among the mixture components. In such a setting, it can be difficult to specify a set of basis functions to represent $h(\cdot)$ or to fit the resulting model that has a high-dimensional parameter space; we therefore propose to use a kernel machine representation.

2.1 *Overview of KMR*

We assume that $h\colon \mathbb{R}^M \to \mathbb{R}$ resides in a function space $\mathcal{H}_K$ with a positive semidefinite reproducing kernel $K\colon \mathbb{R}^M \times \mathbb{R}^M \to \mathbb{R}$. A kernel function $K(\mathbf{z}, \mathbf{z}')$ has two arguments: $\mathbf{z} = (z_1, \ldots, z_M)^{\mathrm{T}}$, which represents the vector of environmental exposures, or mixture components (which we will refer to as an *exposure profile*) for one subject, and $\mathbf{z}' = (z'_1, \ldots, z'_M)^{\mathrm{T}}$, which represents the exposure profile for a second subject. There are two ways to characterize h. One can use a basis-function representation (also called the *primal form*), with $h(\mathbf{z}) = \sum_{l=1}^{L} \phi_l(\mathbf{z})\eta_l$ for some set of basis functions $\{\phi_l\}_{l=1}^{L}$ and coefficients $\{\eta_l\}_{l=1}^{L}$. Alternatively, one can represent h using a positive-definite kernel function $K(\cdot, \cdot)$, termed the *dual form*, with $h(\mathbf{z}) = \sum_{i=1}^{n} K(\mathbf{z}_i, \mathbf{z})\alpha_i$ for some set of coefficients $\{\alpha_i\}_{i=1}^{n}$. Mercer's theorem (Cristianini and Shawe-Taylor, 2000) established that a kernel function $K(\cdot, \cdot)$ used in the dual form for h implicitly specifies a unique function spanned by a particular set of orthogonal basis functions in the primal representation of h. Examples of this correspondence include the linear kernel $K(\mathbf{z}, \mathbf{z}') = 1 + z_1 z'_1 + \cdots + z_M z'_M$, with basis representation $\{z_m\}_{m=1}^{M}$; the quadratic kernel $K(\mathbf{z}, \mathbf{z}') = (1 + z_1 z'_1 + \cdots + z_M z'_M)^2$, with basis representation $\{z_m, z_m z_{m'}\}_{m,m'=1}^{M}$; and the Gaussian kernel $K(\mathbf{z}, \mathbf{z}') = \exp\{-\sum_{m=1}^{M}(z_m - z'_m)^2/\rho\}$ with ρ a tuning parameter, represented by set of radial

basis functions. Operationally, Liu *and others* (2007) showed that model (2.1) with h specified in the dual form can be expressed as the mixed model

$$\begin{aligned} y_i &\sim N(h_i + \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}, \sigma^2) \quad \text{independent; } i = 1, \ldots, n, \\ \mathbf{h} &\equiv (h_1, \ldots, h_n)^{\mathrm{T}} \sim N(\mathbf{0}, \tau\mathbf{K}), \end{aligned} \tag{2.2}$$

where $\mathbf{K}$, referred to as the *kernel matrix*, has (i, j)-element $K(\mathbf{z}_i, \mathbf{z}_j)$.

Choice of kernel. We focus on the Gaussian kernel, which flexibly captures a wide range of underlying functional forms for $h(\cdot)$, although the methods are applicable to a broad choice of kernels. To provide some intuition for KMR using the Gaussian kernel, consider the effect on health of exposure to the profile $\mathbf{z}_i$ for the ith person, given by $h_i = h(\mathbf{z}_i)$. Under model (2.2), we assume $\mathrm{cor}(h_i, h_j) = \exp\{-(1/\rho)\sum_{m=1}^{M}(z_{im} - z_{jm})^2\}$, which implies that two subjects with similar exposures ($\mathbf{z}_i$ "close" to $\mathbf{z}_j$) will have more similar risks (h_i will be close to h_j).

2.2 *Component-wise variable selection*

To allow for variable selection within a Bayesian paradigm, we define the augmented Gaussian kernel function as $K(\mathbf{z}, \mathbf{z}'; \mathbf{r}) = \exp\{-\sum_{m=1}^{M} r_m(z_m - z'_m)^2\}$, where $\mathbf{r} = (r_1, \ldots, r_M)^{\mathrm{T}}$, and we define $\mathbf{K}_{\mathbf{Z},\mathbf{r}}$ to be the $n \times n$ matrix with (i, j)-element equal to $K(\mathbf{z}_i, \mathbf{z}_j; \mathbf{r})$. We assume a "slab-and-spike" prior on the auxiliary parameters,

$$\begin{aligned} r_m \mid \delta_m &\sim \delta_m f_1(r_m) + (1 - \delta_m)P_0, \quad m = 1, \ldots, M, \\ \delta_m &\sim \text{Bernoulli}(\pi), \end{aligned} \tag{2.3}$$

where $f_1(\cdot)$ is a pdf with support on $\mathbb{R}^+$ and P_0 denotes the density with point mass at 0. This approach is analogous to Bayesian variable selection approaches for multiple regression problems (George and McCulloch, 1993) and has been applied in Gaussian process models (Linkletter *and others*, 2006; Savitsky *and others*, 2011). The posterior mean of the indicator δ_m has the natural interpretation as the posterior probability that component m is an important component of the mixture, or the posterior "inclusion probability" of component m. Other kernel functions may be augmented in a similar way. For example, the quadratic kernel may be expanded as $K(\mathbf{z}, \mathbf{z}'; \mathbf{r}) = (1 + r_1 z_1 z'_1 + \cdots + r_m z_M z'_M)^2$.

2.3 *Hierarchical variable selection*

In situations where mixture components are highly correlated, the above formulation that treats components exchangeably may fail because the data may not be able to distinguish among these correlated components. We therefore propose a hierarchical variable selection approach, which incorporates information on the structure of the mixture into the model.

Suppose that the mixture components $z_1, \ldots, z_M$ can be partitioned into groups $\mathcal{S}_g$ $(g = 1, \ldots, G)$ such that within-group correlation is high while across-group correlation is low. For example, a wealth of information about air pollution sources is typically known, which could be used to group the pollutants. We then assume that the indicator variables from the slab-and-spike prior in (2.3) are distributed as

$$\begin{aligned} \boldsymbol{\delta}_{\mathcal{S}_g} \mid \omega_g &\sim \text{Multinomial}(\omega_g, \boldsymbol{\pi}_{\mathcal{S}_g}), \quad g = 1, \ldots, G, \\ \omega_g &\sim \text{Bernoulli}(\pi), \end{aligned} \tag{2.4}$$

where $\boldsymbol{\delta}_{\mathcal{S}_g} = (\delta_m)_{z_m \in \mathcal{S}_g}$ is the vector of indicator variables and $\boldsymbol{\pi}_{\mathcal{S}_g}$ is the corresponding vector of prior probabilities for the mixture components z_m in group $\mathcal{S}_g$. This approach allows at most a single component

from a group (of highly correlated components) to enter into the model at a time. Although this assumes that two components from the same group do not have independent or interactive effects on the health outcome, in the setting of high within-group correlation, such effects would not be identifiable in a more general model.

2.4 *Prior specification, estimation, and prediction*

In Section A of supplementary material available at *Biostatistics* online, we specify prior distributions for each of the parameters above; in Section B we provide details on the Markov chain Monte Carlo sampler used to fit BKMR with component-wise and hierarchical variable selection (B.1), methods for estimating subject-specific health effects (B.2), and methods for predicting health effects at new exposure profiles (to estimate the multivariate exposure-response function; B.3).

3. Simulation studies

We evaluated the ability of BKMR (and compared its performance to frequentist KMR methods) in flexibly estimating the exposure-response function and in identifying mixture components responsible for health effects, under a range of plausible data generating scenarios based on real exposure datasets.

3.1 *Setup*

Data generation. We generated 300 datasets of 100 observations each, $\{y_i, x_i, \mathbf{z}_i\}_{i=1}^{100}$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iM})^{\mathrm{T}}$ represents an exposure profile with M mixture components; x_i is a confounder generated by $x_i \sim N(3\cos z_{i1}, 2)$; and the health outcomes are generated by $y_i \sim N(\beta x_i + h(z_{i1}, \ldots, z_{iQ}), \sigma^2)$, where we assumed that the health outcome depends on a subset of $Q < M$ of the available exposure variables. We set σ^2 to correspond to a realistic signal-to-noise ratio based on the Bangladesh application (Section 4).

We considered two choices for the total number of mixture components ($M = 3, 13$), and we generated exposure data based on empirical distributions of real data. For $M = 3$, each exposure dataset $\{\mathbf{z}_i\}_{i=1}^{100}$ was obtained by resampling 100 rows of the exposure data from our Bangladesh application (Section 4), which consists of arsenic (As), manganese (Mn), and lead (Pb) exposures of Bangladeshi children. For $M = 13$, we considered an exposure dataset that consisted of daily measures, from 1999 to 2011, of air pollution constituents monitored at a central site in Boston. We selected 13 constituents that have been used previously in studies of the health effects of air pollution (listed in Figure 1). The daily component data were standardized by subtracting the median and dividing by the interquartile range (IQR), and outlier values ($\geqslant 5$ IQR away from the median) were removed. The correlation matrix for this data is in Figure 1 of supplementary material available at *Biostatistics* online. We then generated each exposure dataset $\{\mathbf{z}_i\}_{i=1}^{100}$ by resampling 100 rows of this Boston air pollution data.

We considered several exposure-response functions $h(\cdot)$ that varied depending on the number of pollutants included, the degree of correlation of the included pollutants, and the shape of the function. We first considered three $h(\cdot)$ that depended on just one or two of the pollutants (Figure 1): a non-linear function of z_{i1} (h_1), a linear function with main effects of z_{i1} and z_{i2} and their interaction (h_2), and a non-linear function of both z_{i1} and z_{i2} with a synergistic interaction between them (h_3). We considered a scenario where the two included pollutants in h_2 and h_3 were essentially uncorrelated (Mn, Pb for the Bangladesh dataset [cor = 0]; Al, Cu for the Boston dataset [cor = 0.17]) as well as a scenario where the two pollutants were more highly correlated (Mn, As for the Banladesh dataset [cor = 0.58]; Al, Ca for the Boston dataset [cor = 0.68]). Finally, to evaluated BKMR under a more complex setting with a larger number of

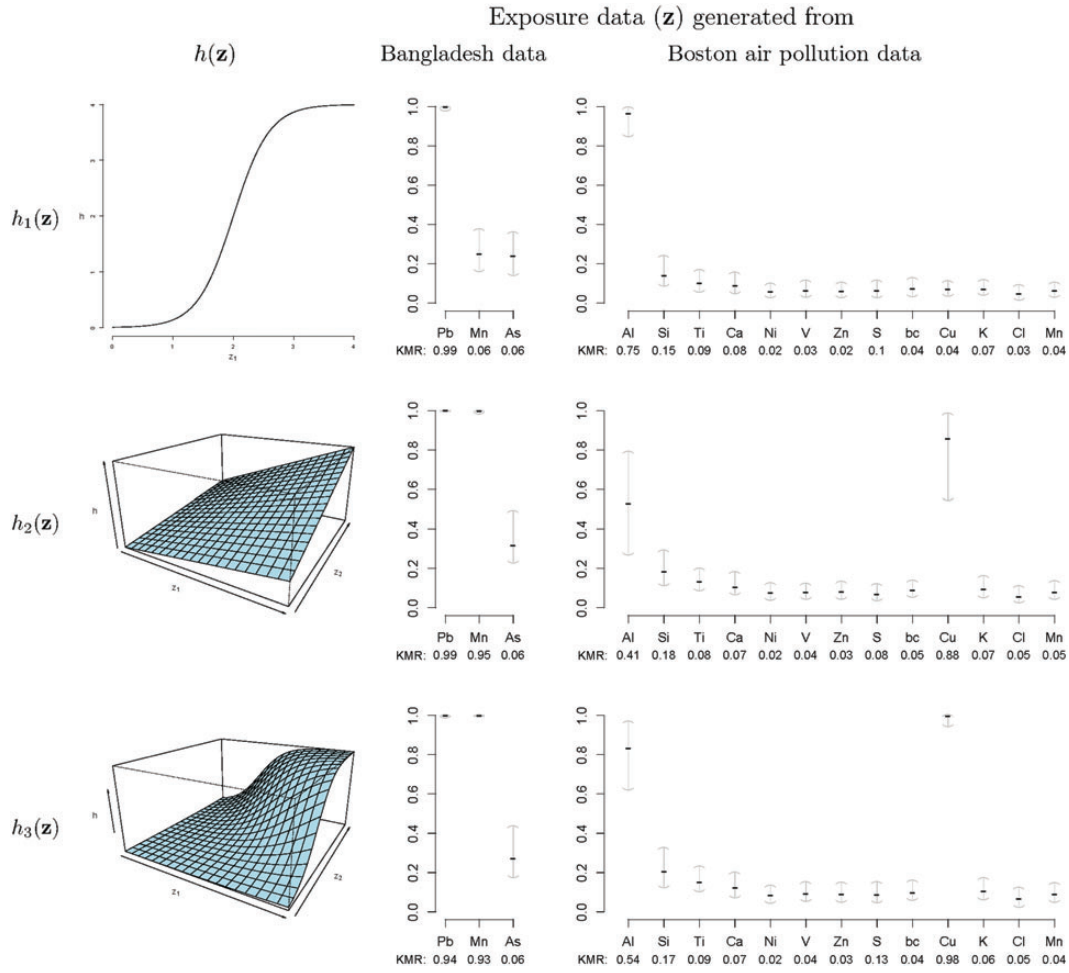

Fig. 1. Median (25%, 75%) of the PIPs from BKMR with component-wise variable selection, across 300 simulated datasets for each of three true $h(\mathbf{z})$ functions. The vector of exposure data $\mathbf{z}$ was generated either based on the Bangladesh data with $M = 3$ mixture components, where the truly associated components were Pb for h_1, and Pb and Mn for h_2 and h_3; or based on the Boston air pollution data with $M = 13$ mixture components, where the truly associated components were Al for h_1, and Al and Cu for h_2 and h_3. The proportion of simulation iterations for which each mixture component had p-value <0.05 under the garrote test for KMR is printed below the x-axis. Pb, lead; Mn, manganese; As, arsenic; Al, aluminum; Si, silicon; Ti, titanium; Ca, calcium; Ni, nickel; V, vanadium; Zn, zinc; S, sulphur; bc, black carbon; Cu, copper; K, potassium; Cl, chlorine.

mixture components, we considered two functions that included 6 of the 13 components from the Boston air pollution dataset:

$$h_4(\mathbf{z}_i) = h_3(\mathrm{Al}_i, \mathrm{Ni}_i) + h_3(\mathrm{S}_i, \mathrm{Cu}_i) + h_3(\mathrm{Cl}_i, \mathrm{Mn}_i), \tag{3.1}$$

$$h_5(\mathbf{z}_i) = h_3(\mathrm{Al}_i, \mathrm{Ni}_i) + h_3(\mathrm{S}_i, \mathrm{Cu}_i) + h_3(\mathrm{Cl}_i, \mathrm{Si}_i). \tag{3.2}$$

These two functions are the same, except that h_4 does not include any very highly correlated pollutants (the largest correlation is 0.49 between Al and S), whereas h_5 includes both Al and Si, which have a correlation of 0.87. Because these higher-dimensional h require more power to detect, we halved the residual standard deviation (SD) σ when compared with the simulation studies for h_3.

Models. First, we fit KMR using a frequentist approach (Liu *and others*, 2007), both without (KMR) and with (KMR-vs) variable selection. To conduct the variable selection, we applied the garrote KMR test from Maity and Lin (2011) to each component z_m ($m = 1, \ldots, M$) sequentially, and then re-fit the KMR including only those components with $p < 0.05$. Secondly, we fit three BKMR models: a model without variable selection (BKMR), a model with component-wise variable selection (BKMR-vs), and for the $M = 13$ components exposure dataset, a model with hierarchical variable selection (BKMR-hvs). For BKMR-hvs, we defined the component groups $\mathcal{S}_1, \ldots, \mathcal{S}_8$ (shown in Figure 2) based on knowledge of Boston air pollution sources. The within-group correlations ranged from 0.68 to 0.87 in $\mathcal{S}_1$ and from 0.45 to 0.8 in $\mathcal{S}_2$ (Figure 1 of supplementary material available at *Biostatistics* online). We ran each MCMC sampler (described in Section B of supplementary material available at *Biostatistics* online) for 10 000 iterations and kept the last 5000 samples. Finally, to quantify the optimal performance achievable if the true pollutants included in the h function were known, we considered an "oracle" model. For h_1, we fit a generalized additive model (GAM) including only z_{i1}, modeled using penalized splines, and a thin-plate regression basis (Wood, 2006); for h_2, we fit a linear model with z_{i1}, z_{i2}, and the interaction term $z_{i1}z_{i2}$; for h_3 we fit a GAM including a bivariate smooth function of z_{i1} and z_{i2}; and for h_4 and h_5 we fit a GAM including independent bivariate smooth functions of the truly included pollutants in equations (3.1) and (3.2), respectively.

3.2 *Results*

Estimating the exposure-response function. We first evaluated the ability of each approach to estimate the subject-specific mixture effects $h_i = h(\mathbf{z}_i)$. Here we focus on results for the univariate and bivariate exposure-response functions under the scenario of uncorrelated exposures (Table 1); the relative performance of the methods for the other scenarios was similar. The approaches with variable selection (KMR-vs, BKMR-vs, and BKMR-hvs) each performed comparably to the oracle model (and outperformed the corresponding models without variable selection) in estimating the h_i for the Bangladesh ($M = 3$) exposure dataset. For the Boston ($M = 13$) air pollution dataset, the Bayesian variable selection approaches outperformed all of the other methods (except the oracle model) for each $h(\cdot)$. Across all scenarios, the Bayesian approaches were better able to capture the uncertainty in the $\hat{h}_i$ when compared with the corresponding frequentist methods, achieving posterior SD estimates that were close to the empirical standard errors and interval coverage closest to the nominal (95%) level. KMR-vs had especially poor coverage, particularly for the $M = 13$ scenarios, suggesting that the two-stage approach to estimating the exposure-response function does not fully account for uncertainty due to variable selection.

Identifying important mixture components. We next evaluated the ability of the methods to identify which mixture component(s) were included in $h(\cdot)$. Figure 1 shows, for the univariate and bivariate exposure-response functions h_1–h_3 under the scenario of uncorrelated exposures, the median (IQR) for the posterior inclusion probabilities (PIPs) under BKMR-vs, as well as the proportion of iterations for which each component was identified as statistically significant under the garrote KMR test. For the Bangladesh ($M = 3$) dataset, the garrote test achieved high power and the nominal type I error rate, and BKMR-vs was able to distinguish between the important versus unimportant components. For the Boston ($M = 13$) dataset, the approaches were able to identify Cu, a component whose correlation with the other pollutants ranged from 0.13 to 0.29, as important in the scenarios where it was included in $h(\cdot)$. On the other hand, for Al, a component highly correlated with several others (cor = 0.87 with Si, 0.7 with Ti, and 0.68 with Ca), the garrote test had lower power and had inflated type I errors with its correlated exposures, especially Si. For BKMR-vs, while the PIPs remained higher for Al than for its correlated exposures, Si also had higher PIPs relative to the other, unimportant components. Compared with the uncorrelated scenario, when the two pollutants included in h_2 and h_3 were more highly correlated, the PIPs remained similar or were reduced (cf. Figure 2 of supplementary material available at *Biostatistics* online to Figure 1). For the

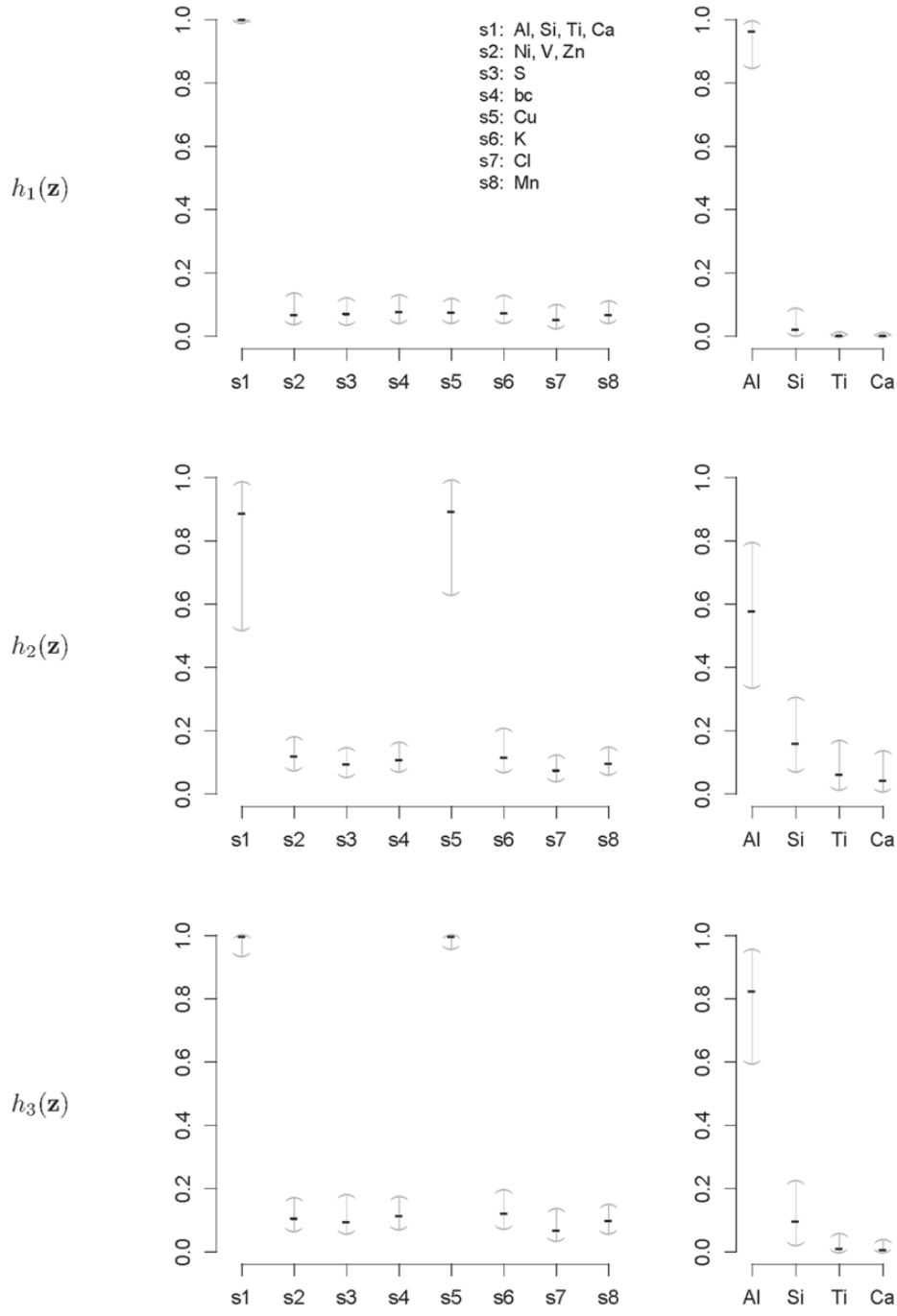

Fig. 2. Median (25%, 75%) of the PIPs from BKMR with hierarchical variable selection, across 300 simulated datasets for each of three true $h(\mathbf{z})$ functions. Exposure data $\mathbf{z}$ were generated based on the Boston air pollution data with $M = 13$ mixture components categorized within eight groups. The truly associated components were Al (one of four pollutants in group 1) for h_1, and Al and Cu (sole pollutant in group 5) for h_2 and h_3. Plots on left show the PIPs for each group, and plots on the right show the conditional PIPs for the components in group 1 given that group 1 was included in the model.

Table 1. *Performance of estimated subject-specific mixture effects,* $h_i = h(\mathbf{z}_i)$, *across* 300 *simulated datasets for each of three true exposure-response functions* $h(\cdot)$, *and two exposure generating models* ($M = 3, 13$)

	Regression of $\hat{h}$ on h			Uncertainty		
	Intercept	Slope	R^2	SD($\hat{h} \mid h$)	SE	Cvg.
$M = 3$						
$h_1(\mathbf{z})$						
Oracle	0.06	0.93	0.93	0.47	0.39	0.90
KMR	0.25	0.74	0.80	0.67	0.50	0.84
KMR-vs (297)	0.09	0.89	0.92	0.48	0.39	0.89
BKMR	0.24	0.76	0.84	0.62	0.49	0.87
BKMR-vs	0.15	0.85	0.92	0.51	0.45	0.90
$h_2(\mathbf{z})$						
Oracle	0.00	1.00	0.97	0.14	0.13	0.95
KMR	0.10	0.91	0.95	0.16	0.15	0.93
KMR-vs	0.09	0.92	0.95	0.16	0.13	0.92
BKMR	0.08	0.93	0.95	0.16	0.16	0.95
BKMR-vs	0.07	0.94	0.95	0.16	0.16	0.96
$h_3(\mathbf{z})$						
Oracle	0.04	0.91	0.90	0.28	0.27	0.95
KMR	0.08	0.84	0.85	0.33	0.27	0.91
KMR-vs	0.08	0.84	0.84	0.33	0.23	0.87
BKMR	0.08	0.84	0.86	0.32	0.27	0.92
BKMR-vs	0.07	0.87	0.89	0.30	0.25	0.93
$M = 13$						
$h_1(\mathbf{z})$						
Oracle	0.11	0.92	0.92	0.50	0.38	0.86
KMR	0.60	0.63	0.72	0.80	0.60	0.82
KMR-vs (273)	0.31	0.81	0.79	0.70	0.38	0.76
BKMR	0.48	0.69	0.73	0.78	0.71	0.91
BKMR-vs	0.25	0.84	0.90	0.56	0.52	0.92
BKMR-hvs	0.26	0.84	0.89	0.56	0.49	0.90
$h_2(\mathbf{z})$						
Oracle	0.00	1.00	0.93	0.24	0.21	0.95
KMR	0.42	0.72	0.76	0.36	0.31	0.92
KMR-vs (295)	0.39	0.73	0.69	0.39	0.20	0.75
BKMR	0.30	0.79	0.72	0.38	0.40	0.97
BKMR-vs	0.30	0.79	0.81	0.33	0.32	0.95
BKMR-hvs	0.32	0.78	0.81	0.33	0.31	0.95
$h_3(\mathbf{z})$						
Oracle	0.05	0.94	0.90	0.44	0.38	0.91
KMR	0.24	0.74	0.77	0.62	0.51	0.91
KMR-vs (298)	0.21	0.76	0.74	0.65	0.33	0.76

continued.

Table 1. *continued.*

	Regression of $\hat{h}$ on h			Uncertainty		
	Intercept	Slope	R^2	SD($\hat{h} \mid h$)	SE	Cvg.
BKMR	0.20	0.77	0.77	0.61	0.59	0.95
BKMR-vs	0.14	0.85	0.87	0.50	0.47	0.94
BKMR-hvs	0.14	0.84	0.86	0.50	0.44	0.93

Summary measures were obtained by regressing the estimated $\hat{h}_i$ on the true h_i and reporting the average intercept, slope, and R^2 across simulation iterations; uncertainty measures were obtained by averaging over the uncertainty measures for the h_i at each iteration and then averaging across all iterations. "SD" denotes the empirical standard error of the estimated $\hat{h}_i$, "SE" denotes the estimated standard error or posterior SD of the $\hat{h}_i$, and "Cvg." denotes the proportion of times that the 95% confidence intervals or posterior credible intervals covered the true h_i. Note that for some iterations no variables satisfied $p < 0.05$ under the garrote kernel test and so KMR-vs was not applicable; the number of iterations for which KMR-vs was fit is given in parentheses beside the method name.

higher-dimensional h_4 and h_5, BKMR-vs was generally able to distinguish the truly associated pollutants from the unassociated pollutants (Figure 2 of supplementary material available at *Biostatistics* online).

Figure 2 shows the PIPs for each group (i.e. the posterior mean of the group indicators ω_g), as well as the conditional PIPs for the components of group $\mathcal{S}_1 = \{\text{Al, Si, Ti, Ca}\}$ (i.e. the posterior mean of $\boldsymbol{\delta}_{\mathcal{S}_1} \mid \omega_1 = 1$) for h_1–h_3 under the scenario of uncorrelated exposures (results for the other scenarios are in Figure 3 of supplementary material available at *Biostatistics* online). Across all scenarios, there is a clear separation between the PIPs for the groups that included one of the components that was truly predictive of health versus those that did not. In addition, for the truly associated pollutants within a multi-pollutant group (group $\mathcal{S}_1$ under h_1–h_5, group $\mathcal{S}_2$ under h_4 and h_5), the group-specific PIPs were considerably larger than the corresponding component-specific PIPs obtained from BKMR-vs. This suggests that by incorporating the structure of the mixture into the model, BKMR-hvs achieves greater power to detect important components in the high correlation setting.

Within the multi-pollutant groups, the truly important components had higher conditional PIPs than the unimportant components. In the scenarios where the exposure-response function did not include correlated components, BKMR-hvs was better able to distinguish the important from the unimportant components when compared with BKMR-vs. When two pollutants from the same group were included in the exposure-response function, there was considerable variability in the conditional PIPs across simulation repetitions (see h_2, h_3, and h_5 in Figure 3 of supplementary material available at *Biostatistics* online). In particular, for each generated dataset, usually one of the two important pollutants within the group had a high conditional PIP while the other pollutants in the group had much lower PIPs. This occurs because the hierarchical variable formulation (Section 2.3) assumes that only one pollutant from each group is included in the exposure-response function. Although this assumption may seem restrictive in that BKMR-hvs is not able to detect independent (or joint) effects of highly correlated pollutants within a group, such effects are typically not well-identified from the data in practice. For example, for BKMR-vs under h_5, Al was identified as important using a threshold of 0.75 (i.e. had PIPs exceeding 0.75) in 59% of simulation repetitions, and Si was identified in 30% of repetitions, but both pollutants were simultaneously identified as important in only 3% of repetitions.

4. APPLICATION TO A STUDY OF METALS MIXTURES AND NEURODEVELOPMENT IN BANGLADESH

Preliminary data from 375 children (ages 1–4 years) were collected as part an ongoing study of metal exposures and neurodevelopment in Bangladesh (NIEHS grant P42 ES016454). A primary outcome

(Y_i; $i = 1, \ldots, 375$) is the (z-scored) motor composite score (MCS), a summary measure of psychomotor development derived as the sum of the fine and gross motor subscales from the Bayley Scales of Infant and Toddler Development (Bailey, 2005). Prenatal exposures ($\mathbf{z}_i$) to As, Mn, and Pb (log transformed) were measured in umbilical cord blood. Exposure levels of Pb and Mn were uncorrelated, and As was inversely correlated with Pb (cor: -0.37) and positively correlated with Mn (cor: 0.58). Covariates ($\mathbf{x}_i$) consisted of gender, age in months at time of neurodevelopmental assessment (modeled using natural cubic splines with 3 degrees of freedom [df]), mother's education, mother's IQ (spline terms with 3 df), an indicator variable for which of two clinics the child visited, and HOME score (a proxy for socioeconomic status).

As a preliminary analysis, we fit linear regression models, adjusted for the covariates $\mathbf{x}_i$. In single-metal models that included As, Mn, and Pb one at a time, as well as in a multi-metal model that included linear main effects of each metal concurrently, none of the metals were significantly associated with MCS (Table 1 of supplementary material available at *Biostatistics* online). To evaluate potential interaction among the three metals, we conducted an F test to compare the fit of the multi-metal model including just main effects of each metal to the larger model that also contained the three pairwise interactions ($p = 0.37$), and to the saturated model that additionally contained the three-way interaction term ($p = 0.60$). Taken together, these results suggested little evidence of an exposure-response association, in the restrictive setting of linear and additive associations.

We then applied BKMR with component-wise variable selection to estimate the joint association of As, Mn, and Pb with MCS in a flexible way, without the need to specify a priori the form of the exposure-response function. We ran the MCMC sampler (described in Section B of supplementary material available at *Biostatistics* online) for 25 000 iterations after a burn in of 25 000 and every fifth sample was kept for inference. The estimated PIPs were 0.68 for Pb, 0.73 for Mn, and 0.77 for As. Figure 3 shows the estimated relationship of Mn and As with MCS for Pb fixed at its median value. This plot suggests an inverted u-shaped relationship for Mn with MCS, but only at middle levels of As exposure. Similar patterns occurred at other levels of Pb (Figure 4 of supplementary material available at *Biostatistics* online). To confirm that our finding of a non-additive and non-linear exposure-response function for Mn and As under BKMR was real and not an artifact of the method, we subsequently fit a GAM with the same covariates as above, together with a main effect of Pb and separate smooth functions (thin-plate regression splines; smoothing parameter estimated using generalized cross validation) of Mn at each tertile of As exposure. We found a similar inverted u relation between Mn and MCS, and the smooth term was only statistically significant at the second tertile of As (Figure 5 of supplementary material available at *Biostatistics* online).

5. Application to a toxicology study of air pollution mixtures and hemodynamics

We considered data from a toxicology study, in which 13 dogs were repeatedly exposed for 5 h to either concentrated ambient particles (CAPs) or filtered air in a cross-over protocol (Bartoli *and others*, 2009). Previous analyses found elevated blood pressure associated with CAPs exposure; our goal was to identify whether particular component(s) of the CAPs are responsible for these observed effects. Let Y_{it} be the average heart rate for dog i at exposure occasion t, $\mathbf{x}_{it}$ be indicator variables for CAPs versus filtered air exposure and for the other experimental conditions (whether the exposure occurred post-occlusion or after prazosin was administered), and $\mathbf{z}_{it}$ be the vector of elemental concentrations for the CAPs components, where we considered the same pollution constituents as in our simulation study (Section 3). Because in this small subsample of days K, Cu, and Mn were also highly correlated with Al, Si, Ti, and Ca (all pairwise correlations among these seven pollutants were >0.76), we included these additional elements in group $\mathcal{S}_1$. After removing several outliers in the elemental concentrations, the dataset consisted of $n = 142$ dog-exposures.

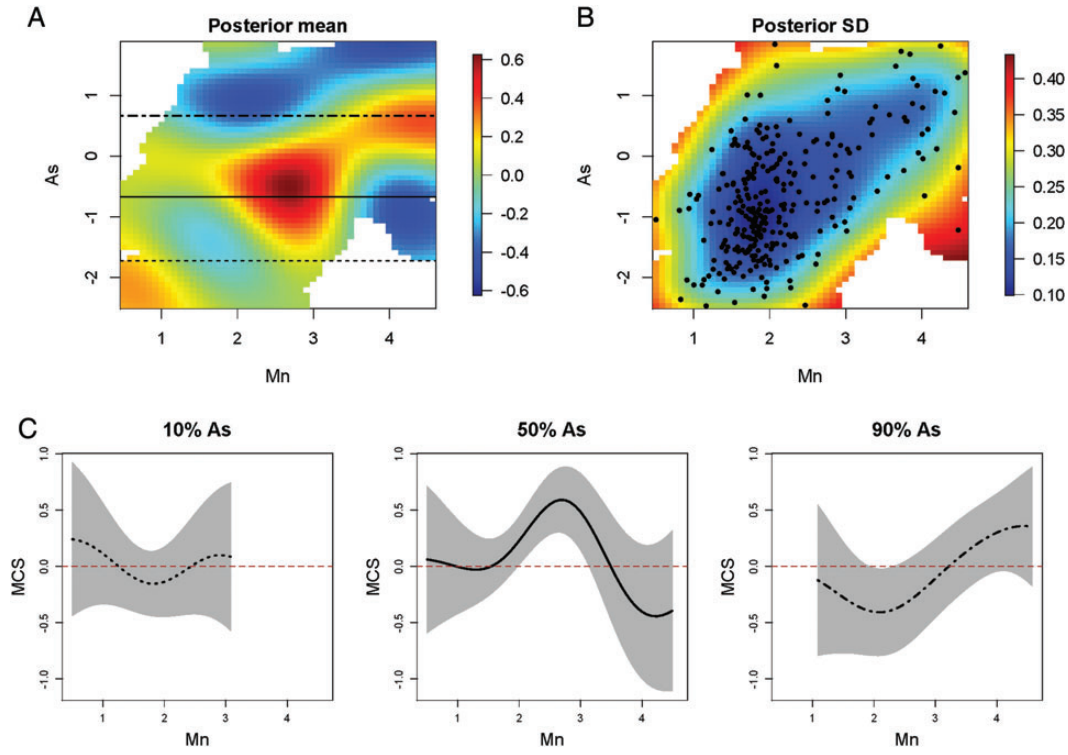

Fig. 3. Relationship of manganese (Mn) and arsenic (As) with the MCS, for lead (Pb) fixed at its median. (A) Posterior mean of the bivariate exposure-response function $\hat{h}$ for Mn and As. Horizontal lines correspond to the 10th, 50th, and 90th percentiles of As. (B) Posterior SD of $\hat{h}$. Points correspond to the observed data points. (C) Relationship of Mn with MCS at three levels of As together with pointwise 95% credible intervals.

We began by fitting linear mixed models (LMMs) of the CAPs components with dog-specific random intercepts, adjusted for the covariates $\mathbf{x}_{it}$. In models that included each component separately, all of the pollutants in $\mathcal{S}_1$ (except Cu) had statistically significant associations with elevated heart rate (none of the other components were associated). However, these associations were no longer significant in the multi-pollutant LMM that included all of the constituents concurrently, and for three of these components the direction of the association was reversed (Table 2 of supplementary material available at *Biostatistics* online).

Because of the longitudinal cross-over design of the study, we extended the BKMR models described in Section 2 to include random (dog-specific) intercepts. We fit BKMR models including all of the 13 pollutants with both component-wise and hierarchical variable selection. We ran the MCMC samplers (described in Section B of supplementary material available at *Biostatistics* online) for 25 000 iterations after a burn in of 25 000 and every fifth sample was kept for inference. We did not find evidence of non-linearity or interaction, so here we focus on the variable selection results. Analogous to the null results from the multi-pollutant LMM, under the component-wise BKMR model we found that each component had a PIP of < 0.4; in contrast, under the hierarchical selection approach group $\mathcal{S}_1$ had a PIP of 0.79 (Figure 4). Given the strong correlations among components in this group, the data did not strongly favor one constituent over the others as driving the observed association between heart rate and this group of elements (the conditional inclusion probabilities ranged from 0.04 for Cu to 0.36 for Si). In this case, our

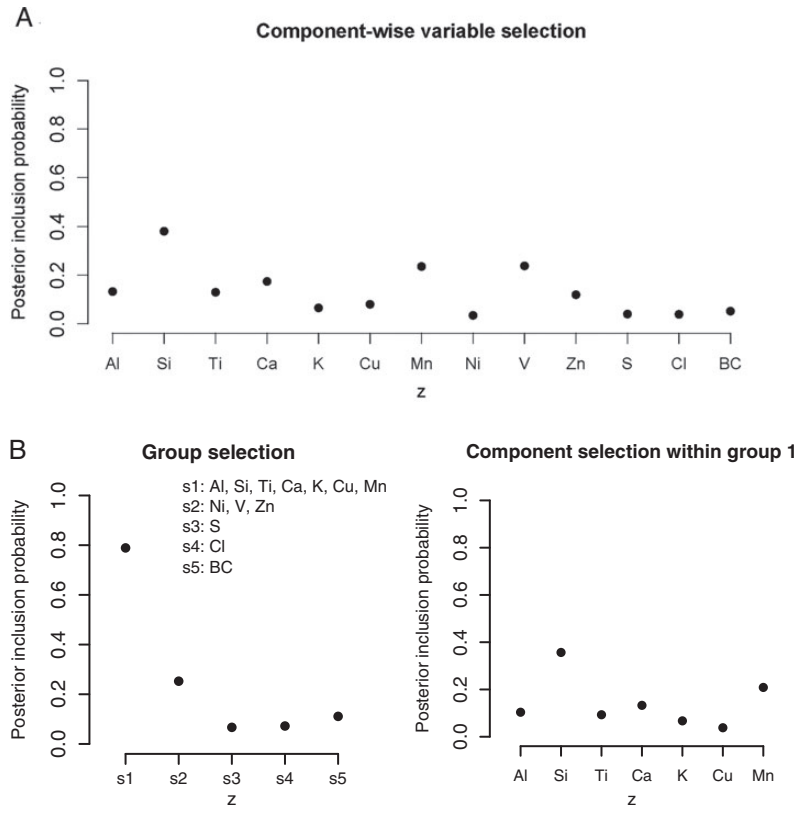

Fig. 4. PIPs for the toxicology application estimated from BKMR with component-wise variable selection (A) and hierarchical variable selection (B). Left panel shows group-specific PIPs and right panel shows conditional inclusion probabilities for the components in group 1.

strong preference is the hierarchical variable selection approach, as it accurately conveys that there is a group of elements that are associated with the outcome but that the data cannot definitively identify a constituent driving this association.

6. Discussion

We have proposed BKMR as a new approach to estimate the health effects of multi-pollutant mixtures. Our simulation studies highlighted two main advantages of BKMR over existing frequentist approaches. First, by conducting variable selection and health effect estimation simultaneously the Bayesian approach was able to more fully capture the uncertainty in the exposure-response function due to selecting which mixture components to include in the model. Secondly, the garrote KMR test (Maity and Lin, 2011) had low power to detect important mixture components in the setting of multiple highly correlated exposures. Because this procedure tests for the effect of a variable Z_1 by conducting a score-based test of the null hypothesis $H_0 : h(z_1, z_2, \ldots, z_M) = h(z_2, \ldots, z_M)$, if the null model already includes a variable that is highly correlated with a truly important variable, then there may not be enough information remaining in the data to detect the true association. Our hierarchical variable selection approach addresses this issue by allowing one component from a group of highly correlated components to enter into the model at a time.

We focused on the Gaussian kernel, although other kernels could be considered. In simulation studies of KMR with $h(\cdot)$ having a complicated functional form, Liu *and others* (2007) found the Gaussian kernel to outperform both the quadratic and ridge regression kernels. Our simulation studies demonstrated that the Gaussian kernel performed well across a range of plausible exposure-response functions for environmental health applications. In future work, Bayesian model selection could be applied to formally evaluate the choice of kernel.

The larger set of 13 predictors in our simulation studies is not particularly large relative to many high-dimensional application areas, such as statistical genomics and other 'omics settings. A few environmental health studies have considered exposures in the hundreds, but these have typically been conducted with the goal of screening for the most important exposures (Patel *and others*, 2010) or classes of exposures (Kioumourtzoglou *and others*, 2013), rather than fully characterizing the form of the exposure-response surface. Thirteen pollutants represents a typical number considered in PM elemental composition studies focusing on the exposure-response relationship. Computationally, the dimension of the exposure vector is not a limiting factor in these models, as the dimension of the exposures gets reduced into the pairwise distance measures contained in the kernel matrix.

It is well known that Bayesian variable selection methods can be highly sensitive to the specification of the mixture prior. In our applications, we found that changing the distribution of the slab part of the prior ($f_1(\cdot)$ in equation (2.3)) led to changes in the values of the PIPs; however, their relative ordering was preserved across prior specifications. This issue is analogous to the variable importance measures produced by a random forests analysis, whose absolute magnitude can be sensitive to tuning parameters, but that the rank ordering of these importance scores across the multiple pollutants are relatively stable (Liaw and Wiener, 2002).

To our knowledge, this work represents the first instance of incorporating structure among pollutants within the kernel machine framework. By grouping highly correlated pollutants together, our approach had greater power to detect associations between these pollutants and health. In some situations, such groups may represent pollution sources, but not in the (relatively) small toxicology application we considered. In future work, we will consider more complex structures, such as overlapping groups, that likely occur in air pollution source apportionment settings. Another useful extension of the model would be to account for exposure measurement error, which may arise from known error in the measured concentrations, or from uncertainty in estimated source contributions from a source apportionment model (Kioumourtzoglou *and others*, 2014) or in predicted exposures obtained from a spatial model addressing misalignment of the pollutant and outcome data (Gryparis *and others*, 2009; Szpiro *and others*, 2011; Szpiro and Paciorek, 2013).

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

This work was supported by a grant from the Health Effects Institute; National Institutes of Health [ES007142, ES016454, ES000002, ES014930, ES013744, ES017437, ES015533, ES022585]; and U.S. Environmental Protection Agency (EPA) [83479801]. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the US EPA. Further, US EPA does not endorse the purchase of any commercial products or services mentioned in the publication.

References

Bailey, N. (2005). *Bayley Scales of Infant and Toddler Development, Administration Manual*, 3rd edition. San Antonio, TX: Harcourt Assessment.

Bartoli, C. R., Wellenius, G. A., Diaz, E. A., Lawrence, J., Coull, B. A., Akiyama, I., Lee, L. M., Okabe, K., Verrier, R. L. and Godleski, J. J. (2009). Mechanisms of inhaled fine particulate air pollution-induced arterial blood pressure changes. *Environmental Health Perspectives* **117**(3), 361–366.

Billionnet, C., Sherrill, D. and Annesi-Maesano, I. (2012). Estimating the health effects of exposure to multi-pollutant mixture. *Annals of Epidemiology* **22**(2), 126–141.

Breiman, L. (2001). Random forests. *Machine Learning* **45**(1), 5–32.

Carlin, D. J., Rider, C. V., Woychik, R. and Birnbaum, L. S. (2013). Unraveling the health effects of environmental mixtures: an NIEHS priority. *Environmental Health Perspectives* **102**(1), A6–A8.

Claus Henn, B., Schnaas, L., Ettinger, A. S., Schwartz, J., Lamadrid-Figueroa, H., Hernández-Avila, M., Amarasiriwardena, C., Hu, H., Bellinger, D. C., Wright, R. O. and Téllez-Rojo, M. M. (2012). Associations of early childhood manganese and lead coexposure with neurodevelopment. *Environmental Health Perspectives* **120**(1), 126–131.

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

Gennings, C., Sabo, R. and Carney, E. (2010). Identifying subsets of complex mixtures most associated with complex diseases: polychlorinated biphenyls and endometriosis as a case study. *Epidemiology* **21**(Suppl 4), S77–S84.

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

Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J. and Coull, B. A. (2009). Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics* **10**(2), 258–274.

Hu, H., Shine, J. and Wright, R. O. (2007). The challenge posed to children's health by mixtures of toxic waste: the Tar Creek Superfund site as a case-study. *Pediatric Clinics of North America* **54**(1), 155–75, x.

Kioumourtzoglou, M.-A., Coull, B. A., Dominici, F., Koutrakis, P., Schwartz, J. and Suh, H. (2014). The impact of source contribution uncertainty on the effects of source-specific pm2.5 on hospital admissions: a case study in Boston, MA. *Journal of Exposure Science and Environmental Epidemiology* **24**(4), 365–371.

Kioumourtzoglou, M.-A., Zanobetti, A., Schwartz, J. D., Coull, B. A., Dominici, F. and Suh, H. H. (2013). The effect of primary organic particles on emergency hospital admissions among the elderly in 3 US cities. *Environmental Health* **12**(19), 20.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News* **2**, 18–22. http://CRAN.R-project.org/doc/Rnews/.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D. and Kenny, Q. Y. (2006). Variable selection for Gaussian process models in computer experiments. *Technometrics* **48**(4), 478–490.

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

Maity, A. and Lin, X. (2011). Powerful tests for detecting a gene effect in the presence of possible gene-gene interactions using garrote kernel machines. *Biometrics* **67**(4), 1271–1284.

Patel, C. J., Bhattacharya, J. and Butte, A. J. (2010). An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLOS ONE* **5**. Article Number: e10746.

Savitsky, T., Vannucci, M. and Sha, N. (2011). Variable selection for nonparametric Gaussian process priors: models and computational strategies. *Statistical Science* **26**(1), 130–149.

SZPIRO, A. A. AND PACIOREK, C. J. (2013). Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics* **24**(8), 501–517.

SZPIRO, A. A., SHEPPARD, L. AND LUMLEY, T. (2011). Efficient measurement error correction with spatially misaligned data. *Biostatistics* **12**(4), 610–623.

THOMAS, D. C., WITTE, J. S. AND GREENLAND, S. (2007). Dissecting effects of complex mixtures: who's afraid of informative priors? *Epidemiology* **18**(2), 186–190.

TIBSHIRANI, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.

WOOD, S. N. (2006). *Generalized Additive Models: An Introduction with R.* Boca Raton, FL: CRC Press.

ZOU, F., HUANG, H., LEE, S. AND HOESCHELE, I. (2010). Nonparametric bayesian variable selection with applications to multiple quantitative trait loci mapping with epistasis and gene-environment interaction. *Genetics* **186**(1), 385–394.

[*Received May 22, 2014; revised November 3, 2014; accepted for publication November 7, 2014*]