Environmental Health

Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression

Jennifer F. Bobb[1,2*], Birgit Claus Henn[3], Linda Valeri[4] and Brent A. Coull[5]

Abstract

Background: Estimating the health effects of multi-pollutant mixtures is of increasing interest in environmental epidemiology. Recently, a new approach for estimating the health effects of mixtures, Bayesian kernel machine regression (BKMR), has been developed. This method estimates the multivariable exposure-response function in a flexible and parsimonious way, conducts variable selection on the (potentially high-dimensional) vector of exposures, and allows for a grouped variable selection approach that can accommodate highly correlated exposures. However, the application of this novel method has been limited by a lack of available software, the need to derive interpretable output in a computationally efficient manner, and the inability to apply the method to non-continuous outcome variables.

Methods: This paper addresses these limitations by (i) introducing an open-source software package in the R programming language, the *bkmr* R package, (ii) demonstrating methods for visualizing high-dimensional exposure-response functions, and for estimating scientifically relevant summaries, (iii) illustrating a probit regression implementation of BKMR for binary outcomes, and (iv) describing a fast version of BKMR that utilizes a Gaussian predictive process approach. All of the methods are illustrated using fully reproducible examples with the provided R code.

Results: Applying the methods to a continuous outcome example illustrated the ability of the BKMR implementation to estimate the health effects of multi-pollutant mixtures in the context of a highly nonlinear, biologically-based dose-response function, and to estimate overall, single-exposure, and interactive health effects. The Gaussian predictive process method led to a substantial reduction in the runtime, without a major decrease in accuracy. In the setting of a larger number of exposures and a dichotomous outcome, the probit BKMR implementation was able to correctly identify the variables included in the exposure-response function and yielded interpretable quantities on the scale of a latent continuous outcome or on the scale of the outcome probability.

Conclusions: This newly developed software, integrated suite of tools, and extended methodology makes BKMR accessible for use across a broad range of epidemiological applications in which multiple risk factors have complex effects on health.

Keywords: Multiple exposures, Mixtures, Exposure-response, Variable selection, Health risk estimation

* Correspondence: jennifer.f.bobb@kp.org
[1]Biostatistics Unit, Kaiser Permanente Washington Health Research Institute, 1730 Minor Ave #1600, Seattle, WA 98101, USA
[2]Department of Biostatistics, University of Washington, Seattle, WA, USA
Full list of author information is available at the end of the article

Bobb *et al. Environmental Health* (2018) 17:67

Page 2 of 10

Background

Estimating the health effects of several concurrent exposures is of increasing interest in epidemiology. For example, in environmental health interest lies in estimating the impacts of multi-pollutant mixtures, such as air pollution [1], toxic waste [2], and persistent organic chemicals [3]. Although studies have traditionally focused on estimating the health impacts of individual exposures, it is increasingly being recognized that populations are exposed to a wide range of factors across multiple domains, including environmental stressors and genetic and psychosocial determinants, and that these factors should be considered in conjunction [4].

A major barrier to studying the joint effects of many exposures concurrently is the lack of established statistical methods and corresponding software. Estimating the health effects of environmental mixtures is challenging because (i) exposures often have nonlinear and non-additive (eg, interactive) relationships with health outcomes, (ii) a high-dimensional vector of exposures may lead to poorly fitting regression models as the number of exposures increases relative to the number of observations in the dataset, and (iii) exposures are often highly correlated. Additionally, there are often several objectives of a multi-exposure health effect analysis, which may include estimating the overall effect of the mixture, identifying individual components that are responsible for the health effects of the mixture, visualizing the exposure-response function, and detecting interactions among pollutants [5]. Several statistical methods have been proposed for estimating the health effects of multiple exposures, including machine learning methods such as random forests [6]; clustering methods and other dimensional reduction methods such as principal components analysis, factor analysis, and structural equation models; and regression penalization methods such as the lasso [7]. However, these methods have typically addressed some but not all of the challenges and/ or scientific objectives described above. Reviews of prior methods and their limitations, as well as systematic comparisons of the performance of selected methods, have been published previously [1, 8–12].

Recently, we developed a new approach for estimating the joint health effects of multivariate exposures, Bayesian kernel machine regression (BKMR), that simultaneously addresses the challenges and scientific objectives described above [13]. First, through use of a kernel function, this approach estimates the multivariable exposure-response function in a flexible way that allows for nonlinear and non-additive effects, while adjusting for covariates including potential confounding factors. Second, the approach simultaneously incorporates variable selection on the (potentially large number of) exposures in a way that controls for multiple testing [14]; this

enables a parsimonious representation of the exposure-response function. Third, we developed a hierarchical variable selection approach that addresses the issue of multicollinearity by first classifying highly correlated exposures into groups, and then simultaneously conducting variable selection on the groups of correlated exposures as well as on the individual exposures within each group. In our prior methodological work describing BKMR [13, 15], we conducted a comprehensive evaluation of the performance of this approach. Through simulation studies based on real-world datasets, we found that (i) BKMR could well estimate exposure-response functions that included both nonlinear and non-additive effects, (ii) BKMR could identify important mixture components through variable selection, and (iii) the hierarchical variable selection approach could detect important groups of highly correlated exposures even in situations where individual components could not be identified. Additionally, BKMR has been applied previously in both toxicological and epidemiological studies leading to scientific insights that were not uncovered using standard regression approaches [15, 16].

Several important gaps limit the applicability of statistical methods for estimating the health effects of multi-pollutant mixtures in environmental health studies. These include a lack of software applying new methodology, data-generating scenarios with complex features (e.g., clustered outcome data), and the need for computationally efficient algorithms that yield correct results in a fraction of the time. A particular challenge of studies estimating the health effects of mixtures is the need to visualize the high-dimensional exposure-response function and to conduct inference in the presence of possible nonlinear and interactive associations of the exposures with the health outcome.

Here we provide several contributions addressing these gaps. First, we introduce an open-source software package (*bkmr*) [17] that implements the new BKMR approach for studying mixtures within the R statistical program [18]. This software provides a general, user-friendly implementation of BKMR, along with a suite of functions for processing model output to enable investigators to address the multifold objectives of a multi-exposure heath effect analysis. Second, we demonstrate methods for characterizing high-dimensional exposure-response functions, including visualizing the exposure-response relationship, and estimating scientifically relevant summaries, such as overall, single-exposure, and interactive health effects. Third, we present extensions to BKMR to enable model fitting to a broader class of applications, including a probit regression implementation for binary outcomes, and the ability to include a random intercept to account for correlated outcome data. Finally, we illustrate how a fast version of BKMR that utilizes a Gaussian predictive process

Bobb *et al. Environmental Health* (2018) 17:67

Page 3 of 10

approach can substantially speed up the model fitting. All of the examples used to illustrate these methods are fully reproducible with the provided R code.

Methods

Overview of BKMR

We first provide a brief overview of BKMR. The kernel machine regression (KMR) model for a continuous outcome is given by

$$Y_i = h(z_{i1}, ..., z_{iM}) + x_i'\beta + \epsilon_i,$$

where Y_i denotes the response for individual i ($i = 1, ..., n$), z_{im} is the mth exposure variable, h denotes the unknown exposure-response function to be estimated, β represents the effect of the covariates (note that x_i is a vector), and the residuals $\epsilon_i \sim N(0, \sigma^2)$ are assumed to be independent and identically (iid) normally distributed with a common variance. As described below, the effect of the exposures of interest modeled through the h function is allowed to be nonlinear and non-additive; the effect of the covariates could be modeled either linearly or more flexibly (e.g., by specifying a spline basis with a fixed number of degrees of freedom [DF] for one or more covariates). Additionally, if any covariates are hypothesized to interact with components of the mixture, then those covariates may be also be included in h.

For studies of multi-pollutant mixtures, the function h may include a large number of exposures of interest, and the relationship between these exposures and the health outcome can be complex, including nonlinear associations of one or more exposures, as well as possible interactions. Even with just a few exposures in the mixture, the combination of nonlinear and non-additive associations can lead to a high-dimension exposure-response relationship. As an illustration, if one were to model each exposure in the mixture using a spline basis with three DF to allow for nonlinearity and also include all of the interaction terms, this would result in a model with 255 parameters in the case of 4 exposures, 1023 parameters in the case of 5 exposures, and more generally, $(1 + DF)^M - 1$ parameters in the case of M exposures. In this setting of a high-dimensional exposure-response function, it can be challenging to specify a set of basis functions (e.g., polynomial or spline terms), and fitting a model including all basis functions and their interactions, as illustrated above, can lead to problems with over-fitting. BKMR addresses this by using a kernel machine representation for h, which regularizes the high-dimensional exposure-response function (details are in [13]). Under the kernel machine representation, rather than directly model the association of the exposures with the health outcome, one instead specifies a kernel function $K(z_i, z_j)$ that induces correlation of

health outcomes among individuals with similar exposure profiles $z = (z_1, ..., z_M)$. In particular, the KMR model assumes that two individuals with similar values of z (i.e. z_i close to z_j) will have similar health risks (i.e., $h_i = h(z_i)$ will be close to $h_j = h(z_j)$).

Operationally, by using the kernel machine representation, the KMR model may be expressed as a mixed-effect model [19], and within a Bayesian context, prior distributions are placed on all of the unknown parameters. The model is fit using Markov chain Monte Carlo (MCMC). Full details of the hybrid Gibbs/Metropolis-Hastings MCMC algorithm is described in Bobb et al. (2015) [13] and the supplemental material thereof.

Incorporating variable selection

To incorporate variable selection, the kernel function may be augmented with auxiliary variables (r_m, for $m = 1, ..., M$), such that when the auxiliary parameter is equal to zero, then the corresponding exposure variable is no longer included in the model (i.e., if r_m is equal to zero then exposure z_m is not selected). Fitting BKMR with component-wise variable selection yields estimates of the posterior inclusion probabilities, which provide measures of variable importance for each exposure. Alternatively, one can apply a hierarchical variable selection approach, in which groups of exposures are specified. In this latter scenario, BKMR estimates the posterior inclusion probability for each pollutant group, as well as posterior inclusion probabilities among pollutants within each group, given that the group was selected into the model. An example of when the hierarchical variable selection approach is useful is to address the issue of multicollinearity by placing highly correlated pollutants into the same group. This approach is evaluated and compared to the component-wise variable selection approach by Bobb et al. (2015) [13].

Extension to clustered outcome data

In the setting of correlated outcome data, through repeated measures within the same person, or through individuals clustered within families or communities, the BKMR model can be extended as $Y_{ij} = h(z_{ij1}, ..., z_{ijM}) + b_i + x_{ij}'\beta + \epsilon_{ij}$, where Y_{ij} is the response for observation j within cluster (e.g., person) i, $b_i \sim N(0, \tau_b^2)$ is a random intercept and $\epsilon_{ij} \sim N(0, \sigma^2)$ are the residual error terms.

Methods for characterizing the exposure-response function

From fitting the BKMR model, one obtains an estimate of the exposure-response function h, which may include nonlinear and non-additive associations. Unless there are very few mixture components, it is not possible to visualize the entire exposure-response function all at once. Therefore, tools are needed to visualize cross-sections of

Bobb *et al. Environmental Health* (2018) 17:67

Page 4 of 10

h. One example of a cross-section of interest is to visualize how a single exposure is related to the outcome when all of the other exposures are fixed to specific level (e.g., median value). Similarly, one could visualize the bivariate relationship of two exposures with the health outcome, while fixing all of the other exposures to a specific level. These cross sections and others can be visualized using the *bkmr* software package and are illustrated below.

In addition to visualizing the exposure-response relationship, inference may be conducted on scientifically relevant summaries of interest. Here we define three such summaries, which quantify overall, single-exposure, and interactive health effects. In particular, we define the *overall effect* as the change in the mean outcome when all of the exposures ($z_1, ..., z_M$) are fixed at their 75th percentile as compared to when all of the exposures are fixed to their 25th percentile, with all of the covariates x held constant. With notation, this is given by $\Delta_{tot}(25, 75)$ $= h(z_1^{75}, ..., z_M^{75}) - h(z_1^{25}, ..., z_M^{25})$, where z_m^p denotes the *p*th percentile of the *m*th exposure variable. A second quantity of interest is a *single-exposure effect*, which we define as the change in the mean outcome when a single exposure is fixed at its 75th percentile as compared to when it is fixed at its 25th percentile, when all of the other exposures are fixed at their median value and all of the covariates x are held constant. For example, for exposure z_1, this is given by $\Delta_1(25, 75 \,|50) = h(z_1^{75}, z_2^{50}, ..., z_M^{50}) - h(z_1^{25}, z_2^{50}, ..., z_M^{50})$, and for other exposures the quantity $\Delta_m(25, 75 \,|50)$ is defined analogously. Quantifying potential interaction is often another major goal of a mixtures health effect analysis. To facilitate this, we define an *interactive effect* as the difference in the single-exposure health effect when all of the other exposures are fixed at their 75th percentile, as compared to when all of the other exposures are fixed at their 25th percentile, given by $\Delta_m(25, 75 \,|75) - \Delta_m(25, 75 \,|25)$. We note that the choice here of using the 25th and 75th percentiles is illustrative; these values may be modified as desired, and the above summaries can be calculated using any choice of threshold. Within a Bayesian framework, inference on the parameters above is conducted by calculating posterior mean estimates and 95% credible intervals for any of the numerical summaries of interest. Inference on other functionals of *h* that set exposures to fixed values may be conducted analogously.

Probit BKMR for binary outcomes

BKMR can be extended to binary outcomes via generalized linear modeling. For reasons of computational efficiency for Bayesian inference, we use probit, rather than logistic, regression. The probit BKMR model is given by

$$\Phi^{-1}(\mu_i) = h(z_{i1}, ..., z_{iM}) + x_i'\boldsymbol{\beta},$$

where Φ is the cumulative distribution function (CDF) for the standard normal distribution (Φ^{-1} is the probit link function) and $\mu_i = P(Y_i = 1)$ is the probability of an event (Y_i is a binary [0/1] variable).

It is well known that the probit model can be expressed using a latent normal random variable formulation. In particular, the probit model above can be expressed as $Y_i^* = h(z_{i1}, ..., z_{iM}) + x_i'\boldsymbol{\beta} + e_i$, where e_i is standard normal and $Y_i = I(Y_i^* > 0)$ is equal to 1 if $Y_i^* > 0$ and is equal to zero otherwise. Under this formulation, extension of the BKMR model from Gaussian outcomes to binary outcomes is relatively straightforward. One can simply apply the MCMC algorithm derived for normally distributed outcomes with an additional step of sampling from the posterior distribution of the latent Y_i^* variables using a truncated normal distribution.

Although probit regression tends to be less common than logistic regression in many environmental health applications, it yields interpretable quantities both on the scale of the latent continuous outcome and on the scale of the outcome probability. In particular, by considering the latent normal formulation above, *h* may be interpreted as the relationship between the exposures and some underlying, continuous latent variable (Y*). For example, if Y is an indicator variable for whether an individual has a particular health outcome, Y* could be interpreted as a latent marker of health status. Additionally, probit model coefficients can be converted into more familiar odds ratios using well-known formulae [20]. In particular, we have $logit(\mu) \approx 1.6 \cdot \Phi^{-1}(\mu)$ so that $\beta_{logit} \approx 1.6 \cdot \beta_{probit}$. This approximation works well provided the probability of the outcome event given the included predictors is not too close to 0 or 1.

Gaussian predictive process

A major computational burden in fitting BKMR is the need to invert an n-by-n matrix (multiple times) at each iteration of the algorithm, where n is equal to the number of observations in the data. One way to reduce the computation time is to employ a Gaussian predictive process [21], an approach originally developed for large spatial datasets and which has been used previously in models with Gaussian process priors [22]. In this approach one specifies a set of points (referred to as 'knots') that covers the exposure space and then computes the projection of each vector of exposures onto the lower dimensional space spanned by the set of knots. Under this approximate approach, rather than needing to invert an n-by-n matrix, the algorithm only needs to invert a square matrix with dimension equal to the number of knots.

Bobb *et al. Environmental Health* (2018) 17:67

Page 5 of 10

Software implementation

The *bkmr* software is implemented as an R (R Development Core Team 2017) package. It has dependencies to the following packages: dplyr, magrittr, nlme, fields, trunccnorm, tidyr, MASS, and tmvtnorm. The R software and these required packages can be obtained from the CRAN website at [https://cran.r-project.org/]. Furthermore, daily builds of package bkmr are provided on the CRAN website [https://cran.r-project.org/web/packages/bkmr/index.html]. It has been published under GPL version 2. Source code is available on GitHub at [https://github.com/jenfb/bkmr].

The package provides a complete framework for applying BKMR to conduct an analysis of the health effects of multiple exposures. The main function (*kmbayes*) implements a MCMC sampler to fit a BKMR model and includes the following features:

- the outcome may be either continuous or binary (specified using the *family* argument)
- option to include a random intercept to account for clustered or repeated measures outcome data (*id* argument)
- option to fit the model with or without variable selection (*varsel*)
- option to apply hierarchical variable selection (*groups*)
- implements a Gaussian predictive process approach to speed up model fitting for large sample sizes (*knots*)
- option to change default settings for the MCMC algorithm (*control.params* argument)

After fitting the BKMR model, a suite of post-processing functions are available, including functions to:

- provide a parsimonious summary of model output (*print* and *summary* methods)
- extract estimates of the posterior inclusion probabilities, which provide measures of variable importance for each exposure (*ExtractPIPs* function)
- extract summaries of posterior distributions of model parameters, including posterior mean, standard deviation, and quantiles (*ExtractEsts* function)
- obtain scientifically relevant summaries of the multivariable exposure-response function (these functions are illustrated in detail through the continuous outcome example below).

Example code illustrating the main kmbayes function is shown in Fig. 1. Additional details on the BKMR implementation are available in the package overview guide [https://jenfb.github.io/bkmr/overview.html].

Practical considerations

Inference based on BKMR is contingent upon convergence of the MCMC algorithm. Several approaches can be used to monitor convergence, including visually inspecting the trace plots of model parameters, or more formal methods such as the Gelman-Rubin diagnostic [23]. The package overview guide [https://jenfb.github.io/bkmr/overview.html] provides details on how to modify the tuning parameters for running the MCMC algorithm in order to speed up convergence.

Additionally, it is good practice to evaluate the sensitivity of results to the choice of prior distribution specification. This can be done in the R package by changing the default settings. Of note, we have found that when conducting BKMR with variable selection, the magnitudes of the posterior inclusion probabilities can be sensitive to the choice of the prior distribution on the r_m parameters (though in our experience the relative ordering of the posterior inclusion probabilities has tended to remain stable) [15]. We therefore recommend varying the specifications of the prior distributions for these r_m parameters; additional guidance is given in the overview guide [https://jenfb.github.io/bkmr/overview.html], including an approach for incorporating prior knowledge on the degree of smoothness of the exposure-response function.

Results

We illustrated the above approaches using two example datasets. For the first example, we applied BKMR to a simulated dataset that was generated as part of the 2015 workshop hosted by the National Institute for Environmental Health Sciences (NIEHS), titled "Statistical Approaches for Assessing Health Effects of Environmental Chemical Mixtures in Epidemiology Studies." The goal of the workshop was to compare statistical methods by applying them to common datasets developed by epidemiologists and toxicologists based on real-world data applications [24]. A key feature of the workshop was that it used simulated datasets generated by scientists who did not develop the statistical methods being compared, which provides an objective benchmark for evaluating the methods' performance. Applying BKMR in this setting illustrates the performance of the method in the context of a highly nonlinear, biologically-based dose-response function. For the second example, we considered a simulated dataset with a larger number of exposures and a dichotomous outcome, in order to illustrate probit BKMR.

Continuous outcome example

For the continuous outcome setting, we used the first simulated data set (Data Set #1) from the NIEHS workshop [25], which included 7 exposure variables ($z_1, ..., z_7$)

Bobb *et al. Environmental Health* (2018) 17:67

Page 6 of 10

```
## load R package
library(bkmr)

## Generate example dataset
dat <- SimData(n = 300)

## Fit BKMR without variable selection
km <- kmbayes(y = dat$y, Z = dat$Z, X = dat$X, verbose = FALSE)

## Fit with variable selection - add more iterations
kmsel <- kmbayes(y = dat$y, Z = dat$Z, X = dat$X, verbose = FALSE,
                 varsel = TRUE, iter = 2000)

## Generate knot matrix; fit with Gaussian predictive process
knots <- fields::cover.design(dat$Z, nd = 50)$design
kmknot <- kmbayes(y = dat$y, Z = dat$Z, X = dat$X, verbose = FALSE,
                  varsel = TRUE, iter = 2000, knots = knots)
```

Fig. 1 Usage example showing R code to fit BKMR with a continuous outcome. Here 'y' denotes the response vector of length n (where n is the number of observations); 'Z' is the n-by-M exposure matrix, where M is the number of exposure variables included in the exposure-response function h; and 'X' is the n-by-P covariate matrix, where P is the number of covariates

and a single covariate x in 500 individuals. We applied BKMR to fit the model $E[Y_i] = h(z_{i1}, ..., z_{i7}) + \beta x_i$, where Y_i denotes the response for individual i, h denotes the unknown exposure-response function to be estimated, and β represents the effect of the covariate. Reproducible code along with complete results from the analysis is available at [https://jenfb.github.io/bkmr/SimData1.html]; here we describe select results.

As mentioned above, several functions are provided for processing the model output. Variable selection yields posterior inclusion probabilities (PIPs), whose values range from 0 to 1 and whose magnitude indicates relative variable importance. In the simulated example, the estimated PIPs were close to 0 for exposures z_3 and z_6 and were 1 for the remaining exposures. To illustrate the methods for visualizing the multivariable exposure-response function, we explored different cross-sections (Fig. 2). For example, Fig. 2a shows the (covariate-adjusted) association of z_7 with the outcome, which indicates a nonlinear relationship with a steeper slope at lower levels of exposure that appears to plateau at higher exposure levels. Figure 2b shows the joint association of z_1 and z_7 with the response at different percentiles of a third exposure (z_5), which is useful for visualizing potential three-way interactions, though in this example the similar pattern in association across levels of z_5 suggests a lack of evidence of three-way interaction.
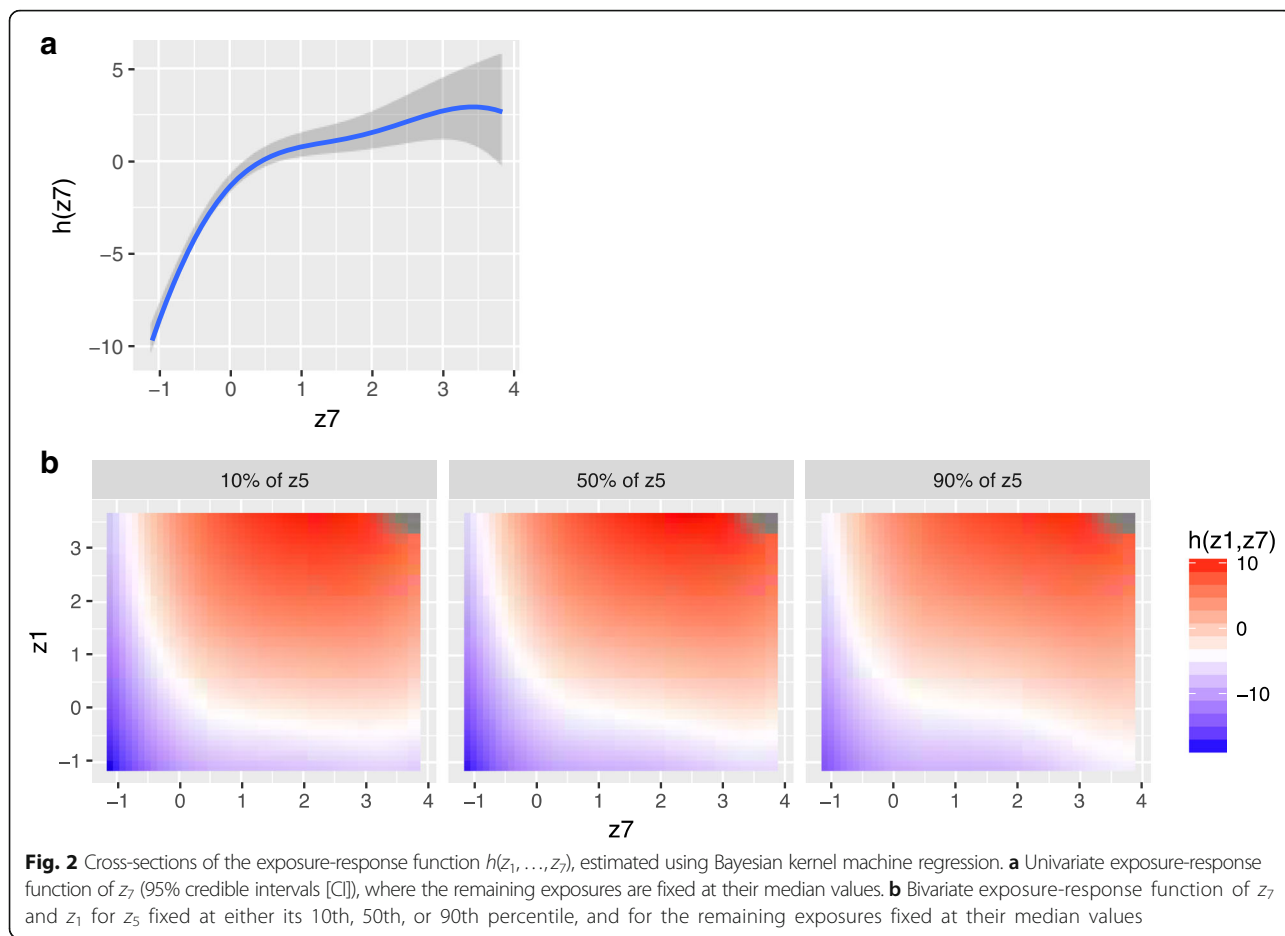
We additionally calculated statistics summarizing the scientifically-relevant features of the exposure-response function described above (Fig. 3). Estimates of the overall effect of the mixture (3a) revealed that increasing levels of joint exposure were associated with higher levels of the outcome. To characterize the contribution of individual exposures to the overall effect, single-exposure effect estimates (3b) suggested that increases in

exposure to z_7, z_1, and z_2 were associated with higher levels of the outcome and that increases in exposure to z_5 and z_4 were associated with lower levels of the outcome. The single-exposure estimate for z_5 was larger in magnitude when all of the remaining exposures were fixed at their 75th percentile as compared to when they were fixed at their 25th percentile, indicating possible interaction of z_5 with one (or more) of the other exposure variables. To further explore this possibility, we calculated interactive effects (3c), which suggested that this interaction is statistically significant.

Comparisons of our results to the true exposure-response function used to generate the simulated dataset [26] demonstrate that BKMR was correctly able to identify which exposures were truly associated with the outcome and the direction of these associations. In addition, we were able to identify the nonlinear exposure-response relationship of the individual predictors, and to well-approximate the full exposure-response function that included both nonlinear and non-additive associations (See [https://jenfb.github.io/bkmr/SimData1.html]). Using the approximate, Gaussian predictive process method led to a reduction in the runtime of 49% when 100 knots were used (from 0.137 to 0.070 s per MCMC iteration) and a reduction in the runtime of 74% when 50 knots were used (to 0.036 s per iteration), without any substantial decrease in accuracy in estimating the exposure-response function in this example. Computations were performed using a 1.7 GHz processor with 8 GB of memory.

Binary outcome example

To illustrate probit BKMR, we simulated a dataset that included 30 exposure variables for a sample size of $n = 200$. The binary outcome depended on quadratic terms of four of the exposures and on a linear interaction term

Bobb *et al. Environmental Health* (2018) 17:67

Page 7 of 10

Fig. 2 Cross-sections of the exposure-response function $h(z_1, \ldots, z_7)$, estimated using Bayesian kernel machine regression. **a** Univariate exposure-response function of z_7 (95% credible intervals [CI]), where the remaining exposures are fixed at their median values. **b** Bivariate exposure-response function of z_7 and z_1 for z_5 fixed at either its 10th, 50th, or 90th percentile, and for the remaining exposures fixed at their median values

between two of these. Reproducible code and detailed results for this example are available at [https://jenfb.github.io/bkmr/ProbitEx.html]; select results are shown in Fig. 4. Posterior inclusion probabilities indicate that BKMR was correctly able to identify the variables included in the exposure-response function (4a) and to identify the quadratic exposure-response function without assuming this relation a priori (4b). As discussed above, the probit BKMR model yields interpretable quantities on the scale of the latent continuous outcome and on the scale of the outcome probability. Here, the u-shaped relationship of z_1 suggests that both higher and lower levels of exposure may be associated with higher levels of the latent continuous outcome as compared to moderate levels of exposure.

We also illustrate how one can use the predicted probabilities from probit BKMR to compute quantities of interest, such as the risk difference (4c). For example, the point estimate (95% posterior credible interval) for the risk difference comparing the probability of the binary outcome when exposure 2 is at its 75th percentile versus its 50th percentile, for all of the remaining exposures fixed at their median value, was 0.42 (0.02, 0.73)

when the single confounder x is fixed at its 25th percentile and was 0.32 (0.01, 0.72) when the confounder x is fixed at its 75th percentile. (The true risk difference was 0.32 and 0.28, respectively.) This indicates evidence of a statistically significant association between increasing levels of exposure 2 from moderate to high with an increased absolute risk of the outcome that persists across levels of confounder x.

Discussion

The *bkmr* software package provides a general, open-source implementation of BKMR, a new and flexible approach for estimating the joint health effects of simultaneous exposure to multiple concurrent risk factors. The model specification can accommodate a broad range of data application scenarios common in environmental health, including continuous or binary outcomes, repeated-measures or clustered outcome data, and highly correlated exposures. A suite of functions is provided to process model output, addressing scientific questions of interest on features of the multivariate exposure-response relationship.
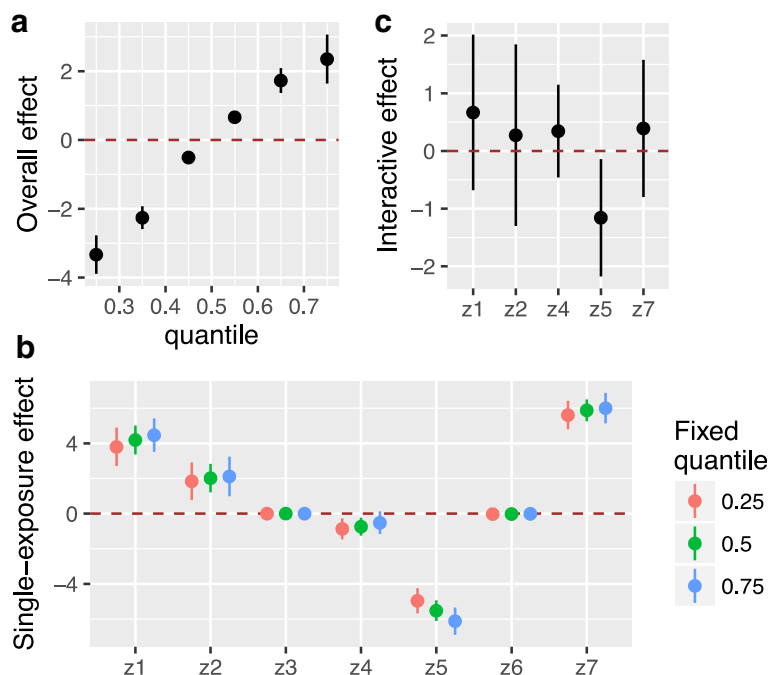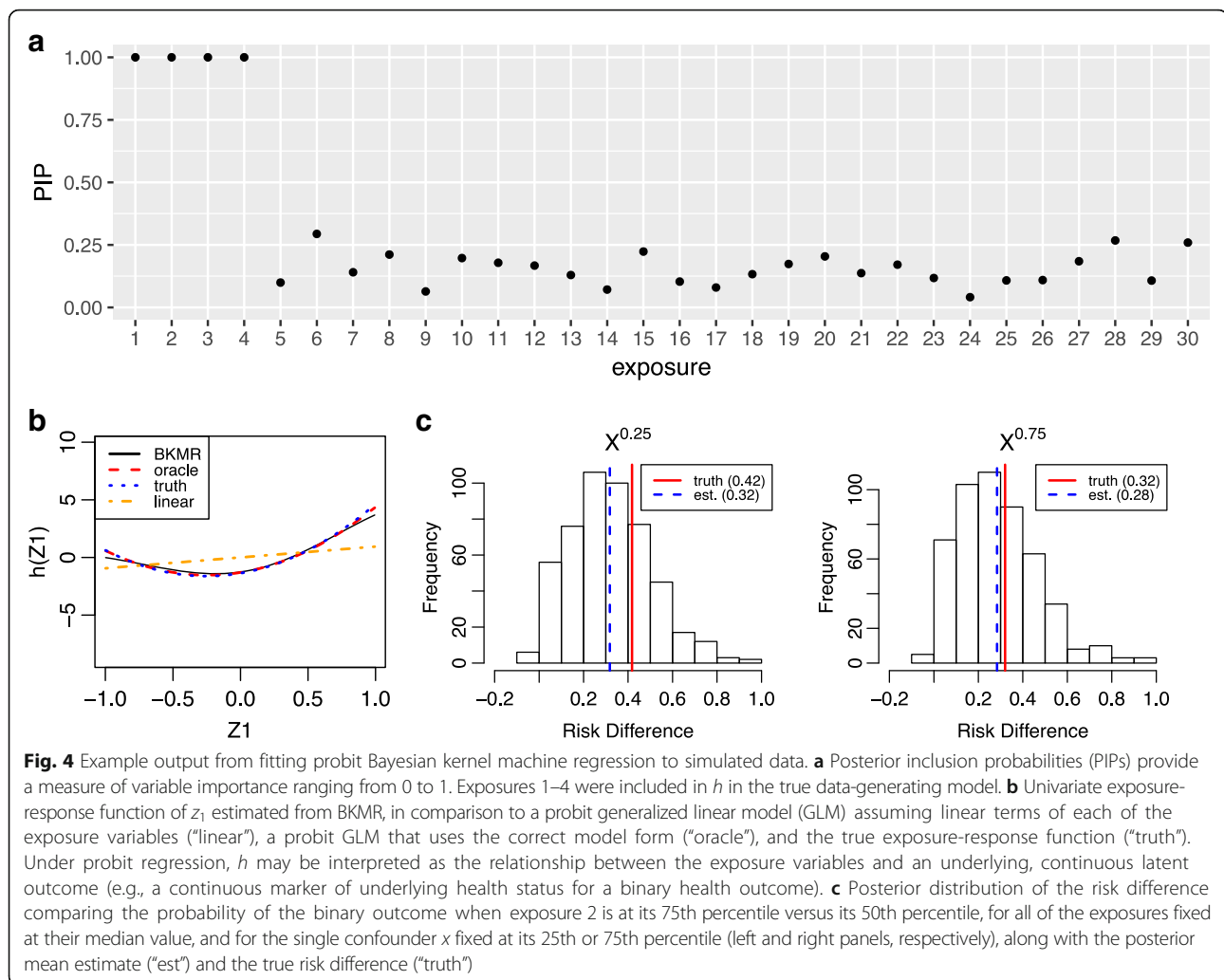
Bobb *et al. Environmental Health* (2018) 17:67

Page 8 of 10

Fig. 3 Numerical summaries of the exposure-response function $h(z_1, \ldots, z_7)$, estimated using Bayesian kernel machine regression. **a** Overall effect of the mixture (95% CI), defined as the difference in the response when all of the exposures are fixed at a specific quantile (ranging from 0.25 to 0.75), as compared to when all of the exposures are fixed at their median value. **b** Single-exposure health effects (95% CI), defined as the change in the response associated with a change in a particular exposure from its 25th to its 75th percentile, where all of the other exposures are fixed at a specific quantile (0.25, 0.50, or 0.75). **c** Interactive effects, defined as the change in the single-exposure health effects when all of the remaining exposures are fixed at their 25th percentile as compared to when they are fixed at their 75th percentile (i.e., red points from Panel **b** subtracted from the corresponding blue points)

A key feature of BKMR is the estimation of the multi-variable exposure-response function, which may often be high-dimensional in studies of the health effects of environmental mixtures. However, it can be challenging to conduct inference in this setting. Accordingly, we proposed several numerical summaries of the exposure-response function to allow investigators to estimate overall effects of the mixture, single-exposure health effects, and interactive effects. Rather than require quantities to correspond to specific parameters of a regression model (e.g., coefficients on main effect or interaction terms), as is often done in statistical modeling, the numerical summaries we proposed can be estimated regardless of the specific form of the regression model. Thus, they are broadly applicable to exposure-response relationships estimated from other (i.e., non-BKMR) statistical models. As with any dimension reduction technique, care is warranted when interpreting summary measures since they could mask potentially complex features of the data. For example, a single-exposure summary that compares health outcomes at high versus low exposure could appear null if there is a u-shaped relationship; likewise, an apparent null overall association may be observed if half of the exposures are positively associated and half are negatively associated with a similar magnitude. It is therefore

recommended to explore a range of summary measures and to visualize different cross-sections of the exposure-response surface, together with the PIP, or variable importance, scores.

The MCMC algorithm we implemented for fitting BKMR employs several tricks to speed up the computation. First, rather than update the subject-specific effects of the mixture h_i within the main function used to fit BKMR, we marginalize the posterior distribution over these parameters [13]. These subject-specific effects are typically not themselves of scientific interest; rather, investigators often desire estimates of the general form of the exposure-response function, which can be visualized and summarized via the post-processing functions described above. Second, the implementation for binary outcomes utilizes the latent normal specification of probit regression, which has computational advantages for Bayesian inference. Third, the software allows for applying a Gaussian predictive process approach [21], originally developed for large spatial datasets, which projects the exposure space onto a smaller number of points ('knots'), leading to efficient computation of health risk estimates.

Several additional functionalities could be added. Allowing for count outcome data via Poisson BKMR

Bobb *et al. Environmental Health* (2018) 17:67

Page 9 of 10

Fig. 4 Example output from fitting probit Bayesian kernel machine regression to simulated data. **a** Posterior inclusion probabilities (PIPs) provide a measure of variable importance ranging from 0 to 1. Exposures 1–4 were included in h in the true data-generating model. **b** Univariate exposure-response function of z_1 estimated from BKMR, in comparison to a probit generalized linear model (GLM) assuming linear terms of each of the exposure variables ("linear"), a probit GLM that uses the correct model form ("oracle"), and the true exposure-response function ("truth"). Under probit regression, h may be interpreted as the relationship between the exposure variables and an underlying, continuous latent outcome (e.g., a continuous marker of underlying health status for a binary health outcome). **c** Posterior distribution of the risk difference comparing the probability of the binary outcome when exposure 2 is at its 75th percentile versus its 50th percentile, for all of the exposures fixed at their median value, and for the single confounder x fixed at its 25th or 75th percentile (left and right panels, respectively), along with the posterior mean estimate ("est") and the true risk difference ("truth")

would enable the model to be applied to time series studies that estimate the joint health effects of day-to-day changes in multiple community-level risk factors (e.g., temperature and air pollution) on daily outcomes (e.g., hospitalization rates) [27]. However, implementation of the MCMC algorithm in this setting requires additional complexity because the computational tricks described above are not applicable. Additionally, the implementation focuses on a particular choice of kernel function for specifying the BKMR model, namely the Gaussian kernel. Our previous simulation studies showed that this specification is relatively flexible, accurately capturing a wide range of underlying forms of the true exposure-response function. However, the ability to specify other kernel functions could be added in the future. Along these same lines, our focus has been on estimating the joint health effects of continuous exposure variables; allowing for exposure-response surfaces that are functions of both categorical and continuous exposures may also be of interest. Finally, beyond estimating

specific interactive-effect summary measures, one may be more broadly interested in detecting whether two groups of exposure variables interact [28–30]. This could be done within the BKMR framework by applying kernel decomposition methods to evaluate whether the kernel function $h(z_1, z_2)$ could be expressed as $h(z_1) + h(z_2)$ for two groups of exposures (z_1 and z_2).

Conclusions

In summary, this newly developed software provides an integrated set of tools for conducting a mixtures health effect analysis. The software and expanded toolbox make BKMR accessible for use across a broad range of epidemiological applications in which a large number of exposures have complex, potentially nonlinear and interactive effects on health.

Abbreviations
BKMR: Bayesian kernel machine regression; CI: Credible interval; DF: Degrees of freedom; KMR: Kernel machine regression; MCMC: Markov chain Monte Carlo; NIEHS: National Institute for Environmental Health Sciences; PIP: Posterior inclusion probability

Bobb *et al. Environmental Health* (2018) 17:67

Page 10 of 10

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information files.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Biostatistics Unit, Kaiser Permanente Washington Health Research Institute, 1730 Minor Ave #1600, Seattle, WA 98101, USA. [2]Department of Biostatistics, University of Washington, Seattle, WA, USA. [3]Department of Environmental Health, Boston University School of Public Health, Boston, MA, USA. [4]Psychiatric Biostatistics Laboratory, McLean Hospital, Belmont, MA, USA. [5]Department of Biostatistics, Harvard T H Chan School of Public Health, Boston, MA, USA.

References

1. Billionnet C, Sherrill D. Annesi-Maesano I, study G: estimating the health effects of exposure to multi-pollutant mixture. Ann Epidemiol. 2012;22:126–41.
2. Hu H, Shine J, Wright RO. The challenge posed to children's health by mixtures of toxic waste: the Tar Creek superfund site as a case-study. Pediatr Clin N Am. 2007;54:155–75.
3. Gennings C, Sabo R, Carney E. Identifying subsets of complex mixtures most associated with complex diseases: polychlorinated biphenyls and endometriosis as a case study. Epidemiology. 2010;21(Suppl 4):S77–S84.
4. Carlin DJ, Rider CV, Woychik R, Birnbaum LS. Unraveling the health effects of environmental mixtures: an NIEHS priority. Environ Health Perspect. 2013; 121:A6–8.
5. Braun JM, Gennings C, Hauser R, Webster TF. What can epidemiological studies tell us about the impact of chemical mixtures on human health? Environ Health Perspect. 2016;124:A6–9.
6. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
7. Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Stat Soc - Series B. 1996;58:267–88.
8. Dominici F, Peng RD, Barr CD, Bell ML. Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. Epidemiology. 2010;21:187–94.
9. Agier L, Portengen L, Chadeau-Hyam M, Basagana X, Giorgis-Allemand L, Siroux V, Robinson O, Vlaanderen J, Gonzalez JR, Nieuwenhuijsen MJ, et al. A systematic comparison of linear regression-based statistical methods to assess Exposome-health associations. Environ Health Perspect. 2016;124: 1848–56.
10. Barrera-Gómez J, Agier L, Portengen L, Chadeau-Hyam M, Giorgis-Allemand L, Siroux V, Robinson O, Vlaanderen J, González JR, Nieuwenhuijsen M, et al. A systematic comparison of statistical methods to detect interactions in exposome-health associations. Environ Health. 2017;16:74.
11. Sun Z, Tao Y, Li S, Ferguson KK, Meeker JD, Park SK, Batterman SA, Mukherjee B. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. Environ Health. 2013;12:85.
12. Stafoggia M, Breitner S, Hampel R, Basagaña X. Statistical approaches to address multi-pollutant mixtures and multiple exposures: the state of the science. Curr Environ Health Rep. 2017;4:481–90.
13. Bobb JF, Valeri L, Claus Henn B, Christiani DC, Wright RO, Mazumdar M, Godleski JJ, Coull BA. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics. 2015;16:493–508.
14. Scott JG, Berger JO. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. Ann Stat. 2010;38:2587–619.
15. Coull BA, Bobb JF, Wellenius GA, Kioumourtzoglou MA, Mittleman MA, Koutrakis P, Godleski JJ. Part 1. Statistical Learning Methods for the Effects of Multiple Air Pollution Constituents. Res Rep Health Eff Inst. 2015:5–50.
16. Valeri L, Mazumdar MM, Bobb JF, Claus Henn B, Rodrigues E, Sharif OIA, Kile ML, Quamruzzaman Q, Afroz S, Golam M, et al. 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. 2017;125:067015.
17. Bobb JF. bkmr: Bayesian Kernel Machine Regression. R package version 0.2. 0. 2017. https://github.com/jenfb/bkmr. Accessed 6 Apr 2017.
18. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2017. https://www.R-project.org/. Accessed 13 Aug 2018.
19. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007;63:1079–88.
20. Amemiya T. Qualitative response models: a survey. JEL. 1981;19:1483–536.
21. Banerjee S, Gelfand AE, Finley AO, Sang H. Gaussian predictive process models for large spatial data sets. J Royal Stat Soc - Series B. 2008;70:825–48.
22. Savitsky T, Vannucci M, Sha N. Variable selection for nonparametric Gaussian process priors: models and computational strategies. Stat Sci. 2011;26:130–49.
23. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Stat Sci. 1992;7:457–72.
24. Taylor KW, Joubert BR, Braun JM, Dilworth C, Gennings C, Hauser R, Heindel JJ, Rider CV, Webster TF, Carlin DJ. Statistical approaches for assessing health effects of environmental chemical mixtures in epidemiology: lessons from an innovative workshop. Environ Health Perspect. 2016;124:A227–9.
25. National Institute of Environmental Health Sciences. Statistical Approaches for Assessing Health Effects of Environmental Chemical Mixtures in Epidemiology Studies. 2015; https://www.niehs.nih.gov/news/events/pastmtg/2015/statistical/. Accessed 20, June 2018.
26. Synthetic Data Set #1. https://www.niehs.nih.gov/about/events/pastmtg/2015/statistical/simulated_dataset_1_answers.pdf. Accessed 10 Aug, 2017.
27. Bobb JF, Dominici F, Peng RD. Reduced hierarchical models with application to estimating health effects of simultaneous exposure to multiple pollutants. J Royal Stat Soc - Series C. 2013;62:451–72.
28. Lampa E, Lind L, Lind PM, Bornefalk-Hermansson A. The identification of complex interactions in epidemiology and toxicology: a simulation study of boosted regression trees. Environ Health. 2014;13:57.
29. Larson NB, Schaid DJ. A kernel regression approach to gene-gene interaction detection for case-control studies. Genet Epidemiol. 2013;37:695–703.
30. Antonelli J MM, Bellinger D, Christiani D, Wright R, Coull BA. Bayesian variable selection for multi-dimensional semiparametric regression models. 2017; https://arxiv.org/abs/1711.11239. Accessed 21 June 2018.