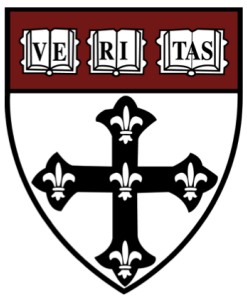
Workshop on Analyzing Mixtures in Environmental Health Studies: Bayesian Kernel Machine Regression

Brent Coull

Harvard T.H. Chan School of Public Health
Departments of Biostatistics and Environmental Health

20 August 2019



Mixtures health effect analysis

Possible objectives of a mixtures analysis could include:

- Detection / estimation of an effect of the overall mixture.
- Identification of pollutant or group of pollutants responsible for observed mixture effects.
- Visualizing the exposure-response function
- Detection of interactions among individual pollutants.

Bayesian Kernel Machine Regression (BKMR) is designed to address all four of these objectives.

Bird's-eye (over)view of existing mixtures methods

Today

- Overall strategy today:
  - This hour: BKMR modeling framework and options implemented in the `bkmr` R package.
  - Later today: Extensions and adaptations of BKMR for modern epidemiological study designs, for which prototype code is freely available (but not in the R package yet) or under development.

Kernel machine regression (KMR)

$$Y_i = h(z_{i1}, ..., z_{iM}) + \boldsymbol{\beta}\mathbf{x}_i + \epsilon_i$$

- Y_i: continuous, normally distributed health endpoint
- $\mathbf{z}_i$: M exposure measures
- $\mathbf{x}_i$: potential confounders
- $\boldsymbol{\beta}$: effects of potential confounders
- $\epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$
- $h\left(\cdot\right)$ is an unknown but smooth function

KMR as a Spatial Mixed Model

Operationally we fit the model

$$Y_i = h_i + \boldsymbol{\beta}\mathbf{x}_i + \epsilon_i, \quad i = 1, \ldots, n.$$

- We tie this to the multi-pollutant exposure by assuming
  - a pair of random effects (h_i, h_j) are correlated
  - the strength of this correlation is determined by the "distance" between a pair in the multivariate exposure space.
- By clever selection of the distance metric (i.e. the "kernel" function), we can accommodate nonlinear and non-additive effects of the multivariate exposure.

Liu, Lin and Ghosh, *Biometrics* 2007

We use the so-called Gaussian kernel, which implies a radial basis representation for $h\left(z_{i1}, \ldots, z_{iM}\right)$:

$$K_\rho(\mathbf{z}_i, \mathbf{z}_j) = \exp\left\{-\frac{1}{\rho}\sum_{m=1}^{M}\left(z_{im} - z_{jm}\right)^2\right\} \equiv K_{i,j}$$

So consider my modest study of n=4:

M-A Kioumourtzoglou, J. Goldsmith, C. Gennings, B. Coull

$$Y_i = h_i + \boldsymbol{\beta}\mathbf{x}_i + \epsilon_i, \;\; i = 1, \ldots, 4$$

$$\begin{pmatrix} h_{MK} \\ h_{JG} \\ h_{CG} \\ h_{BC} \end{pmatrix} \sim MVN\left[\mathbf{0}, \tau\begin{pmatrix} 1 & K_{MK,JG} & K_{MK,CG} & K_{MK,BC} \\ K_{MK,JG} & 1 & K_{JG,CG} & K_{JG,BC} \\ K_{MK,CG} & K_{JG,CG} & 1 & K_{CG,BC} \\ K_{MK,BC} & K_{JG,BC} & K_{CG,BC} & 1 \end{pmatrix}\right]$$

Bayesian Kernel Machine Regression (BKMR)

If there are only a few components driving health effects, it can be useful to embed a variable selection procedure within KMR:

$$Y_i = h\left(\overset{?}{z_{i1}}, \overset{?}{z_{i2}}, \overset{?}{z_{i3}}, \ldots, \overset{?}{z_{iM}}\right) + \boldsymbol{\beta}\mathbf{x}_i + \epsilon_i$$

- Can be learned from the data using the kernel with weights r_m:

$$K_{vs}(\mathbf{z}_i, \mathbf{z}_j) = \exp\left\{-\sum_{m=1}^{M} r_m \left(z_{im} - z_{jm}\right)^2\right\}$$

- $r_m = 0$ corresponds to to no association between Y and m^{th} exposure
- We obtain an estimate of the probability that each exposure is important.

Precise model for r_m provided in technical appendix at end of these slides

BKMR with Hierarchical Variable Selection

- In many analyses, there exists a structure among exposures:
  - Air pollutants from specific sources (XRF)
  - Chemical structure
  - Exposures from multiple time windows, ...

$$Y_i = h\left(\overbrace{Z^1_{i1}, Z^1_{i2}, Z^1_{in_1}}^{?}, \overbrace{Z^2_{i1}, Z^2_{i2}, \ldots, Z^2_{in_2}}^{?}, \ldots,\right) + \boldsymbol{\beta}\mathbf{x}_i + \epsilon_i$$

$$? \quad ? \quad ? \qquad ? \quad ? \qquad ?$$

- Now we obtain
  - an estimate of the probability each group of exposures is important
  - an estimate of the probability that, given a group is important, each exposure in that group is driving that group-outcome association.

Precise model for r_m provided in technical appendix at end of these slides

Estimation

- We choose a Bayesian approach to model fitting*.
- Advantages:
  - Can estimate the importance of each variable (r_m) simultaneously.
  - Uncertainty estimate for just about any quantity of interest.
  - Straightforward to extend to longitudinal data:

$$Y_{it} = h\left(z_{it1}, \ldots, z_{itM}\right) + \boldsymbol{\beta}\mathbf{x}_{it} + b_i + \epsilon_{it}$$

- In the end, we obtain as output
  - $\widehat{h}\left(z_{i1}, \ldots, z_{iM}\right)$ (exposure-response).
  - the posterior probability that variable m is important ($r_m > 0$)

    (or, in *hvs* framework, probability a group is important as well as the probability each variable is the one that drives the group-specific association)

Bobb et al. *Biostatistics* 2015; Coull et al. HEI Report 2015

Bayesian Paradigm to Statistical Inference

- Bayesian paradigm to statistical inference:

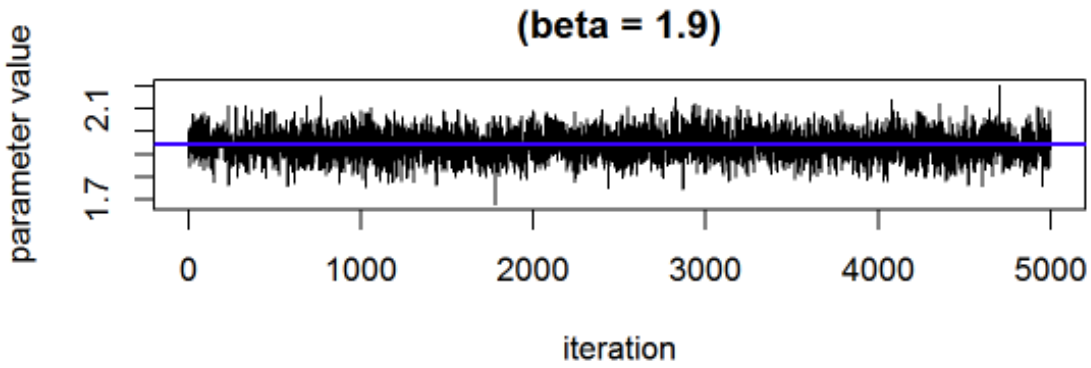
Knowledge about a model parameter (obtained from prior experience and/or data) forms a "belief structure" that assigns relative weight to different values of that parameter.

- These relative weightings are reflected by probability distributions.
- That is, model parameters are treated as random variables.

$$\text{Prior Distribution} + \text{Data} \rightarrow \text{Posterior Distribution}$$

Markov Chain Monte Carlo (MCMC)

- Like most complex Bayesian models, the posterior distributions of the model parameters in BKMR do not have a nice, tractable form.
- We therefore sample values from the posterior distributions using a Markov Chain. Practically, we need to
  - specify how many parameter values we want to sample
  - decide when the chain reaches steady state ("converges")
  - look at plots to diagnose that this has happened.



- Typically, for complicated models like BKMR, we need at least tens of thousands of samples.

Characterization of the Exposure-Response Function

- Unless there are very few mixture components, it is not possible to visualize the entire exposure-response function all at once.
- Estimates $\widehat{h}\left(z_{i1}, \ldots, z_{iM}\right)$ allow us to estimate differences (and associated uncertainty) for any exposure contrasts of interest.
- We have found specific types of plots very useful to provide insights into this multivariate E-R surface:
  - Overall effects
  - Main effect estimates
  - Bivariate E-R surfaces
  - Interaction plots (multiple types)

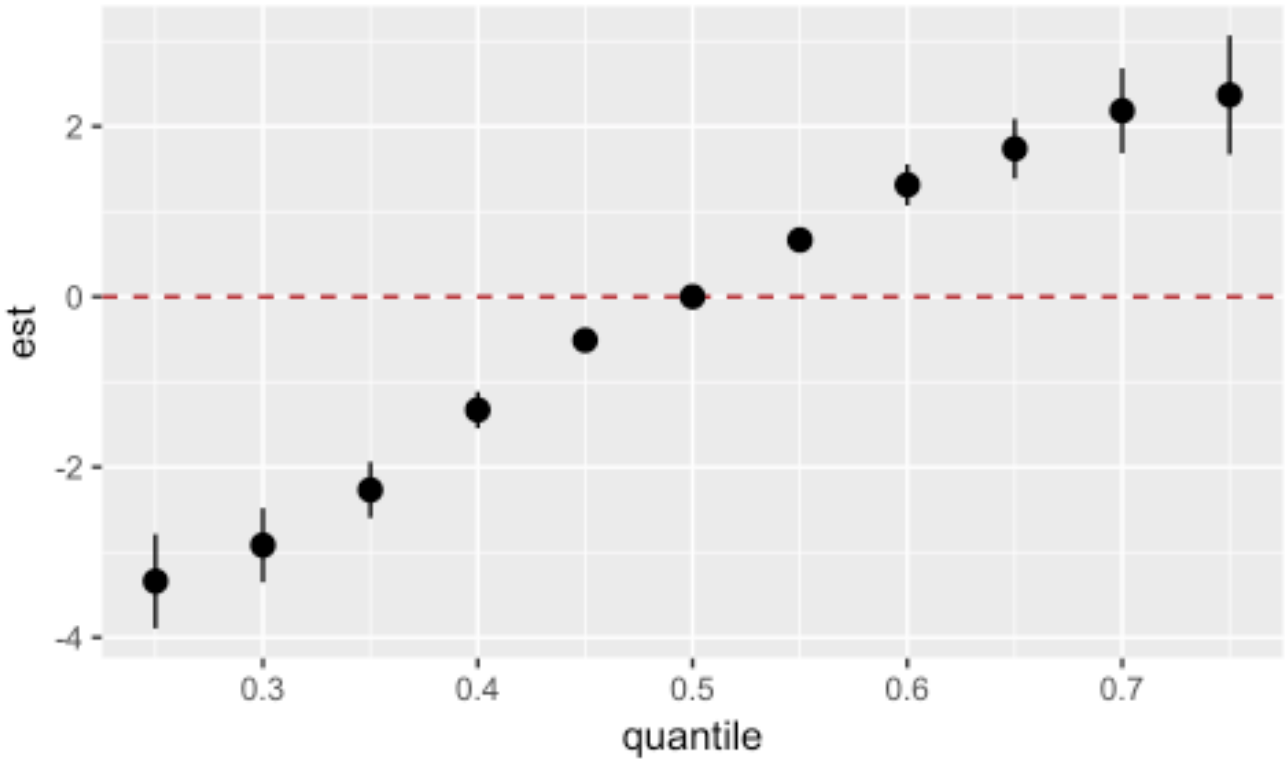
Illustration: 2015 NIEHS Workshop Simulated Data

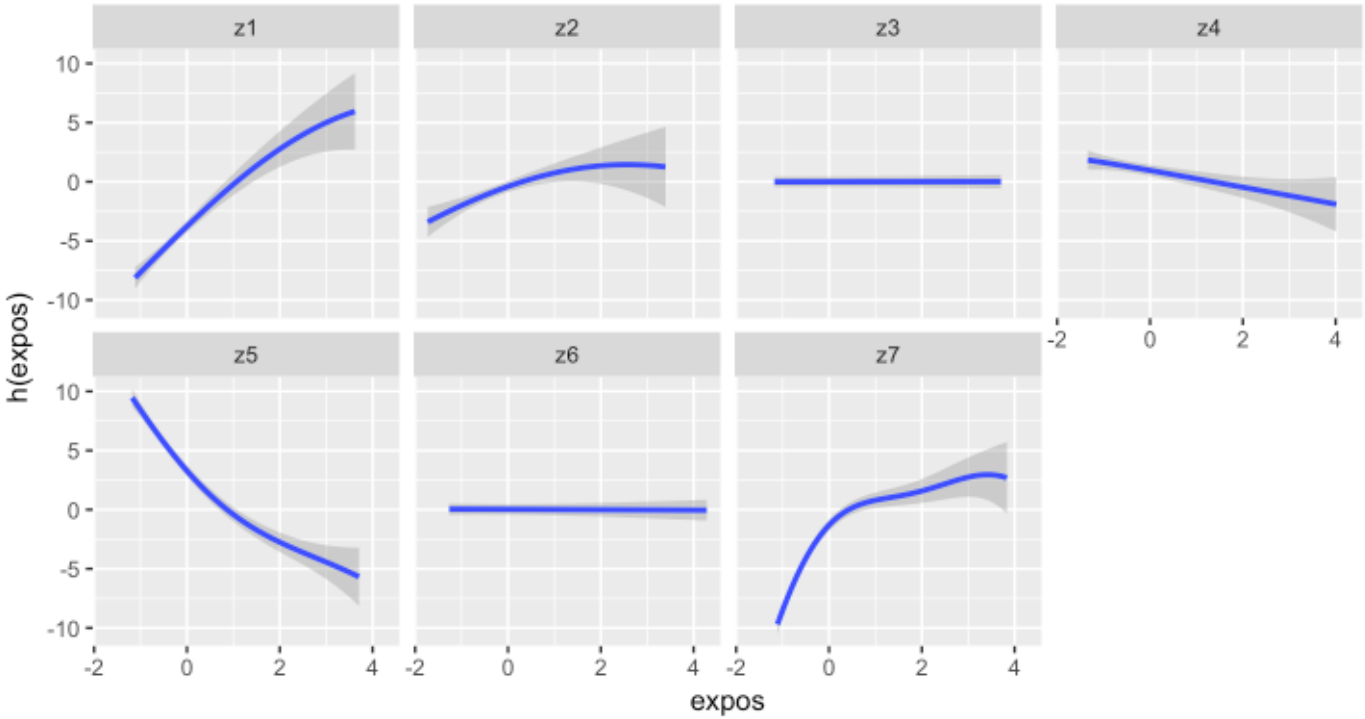
- Rather than write out a formula for each type of contrast, let's inspect some examples visually.
- Consider the first simulated data set provided at the 2015 NIEHS Workshop "Statistical Approaches for Assessing Health Effects of Environmental Chemical Mixtures in Epidemiology Studies"

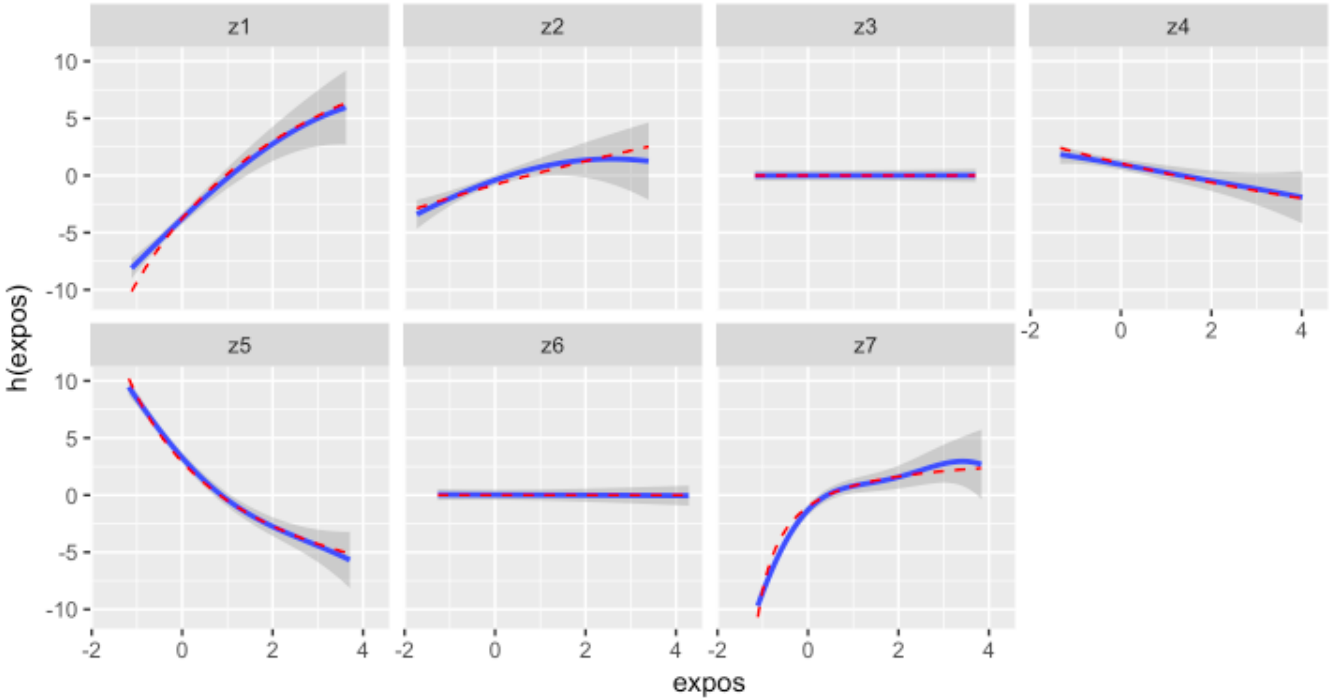
$$h(z_1, z_2, z_4, z_5, z_7) = \frac{\alpha_1 \left(\frac{T}{K_1} + \frac{z_1}{K_1} + \frac{z_2}{K_2} \right)}{1 + \frac{T}{K_1} + \frac{z_1}{K_1} + \frac{z_2}{K_2} + \frac{z_4}{K_4} + \frac{z_5}{K_5}} \left(R_{00} + \frac{\lambda z_7}{K_7 + z_7} \right).$$

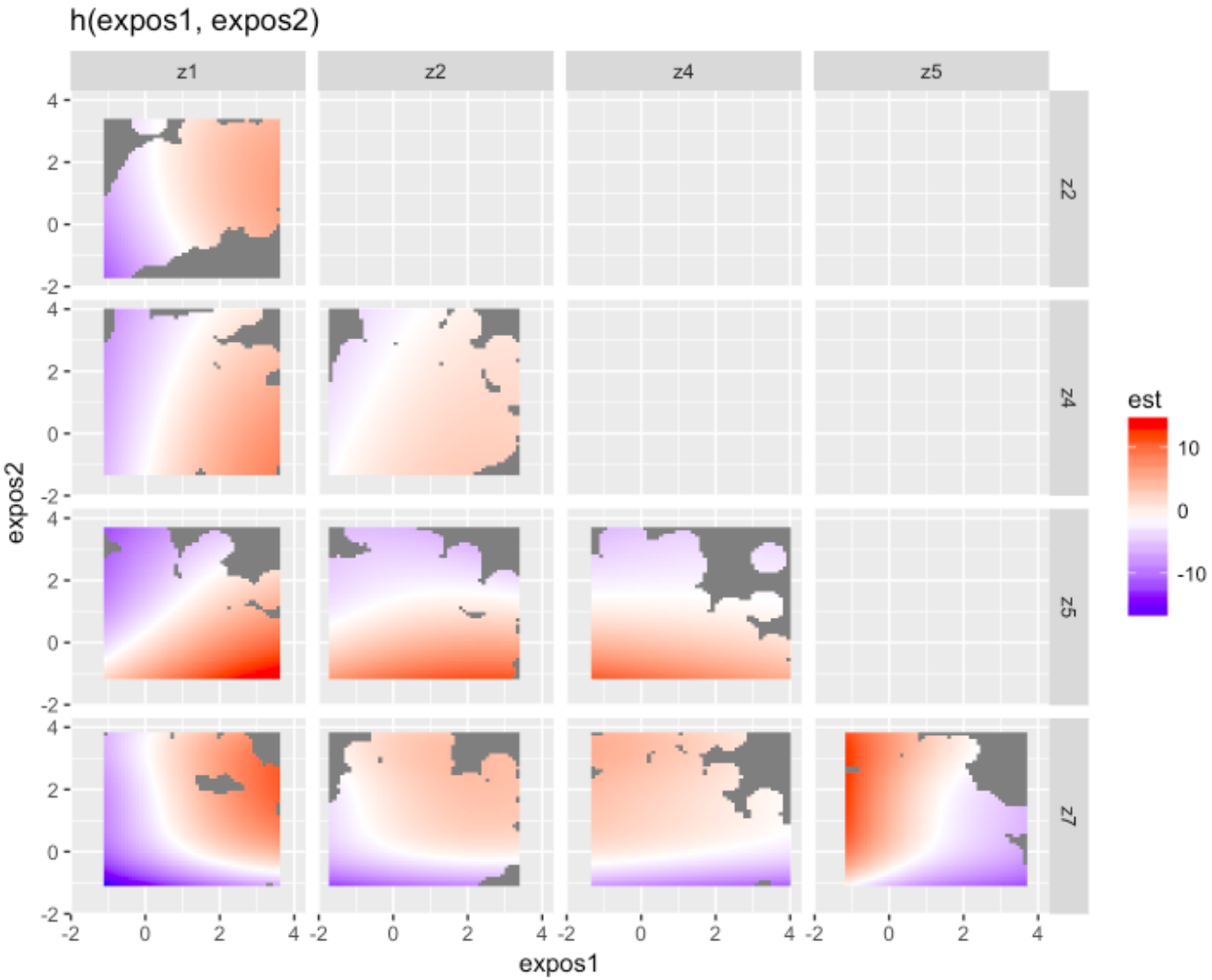
- All code available to obtain data and make these graphics available at
  - https://jenfb.github.io/bkmr/SimData1.html

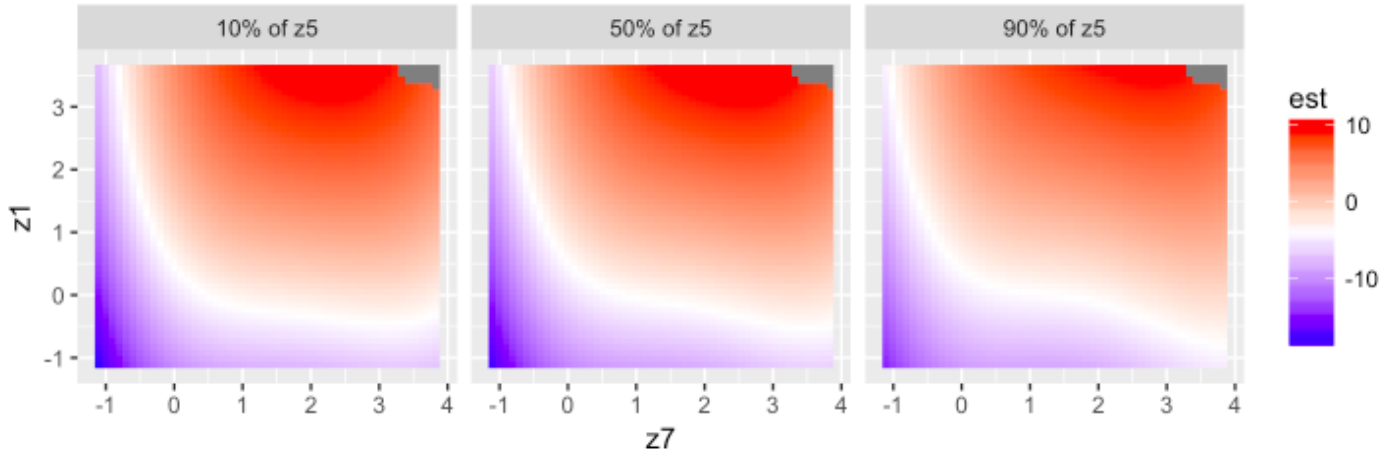
Bobb et al. *Env Health* 2018

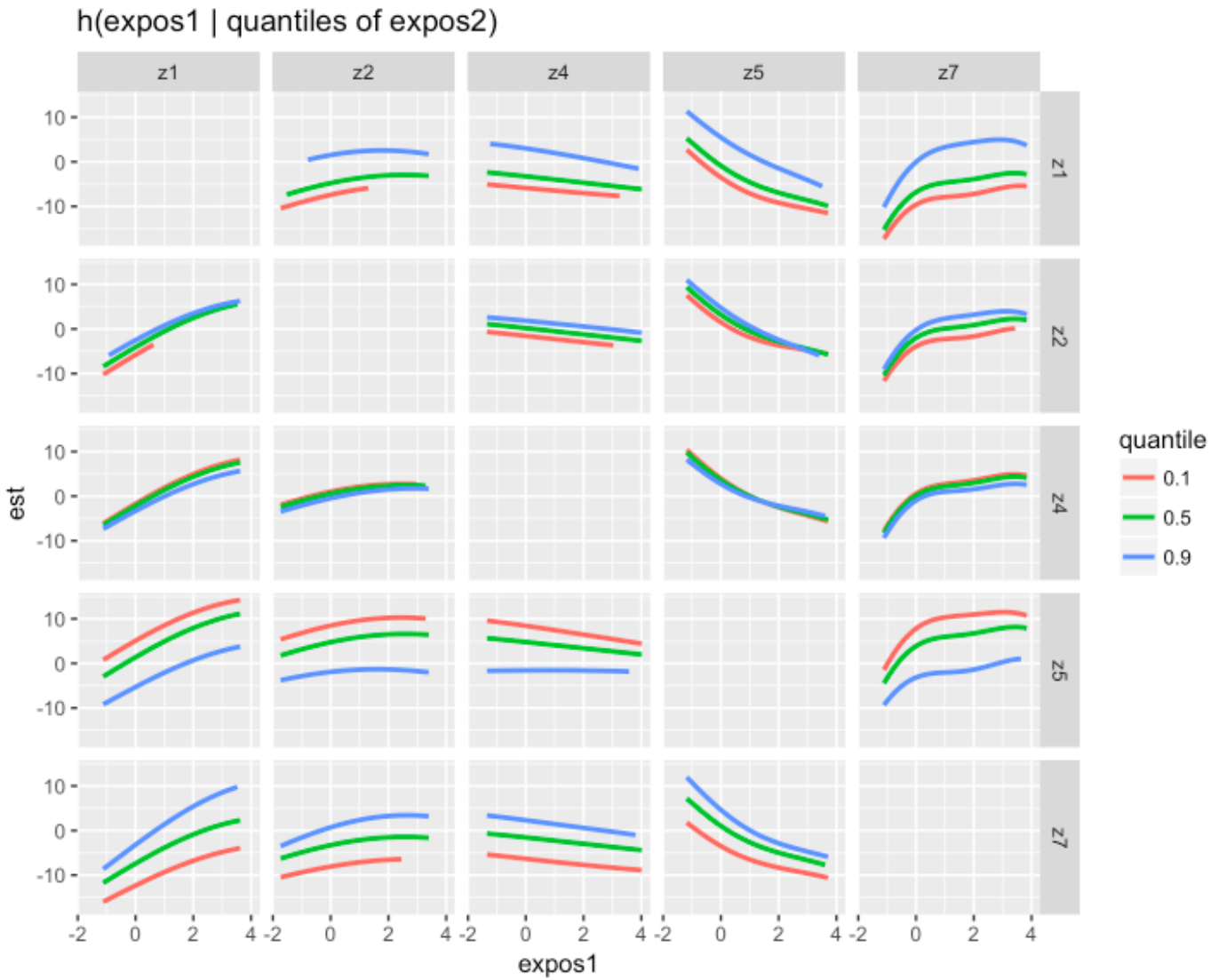

2
0
-2
-4
est
0.3
0.4
0.5
0.6
0.7
quantile

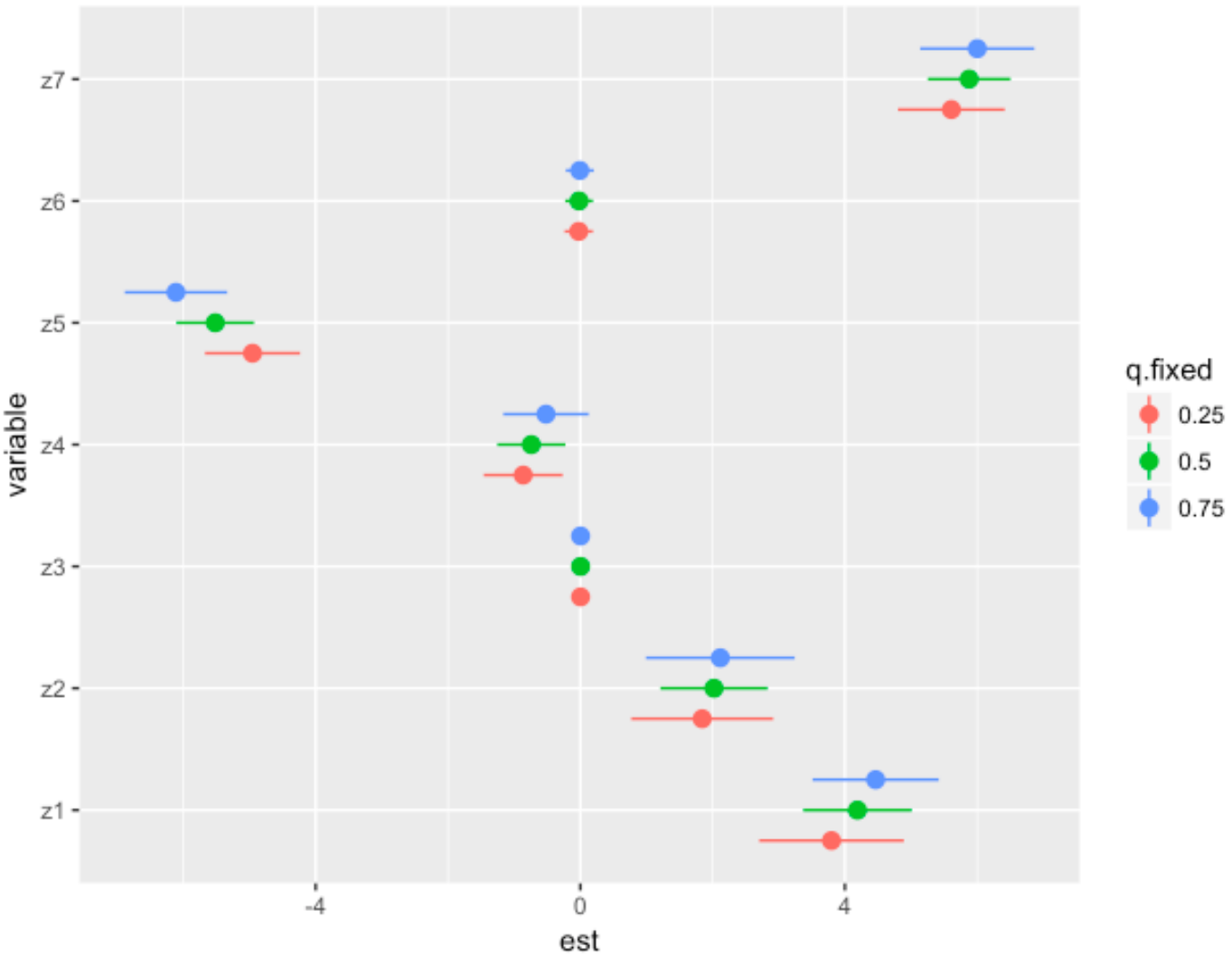

z1
z2
z3
z4
z5
z6
z7
h(expos)
expos
10
5
0
-5
-10
-2
0
2
4

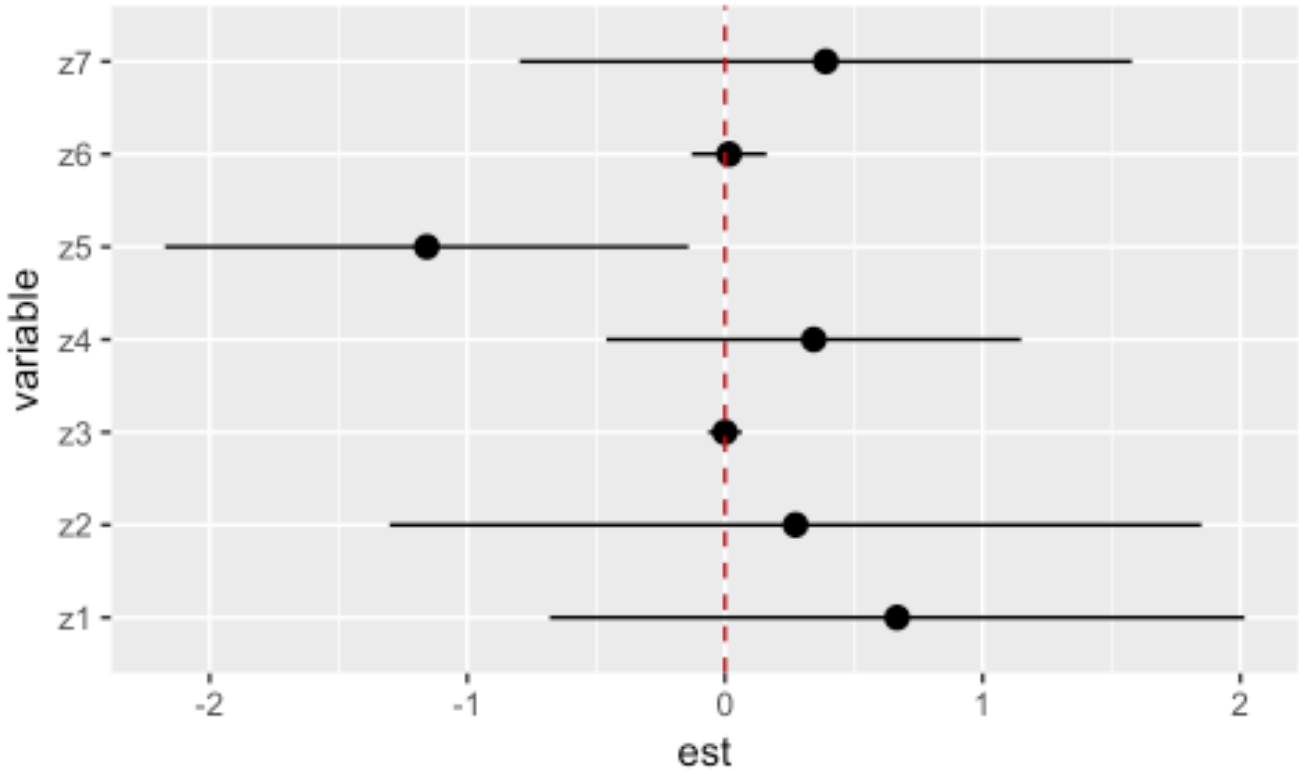

z1
z2
z3
z4
z5
z6
z7
h(expos)
expos
10
5
0
-5
-10
-2
0
2
4


h(expos1, expos2)
z1
z2
z4
z5
z2
z4
z5
z7
expos2
expos1
est
10
0
-10


10% of z5
50% of z5
90% of z5
z1
z7
est
10
0
-10
-1
0
1
2
3
4


h(expos1 | quantiles of expos2)
z1
z2
z4
z5
z7
est
expos1
quantile
0.1
0.5
0.9


z7
z6
z5
z4
z3
z2
z1
variable
-4
0
4
est
q.fixed
0.25
0.5
0.75


z7
z6
z5
z4
z3
z2
z1
variable
-2
-1
0
1
2
est

Strategy for Improving Computational Feasibility: Gaussian predictive process

- BKMR can be slow for large sample sizes (large n).
- Primarily due to the need to invert an n-by-n matrix (multiple times) at each iteration of the algorithm.
- Fast computational approximation: Gaussian predictive process
  - Specify a set of k points (referred to as "knots") that covers the exposure space.
  - Compute the projection of each vector of exposures onto the lower dimensional space spanned by the set of knots.
- End result: The algorithm only needs to invert a square matrix with dimension equal to k, not n.

Computational Savings: BKMR Fits to NHANES Lab Data

Table 1. Run times (hours) for BKMR applied to the NHANES lab data ($n = 1003$, 20,000 MCMC iterations)

Model	Variable Selection	Hier. Variable Selection
Full	7.0	5.6
GPP, $k = 100$	1.4	1.3
GPP, $k = 50$	0.9	0.9

Probit BKMR for Binary Outcomes

- Y_i is binary [0/1] variable
- $\mu_i = P(Y_i = 1)$

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

- Φ is the CDF for the std normal distribution (Φ^{-1} = probit link)
- All other terms remain the same as BKMR for a continuous outcome.

Probit BKMR: Interpretation

- Probit regression tends to be less common than logistic regression
- One reason: interpretability. Logistic regression coefficients = log odds ratios.
- In Bayesian models, probit typically much more computationally feasible.
- Probit model coefficients approximate logistic coefficients.
- That is, since $\text{logit}(\mu) \approx 1.6\Phi^{-1}(\mu)$, we have

$$\beta_{\text{logit}} \approx 1.6\beta_{\text{probit}},$$

  provided $P(Y = 1|\mathbf{z}, \mathbf{x})$ not too close to 0 or 1.
- R code also allows contrasts on probability scale (risk differences)

Data Analysis: Practical Considerations

- PIP values can be sensitive to priors (although relative importance tends to be stable)
- Assess sensitivity of conclusions to choice of values for prior values for tuning parameters
- The algorithm is more stable when it isn't dealing with exposures on vastly different scales. We typically center and scale both the outcome and the exposures (and continuous confounders).
- Be wary of exposure outliers (think GAMs with smooth terms). We typically log exposure concentrations.
- I like to verify detected patterns with parametric regression models.

bkmr R package

- The main function (kmbayes) implements a MCMC sampler to fit a BKMR model and allows for:
  - 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)
  - 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)

bkmr R package

bkmr contains a suite of post-processing functions, including functions to obtain

- trace plots from the MCMC sampler (TracePlot function)
- a parsimonious summary of model output (print and summary methods)
- posterior inclusion probabilities (ExtractPIPs function)
- summaries of posterior distributions of model parameters, including posterior mean, standard deviation, and quantiles (ExtractEsts function)
- scientifically relevant summaries of the exposure-response function

Summary

- BKMR is a flexible approach for estimating the joint health effects of simultaneous exposure to multiple concurrent risk factors.
- It's flexibility is useful in the presence of a complex non-linear or non-additive exposure-response function.
- In simpler scenarios this flexibility yields an approach that is not as powerful as more structured approaches.
- At the very least can be used as a complementary approach that checks the assumptions of models based on stronger assumptions.

Summary (cont.)

- The `bkmr` software package has been built to implement BKMR for range of data applications involving:
  - binary outcomes
  - repeated measures or clustered data
  - highly correlated exposures
- Numerical summaries based on $\widehat{h}(z_{i1}, \ldots, z_{iM})$ allow for inspection of
  - overall effects of the mixture
  - single-exposure health effects
  - interactive effects
- Several tricks to speed up computation:
  - Probit model for binary outcomes
  - Gaussian predictive process for large n

Acknowledgments

- Jennifer F. Bobb
- Katrina Devick
- Lizzy Gibson
- Linda Valeri
- Birgit Claus Henn
- Maitreyi Mazumdar
- David Bellinger
- David Christiani
- Robert Wright
- Marianthi-Anna Kioumourtzoglou
- Petros Koutrakis
- John Godleski

Bayesian framework for variable selection

$$K_{vs}(\mathbf{z}_i, \mathbf{z}_j) = \exp\left\{-\sum_{m=1}^{M} r_m \left(z_{im} - z_{jm}\right)^2\right\}$$

"Spike-and-slab" prior for r_m:

$$r_m \mid \delta_m \sim \delta_m \text{Gamma}(\mu, v^2) + (1 - \delta_m) P_0$$
$$\delta_m \sim \text{Bernoulli}(\pi)$$

and P_0 denotes the density with point mass at 0.

- Posterior inclusion probability: $P(\delta_m = 1 \mid y)$

Hierarchical (group-specific) variable selection

$$K_{vs}(\mathbf{z}_i, \mathbf{z}_j) = \exp\left\{-\sum_{m=1}^{M} r_m \left(z_{im} - z_{jm}\right)^2\right\}$$

"Spike-and-slab" prior for r_m:

$$r_m \mid \delta_m \sim \delta_m \text{Gamma}(\mu, v^2) + (1 - \delta_m) P_0$$

$$\delta_m = \omega_g \omega_{m|g},$$

- $\omega_g \sim \text{Bernoulli}(\pi_g)$
- $(\omega_{1|g}, \ldots, \omega_{m_g|g}) \sim \text{Multinomial}(\pi_{1|g}, \ldots, \pi_{m_g|g})$

Metal Mixtures and Neurodevelopment in Bangladesh

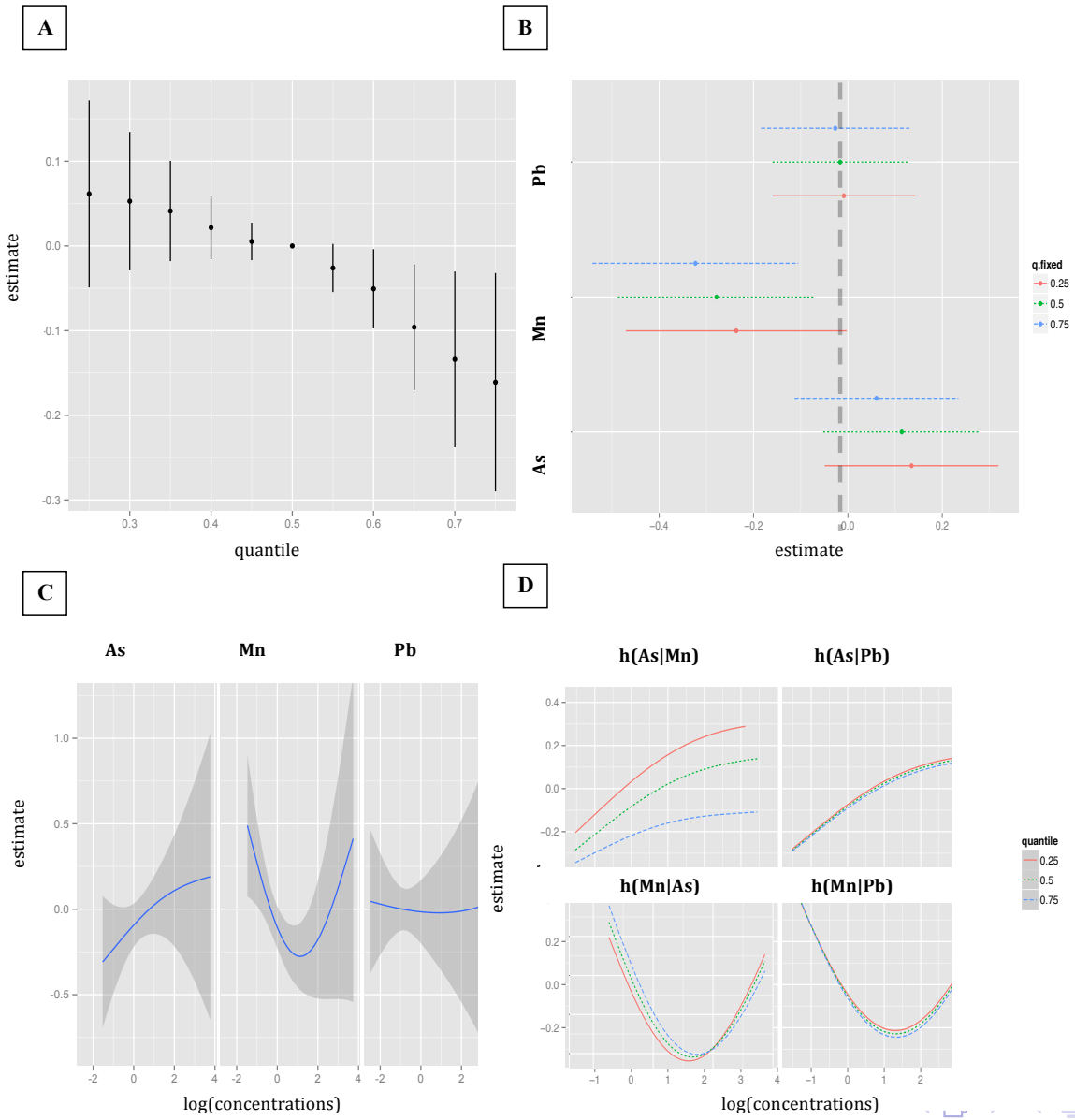
$$Y_i = h(As_i, Mn_i, Pb_i) + \boldsymbol{\beta}\mathbf{x}_i + \epsilon_i$$

- 409 children from Pabna clinic in Bangladesh
- **Outcome:** Bayley Cognitive Development Score at 20-40 months of age; z-scored
- **Exposures:** arsenic (As), manganese (Mn), and lead (Pb) measured in cord blood; log-transformed and standardized
- **Covariates:** infant sex, mother's IQ, "homescore" (SES proxy), clinic, age, and mother's education

Valeri et al. *EHP* 2017

Bangladesh Results

Figure 2.



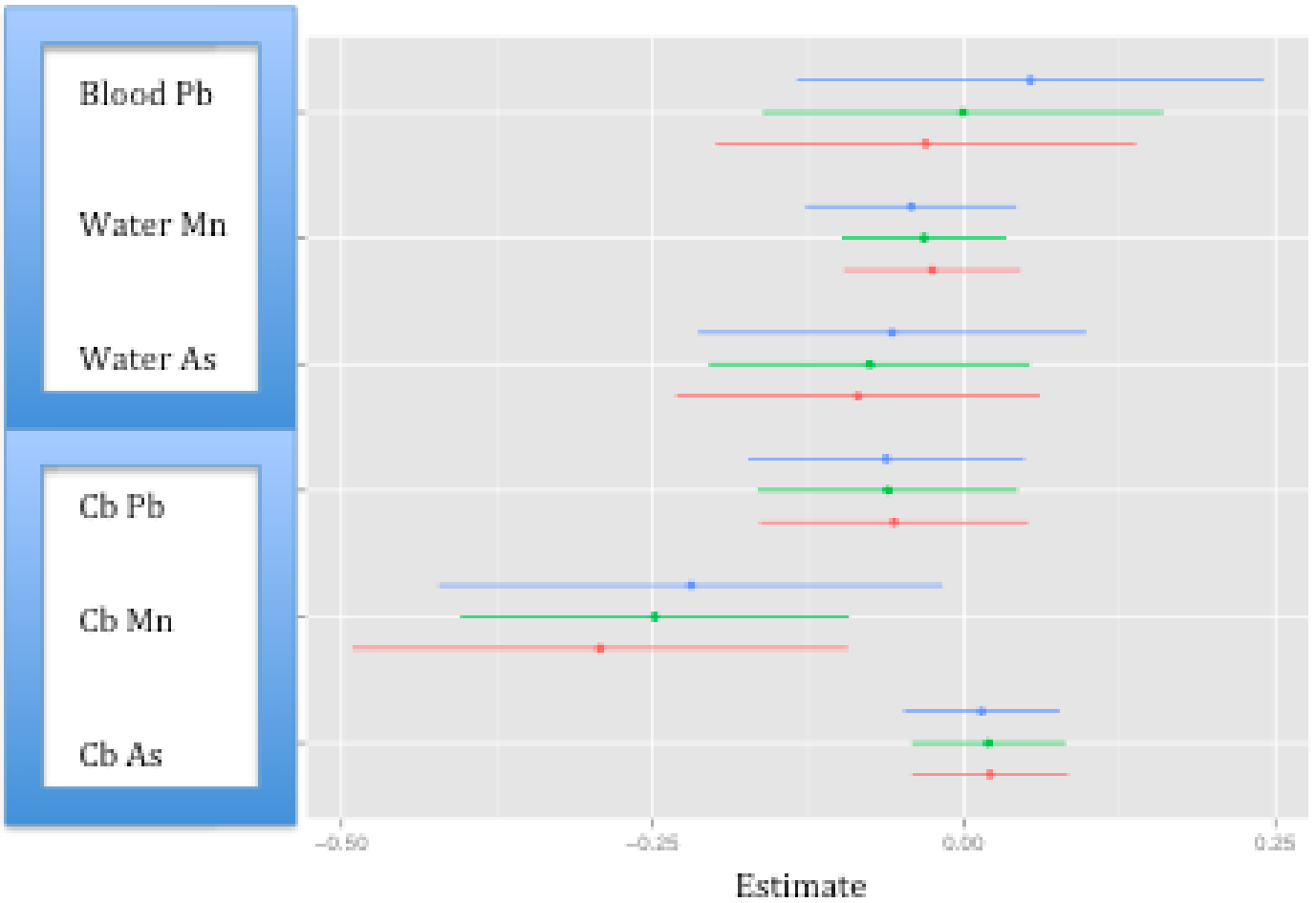
BKMR with Hierarchical Variable Selection

Q: Are prenatal exposures still associated with outcome after controlling for concurrent exposures?

$$Y_i = h\left(\overbrace{As_i^{CB}, Mn_i^{CB}, Pb_i^{CB}}^{?}, \overbrace{As_i^{W}, Mn_i^{W}, Pb_i^{B}}^{?}\right) + \boldsymbol{\beta}\mathbf{x}_i + \epsilon_i$$

- We obtain
  - an estimate of the probability that prenatal, concurrent exposures are important
  - an estimate of the probability that, given exposures at a given time are important, each metal in that group is driving that time-outcome association.

Bangladesh Data: Critical Window Analysis



Post-natal exposure: posterior inclusion probability = 28%

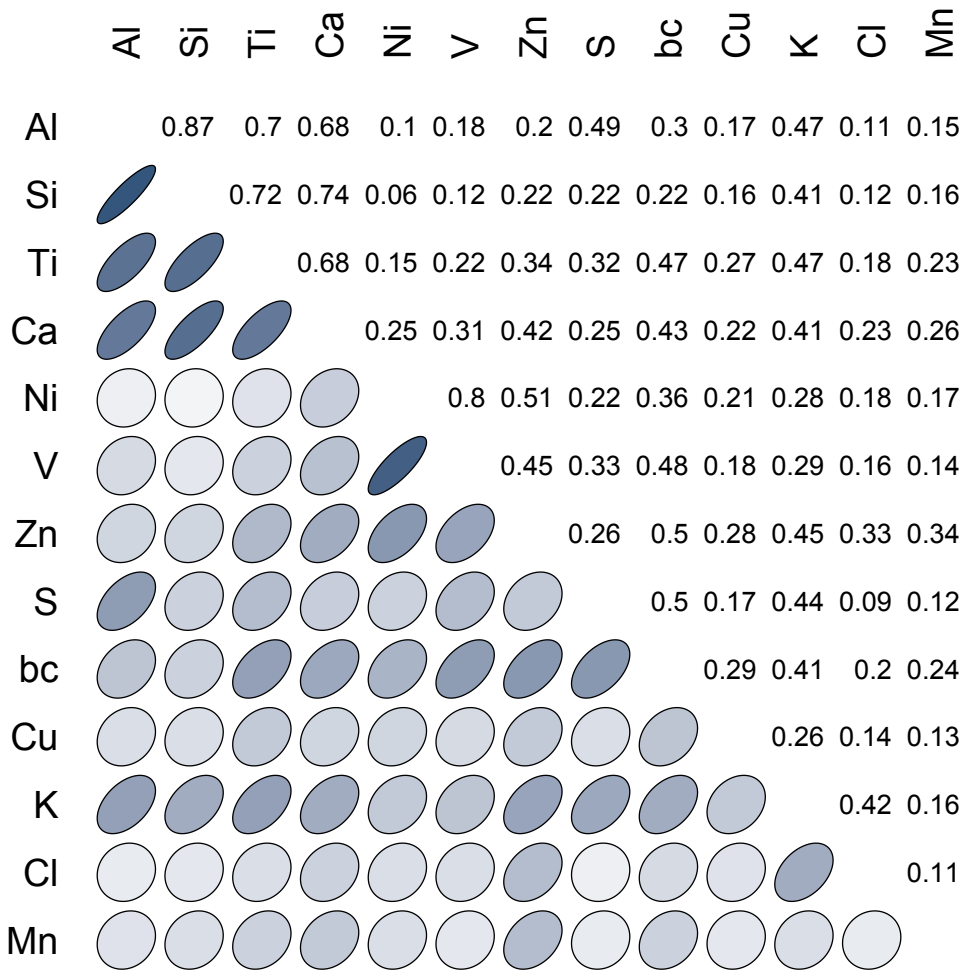
Pre-natal exposure: posterior inclusion probability = 99%

Manganese most neurotoxic component (Driven by high concentrations of Mn in Pabna Clinic)

Hierarchical Variable Selection: Effect of CAPs composition in canines

- Blood pressure and heart rate in canines exposed to concentrated air particles (CAPS; Bartoli et al. *EHP* 2009)
  - Cross-over study (CAPS, Filtered Air)
  - Some dogs had a balloon occluder implanted on their left anterior descending coronary artery
  - Additional experiments were conducted with prazosin in 8 of the original 13 animals
  - Total dataset of n=142 dog-exposures from 13 animals
- XRF analyses of CAPS exposures yielded concentrations of: (Al, Si, Ti, Ca, K, Cu, Mn, Ni, V, Zn, S, Cl, BC).

Boston data: XRF concentration correlations



Canine CAPs exposure study: model

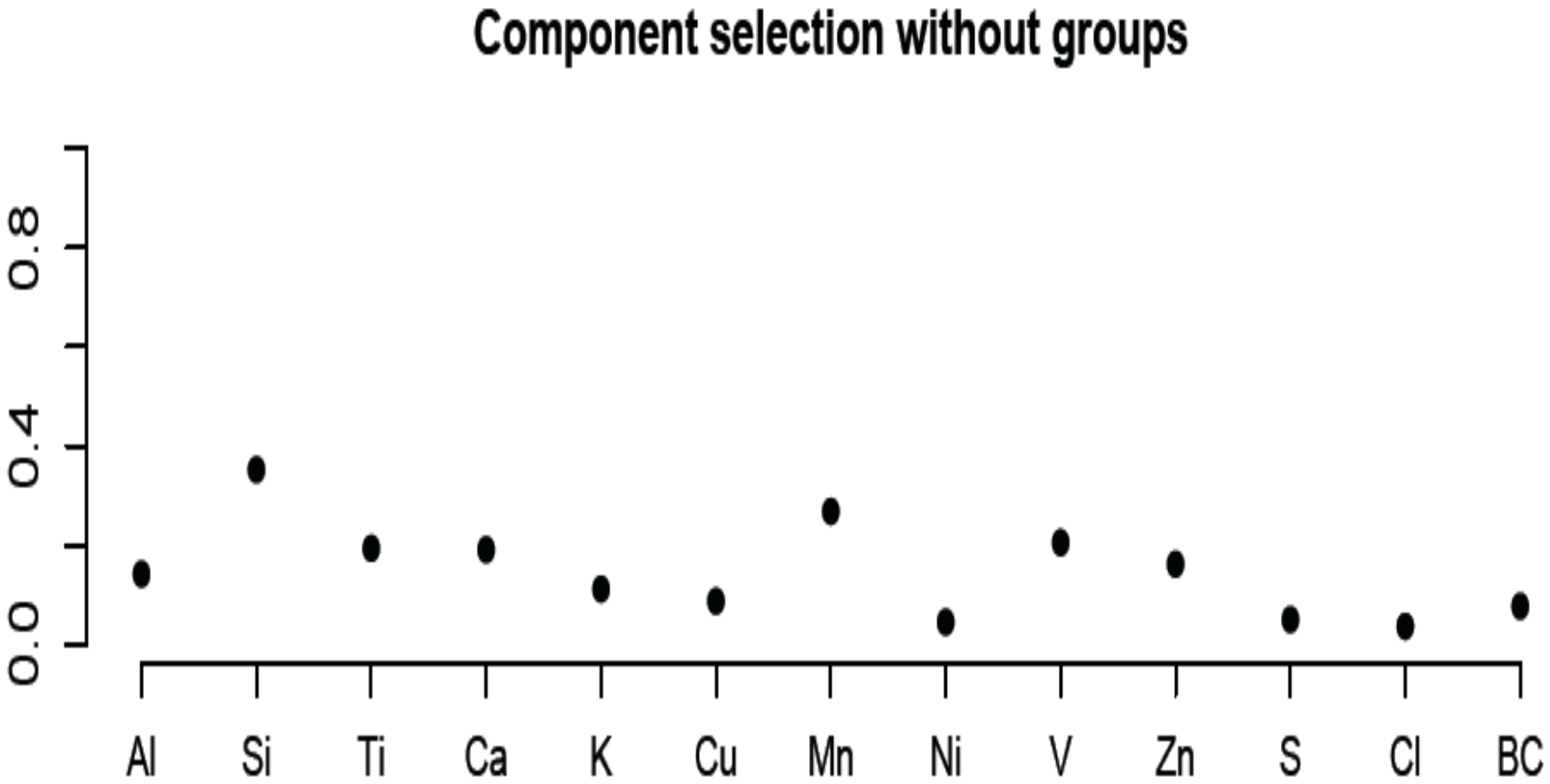
Applied to average heart rate for a given exposure, we fit the model

$$HR_{it} = h\left(Al_{it}, Si_{it}, Ti_{it}, Ca_{it}, K_{it}, \ldots, S_{it}, Cl_{it}, BC_{it}\right) + \boldsymbol{\beta}\mathbf{x}_{it} + b_i + \epsilon_{it}$$

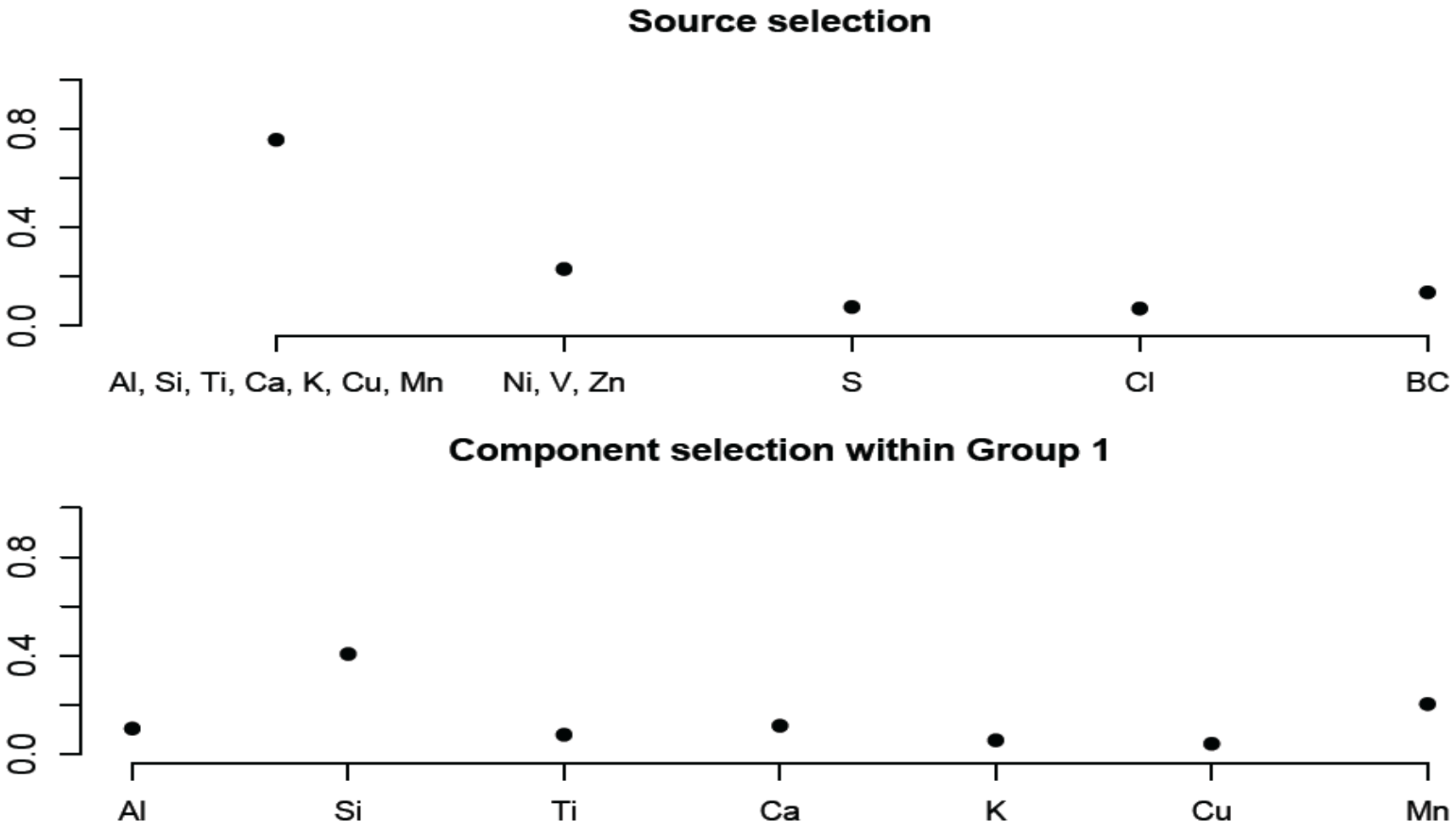
to each outcome, where

- $\mathbf{x}_{it} = [I(\text{Occlusion})_{it}, I(\text{Prazosin})_{it}]^{\mathsf{T}}$
- $b_i \overset{iid}{\sim} N(0, \sigma_b^2)$

Canine CAPs exposure study: Variable selection



Canine CAPS study: Group-specific selection



```
## install bkmr for fitting and ggplot2 for plotting results
install.packages("bkmr")
install.packages("ggplot2")

## load packages
library(bkmr)
library(ggplot2)
```

```
## 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)

## Fit with hierarchical variable selection
kmsel <- kmbayes(y = dat$y, Z = dat$Z, X = dat$X, verbose = FALSE,
groups=c(rep(1,times=9),rep(2,times=2),rep(3,times=3),rep(4,times=4)),
                 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)

## Fit a probit model for a binary response and tweak updating strategy
## for r_m parameters
fitpr <- kmbayes(y = dat$y, Z = dat$Z, X = dat$X, verbose = FALSE,
   varsel = TRUE,  iter = 10000, family = "binomial",
   control.params = list(r.jump2 = 0.5))
```