Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2012)
61, *Part* 1, *pp.* 47–66

Variable selection for high dimensional Bayesian density estimation: application to human exposure simulation

Brian J. Reich and Eric Kalendra,

North Carolina State University, Raleigh, USA

Curtis B. Storlie

Los Alamos National Laboratory, USA

and Howard D. Bondell and Montserrat Fuentes

North Carolina State University, Raleigh, USA

[Received March 2010. Final revision April 2011]

Summary. Numerous studies have linked ambient air pollution and adverse health outcomes. Many studies of this nature relate outdoor pollution levels measured at a few monitoring stations with health outcomes. Recently, computational methods have been developed to model the distribution of personal exposures, rather than ambient concentration, and then relate the exposure distribution to the health outcome. Although these methods show great promise, they are limited by the computational demands of the exposure model. We propose a method to alleviate these computational burdens with the eventual goal of implementing a national study of the health effects of air pollution exposure. Our approach is to develop a statistical emulator for the exposure model, i.e. we use Bayesian density estimation to predict the conditional exposure distribution as a function of several variables, such as temperature, human activity and physical characteristics of the pollutant. This poses a challenging statistical problem because there are many predictors of the exposure distribution and density estimation is notoriously difficult in high dimensions. To overcome this challenge, we use stochastic search variable selection to identify a subset of the variables that have more than just additive effects on the mean of the exposure distribution. We apply our method to emulate an ozone exposure model in Philadelphia.

Keywords: Air pollution; Bayesian non-parametrics; High dimensional data; Kernel stick breaking prior; Stochastic computer models

1. Introduction

Numerous studies have linked ambient air pollution (e.g. ozone or particulate matter) and adverse health outcomes (e.g. asthma, birth defects and mortality). Many studies of this nature relate outdoor pollution levels measured at a few monitoring stations with counts of health outcomes (e.g. Schwartz (1994) and Pope *et al.* (1995)). A limitation of this approach is that the measurements from monitoring stations are used to represent the pollution exposure for every person near the station. However, the amount of pollution that enters the body varies considerably from person to person depending on the individual's daily activity and living conditions. As a result of ignoring this variation, the estimated association between ambient concentration

Address for correspondence: Brian J. Reich, Department of Statistics, North Carolina State University, 2311 Stinson Drive, Box 8203, Raleigh, NC 27695-8203, USA.
E-mail: reich@stat.ncsu.edu

© 2011 Royal Statistical Society 0035–9254/12/61047

and the health outcome often varies from location to location (Dominici *et al.*, 2002; Fuentes *et al.*, 2006) or season to season (Lee and Shaddick, 2007) because human activity and living conditions vary with space and time.

Ideally, daily personal exposure would be monitored for every subject in the study. However, this is usually cost prohibitive. An alternative is to model individual exposure by using a computer simulation. For example, the Environmental Protection Agency has developed a stochastic computer model called the air pollution exposure model, APEX (Binkowski and Roselle, 2003). APEX generates a large number of hypothetical individuals to represent the population of interest. APEX then tracks the individuals through space and time to compute their hourly exposure in various microenvironments (e.g. outdoors, indoors or in vehicle). Although APEX does not actually measure exposures for these individuals, it uses information about human activity patterns, census data, meteorology, housing information, physical properties of the pollutant and diurnal pollution cycles to predict exposure. The simulated individuals are used to estimate the population's exposure distribution. This model has been compared and validated by using observed exposure (e.g. US Environmental Protection Agency (2007)). Using the stochastic model, we can replace the single ambient concentration predictor in the health model with a summary of the exposure distribution; for example the mean or the percentage of the exposures above a threshold could be used as predictors of the number of events in the population on a given day. Also, for studies with individual health data, this approach may be used to estimate each individual's daily exposure by using the individual's characteristics such as type of housing and employment status.

Although it shows great promise, this stochastic simulation approach is computationally expensive and as a result has only been used for local studies (Holloman *et al.*, 2004; Shaddick *et al.*, 2008; Calder *et al.*, 2008; Reich *et al.*, 2009). There is great interest in extending this methodology to the national level. This would require simulating the exposure of hundreds of individuals for each geopolitical unit (e.g. county) and each day of the study (e.g. a few years), which is infeasible with current computing power. In this paper, we propose a novel approach to alleviating this difficulty; we develop a statistical emulator. An emulator is a statistical approximation to a complex computer model. Ideally, the emulator should capture the important features of the computer model and be able to predict a new observation in negligible computing time.

The goal of this paper is to develop an emulator for APEX to be used in a future health study. Building an emulator for APEX poses two major challenges:

(a) APEX is a *stochastic* computer model and
(b) APEX has a large number of inputs.

Although our motivation is APEX, the approach that is developed here could be generalized to other stochastic computer models with similar features, such as models for computer networks (Waupotitsch *et al.*, 2006) or vehicle traffic (Galli *et al.*, 2009). Also, in the process of developing the emulator, we study the environmental and demographic factors that affect exposure. APEX is a complex stochastic non-linear model, and so it is not obvious which inputs are the most influential. Simulation models are being used to study the effects of emission control strategies on personal exposure. Therefore, providing a tool to test for complex relationships between the inputs and the exposure distribution could be very useful in this context.

There is a rich literature about developing statistical emulators for complex computer model output (e.g. Sacks *et al.* (1989) and Fang *et al.* (2006)). The vast majority of these methods deal with deterministic computer models that return the same value for every run with the same inputs. Typically the response surface is modelled by using non-parametric regression, e.g.

splines or Gaussian processes (Sacks *et al.*, 1989; Chen *et al.*, 2006; Fang *et al.*, 2006). As mentioned, developing an emulator for APEX is uniquely challenging because APEX is a stochastic model in that the output is an entire exposure distribution, rather than a scalar. The literature on statistical emulators for stochastic computer models is limited. Iooss *et al.* (2008) recently analysed stochastic computer model output by modelling the distribution's mean and variance using generalized additive models (Hastie and Tibshirani, 1990). Although this may capture many of the important features of the stochastic model, a more flexible model is certainly desirable. For example, in APEX the covariates clearly affect the skewness of the exposure distribution; see Section 5. Sufficiently modelling this tail behaviour could be crucial for the health analysis if, say, the appropriate predictor for the health outcome was the proportion of the population above a threshold.

We develop a Bayesian density estimation method to emulate APEX output. To do this, we extend the kernel stick breaking model of Reich and Fuentes (2007) and Dunson and Park (2008). The kernel stick breaking model specifies the conditional distribution of the outcome given the predictors as a potentially infinite mixture of normal distributions, with the mean of the Gaussian mixture components and the mixture probabilities dependent on the covariates. This model is well suited for the APEX data because it allows not only the mean and variance but also the skewness and all other properties of the distribution to vary smoothly across covariate space. Thus, all of the important properties of the exposure distribution can be carried to the health analysis.

A limitation of the kernel stick breaking approach, and all other density estimation methods, is the so-called 'curse of dimensionality'. To estimate the density at a given point, most density estimates use data in a small window around the point (at least implicitly). For the APEX simulator there are approximately 20 inputs. Even with thousands of model runs, the amount of data in any region of the 20-dimensional covariate space is too small to yield a reliable estimate of the density. Therefore, despite the fact that all the inputs are used in the simulation model, emulation may be improved by excluding inputs with small effects in favour of a parsimonious statistical model.

We reduce the dimension of the covariate space by using Bayesian variable selection. Bayesian variable selection for simple linear regression is a well-studied problem (e.g. George (2000)). In simple linear regression the covariates affect only the mean response and appear in the mean only as a linear combination. A more flexible approach is non-parametric regression which allows the mean to be a smooth surface in covariate space. There are several methods for variable selection in this context (Shively *et al.*, 1999; Gustafson, 2000; Wood *et al.*, 2002; Linkletter *et al.*, 2006; Reich *et al.*, 2008). However, these approaches still only model the mean response and thus ignore important effects from variables that affect the variance or higher moments.

In this paper we propose a variable selection method that not only searches for mean effects but also more generally aims to identify variables with any effect on the conditional distribution of the response. Chung and Dunson (2009) recently proposed to perform variable selection within the stick breaking framework via a probit link. We separate variable selection into two pieces. Covariates are selected to enter the model as

(a) a linear term affecting the mean of the response and/or
(b) a term in the kernel stick breaking density used to model the residual exposure distribution.

This separation is crucial to alleviating the curse of dimensionality because most of the predictors in APEX indeed affect the exposure distribution. However, the effect from most of the predictors can be modelled effectively as a linear change in the mean, leaving a manageable number of variables for residual density estimation.

We also illustrate how our variable selection method can be used as an exploratory tool. With a large number of covariates, searching for non-standard effects such as non-linear mean relationships, variance inflation, missing interactions and increased tail probability is very challenging. Our simulation study shows that the kernel stick breaking model is effective at identifying these relationships. Therefore, our approach is to begin with a linear, main-effects-only model, and then to conduct further exploration for the subset of variables included in the mixture of normals component of the conditional density model. For example, there are an exorbitant number of plausible interactions for the APEX data so including them all in the candidate pool would be overwhelming. Our main-effects-only model identifies four variables as having non-standard effects. Refitting with the six two-way interactions between these four predictors reveals several statistically significant and scientifically meaningful interactions.

The paper proceeds as follows. Section 2 describes APEX and the simulated data and Section 3 develops the statistical emulator. A brief simulation study in Section 4 illustrates the flexibility of our non-parametric model. We analyse APEX output in Section 5. Section 6 concludes.

2. Data description

In this section we give a brief description of APEX; a full description can be found at `http://www.epa.gov/ttn/fera/apex_download.html`. In Section 5 we analyse ozone exposure, although the model below can be used for other pollutants such as carbon monoxide or particulate matter. APEX estimates the population distribution of exposures by simulating personal exposures for hypothetical individuals chosen to represent the study population in terms of age, gender, employment, housing volume, smoking status, etc. The activities of the hypothetical individuals are generated by randomly selecting a diary from the Environmental Protection Agency's consolidated human activity database. This database contains personal diaries of over 22 000 individuals from exposure studies conducted around the USA. The diaries describe the activity pattern of the individual throughout the day and are selected to match the hypothetical individuals on the basis of personal characteristics, season, day of the week and average daily temperature.

APEX tracks the individuals throughout the day and computes their hourly exposure on the basis of the hourly ambient concentration and individuals' current environment. APEX computes exposure for several environments, including residence, bars and restaurants, schools, day care centres, offices, shopping centres, outdoors and vehicles. The exposure for an individual on a given day is then

$$E = \sum_{h=1}^{24} \sum_{j=1}^{N} C_{hj} t_{hj} / 24, \tag{1}$$

where N is the number of environments in the simulation, C_{hj} is the concentration in environment j at hour h, and t_{hj} is the time spent in environment j during hour h. The concentration C_{hj} in the indoor environments is computed by using the differential equation

$$\frac{\mathrm{d}C_{hj}}{\mathrm{d}t} = \mathrm{AER}_j * (C_h^{\mathrm{ambient}} - C_{hj}) - \mathrm{DR} * C_{jk} + C_{hj}^{\mathrm{source}}, \tag{2}$$

where AER_j is the air exchange rate for environment j, C_h^{ambient} is the outdoor concentration during hour h, DR is the decay rate and C_{hj}^{source} is the added concentration due to point sources in environment j, e.g. cooking. The three terms in equation (2) represent respectively the transport of material in and out of the environment, removal of a pollutant from the microenvironment due to deposition, filtration and chemical degradation, and emissions from sources of a pollu-

tant inside the microenvironment. The concentration in the outdoor microenvironment is taken to be $C_{hj} = C_h^{\text{ambient}}$. Since we are interested in exposure to outdoor pollution, we exclude the third term.

To demonstrate our method, we use APEX to generate 5000 ozone exposure observations for residents in the City of Philadelphia in the summer (June–August) of 2001. The inputs include daily temperature and hourly ambient ozone in 498 districts; hourly ozone levels are modelled by using the deterministic CMAQ model (Binkowski and Roselle, 2003). The characteristics of the individuals are generated according to census variables such as age, race and gender distribution, which are freely available for major metropolitan areas. AER_j- and DR-values are drawn for each subject from the default uncertainty distributions to represent a reasonable range of values and are held constant throughout the day for each subject. We use the individual-specific AER_j and DR as predictors for exposure. We also use as a predictor the physical activity index, i.e. the time-averaged metabolic equivalent of task over the day, as a one-number summary of the physical activity diary. In many settings these predictors may not be known exactly. However, we include them to study the model's sensitivity to these factors. The resulting emulator could still be used in the absence of these variables by placing an uncertainty distribution on them and calculating the marginal exposure distribution by using numerical integration over the conditional distribution that is developed in Section 5, or by simply refitting with, say, the mean and variance of the uncertainty distributions as predictors.

3. Variable selection for Bayesian density estimation

In this section we propose a fully Bayesian method for variable selection in non-parametric density estimation. Our method builds on the kernel stick breaking model which we describe in Section 3.1. Section 3.2 proposes a stochastic search variable selection model to search for important subsets of the predictors to describe the conditional density of the outcome. In Section 3.3 we make recommendations for how to use the density estimates in a future health study. Computing details are given in Section 3.4.

3.1. Kernel stick breaking model

The kernel stick breaking model is an extension of the ordinary stick breaking model of Sethuraman (1994), which we describe below. For general Bayesian modelling, the stick breaking prior offers a way to model a distribution as an unknown quantity to be estimated from the data. The stick breaking prior for the unknown distribution F is the infinite mixture of normal distributions

$$F \stackrel{\mathcal{D}}{=} \sum_{k=1}^{\infty} p_k N(\mu_k, \sigma_k), \tag{3}$$

where p_k are the mixture probabilities and $N(\mu, \sigma)$ is the normal density with mean μ and standard deviation σ. The mixture probabilities 'break the stick' into an infinite number of pieces so the sum of the pieces is 1, i.e. $\Sigma_{k=1}^{\infty} p_k = 1$. This constraint is satisfied stochastically by introducing latent variables $v_k \sim^{\text{IID}} \text{beta}(1, D)$, where $D > 0$ is a hyperparameter. The first mixture probability is modelled as $p_1 = v_1$. Subsequent mixture probabilities are

$$p_k = v_k\left(1 - \sum_{j=1}^{k-1} p_j\right) = v_k \prod_{j=1}^{k-1} (1 - v_j), \tag{4}$$

where $1-\Sigma_{j=1}^{k-1} p_j$ is the probability not accounted for by the first $k-1$ mixture components, and v_k is the proportion of the remaining probability assigned to the kth component.

The kernel stick breaking model allows the density of the response y to depend on the predictors $\mathbf{x}=(x_1,\ldots,x_p)'$. We assume that the response is scaled to have mean 0 and unit variance. Let

$$y=\mathbf{x}'\boldsymbol{\beta}+\varepsilon \qquad \varepsilon\sim F(\varepsilon|\mathbf{x}), \tag{5}$$

where $\boldsymbol{\beta}$ is the vector of regression coefficients in the linear mean. Following Reich and Fuentes (2007) and Dunson and Park (2008), the conditional distribution of ε given $\mathbf{x}$ is modelled as the infinite mixture

$$F(\varepsilon|\mathbf{x})=\sum_{k=1}^{\infty} p_k(\mathbf{x})\, N(\mu_k,\sigma_k), \tag{6}$$

where $p_k(\mathbf{x})$ are the mixture weights with $\Sigma_{k=1}^{\infty} p(\mathbf{x})=1$ for all $\mathbf{x}$. The means and variances have priors $\mu_k \sim^{\mathrm{IID}} F_0$ and $\sigma_k \sim^{\mathrm{IID}} U(0,\sigma_{\max})$ respectively. The kernel stick breaking weights allow the density to be different in different regions of the covariate space. These densities are tied together by the base distribution F_0. We take the base distribution to be $\mu_k \sim^{\mathrm{IID}} N(0,\tau)$, where τ is a hyperparameter to be estimated by the data.

The mixture probabilities vary with $\mathbf{x}$ through a series of kernel functions $w_k(\mathbf{x})\in[0,1]$. The mixture probabilities are $p_1(\mathbf{x})=v_1\, w_1(\mathbf{x})$ and

$$p_k(\mathbf{x})=v_k\, w_k(\mathbf{x}) \prod_{j=1}^{k-1}\{1-v_j\, w_j(\mathbf{x})\} \tag{7}$$

for $k>1$. Here $\Pi_{j=1}^{k-1}\{1-v_j\, w_j(\mathbf{x})\}$ is the proportion of the stick attributed to the first $k-1$ terms and $v_k\, w_k(\mathbf{x})$ is the proportion of the remaining stick attributed to component k for an observation with covariates $\mathbf{x}$. Since in most cases y's density is a fairly smooth function of $\mathbf{x}$, we use squared exponential kernels (although other kernels are possible), i.e.

$$w_k(\mathbf{x})=\exp\{-(\mathbf{x}-\boldsymbol{\psi}_k)'\Sigma(\mathbf{x}-\boldsymbol{\psi}_k)\}, \tag{8}$$

where $\boldsymbol{\psi}_k=(\psi_{k1},\ldots,\psi_{kp})'$ is the kernel's centre and Σ is the $p\times p$ matrix that controls its spread and shape. To facilitate prior specification, we scale the predictors so that $x_j\in[0,1]$ for $j=1,\ldots,p$ and then assume that the knots have priors $\psi_{kj}\sim^{\mathrm{IID}} U(0,1)$. As before $v_k \sim^{\mathrm{IID}}$ beta$(1,D)$.

In equation (6), the conditional mean and variance of y are

$$E(y|\mathbf{x})=\mu(\mathbf{x})=\mathbf{x}'\boldsymbol{\beta}+\sum_{k=1}^{\infty} p_k(\mathbf{x})\mu_k, \tag{9}$$

$$V(y|\mathbf{x})=\sigma^2(\mathbf{x})=\sum_{k=1}^{\infty} p_k(\mathbf{x})\left[\sigma_k^2+\left\{\mu_k-\sum_{l=1}^{\infty} p_l(\mathbf{x})\mu_l\right\}^2\right]. \tag{10}$$

Therefore, although the parametric portion of the mean model is simply $\mathbf{x}'\boldsymbol{\beta}$, the kernel stick breaking model for the mean can accommodate more complicated mean structures, such as non-linearity and interaction effects. Also, as shown by equation (10), varying the probabilities with $\mathbf{x}$ gives a rich class of models for the variance and higher moments of $\mathbf{y}$ as a function of $\mathbf{x}$.

3.2. Bayesian variable selection

We perform variable selection separately on the parametric mean $\mathbf{x}'\beta$ and the residual distri-

bution $F(\varepsilon|\mathbf{x})$. In both cases we use stochastic search variable selection. Following George and McCulloch (1993, 1997), let

$$\beta_j = \pi_{1j}\theta_j,$$

where $\pi_{1j} \sim$ Bern(0.5) and $\theta_j \sim N(0,\sigma_\theta^2)$. If $\pi_{1j} = 0$ then $\beta_j = 0$ and x_j is removed from the parametric mean. In contrast, if $\pi_{1j} = 1$, x_j is included and its coefficient $\beta_j = \theta_j$ has a vague normal prior. The posterior mean of π_{1j} represents the posterior probability that the linear mean depends on x_j.

For the non-parametric component, we assume that Σ is diagonal with diagonal elements $\alpha_1, \ldots, \alpha_p$. Defining $\rho_j = \exp(-\alpha_j)$, equation (8) can be written

$$w_k(\mathbf{x}) = \prod_{j=1}^{p} \rho_j^{(x_j - \psi_{kj})^2}. \tag{11}$$

Variable selection is performed by giving ρ_j prior mass at 1; if $\rho_j = 1$ then x_j does not appear in the kernels $w_k(\mathbf{x})$ and thus x_j does not appear in $F(\varepsilon|\mathbf{x})$. This interpretation also holds for binary covariates. The log-ratio of the weight for term k for $x_j = 1$ compared with $x_j = 0$ is $2(\frac{1}{2} - \psi_{kj})\log(\rho_j)$. Therefore, the stick breaking weights vary by x_j if and only if ρ_j is less than 1, and the knot ψ_{kj} controls the value of x_j that is favoured by term k.

We assume that

$$\rho_j = 1 - \pi_{2j}\gamma_j, \tag{12}$$

where $\pi_{2j} \sim$ Bern(0.5) and $\gamma_j \sim^{\text{IID}} U(0,1)$. If $\pi_{2j} = 0$ then $\rho_j = 1$ and x_j is removed from the model for the residual distribution. If $\pi_{2j} = 1$ then $\rho_j < 1$ and x_j is included in the residual model.

3.3. Incorporating the exposure distribution in a future health model

In this section we suggest how the exposure distribution could be used in a future health study. Here we focus on time series analysis of count data. Denote the z_t as the number of events in a given region on day t, and C_t as the ambient concentration on day t, or perhaps day $t - l$ for some lag l. A natural model for the counts is $z_t \sim \text{Poisson}\{\lambda_t \exp(\beta C_t)\}$, where λ_t includes population size and confounders such as meteorology and time trends.

One approach would be to replace the ambient concentration with an estimate of the mean of the exposure distribution, e.g. the posterior median of equation (9). Although there may be some value in using the exposure simulator to provide an improved one-number summary of exposure compared with ambient concentration, we feel that a greater contribution is to model the entire exposure distribution of the population, and to use this distribution as a functional predictor. If the exposure e_{ti} for each member of the population of interest was known, we might model the expected number of events as $\Sigma_i \lambda_t \exp(\beta e_{ti})$. To approximate this sum, Reich *et al.* (2009) assumed that the exposure distribution on day t, $q_t(e)$, is Gaussian with mean μ_t and standard deviation σ_t, and the expected number of events becomes proportional to

$$\int \lambda_t \exp(\beta e)\, q_t(e)\, \mathrm{d}e = \lambda_t \exp(\mu_t \beta + \sigma_t^2 \beta^2). \tag{13}$$

In addition, they accounted for uncertainty in the exposure distribution by placing priors (derived from several model simulations) on μ_t and σ_t.

A similar approach could be taken for a non-Gaussian exposure distribution. The emulator proposed is a mixture of normal distributions, and the integral in equation (13) for the expected counts can be conveniently expressed as

$$\lambda_t \sum_k p_{kt} \exp(\mu_{kt}\beta + \sigma_{kt}^2 \beta^2), \tag{14}$$

where p_{kt}, μ_{kt} and σ_{kt} are the mixture weight, mean and standard deviation for the kth mixture component. Alternatively, it is possible to approximate the convolution integral in equation (13) by using, say, the quintiles of the exposure distribution $q_{1t}, \ldots, q_{5t}$, i.e.

$$\lambda_t \sum_{k=1} \exp(q_{kt}\beta), \tag{15}$$

where $q_{1t}, \ldots, q_{5t}$ have priors based on several model simulations.

The analysis would proceed similarly for a study with individual level data. A common model for individual binary responses, y_i, is the logistic regression model $\text{logit}\{P(y_i = 1)\} = \beta_0 + \beta_1 E_i$, where E_i is the exposure for subject i. In this case, E_i would be taken to be unknown, with the APEX exposure distribution as its prior. This errors-in-covariates model would account for both uncertainty in the exposure distribution, as well as uncertainty in the exposure of each individual relative to this distribution.

3.4. Computational details

Markov chain Monte Carlo (MCMC) sampling is carried out in R (R Development Core Team, 2006). To facilitate MCMC sampling, we introduce latent group indicators $g_1, \ldots, g_n$ and reformulate model (6) as

$$y_i | g_i \sim N\left(\sum_{j=1}^{p} x_{ij}\pi_{1j}\theta_j + \mu_{g_i}, \sigma_{g_i} \right), \tag{16}$$

$$g_i \sim \text{categorical}\{p_1(\mathbf{x}_{(i)}), p_2(\mathbf{x}_{(i)}), \ldots\}, \tag{17}$$

where $\mathbf{x}_{(i)}$ is the vector of predictors for observation i. The mean parameters π_{1j} and θ_j have conjugate priors and are updated individually by using Gibbs sampling. Given $N = \max\{g_1, \ldots, g_n\}$, we need to update (μ_k, σ_k, v_k) only for $k = 1, \ldots, N$. The remaining terms do not enter the posterior except through their priors. The μ_k are updated by using Gibbs sampling and σ_k and v_k are updated individually by using Metropolis sampling with Gaussian candidate distributions. Candidates with zero probability are simply rejected. The variable indicators π_{2j} are updated separately by using Gibbs sampling.

The group indicators g_i are also updated by using Metropolis sampling. Candidates g_i are generated from the prior $g_i \sim \text{categorical}\{p_1(\mathbf{x}_{(i)}), p_2(\mathbf{x}_{(i)}), \ldots\}$. Following Papaspiliopoulos and Roberts (2008), we generate the candidate by first drawing $w \sim U(0, 1)$. If $w < \Sigma_{l=1}^{N} p_l(\mathbf{x}_{(i)})$, we take $\min\{g | w < \Sigma_{l=1}^{g} p_l(\mathbf{x}_{(i)})\}$ as the candidate. If $w \geqslant \Sigma_{k=1}^{N} p_k(\mathbf{x}_{(i)})$, we increase N, drawing the corresponding (μ_N, σ_N, v_N) from their priors, until $w < \Sigma_{l=1}^{N} p_l(\mathbf{x}_{(i)})$, and use the new N as the candidate for g_i.

An alternative sampling approach is to replace model (6)'s infinite mixture with a finite mixture of m components by defining the probability for the mth term as $p_m(\mathbf{x}_{(i)}) = 1 - \Sigma_{j=1}^{m-1} p_j(\mathbf{x}_{(i)})$ for all i. We use this alternative approach with $m = 50$ for the analysis in Section 5. To monitor the validity of this approximation, we inspect the posterior samples of $p_m(\mathbf{x}_{(i)})$. For these data, the posterior mean of $p_m(\mathbf{x}_{(i)})$ is less than 0.001 for all i.

The MCMC algorithm for the full model with $n = 5000$ responses in Section 5 runs in a few hours on an ordinary personal computer. APEX runs for Philadelphia required several days of computing. In addition to this difference, the advantage of an emulator is that it should be possible to run the full APEX model and MCMC emulation algorithm on a subset of days and locations, and then immediately to extrapolate to the complete data set by using the estimated exposure distribution, making a national study feasible. Extrapolation to new locations

will become more complicated as APEX incorporates more local features such as commuting patterns and land use variables.

4. Simulation study

We conduct a brief simulation study to evaluate the ability of the kernel stick breaking model to identify several types of non-standard features in the data. We simulate data under five designs, which are described in Sections 4.1–4.5. Each design has $n = 200$ observations and $p = 10$ covariates $x_1, \ldots, x_{10}$ generated independently from the $U(0, 1)$ distribution. We compare three models which are all special cases of Section 3's full kernel stick breaking model as follows:

(a) model 1, a linear regression model with normal errors, $F(\varepsilon|\mathbf{x}) = N(0, \sigma)$ in equation (5);
(b) model 2, a linear regression model with non-parametric errors, $w_j(\mathbf{x}) = 1$ for all j and $\mathbf{x}$ in equation (7) so $F(\varepsilon|\mathbf{x}) = F(\varepsilon)$;
(c) model 3, a full kernel stick breaking model.

Model 1 is the usual Gaussian linear regression model. Model 2 is more flexible because it does not assume that the residuals are Gaussian. However, model 2's residual distribution does not depend on $\mathbf{x}$, so the effect of $\mathbf{x}$ is linear in the mean and the covariates do not affect the higher moments.

For each design we generate $S = 50$ data sets. For each simulated data set and each of the models we compute each covariate's linear mean inclusion probability ($P(\beta_j \neq 0|y)$) and its kernel bandwidth inclusion probability ($P(\rho_j < 1|y)$). Table 1 gives the mean (with standard deviation in parentheses) of the S inclusion probabilities for each model and each covariate.

We use a vague, yet proper, prior for the hyperparameters. We take $\sigma_{\max} = 10$, $D \sim$gamma(0.1, 0.1) and $\tau^{-2} \sim$gamma(0.1,0.1), where the gamma priors are parameterized to have mean 1 and variance 10. We assume that the prior standard deviation of the regression parameters in the linear mean is $\sigma_\theta = 10$.

4.1. Linear model

The first design is the usual parametric linear model with

$$y = 2.5x_1 + 2.0x_2 + 1.5x_3 + 1.0x_4 + 0.5x_5 + \varepsilon,$$

where ε has a standard normal distribution. The results for this simulation are given in Table 1, part (a). As expected, the Gaussian linear model (model 1) identifies the highest proportion of the truly important linear regression coefficients. However, the non-Gaussian models (models 2 and 3) give nearly identical inclusion probabilities, so it seems that the added flexibility in the residual distribution leads to only a small sacrifice in variable selection for Gaussian data.

On average, the inclusion probabilities for the unimportant linear regression coefficients (variables 6–10) are around 0.05 for all three methods. The inclusion probabilities for these variables exceed 0.5 for at most one of the 50 data sets, so the type I error is even less than 0.05 (assuming that a variable is deemed important if it is included with probability higher than 0.5). Model 3 also performs variable selection for the residual density estimation. In this case none of the predictors should be included in the kernel stick breaking portion of the model. The average probability that variables are included in the stick breaking portion of the model (i.e. $P(\rho_j < 1)$) is around 0.25, and these probabilities exceed 0.50 less than 5% of the time (the results are not shown). It appears that the Bayesian model is well calibrated.

Table 1. Means (and standard deviations in parentheses) of the posterior inclusion probabilities for the simulation study†

Variance	*Linear mean parameters, β_j*			*Kernel bandwidths, ρ_j, model 3*
	Model 1	*Model 2*	*Model 3*	
(a) Design 1: Gaussian linear model				
1	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.26 (0.10)
2	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.26 (0.10)
3	1.00 (0.00)	1.00 (0.06)	0.97 (0.16)	0.27 (0.15)
4	0.87 (0.21)	0.87 (0.23)	0.83 (0.27)	0.30 (0.16)
5	0.33 (0.37)	0.32 (0.37)	0.30 (0.34)	0.26 (0.12)
6–10	0.05 (0.05)	0.05 (0.06)	0.06 (0.08)	0.24 (0.09)
(b) Design 2: heteroscedastic model				
1	0.97 (0.12)	0.98 (0.11)	0.98 (0.06)	0.26 (0.10)
2	0.98 (0.06)	0.98 (0.09)	0.99 (0.06)	0.23 (0.07)
3	0.97 (0.13)	0.96 (0.12)	0.74 (0.38)	0.68 (0.28)
4	0.04 (0.07)	0.03 (0.03)	0.04 (0.07)	0.51 (0.27)
5–10	0.05 (0.08)	0.04 (0.07)	0.05 (0.08)	0.22 (0.06)
(c) Design 3: non-linear model				
1	0.69 (0.34)	0.66 (0.35)	0.61 (0.40)	0.28 (0.12)
2	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.20 (0.16)
3	0.10 (0.18)	0.11 (0.20)	0.12 (0.21)	0.93 (0.20)
4–10	0.07 (0.09)	0.07 (0.10)	0.08 (0.12)	0.22 (0.07)
(d) Design 4: interaction model				
1	1.00 (0.00)	1.00 (0.00)	0.94 (0.21)	0.44 (0.28)
2	0.10 (0.17)	0.10 (0.17)	0.10 (0.20)	0.60 (0.26)
3–10	0.06 (0.08)	0.06 (0.08)	0.07 (0.11)	0.12 (0.09)
(e) Design 5: higher order model				
1	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.26 (0.10)
2	0.84 (0.29)	0.91 (0.23)	0.94 (0.18)	0.28 (0.13)
3	0.05 (0.28)	0.13 (0.24)	0.11 (0.15)	0.94 (0.16)
4–10	0.05 (0.07)	0.04 (0.07)	0.04 (0.05)	0.27 (0.10)

†For the linear mean parameters we report the mean and standard deviation of $P(\beta_j \neq 0)$ and for the kernel stick breaking parameters we report the mean and standard deviation of $P(\rho_j < 1)$. The models are 1, the parametric linear regression model, 2, the linear regression model with non-Gaussian errors, and 3, the full kernel stick breaking model.

4.2. Heteroscedastic model

The heteroscedastic model is

$$y = x_1 + x_2 + x_3 + (x_3 x_4 + 0.5)\varepsilon,$$

where ε has a standard normal distribution. In this model x_3 affects both the mean and the variance. Note that in Table 1, part (b), model 3 regularly includes x_3 in both the mean (the average inclusion probability is 0.74) and in the kernel stick breaking (the average inclusion probability is 0.68) portions of the model. The fourth variable appears only in the variance, and model 3 includes this covariate more often in the mixture model than the linear mean. The effect of x_4 is completely ignored by models 1 and 2.

4.3. Non-linear model

The third simulation design is

$$y = x_1 + \log(x_2) + 10(x_3 - 0.5)^2 + \varepsilon,$$

where ε has a standard normal distribution. In this design, x_2 and x_3 both have non-linear relationships with the outcome. The kernel stick breaking model can identify x_3's non-linearity. The average inclusion probability for x_3 in the residual distribution is 0.93. However, in the range $x_2 \in (0, 1)$, $\log(x_2)$ is only slightly non-linear and the kernel stick breaking model prefers to include x_2 in only the linear mean term. The average inclusion probability for x_2 in the residual distribution is only 0.20, so the model cannot detect this non-linear relationship.

4.4. Interaction model

The interaction model is

$$y = 2x_2\, I(x_1 < 0.5) - 2x_2\, I(x_1 > 0.5) + \varepsilon,$$

where ε has a standard normal distribution. Although this density is discontinuous in covariate space, we hope to show that our model is sufficiently flexible to accommodate this deviation from the assumptions. None of the three models contain the interaction between x_1 and x_2 in their parametric mean term. All the models can identify the main effect for x_1, but x_2 is rarely included in the mean term because its effect is only apparent conditioned on x_1. Model 3 can accommodate the missing interaction in the residual model; on average x_2 is included in the residual distribution with probability 0.60.

4.5. Higher order model

The final simulation design is

$$y = 2x_1 + x_2 + I(x_3 < 0.5)U + I(x_3 > 0.5)t_{2.5}/\sqrt{5}$$

where U has a $U(-0.5\sqrt{12}, 0.5\sqrt{12})$ distribution and $t_{2.5}$ has t-distribution with 2.5 degrees of freedom. In this simulation the errors are not Gaussian and model 1's assumptions are violated. As a result, the semiparametric linear model's (model 2) inclusion probabilities for the parameters in the linear mean (x_1 and x_2) are higher on average than in the parametric model. This design deviates from the usual linear model because the residual distribution's tail behaviour (but not its mean, variance or skewness) depends on x_3. The average inclusion probability for x_3 in model 3's residual distribution is 0.94. Also, the kernel stick breaking model which correctly characterizes the residual distribution's dependence on x_3 has the highest average inclusion probabilities for x_1 and x_2 in the linear mean.

We also repeated the final simulation design with correlated predictors. We generated Gaussian variables z_j with mean 0 and auto-regressive correlation $\mathrm{cov}(z_j, z_k) = 0.7^{|j-k|}$, and then transformed the predictors to the unit interval by using the probit link $x_j = \Phi(z_j)$, where Φ is the standard normal distribution function. The inclusion probabilities in the linear mean for x_1 and x_2 changed from 1.00 and 0.94 with uncorrelated predictors to 1.00 and 0.73 with correlated predictors, and the inclusion probability for x_3 is the kernel bandwidth changed from 0.94 with uncorrelated predictors to 0.91 with correlated predictors. The inclusion probabilities for the other predictions were similar to the uncorrelated predictors case. As with most variable selection methods, the results were quite sensitive to collinearity. However, even with moderate correlation the method remains fairly effective.

In summary, this simulation study shows that the kernel stick breaking model is very competitive with the parametric model even when the data are generated from a Gaussian linear model. The kernel stick breaking model is also effective at identifying several types of effects, including non-linear mean relationships, variance inflation, missing interactions and increased tail probability. Simulation design 5 also demonstrates that properly modelling the residual distribution can improve the likelihood of selecting important predictors in the linear mean.

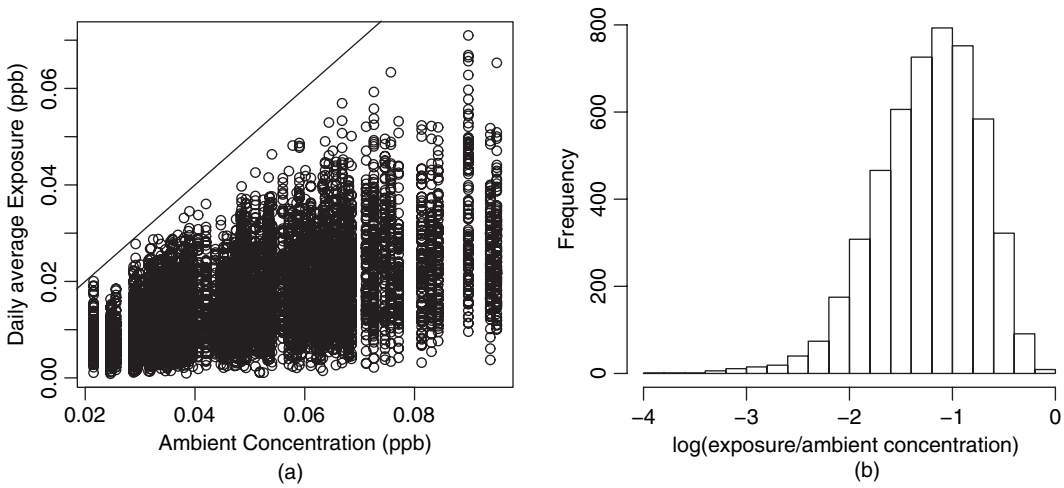

Fig. 1. Plots of the APEX data: (a) average daily exposure *versus* daily average ambient concentration; (b) log-ratio of exposure to ambient concentration

Table 2. Posterior summaries for the main effects model†

Parameter	*Overall inclusion probability*	*Linear term* β_j		*Kernel bandwidths* $P(\rho<1)$
		$P(\beta_j \neq 0)$	*95% interval*	
Weekend	1.00	1.00	(0.06, 0.10)	0.05
Physical activity index	1.00	1.00	(0.30, 0.50)	0.22
Temperature	1.00	0.00	(0.00, 0.00)	1.00
Ambient concentration	1.00	1.00	(−0.20, −0.10)	0.15
No air-conditioning in home	1.00	0.01	(0.00, 0.00)	1.00
Gender (male ≡ 1)	1.00	1.00	(0.10, 0.15)	0.04
Employed	1.00	0.01	(0.00, 0.00)	1.00
Age ⩽ 4 years	0.37	0.32	(−0.11, 0.00)	0.06
Age 5–18 years	0.12	0.00	(0.00, 0.00)	0.12
Age ⩾ 65 years	0.04	0.00	(0.00, 0.00)	0.04
AER residence	1.00	1.00	(0.47, 0.77)	1.00
AER bar or restaurant	0.08	0.00	(0.00, 0.00)	0.08
AER school × age 5–18	1.00	1.00	(0.17, 0.34)	0.06
AER child care × age ⩽ 4	0.06	0.01	(0.00, 0.00)	0.05
AER office × employed	1.00	1.00	(0.21, 0.41)	0.05
AER shop	0.04	0.00	(0.00, 0.00)	0.04
Decay rate	1.00	1.00	(−0.19, −0.08)	0.07

†'Overall inclusion probability' is the probability that the variable is included in the mean or residual model component (i.e. $\beta_j \neq 0$ or $\rho < 1$).

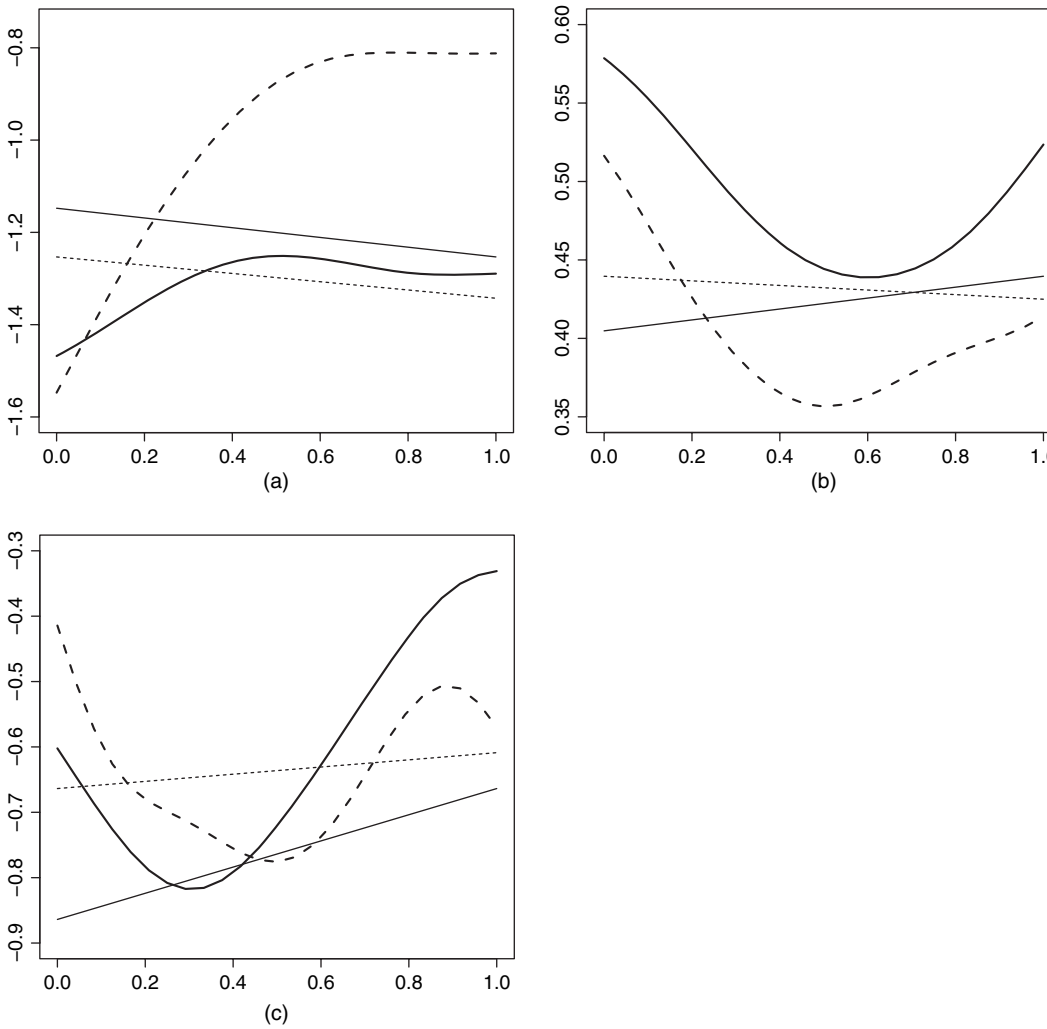

Fig. 2. Estimated moments of the log-exposure ratio by four important covariates (———, no air-conditioning; -------, employment; ———, temperature; – – –, AER-residual): (a) mean; (b) standard deviation; (c) skewness

5. Analysis of the APEX data

To illustrate the method proposed, we use the data that were described in Section 2 to build a statistical emulator for personal ozone exposure. We analyse $n = 5000$ simulated exposures, using the same priors as in Section 4. Fig. 1(a) plots the average daily exposures against the ambient concentrations. Although exposure generally increases linearly with ambient concentration, there is considerable variation in the ratio of exposure to ambient concentration due to human activity patterns and other factors that were discussed in Section 2. To account for the natural multiplicative relationship between concentration and exposure, we analyse the logarithm of the ratio of exposure to ambient concentration, plotted in Fig. 1(b). The log-ratio is slightly left skewed.

Table 2 gives the inclusion probabilities from fitting the full kernel stick breaking model with the main effects linear trend. Six predictors are included (in either the mean and residual

models, 'Overall inclusion probability') with probability less than 0.5: air exchange rates in bars or restaurants, child care facilities and shopping centres, and the three dummy variables for age. These variables are in fact included in the stochastic exposure model and certainly play some role in the simulation. However, it appears that their effects are not sufficiently large to warrant the added model complexity. Excluding these variables from the emulator provides a simpler model in terms of both mathematical and practical complexity, as less data must be collected to emulate APEX.

Several predictors are included in the linear mean but not residual distribution models. These predictors are adequately modelled by using simple linear regression. The exposure ratio is higher for males and on weekends. The exposure ratio also has the expected relationships with physical activity, ozone decay rate and the air exchange rates for schools and offices. Ambient concentration is included with probability 1 and its posterior median coefficient is -0.15. Ambient concentration is also included as a fixed offset, so this reflects the known non-linear relationship between exposure E and ambient concentration C,

$$E \propto C \exp(-0.15C). \tag{18}$$

The model identifies air-conditioning in the home, temperature, employment status and residential air exchange rate as important predictors with non-standard effects. These variables are all included in the stick breaking component of the model with probability near 1. To illustrate how the exposure distribution depends on each of these variables, Fig. 2 presents the mean, standard deviation and skewness of the exposure distribution for a range of values for these three covariates. To create this plot for one covariate, all other covariates are fixed at their medians (rather than means, since many of the covariates are binary), and the moments are calculated on a grid of values for the covariate of interest by using formulae such as equation (9). The moments for the binary variables air-conditioning and employment status are calculated only at the end points and, for comparative purposes, all covariates are scaled to the unit intervals.

Fig. 2 shows that residential air exchange rate has the most dramatic effect on the exposure distribution. The strong relationship between residential air exchange rate and exposure is also apparent in Fig. 3(a)'s plot of the raw data. As the residential air exchange rate increases, the mean exposure increases because more ozone penetrates into the residence. The mean function is non-linear and plateaus when the air exchange rate reaches 3 (scaled to about 0.5 in Fig. 2). Fig. 2 also shows that the standard deviation decreases with air exchange rate. With a large exchange rate the variability due to human activity becomes less relevant because the indoor and outdoor environments have similar ozone levels. Residential air exchange rate also has a dramatic effect on the skewness of the exposure distribution. The posterior mean densities in Fig. 3(b) are all left skewed since density is essentially bounded at zero because exposure rarely exceeds the ambient concentration. The right-hand tail is more compressed for large air exchange rates because more mass is near the upper bound.

Employment status affects both the mean and the variance of the outcome; the sample mean (with standard deviation in parentheses) of the log-exposure ratio is -1.26 (0.47) for employed people and -1.22 (0.51) for unemployed people. Fig. 4 shows that employment status also affects the shape of the distribution. Both densities in Fig. 4 are skewed to the left, but the unemployment density has more mass near zero and thus a higher probability of exposure approaching the ambient concentration. It may be that unemployed people are more likely to spend considerable time outdoors.

Residential air-conditioning is thought to be a major driver of the exposure model and is currently the subject of research at the Environmental Protection Agency. It is believed that

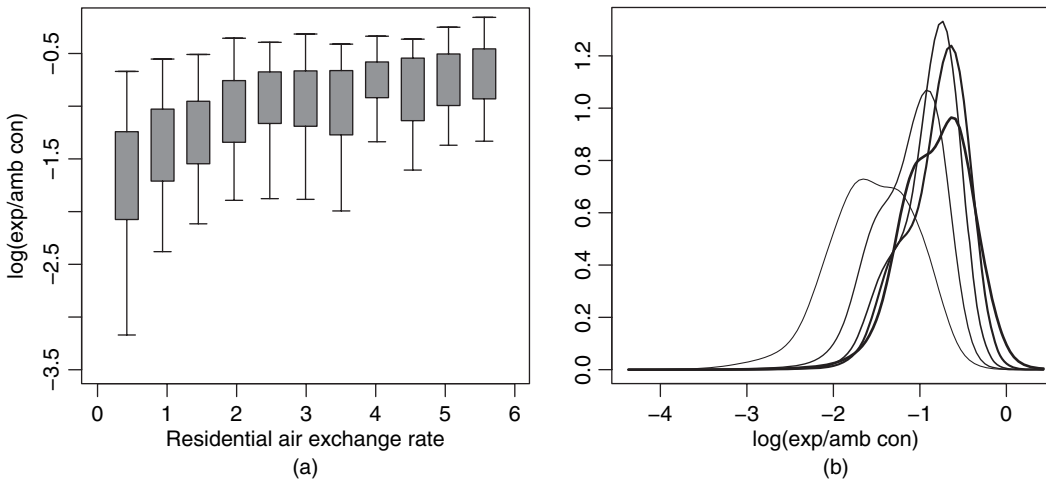

Fig. 3. (a) Raw data and (b) posterior mean density of the log-exposure ratio by residential air exchange rate: in (b) several densities are plotted for residual air exchange rate varying from 0.42 (———) to 5.56 (———)

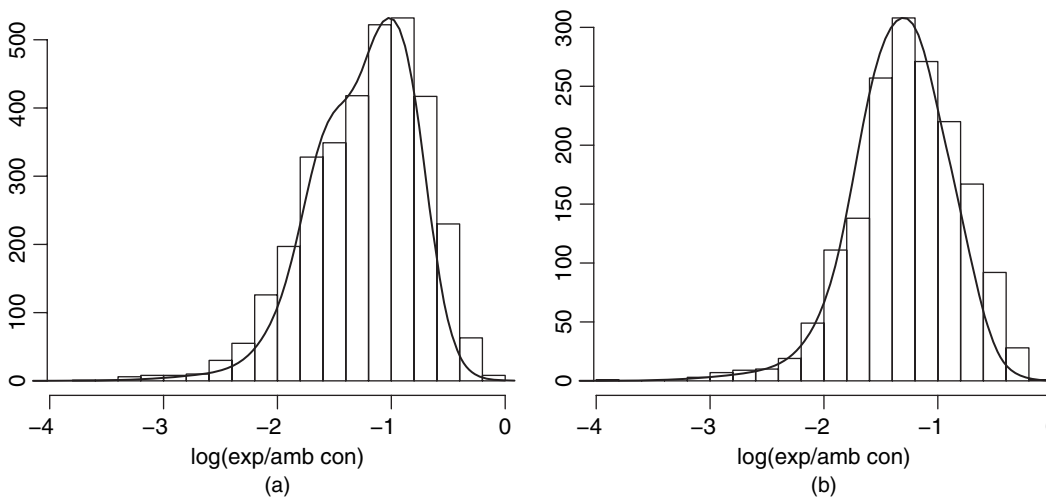

Fig. 4. Data *versus* posterior mean conditional density for exposure for (a) unemployed and (b) employed people

people with air-conditioning are exposed to less ozone because the air-conditioning system filters the ozone and prevents the outdoor ozone from penetrating the home. Indeed, for these data the mean response is larger for people without (−1.02) than with (−1.27) air-conditioning. Surprisingly air-conditioning is not included in the linear mean term (Table 2). Air-conditioning is, however, included in the stick breaking component of the model with probability 1, but Fig. 2 shows that with the other predictors fixed at their medians air-conditioning has a positive effect on the mean response. The simulation in Section 4.4 illustrates that the kernel stick breaking model can be used to identify missing interactions, and the conflicting results for air-conditioning suggest that interactions should be added to our model.

Table 3 gives the inclusion probabilities for the model that includes all two-way interactions

between the four variables identified as having non-standard effects (air-conditioning, employment status, temperature and residential air exchange rate). These effects are added to the linear mean model only; thus the kernel stick breaking model is the same as the previous fit. Two of the six interactions are included with probability greater than 0.5. We note that the non-significant interactions in the mean do not necessarily imply additive effects for these predictors because of potential non-linearity in the kernel stick breaking residual component. Residential air exchange rate interacts in the linear mean with air-conditioning and employment status; residential air exchange rate has less effect for homes with air-conditioning because air-conditioning prevents ozone from penetrating, and residential air exchange rate has less effect for employed people because they spend less time in their residences. Including these interactions reduces the probability that employment status is included in the kernel stick breaking portion of the model from 1.00 to 0.82.

Despite the addition of the interactions, temperature and air-conditioning remain in the kernel stick breaking portion of the model with probability 1.00. Fig. 5(a) plots the response by temperature and air-conditioning status. The effect of air-conditioning is the strongest (on the mean and variance) when the temperature is between 70 and 80 °F. People without air-conditioning are most likely to open their windows in this temperature range, creating the greatest contrast between the two groups. The effect on air-conditioning is smaller for very high temperatures because people without air-conditioning may close both windows and blinds to stay cool. The kernel stick breaking model identifies this complicated relationship between expo-

Table 3. Posterior summaries for the model with interactions†

Parameter	*Overall inclusion probability*	*Linear term β_j*		*Kernel bandwidths $P(\rho<1)$*
		$P(\beta_j \neq 0)$	*95% interval*	
Weekend	1.00	1.00	(0.06, 0.11)	0.05
Physical activity index	1.00	1.00	(0.29, 0.48)	0.13
Temperature	1.00	0.30	(−0.33, 0.00)	1.00
Ambient concentration	1.00	1.00	(−0.21, −0.12)	0.25
No air-conditioning in home	1.00	1.00	(0.14, 0.31)	1.00
Gender (male ≡ 1)	1.00	1.00	(0.10, 0.15)	0.02
Employed	0.82	0.00	(0.00, 0.00)	0.82
Age ⩽ 4 years	0.63	0.61	(−0.25, 0.00)	0.03
Age 5–18 years	0.04	0.01	(0.00, 0.00)	0.03
Age ⩾ 65 years	0.31	0.01	(0.00, 0.00)	0.31
AER residence	1.00	1.00	(0.63, 0.99)	1.00
AER bar or restaurant	0.09	0.04	(−0.04, 0.00)	0.05
AER school × age 5–18	1.00	1.00	(0.20, 0.37)	0.05
AER child care × age ⩽ 4	0.61	0.59	(0.00, 0.30)	0.03
AER office × employed	1.00	1.00	(0.24, 0.44)	0.04
AER shop	0.04	0.00	(0.00, 0.00)	0.04
Decay rate	1.00	1.00	(−0.19, −0.08)	0.06
Temperature × no air-conditioning	0.03	0.03	(0.00, 0.00)	0.00
Temperature × employed	0.49	0.49	(−0.19, 0.00)	0.00
Temperature × AER residence	0.30	0.30	(0.00, 0.72)	0.00
No air-conditioning × employed	0.01	0.01	(0.00, 0.00)	0.00
No air-conditioning × AER residence	1.00	1.00	(−0.59, −0.26)	0.00
Employed × AER residence	0.57	0.57	(−0.26, 0.00)	0.00

†'Overall inclusion probability' is the probability that the variable is included in the mean or residual model component (i.e. $\beta_j \neq 0$ or $\rho < 1$).

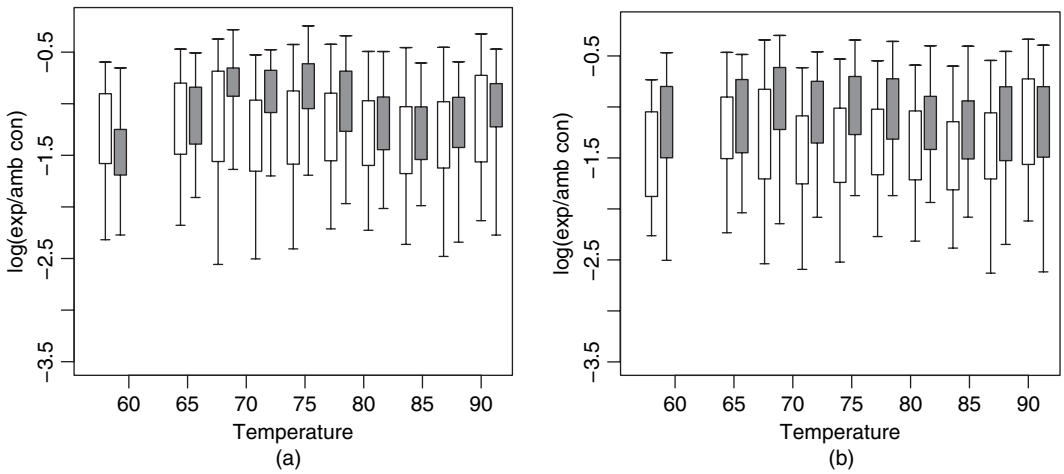

Fig. 5. (a) Plot of the raw data by temperature and air-conditioning (■, no air-conditioning; □, air-conditioning) and (b) temperature and residential air exchange rate (■, residential air exchange rate above the median residential air exchange rate)

sure, temperature and air-conditioning. Similarly, Fig. 5(b) shows that the effect of residential air exchange rates is the largest when the temperature is between 70 and 80 °F. Also, the mean response is between -1.25 and -1.28 for all combinations of air-conditioning and employment, except for the group without air-conditioning or employment, which has mean response -0.98. Therefore, it appears that air-conditioning has a protective effect only for the unemployed.

Including the interactions also affects the other variables in the model. Air-conditioning status is now included in the linear mean, as expected. Also the child care air exchange rate and the indicator of age less than 4 years are included in the linear mean term, providing evidence of a protective effect for young children. This may be because the difference between air quality in homes and child care facilities is only substantial after accounting for both residential air exchange rate and air-conditioning.

To inspect the fit of the final model we use fivefold cross-validation. We fit the final model described above and linear regression model with the same variables included in the mean, but Gaussian residuals, independent over the covariates. Fig. 6(a) plots the posterior predictive means against the observed data. There is a clear correlation, but considerable variation in the observations, emphasizing the need to consider the distribution of exposure across the population rather than a simple one-number summary. To evaluate whether the posterior predictive distribution fits the data well, we compute the probability inverse transform diagnostic PIT that was discussed in Gneiting *et al.* (2007). For each observation we compute $\mathrm{PIT}_i = P(y^* < y_i)$, where y_i is the observation and y_i^* follows the posterior predictive distribution. Assuming that the model is correct, PIT_i should approximately follow a $U(0, 1)$ distribution. Fig. 6(b) shows that the PIT-diagnostics more closely resemble uniform than those from the linear regression model.

6. Discussion

In this paper we present a method for variable selection with Bayesian conditional density estimation. We alleviate the curse of dimensionality by using stochastic search variable selection to identify a subset of covariates that have more than just additive effects on the mean of

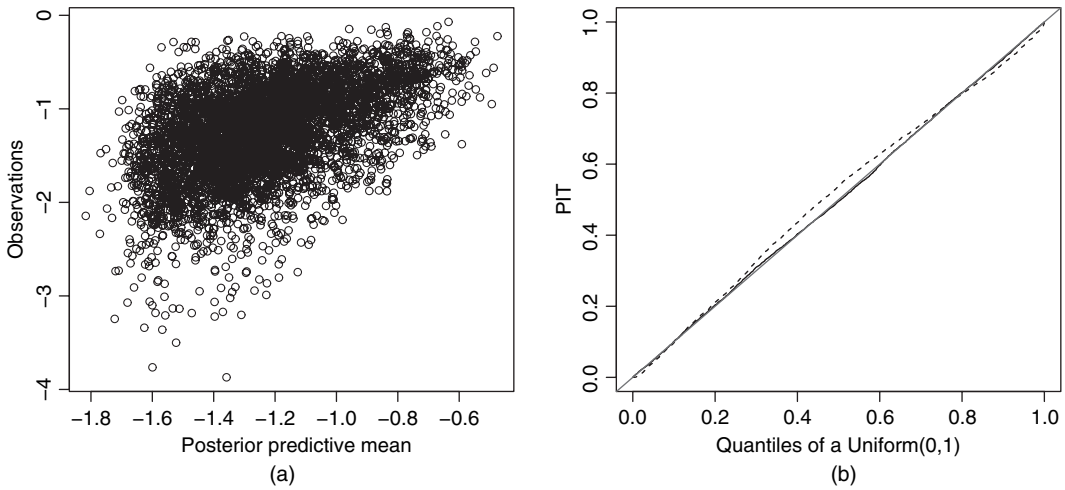

Fig. 6. Summary of the cross-validation: (a) testing data *versus* the posterior predictive mean; (b) PIT-statistics (*QQ*-plot) for the Gaussian (———) and non-Gaussian (- - - - -) models (⁄, 45° line)

the conditional density. We use our approach to build an emulator for a pollution exposure model to be used in a future large-scale study of the effect of air pollution on human health.

A strength of our approach is its flexibility; as we show in the simulation study we can identify several complicated effects. Identifying variables with complicated effects aids in the model building process, as we can focus on building a parametric model for a few variables rather than high dimensional exploratory analysis. This is demonstrated by our analysis of the APEX simulator. After an initial fit with the kernel stick breaking model we add several interactions. It is possible that after a few more iterations of this process we could postulate a model that did not include any predictors in the non-parametric part. In this case, the kernel stick breaking model serves as a guide to model building and as verification that the parametric model captures the important features of the data.

A future extension of this work would be to combine the APEX exposure simulator with field data to validate and/or improve the estimate of the exposure distribution by identifying and accounting for systematic biases in the APEX model. The calibration–validation problem was discussed for deterministic models in Bayarri *et al.* (2007). A simultaneous model for stochastic APEX simulator and field data would be to assume that the two sources of data shared some features, e.g. the mixture probabilities and variances, but had different regression coefficients and mixture means that were given multivariate priors to borrow strength across the two sources of data.

Acknowledgements

The authors thank the National Science Foundation (Reich, DMS-0354189; Bondell, DMS-0705968; Fuentes, DMS-0706731 and DMS-0353029), Sandia National Laboratories (Storlie, Sandia University Research Program grant 22858), the Environmental Protection Agency (Fuentes, R833863) and National Institutes of Health (Fuentes, 5R01ES014843-02) for partial support of this work. The authors also thank John Langstaff of the US Environmental Protection Agency for his help with the APEX model and interpreting the results.

References

Bayarri, M. H., Berger, J. O., Paulo, R., Sacks, J., Cafeo, J. A., Cavendish, J., Lin, C. H. and Tu, J. (2007) A framework for validation of computer models. *Technometrics*, **49**, 138–154.

Binkowski, F. S. and Roselle, S. J. (2003) Models-3 community multiscale air quality (CMAQ) model aerosol component, 1: model description. *J. Geophys. Res.*, **108**, 41–83.

Calder, C. A., Holloman, C. H., Bortnik, S. M., Strauss, W. J. and Morara, M. (2008) Relating ambient particulate matter concentration levels to mortality using an exposure simulator. *J. Am. Statist. Ass.*, **103**, 137–148.

Chen, V. C. P., Tsui, K. L., Barton, R. R. and Meckesheimer, M. (2006) A review on design, modeling and applications of computer experiments. *IIE Trans.*, **38**, 273–291.

Chung, Y. and Dunson, D. B. (2009) Nonparametric Bayes conditional distribution modeling with variable selection. *J. Am. Statist. Ass.*, **104**, 1646–1660.

Dominici, F., Daniels, M., Zeger, S. L. and Samet, J. M. (2002) Air pollution and mortality: estimating regional and national dose response relationships. *J. Am. Statist. Ass.*, **97**, 100–111.

Dunson, D. B. and Park, J. H. (2008) Kernel stick-breaking processes. *Biometrika*, **95**, 307–323.

Fang, K. T., Li, R. and Sudjianto, A. (2006) *Design and Modeling for Computer Experiments*. Boca Raton: Chapman and Hall–CRC.

Fuentes, M., Song, H., Ghosh, S. K., Holland, D. M. and Davis, J. M. (2006) Spatial association between speciated fine particles and mortality. *Biometrics*, **62**, 855–863.

Galli, E., Cuéllar, L., Eidenbenz, S., Ewers, M., Mniszewski, S. and Teuscher, C. (2009) ActivitySim: large-scale agent-based activity generation for infrastructure simulation. In *Proc. 2009 Spring Simulation Multiconf.*, pp. 1–9. San Diego: Society for Computer Simulation International.

George, E. I. (2000) The variable selection problem. *J. Am. Statist. Ass.*, **95**, 1304–1308.

George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *J. Am. Statist. Ass.*, **88**, 881–889.

George, E. I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statist. Sin.*, **7**, 339–373.

Gneiting, T., Balabdaoui, F. and Raftery, A. E. (2007) Probabilistic forecasts, calibration and sharpness. *J. R. Statist. Soc.* B, **69**, 243–268.

Gustafson, P. (2000) Bayesian regression modeling with interactions and smooth effects. *J. Am. Statist. Ass.*, **95**, 745–763.

Hastie, T. and Tibshirani, R. (1990) *Generalized Additive Models*. New York: Chapman and Hall.

Holloman, C. H., Bortnik, S., Morara, M., Strauss, W. and Calder, C. (2004) A Bayesian hierarchical approach for relating PM2.5 exposure to cardiovascular mortality in North Carolina. *Environ. Hlth Perspect.*, **112**, 1282–1288.

Iooss, B., Ribatet, M. and Marrel, A. (2008) Global sensitivity analysis of stochastic computer models with generalized additive models. (Available from `http://fr.arxiv.org/abs/0802.0443v3`.)

Lee, D. and Shaddick, G. (2007) Time-varying coefficient models for the analysis of air pollution and health outcome data. *Biometrics*, **63**, 1253–1261.

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

Papaspiliopoulos, O. and Roberts, G. (2008) Retrospective MCMC for Dirichlet process hierarchical models. *Biometrika*, **95**, 169–186.

Pope, C. A., Dockery, D. and Schwartz, J. (1995) Review of epidemiological evidence of health effects of particulate air pollution. *Inhaln Toxicol.*, **47**, 1–18.

R Development Core Team (2006) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Reich, B. J. and Fuentes, M. (2007) A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Ann. Appl. Statist.*, **1**, 249–264.

Reich, B. J., Fuentes, M. and Burke, J. (2009) Analysis of the effects of ultrafine particulate matter while accounting for human exposure. *Environmetrics*, **20**, 131–146.

Reich, B. J., Storlie, C. S. and Bondell, H. D. (2008) Bayesian variable selection for nonparametric regression: application to deterministic computer codes. *Technometrics*, **51**, 110–120.

Sacks, J., Welch, W. L., Mitchell, T. J. and Wynn, H. P. (1989) Design and analysis of computer experiments. *Statist. Sci.*, **4**, 409–435.

Schwartz, J. (1994) Air pollution and daily mortality: a review and meta analysis. *Environ. Res.*, **64**, 36–52.

Sethuraman, J. (1994) A constructive definition of Dirichlet priors. *Statist. Sin.*, **4**, 639–650.

Shaddick, G., Lee, D., Zidek, J. V. and Salway, R. (2008) Estimating exposure response functions using ambient pollution concentrations. *Ann. Appl. Statist.*, **4**, 1249–1270.

Shively, T., Kohn, R. and Wood, S. (1999) Variable selection and function estimation in nonparametric regression using a data-based prior (with discussion). *J. Am. Statist. Ass.*, **94**, 777–806.

US Environmental Protection Agency (2007) Ozone population exposure analysis for selected urban areas. *Publication 452R07010*. US Environmental Protection Agency, Research Triangle Park.

Waupotitsch, R., Eidenbenz, S., Smith, J. P. and Kroc, L. (2006) Multi-scale integrated information and telecommunications system (MIITS): first results from a large-scale end-to-end network simulator. In *Proc. 38th Conf. Winter Simulation, Monterey*, pp. 2132–2139.

Wood, S., Kohn, R., Shively, T. and Jiang, W. (2002) Model selection in spline nonparametric regression. *J. R. Statist. Soc.* B, **64**, 119–139.