

# Variable selection for high-dimensional Bayesian density estimation: Application to human exposure simulation

Brian J. Reich<sup>a1</sup>, Eric Kalendra<sup>a</sup>, Curtis B. Storlie<sup>b</sup>, Howard D. Bondell<sup>a</sup>, Montserrat Fuentes<sup>a</sup>

<sup>a</sup> Department of Statistics, North Carolina State University

<sup>b</sup> Statistical Sciences Group, Los Alamos National Laboratory

April 15, 2011

## Abstract

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. While these methods show great promise, they are limited by the computational demands of the exposure model. In this paper 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. That is, 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.

**Key words:** Air pollution; Bayesian nonparametrics; high dimensional data; kernel stick-breaking prior; stochastic computer models.

---

<sup>1</sup>Corresponding author, email: reich@stat.ncsu.edu. The authors thank the National Science Foundation (Reich, DMS-0354189, Bondell, DMS-0705968, Fuentes, DMS-0706731, DMS-0353029), Sandia National Laboratories (Storlie, SURP 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 EPA for his help with the APEX model and interpreting the results.

# 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; 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 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 using a computer simulation. For example, the EPA 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, in-vehicle). Although APEX does not actually measure exposure 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 using observed exposure (for example, US EPA, 2007). Using the stochastic model, we can replace the single ambient concentration predictor in the health model

with a summary of the exposure distribution, e.g., 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 using the individual's characteristics such as housing type 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; Shaddock 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 the APEX model to be used in a future health study. Building an emulator for APEX poses two major challenges: (1) APEX is a *stochastic* computer model, and (2) APEX has a large number of inputs. Although our motivation is the APEX model, the approach developed here could be generalized to other stochastic computer models with similar features, such as models for a 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; 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 modeled using nonparametric 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 the APEX model 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 analyzed stochastic computer model output by modeling the distribution’s mean and variance using generalized additive models (Hastie and Tibshirani, 1990). While this may capture many of the important features of the stochastic model, a more flexible model is certainly desirable. For example, in the APEX model the covariates clearly affect the skewness of the exposure distribution; see Section 5. Sufficiently modeling this tail behavior 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 normals, 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 “the 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 of the inputs are used in the simulation model, emulation may be improved by excluding inputs with small effects in favor of a parsimonious statistical model.

We reduce the dimension of the covariate space 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 only affect the mean response, and only appear in the mean as a linear combination. A more flexible approach is nonparametric 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., 2009). 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 more generally aims to identify variables with any effect on the conditional distribution of the response. Chung and Dunson (2010) 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 (1) a linear term affecting the mean of the response and/or (2) 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 the APEX model indeed affect the exposure distribution. However, the effect from most of the predictors can be modeled 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 conduct further exploration for the subset of variables included in the mixture of normals component of the conditional density model. For example, there are 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 the APEX model and the simulated data and Section 3 develops the statistical emulator. A brief simulation study in Section 4 illustrates the flexibility of our nonparametric model. We analyze the APEX model output in Section 5. Section 6 concludes.

## 2 Data description

In this section we give a brief description of the APEX model; a full description can be found at [http://www.epa.gov/ttn/fera/apex\\_download.html](http://www.epa.gov/ttn/fera/apex_download.html). In Section 5 we analyze 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 exposure 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 EPA’s Consolidated Human Activity Database (CHAD). CHAD contains personal diaries of over 22,000 individuals from exposure studies conducted around the US. The diaries describe the activity pattern of the individual throughout the day and are selected to match the hypothetical individuals based on personal characteristics, season, day of the week, and average daily temperature.

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

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

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 using the differential equation

$$\frac{dC_{hj}}{dt} = AER_j * (C_h^{ambient} - C_{hj}) - DR * C_{jk} + C_{hj}^{source}, \quad (2)$$

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 (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 pollutant inside the microenvironment. The concentration in the outdoor microenvironment is taken to be  $C_{hj} = C_h^{ambient}$ . 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 is modeled 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 a 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 (PAI), i.e., the time-averaged METS (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 using numerical integration over the conditional distribution 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 nonparametric 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 The 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 modeling, 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 normals

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

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

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

where  $1 - \sum_{j=1}^{k-1} p_j$  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  $k^{\text{th}}$  component.

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

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

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  $\varepsilon$  given  $\mathbf{x}$  is modeled as the infinite mixture

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

where  $p_k(\mathbf{x})$  are the mixture weights with  $\sum_{k=1}^{\infty} p_k(\mathbf{x}) = 1$  for all  $\mathbf{x}$ . The means and variances have priors  $\mu_k \stackrel{iid}{\sim} F_0$  and  $\sigma_k \stackrel{iid}{\sim} U(0, \sigma_{max})$ , 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  $\mu_k \stackrel{iid}{\sim} N(0, \tau)$ , where  $\tau$  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

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

for  $k > 1$ . Here  $\prod_{j=1}^{k-1} (1 - v_j w_j(\mathbf{x}))$  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.,

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

where  $\boldsymbol{\psi}_k = (\psi_{k1}, \dots, \psi_{kp})'$  is the kernel's center and  $\Sigma$  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  $j = 1, \dots, p$  and then assume the knots have priors  $\psi_{kj} \stackrel{iid}{\sim} \text{Unif}(0,1)$ . As before  $v_k \stackrel{iid}{\sim} \text{Beta}(1, D)$ .

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

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

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

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 (10) varying the probabilities with  $\mathbf{x}$  gives a rich class 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  $\mathbf{x}'\boldsymbol{\beta}$  and the residual distribution  $F(\varepsilon|\mathbf{x})$ . In both cases we use stochastic search variable selection. Following George and McCulloch (1993, 1997), let

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

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

For the nonparametric component, we assume  $\Sigma$  is diagonal with diagonal elements  $\alpha_1, \dots, \alpha_p$ .

Defining  $\rho_j = \exp(-\alpha_j)$ , (8) can be written

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

Variable selection is performed by giving  $\rho_j$  prior mass at one; 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 to  $x_j = 0$  is  $2(1/2 - \psi_{kj}) \log(\rho_j)$ . Therefore, the stick-breaking weights vary by  $x_j$  if and only if  $\rho_j$  is less than one, and the knot  $\psi_{kj}$  controls the value of  $x_j$  that is favored by term  $k$ .

We assume

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

where  $\pi_{2j} \sim \text{Bern}(0.5)$  and  $\gamma_j \stackrel{iid}{\sim} \text{Unif}(0,1)$ . 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  $\lambda_t$  includes population size and confounders such as meteorology, time trends, etc.

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 (9). While there may be some value in using the exposure simulator to provide an improved one-number summary of exposure compared to ambient concentration, we feel a greater contribution is to model the entire exposure distribution of the population, and 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  $\sum_i \lambda_t \exp[\beta e_{ti}]$ . To approximate this sum, Reich et al (2009) assume that the exposure distribution

on day  $t$ ,  $q_t(e)$ , is Gaussian with mean  $\mu_t$  and standard deviation  $\sigma_t$ , and the expected number of events becomes proportional to

$$\int \lambda_t \exp(\beta e) q_t(e) de = \lambda_t \exp(\mu_t \beta + \sigma_t^2 \beta^2). \quad (13)$$

In addition, they account for uncertainty in the exposure distribution by placing priors (derived from several model simulations) on  $\mu_t$  and  $\sigma_t$ .

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

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

where  $p_{kt}$ ,  $\mu_{kt}$  and  $\sigma_{kt}$  are the mixture weight, mean and standard deviation for the  $k^{th}$  mixture component. Alternatively, it is possible to approximate the convolution integral in (13) using, say, the quintiles of the exposure distribution  $q_{1t}, \dots, q_{5t}$ , i.e.,

$$\lambda_t \sum_{k=1}^5 \exp(q_{kt} \beta). \quad (15)$$

where  $q_{1t}, \dots, q_{5t}$  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

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

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

$$g_i \sim \text{Categorical}(p_1(\mathbf{x}_{(i)}), p_2(\mathbf{x}_{(i)}), \dots), \quad (17)$$

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

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

An alternative sampling approach is to replace (6)'s infinite mixture with a finite mixture of  $m$  components by defining the probability for the  $m^{\text{th}}$  term as  $p_m(\mathbf{x}_{(i)}) = 1 - \sum_{j=1}^{m-1} p_j(\mathbf{x}_{(i)})$  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 for the full model with  $n = 5,000$  responses in Section 5 runs in a few hours on an ordinary PC. The APEX model 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 extrapolate to the complete data set 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, described in Sections 4.1-4.5 respectively. Each design has  $n = 200$  observations and  $p = 10$  covariates  $x_1, \dots, x_{10}$  generated independently from the Uniform(0,1) distribution. We compare three models which are all special cases of Section 3's full kernel stick-breaking model as follows:

1. Linear regression model with normal errors:  $F(\varepsilon|\mathbf{x}) = N(0, \sigma)$  in (5).
2. Linear regression model with nonparametric errors:  $w_j(\mathbf{x}) = 1$  for all  $j$  and  $\mathbf{x}$  in (7) so  $F(\varepsilon|\mathbf{x}) = F(\varepsilon)$ .
3. Full kernel stick-breaking model

Model 1 is the usual Gaussian linear regression model. Model 2 is more flexible because it does not assume 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 (standard deviation) of the  $S$  inclusion probabilities for each model and each covariate.

We use vague, yet proper prior for the hyperparameters. We take  $\sigma_{max} = 10$ ,  $D \sim \text{Gamma}(0.1, 0.1)$ , and  $\tau^{-2} \sim \text{Gamma}(0.1, 0.1)$ , where the gamma priors are parameterized to have mean 1 and variance 10. We assume 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

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

where  $\varepsilon$  has a standard normal distribution. The results for this simulation are given in Table 1a. 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 exceeds 0.5 for at most 1 of the 50 data sets, so the type I error is even less than 0.05 (assuming 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 (not shown). It appears the Bayesian model is well-calibrated.

## 4.2 Heteroskedastic model

The heteroskedastic model is

$$y = x_1 + x_2 + x_3 + (x_3 \cdot x_4 + .5) \cdot \varepsilon,$$

where  $\varepsilon$  has a standard normal distribution. In this model  $x_3$  affects both the mean and variance. Notice in Table 1b that Model 3 regularly includes  $x_3$  in both the mean (average inclusion probability is 0.74) and in the kernel stick-breaking (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  $\varepsilon$  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 is able to identify  $x_3$ 's nonlinearity. 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 only include  $x_2$  in the linear mean term. The average inclusion probability for  $x_2$  in the residual distri-

bution is only 0.20, so the model is not able to detect this non-linear relationship.

#### 4.4 Interaction model

The interaction model is

$$y = 2 * x_2 I(x_1 < .5) - 2 * x_2 I(x_1 > .5) + \varepsilon,$$

where  $\varepsilon$  has a standard normal distribution. Although this density is discontinuous in covariate space, we hope to show our model is flexible enough 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 of the models are able to 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

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

where  $U$  has a  $\text{Uniform}(-0.5 * \sqrt{12}, 0.5 * \sqrt{12})$  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 the parametric model. This design deviates from the usual linear model because the residual distribution's tail behavior (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 zero and autoregressive correlation  $\text{Cov}(z_j, z_k) = 0.7^{|j-k|}$ , and then transformed the predictors to the unit interval using the probit link  $x_j = \Phi(z_j)$ , where  $\Phi$  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 somewhat 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 modeling the residual distribution can improve the likelihood of selecting important predictors in the linear mean.

## 5 Analysis of the APEX data

To illustrate the proposed method, we use the data described in Section 2 to build a statistical emulator for personal ozone exposure. We analyze  $n = 5,000$  simulated exposures, using the same priors as in Section 4. Figure 1a plots the average daily exposures against the ambient concentrations. While 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 discussed in Section 2. To account for the natural multiplicative relationship between concentration and exposure, we analyze the log of the ratio of exposure to ambient concentration, plotted in Figure 1b. 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 in prob”) with probability less than 0.5: air exchange rates in bars/restaurants, child care facilities, and shopping centers, and the three dummy variable for age. These variables are in fact included in the stochastic exposure model, and certainly play some role in the simulation. However, it appears their effects are not large enough 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 modeled 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 one 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),

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

The model identifies air conditioning in the home, temperature, employment status, and residential AER as important predictors with non-standard effects. These variables are all included in the stick-breaking component of the model with probability near one. To illustrate how the exposure distribution depends on each of these variables, Figure 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 using formulas such as those in (9). The moments for the binary variables air conditioning and employment status are calculated only at the endpoints and, for comparative purposes, all covariates are scaled to the unit intervals.

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

Employment status affects both the mean and variance of the outcome; the sample mean (standard deviation) of the log exposure ratio is -1.26 (0.47) for employed people and -1.22 (0.51) for unemployed people. Figure 4 shows that employment status also affects the shape of the distribution. Both densities in Figure 4 are skewed left, but the unemployed 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 a lot of 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 EPA. It is believed that 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 one, but Figure 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 AER). These effects are added to the linear mean model only, thus the kernel stick-breaking model is 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 AER interacts in the linear

mean with air conditioning and employment status; residential AER has less effect for homes with air conditioning because air conditioning prevents ozone from penetrating, and residential AER 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. Figure 5a 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 degrees (F). People without air conditioning are most likely to open their windows in this temperature window creating the greatest contrast between the two groups. The effect on air conditioning is smaller for very high temperature because people without air conditioning may close both windows and blinds to stay cool. The kernel stick-breaking model identifies this complicated relationship between exposure, temperature, and air conditioning. Similarly, Figure 5b shows that the effect of residential air exchange rates is the largest when the temperature is between 70 and 80 degrees (F). Also, the mean response is between -1.25 and -1.28 for all combinations of AC and employment, except for the group without air conditioning or employment which has mean response -0.98. Therefore, it appears air conditioning only has a protective effect 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 AER and the indicator of age less than 4 are included in the linear mean term, providing evidence of a protective effect for young children. This may be due to the fact that the difference between air quality in homes and child care facilities is only be substantial after accounting for both residential air exchange rate and air conditioning.

To inspect the fit of the final model we use five-fold cross-validation. We fit the final model described above and linear regression model with same variables included in the mean, but Gaussian residuals, independent over the covariates. Figure 6a plots the posterior predictive means versus 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 (PIT) diagnostic discussed in Gneiting, Balabdaoui, and Raftery (2007). For each observation we compute  $PIT_i = P(y^* < y_i)$ , where  $y_i$  is the observation and  $y_i^*$  follows the posterior predictive distribution. Assuming the model is correct, the  $PIT_i$  should approximately follow a Uniform(0,1). Figure 6b 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 using stochastic search variable selection to identify a subset of covariates that have more than just additive effects on the mean of 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 nonparametric 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 is discussed for deterministic models in Bayarri et al. (2007). A simultaneous model for stochastic APEX simulator and field data would be to assume the two data sources 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 data sources.

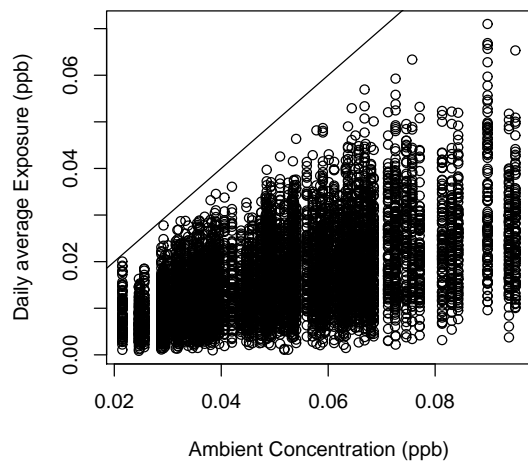
## References

- Bayarri MH, Berger JO, Paulo R, Sacks J, Cafeo JA, Cavendish J, Lin CH, Tu J (2007). A Framework for Validation of Computer Models. *Technometrics*, **49**, 138–154.
- Binkowski FS, Roselle SJ (2003). Models-3 community multiscale air quality (CMAQ) model aerosol component, 1. Model description. *Journal of Geophysical Research*, **108**, 4183.
- Calder CA, Holloman CH, Bortnik SM, Strauss WJ, Morara M (2008). Relating Ambient Particulate Matter Concentration Levels to Mortality Using an Exposure Simulator. *Journal of the American Statistical Association*, **103**, 137–148.
- Chen VCP, Tsui KL, Barton RR, and Meckesheimer M (2006). A review on design, modeling and applications of computer experiments. *IIE Transactions*, **38**, 273-291.
- Chung Y, Dunson DB (2009). Nonparametric Bayes Conditional Distribution Modeling with Variable Selection. In press, *Journal of the American Statistical Association*.
- Dominici F, Daniels M, Zeger SL, Samet JM (2002). Air pollution and mortality: estimating regional and national dose response relationships. *Journal of the American Statistical Association*, **97**, 100–111.
- Dunson DB, Park JH (2008). Kernel stick-breaking processes. *Biometrika*, **95**, 307–323.
- Efroymson RE, Murphy DL (2001). Ecological risk assessment of multimedia hazardous air pollutants: estimating exposure and effects. *Sci Total Environ*, **274**, 219–230.
- Fang KT, Li R, Sudjianto A (2006). Design and modeling for computer experiments. *Chapman & Hall/CRC*.

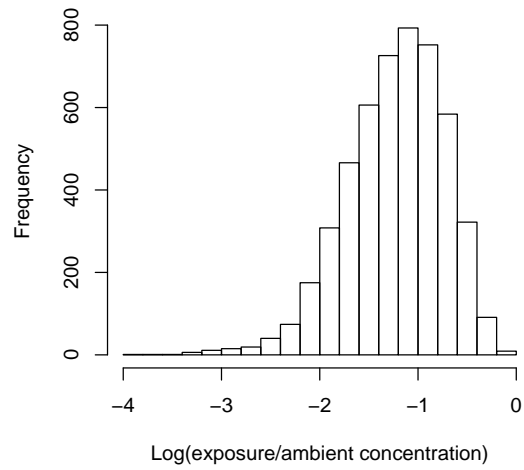
- Ferguson TS (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.
- Ferguson TS (1974). Prior distribution on spaces of probability measures. *The Annals of Statistics*, **2**, 615–629.
- Fuentes M, Song H, Ghosh SK, Holland DM, Davis JM (2006). Spatial association between speciated fine particles and mortality. *Biometrics*, **62**, 855–863.
- Galli E, Cuéllar L, Eidenbenz S, Ewers M, Mniszewski S, Teuscher C (2009). ActivitySim: large-scale agent-based activity generation for infrastructure simulation. SpringSim '09: Proceedings of the 2009 Spring Simulation Multiconference. Society for Computer Simulation International, San Diego, California, 1–9.
- George EI, McCulloch RE (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.
- George EI, McCulloch RE (1997). Approaches for Bayesian variable selection, *Statistica Sinica*, **7**, 339–373.
- George EI (2000). The variable selection problem. *Journal of the American Statistical Association*, **95**, 1304–1308.
- Gneiting T, Balabdaoui F, Raftery AE (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, **69**, 243–268.
- Gustafson P (2000). Bayesian regression modeling with interactions and smooth effects. *Journal of the American Statistical Association*, **95**, 745–763.
- Hastie T, Tibshirani R (1990). Generalized additive models. *Chapman & Hall*.
- Holloman CH, Bortnik S, Morara M, Strauss W, Calder C (2004). A Bayesian hierarchical approach for relating PM<sub>2.5</sub> exposure to cardiovascular mortality in North Carolina. *Environmental Health Perspectives*, **112**, 1282–1288.
- Iooss B, Ribatet M, Marrel A. Global sensitivity analysis of stochastic computer models with generalized additive models. Available at URL: <http://fr.arxiv.org/abs/0802.0443v1>.
- Lee D, 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, Ye KQ (2006). Variable Selection for Gaussian Process Models in Computer Experiments. *Technometrics*, **48**, 478–490.
- Papaspiliopoulos O, Roberts G (2008). Retrospective MCMC for Dirichlet process hierarchical models. *Biometrika*, **95**, 169–186.
- Pope CA, Dockery D, Schwartz J (1995). Review of epidemiological evidence of health effects of particulate air pollution, *Inhalation Toxicology*, **47**, 1–18.
- R Development Core Team (2006). R: A Language and Environment for Statistical Computing. <http://www.R-project.org>.
- Reich BJ, Fuentes M, Burke J (2009). Analysis of the effects of ultrafine particulate matter while accounting for human exposure. *Environmetrics*, **20**, 131–146.
- Reich BJ, Fuentes M (2007). A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *The Annals of Applied Statistics*, **1**, 249–264.

- Reich BJ, Storlie CS, Bondell HD (2008). Bayesian variable selection for nonparametric regression: Application to deterministic computer codes. *Technometrics*, **51**, 110–120.
- Sacks J, Welch WL, Mitchell TJ, Wynn HP (1989). Design and analysis of computer experiments. *Statistical Science*, **4**, 409–435.
- Sethuraman J (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, **4**, 639–650.
- Shaddick G, Lee D, Zidek JV, Salway R (2008). Estimating exposure response functions using ambient pollution concentrations. *Annals of Applied Statistics*, **4**, 1249–1270.
- Shively T, Kohn R, Wood S (1999). Variable selection and function estimation in nonparametric regression using a data-based prior (with discussion). *Journal of the American Statistical Association*, **94**, 777–806.
- Schwartz J (1994). Air pollution and daily mortality: a review and meta analysis. *Environmental Research*, **64**, 36–52.
- United States Environmental Protection Agency (2007). Ozone population exposure analysis for selected urban areas. Publication number 452R07010.
- Waupotitsch R, Eidenbenz S, Smith JP, Kroc L (2006). Multi-scale integrated information and telecommunications system (MIITS): first results from a large-scale end-to-end network simulator. WSC '06: Proceedings of the 38th conference on Winter simulation. Winter Simulation Conference, Monterey, California, 2132–2139.
- Wood S, Kohn R, Shively T, Jiang W (2002). Model selection in spline nonparametric regression. *Journal of the Royal Statistical Society: Series B*, **64**, 119–140.

Figure 1: Plots of the APEX data. Panel (a) plots the average daily exposure against the daily average ambient concentration, Panel (b) plots the log ratio of exposure to ambient concentration.



(a)



(b)

Table 1: Mean (sd) of the posterior inclusion probabilities for the simulation study. For the linear mean parameters we report the mean (sd) of  $P(\beta_j \neq 0)$  and for the kernel stick-breaking parameters we report the mean (sd) 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.

(a) Design 1: Gaussian linear model

Var	Linear mean parameters, $\beta_j$			Kernel bandwidths, $\rho_j$
	Model 1	Model 2	Model 3	Model 3
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: Heteroskedastic model

Var	Linear mean parameters, $\beta_j$			Kernel bandwidths, $\rho_j$
	Model 1	Model 2	Model 3	Model 3
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

Var	Linear mean parameters, $\beta_j$			Kernel bandwidths, $\rho_j$
	Model 1	Model 2	Model 3	Model 3
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

Var	Linear mean parameters, $\beta_j$			Kernel bandwidths, $\rho_j$
	Model 1	Model 2	Model 3	Model 3
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

Var	Linear mean parameters, $\beta_j$			Kernel bandwidths, $\rho_j$
	Model 1	Model 2	Model 3	Model 3
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)

Table 2: Posterior summaries for the main effects model. “Overall in. prob.” is the probability that the variable is included in the mean or residual model component (i.e.,  $\beta_j \neq 0$  or  $\rho < 1$ ).

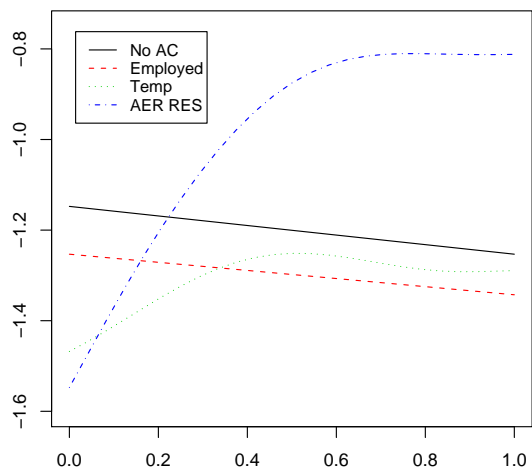
Parameter	Overall in. prob.	Linear term, $\beta_j$		Kernel bandwidths
		$P(\beta_j \neq 0)$	95% Interval	$P(\rho < 1)$
Weekend	1.00	1.00	( 0.06, 0.10)	0.05
Physical activity index	1.00	1.00	(0.30, 0.50)	0.22
Temperature	1.00	0.00	( 0.00, 0.00)	1.00
Ambient Concentration	1.00	1.00	(-0.20, -0.10)	0.15
No AC 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 $\leq 4$	0.37	0.32	(-0.11, 0.00)	0.06
Age 5 – 18	0.12	0.00	( 0.00, 0.00)	0.12
Age $\geq 65$	0.04	0.00	( 0.00, 0.00)	0.04
AER Residence	1.00	1.00	( 0.47, 0.77)	1.00
AER Bar/Rest	0.08	0.00	( 0.00, 0.00)	0.08
AER School $\times$ Age 5 – 18	1.00	1.00	( 0.17, 0.34)	0.06
AER child care $\times$ Age $\leq 4$	0.06	0.01	( 0.00, 0.00)	0.05
AER Office $\times$ 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

Table 3: Posterior summaries for the model with interactions. “Overall in. prob.” is the probability that the variable is included in the mean or residual model component (i.e.,  $\beta_j \neq 0$  or  $\rho < 1$ ).

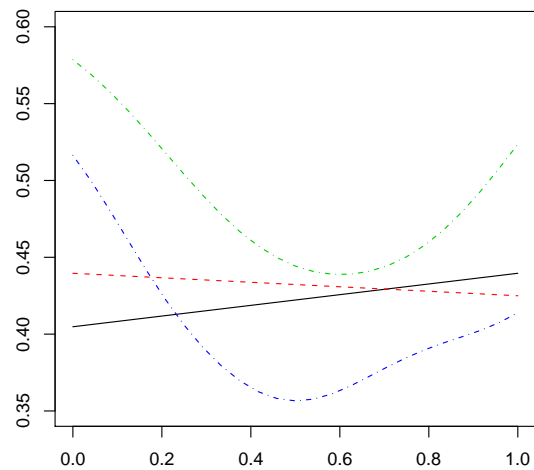
Parameter	Overall in. prob.	Linear term, $\beta_j$		Kernel bandwidths
		$P(\beta_j \neq 0)$	95% Interval	$P(\rho < 1)$
Weekend	1.00	1.00	( 0.06, 0.11)	0.05
Physical activity index	1.00	1.00	( 0.29, 0.48)	0.13
Temperature	1.00	0.30	(-0.33, 0.00)	1.00
Ambient Concentration	1.00	1.00	(-0.21, -0.12)	0.25
No AC 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 $\leq 4$	0.63	0.61	(-0.25, 0.00)	0.03
Age 5 – 18	0.04	0.01	( 0.00, 0.00)	0.03
Age $\geq 65$	0.31	0.01	( 0.00, 0.00)	0.31
AER Residence	1.00	1.00	( 0.63, 0.99)	1.00
AER Bar/Rest	0.09	0.04	(-0.04, 0.00)	0.05
AER School $\times$ Age 5 – 18	1.00	1.00	( 0.20, 0.37)	0.05
AER child care $\times$ Age $\leq 4$	0.61	0.59	(0.00, 0.30)	0.03
AER Office $\times$ 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
Temp $\times$ no AC	0.03	0.03	( 0.00, 0.00)	0.00
Temp $\times$ Employ	0.49	0.49	(-0.19, 0.00)	0.00
Temp $\times$ AER Res	0.30	0.30	(0.00, 0.72)	0.00
no AC $\times$ Employ	0.01	0.01	( 0.00, 0.00)	0.00
no AC $\times$ AER Res	1.00	1.00	(-0.59, -0.26)	0.00
Employ $\times$ AER Res	0.57	0.57	(-0.26, 0.00)	0.00

Figure 2: Estimated moments of the log exposure ratio by four important covariates.

(a) Mean



(b) Standard deviation



(c) Skewness

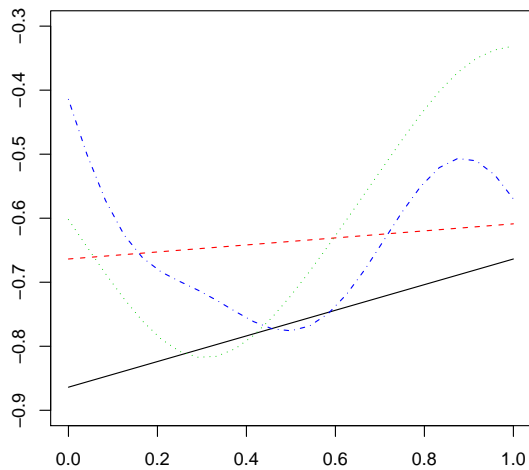


Figure 3: Raw data and posterior mean density of log exposure ratio by residential AER. In Panel (b) several densities are plotted for residual AER varying from 0.42 to 5.56, with line thickness representing the residual AER.

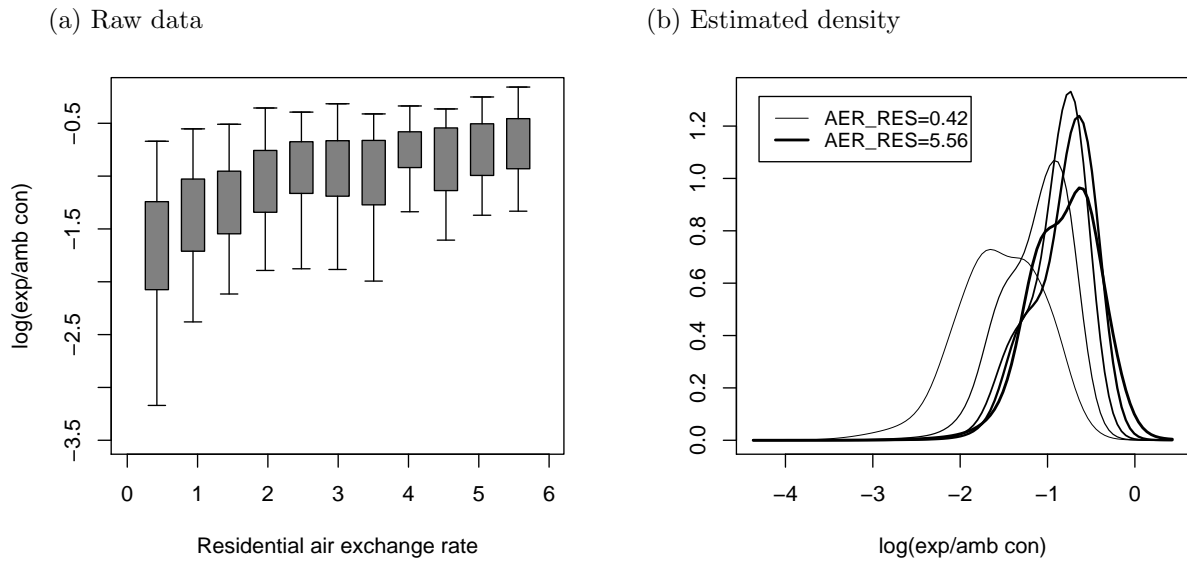


Figure 4: Data versus posterior mean conditional density for exposure for employed and unemployed people.

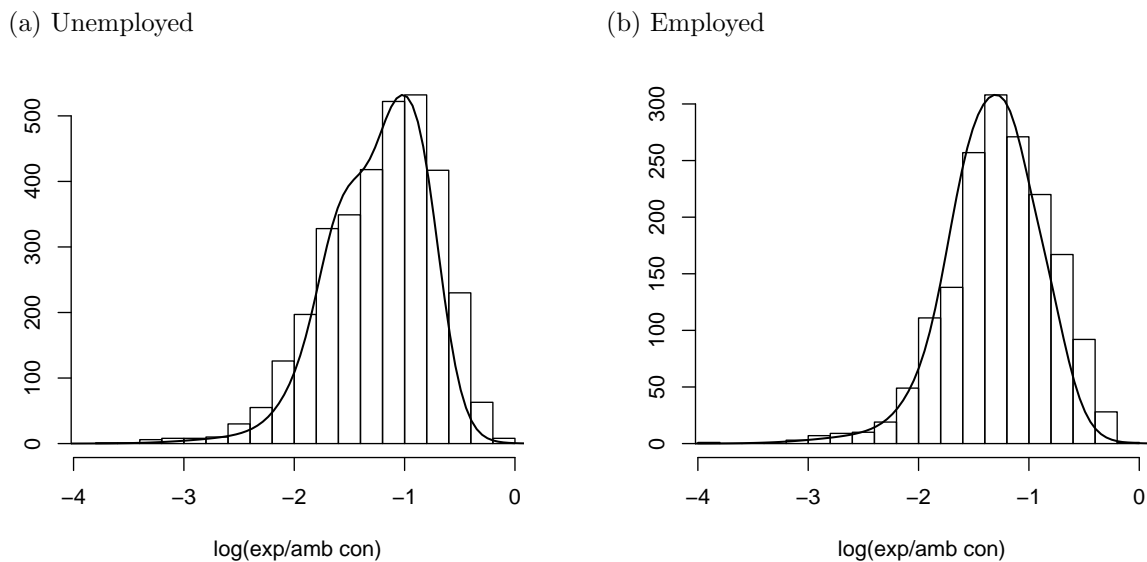


Figure 5: Plot of the raw data by temperature and air conditioning (shaded boxes are no air conditioning, unshaded boxes are air conditioning) and temperature and residential air exchange rate (shaded boxes are residential AER above the median residential AER).

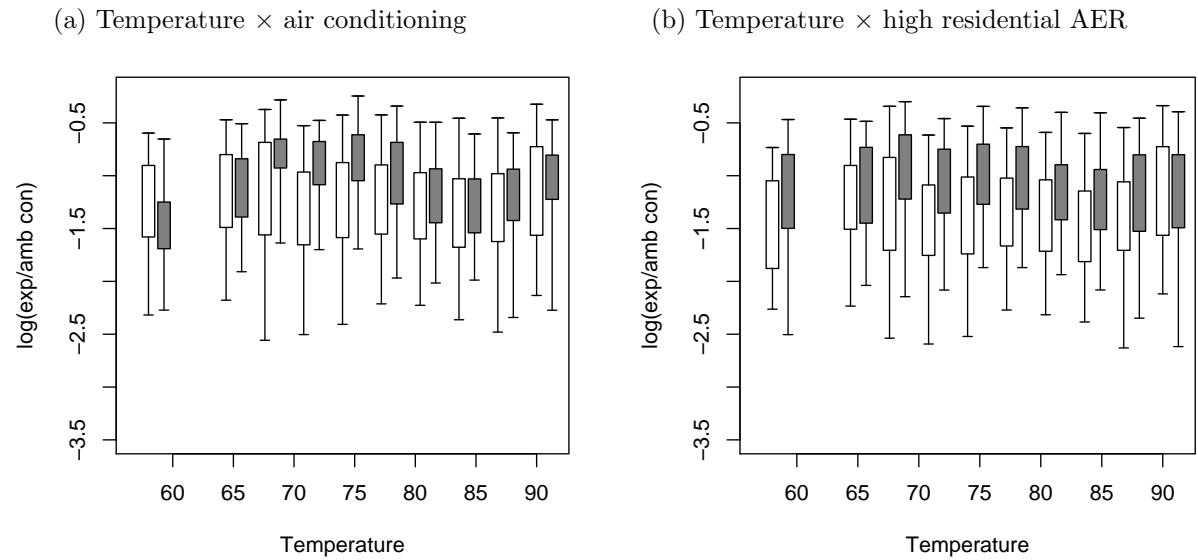


Figure 6: Summary of cross-validation. Panel (a) plots the testing data versus the posterior predictive mean, and Panel (b) plots the PIT statistics for the Gaussian and non-Gaussian models.

