Atmospheric Environment 77 (2013) 464–473

Contents lists available at SciVerse ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Addressing extrema and censoring in pollutant and exposure data using mixture of normal distributions☆

Shi Li [a], Stuart Batterman [b,*], Feng-Chiao Su [b], Bhramar Mukherjee [a]

[a] Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, SPH II, Ann Arbor, MI 48109, USA
[b] Department of Environmental Health Sciences, School of Public Health, University of Michigan, 1420 Washington Heights, SPH II, Ann Arbor, MI 48109, USA

HIGHLIGHTS

- Fitting distributions of volatile organic compound concentrations.
- Finite mixture of normals and Dirichlet process mixture of normals.
- Superior performance compared to the traditional single normal distribution.
- Robustness and ability to characterize uncertainty for model parameters.
- Implemented via Relationship between Indoor, Outdoor and Personal Air study.

ARTICLE INFO

Article history:
Received 20 November 2012
Received in revised form
30 April 2013
Accepted 3 May 2013

Keywords:
Air pollution
Density estimation
Dirichlet process mixture
Limits of detection
Mixture of normal
Volatile organic compounds

ABSTRACT

Background: Volatile organic compounds (VOC), which include many hazardous chemicals, have been used extensively in personal, commercial and industrial products. Due to the variation in source emissions, differences in the settings and environmental conditions where exposures occur, and measurement issues, distributions of VOC concentrations can have multiple modes, heavy tails, and significant portions of data below the method detection limit (MDL). These issues challenge standard parametric distribution models needed to estimate the exposures, even after log transformation of the data.
Methods: This paper considers mixture of distributions that can be directly applied to concentration and exposure data. Two types of mixture distributions are considered: the traditional finite mixture of normal distributions, and a semi-parametric Dirichlet process mixture (DPM) of normal distributions. Both methods are implemented for a sample data set obtained from the Relationship between Indoor, Outdoor and Personal Air (RIOPA) study. Performance is assessed based on goodness-of-fit criteria that compare the closeness of the density estimates with the empirical density based on data. The goodness-of-fit for the proposed density estimation methods are evaluated by a comprehensive simulation study.
Results: The finite mixture of normals and DPM of normals have superior performance when compared to the single normal distribution fitted to log-transformed exposure data. The advantages of using these mixture distributions are more pronounced when exposure data have heavy tails or a large fraction of data below the MDL. Distributions from the DPM provided slightly better fits than the finite mixture of normals. Additionally, the DPM method avoids certain convergence issues associated with the finite mixture of normals, and adaptively selects the number of components.
Conclusions: Compared to the finite mixture of normals, DPM of normals has advantages by characterizing uncertainty around the number of components, and by providing a formal assessment of

Abbreviations: VOC, volatile organic compounds; MDL, method detection limit; DPM, Dirichlet process mixture; RIOPA study, Relationship between Indoor, Outdoor and Personal Air study; GEV, generalized extreme value; EM, expectation maximization; MLE, maximum likelihood estimation; AIC, Akaike information criterion; BIC, Bayesian information criterion; CDF, cumulative distribution function; MSE, mean squared error; MAE, mean absolute error; NHANES, National Health and Nutrition Examination Survey.

* Corresponding author. Department of Environmental Health Sciences, School of Public Health, University of Michigan, Room 6075, SPH2, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA. Tel.: +1 734 763 2417; fax: +1 734 763 8095.
E-mail address: stuartb@umich.edu (S. Batterman).

1352-2310/$ – see front matter
http://dx.doi.org/10.1016/j.atmosenv.2013.05.004

uncertainty for all model parameters through the posterior distribution. The method adapts to a spectrum of departures from standard model assumptions and provides robust estimates of the exposure density even under censoring due to MDL.

1. Introduction

Volatile organic compounds (VOCs) have been used extensively in personal, commercial and industrial products (MDE, 2010; Ling et al., 2011; Weschler, 2011; USEPA, 2012b), and these chemicals are widely found in air in indoor, outdoor and occupational settings. Many VOCs are hazardous, and exposure through inhalation has been associated with a variety of acute and chronic health effects, such as respiratory disease and cancer (Kim and Bernstein, 2009; USEPA, 2012a,b). While concentrations of VOCs in environmental settings are generally much lower than those in occupational settings (Rappaport and Kupper, 2004), moderate and sometimes high concentrations and exposures can be encountered among the general population during certain activities, such as filling vehicles with gasoline and home renovations, in hobbies such as furniture restoration, small engine repair and gun cleaning, and using cleaners, pesticides, pest repellants and air fresheners in poorly ventilated spaces (Batterman et al., 2006; Jia et al., 2008a; D'Souza et al., 2009; Jia and Batterman, 2010; USEPA, 2012b).

The high concentrations found for a portion of the population, along with the much lower concentrations for the bulk of the population, typically results in highly right skewed concentration distributions (Jia et al., 2008b). Extreme value theory and other techniques can model the upper percentiles of VOC concentration distributions, and generalized extreme value (GEV) distributions have been shown to fit VOC data much more closely than lognormal or other types of distributions (Jia et al., 2008b; Batterman et al., 2011; Su et al., 2012). Most data sets also contain many low observations, often including measurements that fall below the method detection limit (MDL). These "non-detects," which represent left-censored data, can be treated by substitution, single or multiple imputation, regression on order statistics (modeling using probability plots of known distributions to estimate summary statistics), and laboratory-generated data (using the original data without replacement) (Antweiler and Taylor, 2008). The extent of data below MDLs can significantly affect the quality of the results (Lubin et al., 2004; Antweiler and Taylor, 2008). The statistical issues associated with the analysis of data with MDL issues are well-known (Taylor et al., 2001; Krishnamoorthy et al., 2009).

Due to the variation in source emissions, differences in the settings and environmental factors where exposures occur, and the measurement issues just noted, distributions of VOC concentrations can have multiple modes, heavy tails, and significant portions of data falling below the MDL that are replaced by a single value. These issues, which can be encountered in exposure and as well as other types of data sets, challenge standard parametric distribution models. While GEV distributions can fit the upper portions of distributions, they do not represent the full distribution of the data. Information on the full distributions of exposure levels is needed to establish exposure/risk guidelines, to estimate health risks and uncertainty estimates across a population (Su et al., 2012), and to facilitate probabilistic analyses (Hammonds et al., 1994).

Mixture distributions, which extend parametric families of distributions to fit datasets that are not adequately fit by a single common distribution, provide a flexible and powerful approach of representing the distribution of a random variable (Titterington et al., 1985; McLachlan and Basford, 1988; McLachlan and Peel, 2000). As examples, a finite mixture of normals applies a set of 'mixing weights' to a specified and finite number of component distributions, while a nonparametric Dirichlet process mixture (DPM) of normals relaxes the need to pre-specify the number of component distributions and is potentially advantageous in terms of handling smoothing, modality and uncertainty (Escobar, 1994; Mueller and Quintana, 2004). Mixture of normals have been extensively used in a variety of important and practical situations, although environmental applications have been very limited (Burmaster and Wilson, 2000; Razzaghi and Kodell, 2000; Taylor et al., 2001; Chu et al., 2005).

This paper evaluates the applicability of mixture of normal distribution method to environmental data, specifically, air pollution concentration and exposure data. Both the traditional finite mixture of normal and the nonparametric DPM of normals are evaluated using a VOC exposure dataset that includes seasonal measurements for approximately 300 individuals, which was collected as part of the Relationship between Indoor, Outdoor and Personal Air (RIOPA) study. Goodness-of-fit for the density estimation methods are evaluated by a comprehensive simulation study.

2. Materials and methods

2.1. VOC measurements

The RIOPA study was designed to evaluate contributions of outdoor and indoor sources to personal exposures of air pollutants, including VOCs and $PM_{2.5}$, among residents of three cities (Elizabeth, NJ, Houston, TX and Los Angeles, CA) selected to reflect potential differences in emissions and other factors likely to influence exposures (Weisel et al., 2005a). Sampling was conducted in two seasons for approximately 100 adults (and a smaller number of children) in each city from summer 1999 through spring 2001. Indoor, outdoor and personal (worn by participants) measurements were obtained using passive samplers for 48 h periods, and 18 VOCs were measured using gas chromatography and mass spectrometry. Analytical work was performed by two laboratories. The RIOPA study represents one of the larger VOC studies in the USA that collected personal samples, which are generally considered to provide exposure estimates that are more accurate than indoor or outdoor samples.

Three VOCs (chloroform, 1,4-dichlorobenzene (1,4-DCB) and styrene) were selected to evaluate mixture distributions. These VOCs differ in terms of their distributions, detection frequencies and other properties. Personal samples for adults were selected, primarily because the sample size for the adult cohort ($n = 544$ for each VOC) was largest, and because the personal samples should best reflect exposure. The two laboratories used to analyze samples had different MDLs. Since the use of two laboratories is somewhat unusual, all data under MDLs were replaced with a single value using $0.5 \times$ the higher MDL. Because the VOC data in RIOPA had many extreme values (Su et al., 2012), the density estimation methods were implemented using logarithms of the concentration value, as described next.

2.2. Finite mixture of normal distributions

Finite mixture distributions are commonly used to identify and model sub-populations within an overall population. Rather than

identifying the sub-population that an individual observation belongs to, these models assume that the observed data randomly arise from distributions with certain probabilities. Let $Y = (Y_1,\ldots,Y_n)$ be a random sample of size n from the overall population with the probability density function of Y_i given as $f(y_i)$. Y is assumed to have arisen from a mixture of an initially specified number of distributions. A K-component mixture of distributions supposes that the density of Y_i can be written as

$$f(y_i) = \sum_{k=1}^{K} \lambda_k f_k(y_i), \quad (1)$$

where f_k is the component density of the k-th cluster, and λ_k is the corresponding weight with the constraint that $0 \le \lambda_k \le 1$and $\sum_{k=1}^{K}\lambda_k = 1$. In many applications, component densities f_k are assumed to be standard parametric families, such as normal distribution $N(\mu_k, \sigma_k^2)$, then

$$f(y_i) = \sum_{k=1}^{K} \lambda_k N\left(\mu_k, \sigma_k^2\right). \quad (2)$$

The finite mixture of normals represented by Eq. (2) is a potential choice for handling concentration and exposure data that can have multiple modes and heavy tails. Such normal mixtures are popular choices with attractive properties (Titterington et al., 1985). Since the mixtures are constructed as a linear combination of normal distributions, they are computationally and analytically tractable, well behaved in the limiting case, and scalable to higher dimensions.

Mixture distributions can be fitted using many techniques, e.g., graphical methods, the method of moments, maximum likelihood estimation (MLE) and Bayesian approaches (Redner and Walker, 1984; Titterington et al., 1985; McLachlan and Peel, 2000). Since closed forms of MLEs of Eq. (1) are not available, mixture distributions are commonly fitted using expectation maximization (EM) type algorithms (Dempster et al., 1977; Meng and Pedlow, 1992; McLachlan and Krishnan, 1997). We used the EM algorithm and considered a constrained maximum likelihood method to estimate Eq. (2) with a further constraint that the location of the first cluster (generally the lowest) is under the MDL, i.e., $\mu_1 \le$ MDL. This constraint ensures that a fitted cluster covers the MDL, which allows it to be interpreted as the sub-population of the data below the MDL.

An important issue in fitting finite mixture distributions is selection of the number of components K. Criteria based on penalized likelihood, such as the Akaike information criterion (AIC), have been applied successfully to mixture distributions (McLachlan and Peel, 2000). While this criterion generally favors larger K, there is considerable practical support for its use due to simplicity (Fraley and Raftery, 1998). The Bayesian information criterion (BIC) appears attractive due to their statistical properties as well as the simplicity of implementation. Though the BIC always leads to a smaller (or equal) number of components than AIC, the BIC can also lead to an overestimate of the number of clusters regardless the clusters' separation (Biernacki et al., 2000). In general, with limited amount of data, a corrected version of AIC such as AICc (Hurvich and Tsai, 1989) may be preferable. For these finite mixture distributions, we fitted model (2) with $K = 2$ to 5 clusters, and selected the optimal model based on AICc. This analysis was conducted for each of the three VOCs.

As a benchmark for comparison, we also fitted the traditional normal distribution, which is a special case of mixture of normals with $K = 1$. (As noted earlier, log-transformed VOC data were used in all cases.)

The finite mixture of normals was implemented using the 'mixtools' package (Benaglia et al., 2009) in R (R Foundation for Statistical Computing, Vienna, Austria). This package fits the finite mixture of normals using EM algorithms through the function *normalmixEM*.

2.3. Dirichlet process mixture of normal distributions

Bayesian density estimation methods using DPM of normal densities have several practical advantages, including optimally trading off local versus global smoothing, assessing modality, and propagating uncertainty on inferences regarding the number of components and thus uncertainty about the density estimate (Ferguson, 1983; Escobar, 1994; Mueller and Quintana, 2004). Instead of pre-specifying the number of clusters, these models allow the number of clusters to be chosen in a data-adaptive way. Let $Y_i \sim N(\mu_i, \sigma_i^2)$ and let $(\mu_i, \sigma_i^2) = \theta_i$. The DPM of normal distributions assumes that these normal parameters θ_i follow a random distribution G generated from Dirichlet process (Ferguson, 1973), which can be represented as:

$$\theta_i | G \sim G \text{ i.i.d and } G|\alpha,\ \ G_0 \sim \mathrm{DP}(\alpha G_0). \quad (3)$$

DP(α G_0) is a Dirichlet process with concentration parameter α and base distribution G_0, which is also known as the prior expectation of G. The precision parameter α determines the concentration of the prior for G around G_0. Blackwell and Macqueen provided the following representation for the leave-one-out conditional distributions (Blackwell and MacQueen, 1973):

$$\theta_i \Big| \theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_n \sim \frac{\alpha}{n-1+\alpha} G_0 + \frac{1}{n-1+\alpha} \sum_{j \ne i}^{n} I_{\theta_j}(\cdot) \quad (4)$$

In this approach, $\theta = (\theta_1,\ldots,\theta_n)$ will be reduced to certain K distinct values ($K < n$) with positive probability. From Eq. (4), two well-known extreme cases of the DPM can be derived. As $\alpha \to \infty$, the DPM reduces to a parametric model, namely $\theta_i \sim G_0$ independent and identically distributed (n clusters), whereas $\alpha \to 0$ implies a common parametric model, namely $\theta_1 = \cdots = \theta_n = \theta^*$ with $\theta^* \sim G_0$ (1 cluster). The baseline distribution G_0 is chosen to be the conjugate normal-inverse-gamma distribution. Hyperpriors could be used on this normal-inverse-gamma distribution to complete the model specification.

The DPM of normals does not require specification of the number of clusters as needed for parametric mixture distributions, such as the finite mixture of normals discussed previously. In practice, suitable values of K will typically be small relative to the sample size n. The implicit prior distribution on K is stochastically increasing with α and is related to the prior distribution on α (Antoniak, 1974). For moderately large n, $E(K|\alpha,n) \approx \alpha \log(1 + n/\alpha)$ (Antoniak, 1974). A formal assessment of uncertainty regarding the number of components K can be obtained through generated draws from the posterior distribution of K as a part of the Bayesian computation scheme.

For the VOC data, the precision parameter α was chosen to follow a Gamma prior distribution, and a sensitivity analysis was conducted with respect to choice of the Gamma parameters. Given the sample size in the test dataset ($n = 544$), for prior information, $\alpha \sim$ Gamma(0.3,0.4) favors $K = 1$–3 clusters; $\alpha \sim$ Gamma(1.2,2.5) favors 1–5 clusters; $\alpha \sim$ Gamma(2,1.5) favors 2–10 clusters; and $\alpha \sim$ Gamma(5,2) favors 5–20 clusters. A sensitivity analysis was conducted on these prior specifications.

Computational methods were followed that allowed the evaluation of posterior distributions for all model parameters and the number of components, and also the resulting predictive distributions (Escobar and West, 1995). Density estimation using DPM can

be directly implemented using the DP package (Jara, 2007; Jara et al., 2011) in R (R Foundation for Statistical Computing, Vienna, Austria), which provides posterior draws of all model parameters under a DPM using Markov chain Monte Carlo methods. A sample code is presented in the Appendix.

2.4. Goodness-of-fit criteria

Goodness-of-fit for the density estimation methods was determined by comparing the estimated cumulative distribution function (CDF) $\widehat{F}_{\text{est}}$ to the empirical CDF $\widehat{F}_{\text{emp}}$ based on the observed data. Although all observed/generated data were used to estimate the CDF by each method, goodness-of-fit was evaluated using only the data above the MDL. Both the mean squared error ($\text{MSE} = \sum_{i,y_i>\text{MDL}}[\widehat{F}_{\text{emp}}(y_i) - \widehat{F}_{\text{est}}(y_i)]^2 / \sum_i I(y_i > \text{MDL})$), and the mean absolute error ($\text{MAE} = \sum_{i,y_i>\text{MDL}}|\widehat{F}_{\text{emp}}(y_i) - \widehat{F}_{\text{est}}(y_i)| / \sum_i I(y_i > \text{MDL})$) were considered. The estimated proportion of observations above the MDL, which is often termed the detection frequency, for empirical and estimated distributions was compared.

2.5. Simulation study

For further evaluation of the mixture distributions, several forms of underlying true distributions and varying amounts of left-censored data (below MDL) were considered as true generation models. Three methods were compared: a single normal distribution; a finite mixture of normals; and DPM of normals. Two underlying distributions with features similar to the three VOC samples from the RIOPA study were selected: a normal $N(0, 2^2)$ and a mixture specified as 1/2 Gamma(3,1.5) + 1/2 Uniform(−3,8). The former is symmetric and the latter is right skewed with heavy tails, and both have multiple modes when data under MDL were replaced by 0.5 MDL. The proportion of data below the MDL, P_0, was set to 15%, 30% and 50% in separate simulations. Goodness-of-fit measures (MSE and MAE described above) were calculated for each method, target distribution, and choice of P_0. A dataset of size $n = 1000$ was generated for each simulation under each setting. The average values of MSE and MAE across 500 simulations are reported.

For the finite mixture of normals, the number of components K was based on the smallest AICc. A convergence problem was encountered when P_0 was high (in the range of 30–50%), possibly because the censored data were set to a single value (0.5 MDL), which resulted in a very small variance of the first (lowest) cluster. Additionally, the MLE method for finite mixture models is susceptible to other problems, e.g., nonunique solutions (Redner and Walker, 1984; Titterington et al., 1985; McLachlan and Peel, 2000). Thus, data below the MDL was replaced by uniformly generated pseudo-data from Uniform (0, MDL) if the finite mixture of normals did not converge. In contrast, all of the single normal and DPM method simulations converged.

3. Results

3.1. Descriptive analysis

The distributions of the sample data sets are shown as histograms in panel A of Figs. 1–3 for chloroform, 1,4-DCB and styrene, respectively. (Fitted density plots for the three density estimation methods are also shown in these figures.) These VOCs show distinct features. For chloroform, 17% of observations fell below the MDL, and observations exceeding the MDL were approximately lognormally distributed. For 1,4-DCB, a larger amount of data, 34%, was below the MDL and the remainder was highly right skewed even after log transformation and contained a number of extreme values. For styrene, most of the data, 66%, fell below the MDL and, like 1,4-DCB, was highly right skewed after log transformation with extreme values. These three VOCs were selected to demonstrate the

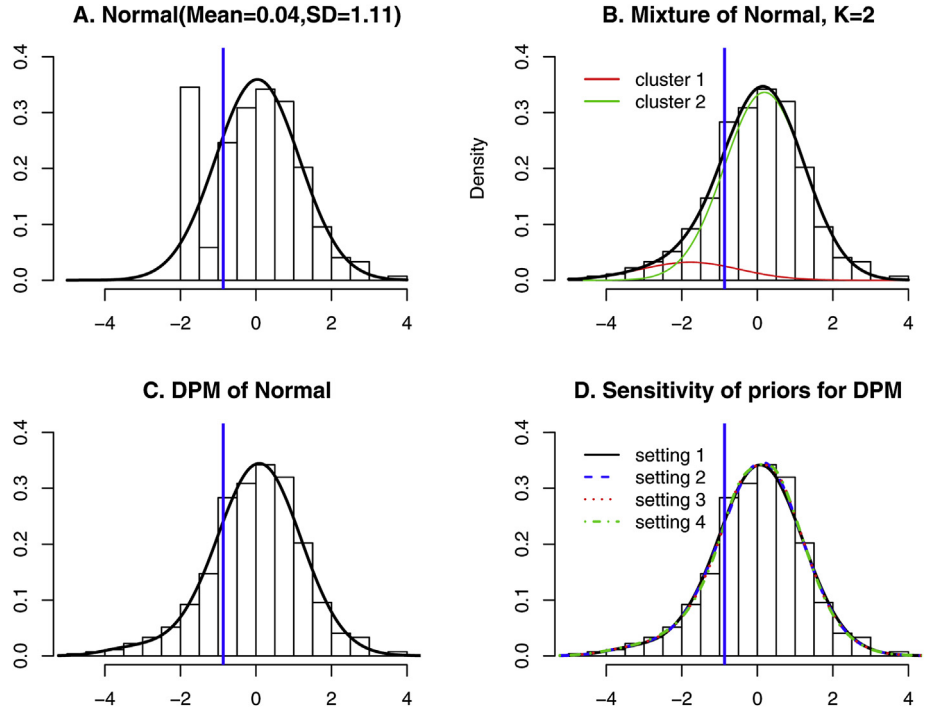

Fig. 1. Fitted density plots for chloroform (log scale) for three models: normal, finite mixture of normals (with the smallest AICc and components indicated), and Dirichlet process mixture (DPM) of normals. Vertical line shows MDL. Sensitivity of priors for DPM, settings 1: $\alpha \sim$Gamma(0.3,0.4); settings 2: $\alpha \sim$Gamma(1.2,2.5); settings 3: $\alpha \sim$Gamma(2,1.5); settings 4: $\alpha \sim$Gamma(5,2).

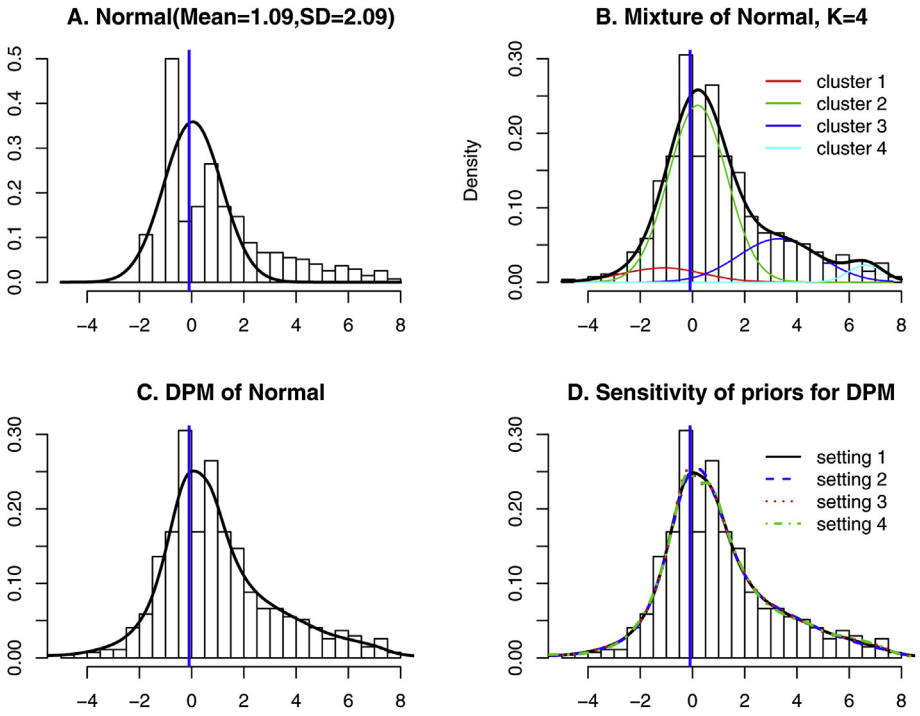

Fig. 2. Fitted density plots for 1,4-DCB (log scale). Otherwise as Fig. 1.

sensitivity and performance of the mixture models across a broad range of levels of censoring.

The selected VOCs differ with respect to sources, exposures, and health effects. Chloroform is considered a probable human carcinogen (causing renal tumors) based on an adequate data of animal studies (USEPA, 2012a). Most inhalation, ingestion and dermal exposures result indoors through contact with chlorinated water, e.g., showering, drinking, and swimming, from which chloroform is released as a by-product of chlorine disinfection (ATSDR, 1997). The median chloroform indoor (0.94 μg m^{-3}) and personal (1.04 μg m^{-3}) concentrations in RIOPA considerably exceeded outdoors levels, which were mostly at the MDL (replaced with 0.21 μg m^{-3}), thus outdoor levels of chloroform provided only a negligible contribution to personal exposure. (MDLs in the two

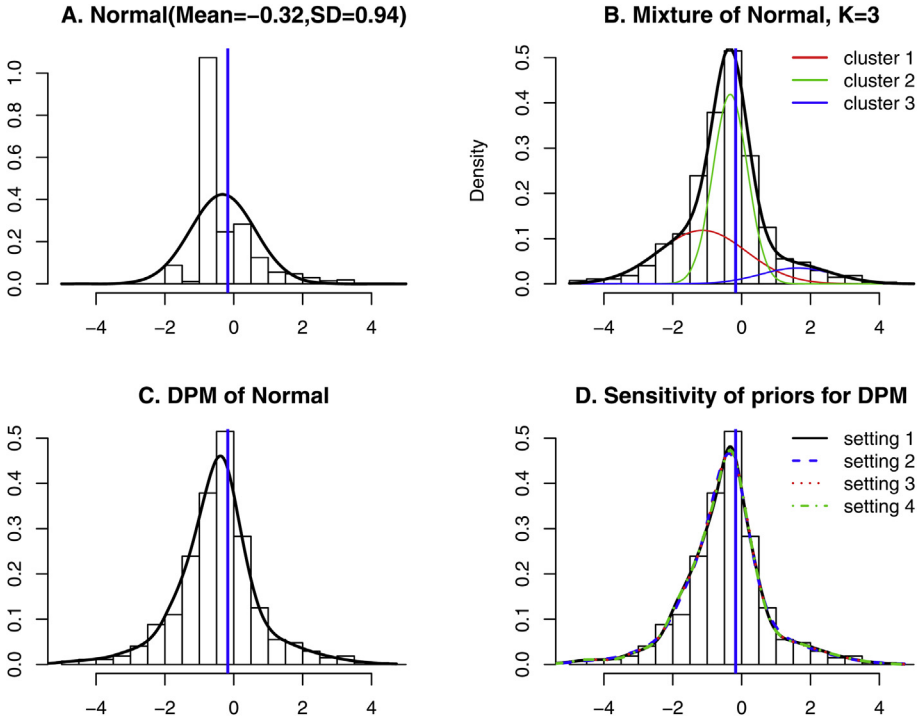

Fig. 3. Fitted density plots for styrene (log scale). Otherwise as Fig. 1.

laboratories used in RIOPA were 0.28 and 0.42 μg m^{-3}.) The unit risk factor (URF) for chloroform in air, 2.3×10^{-5} (μg m^{-3})$^{-1}$, corresponds to a one-in-a-million cancer risk level for long term exposure (USEPA, 2012a). Of the personal samples collected in RIOPA, 9.6% exceeded a cancer risk level of 10^{-4}, and the maximum chloroform level (46.5 μg m^{-3}) corresponds to of 1.1×10^{-3}. Notably, RIOPA participants did not wear their samplers during showering and swimming activities, and thus the highest RIOPA measurements are likely biased downwards from true exposures. The RIOPA measurements may be compared to personal exposure measurements in the National Health and Nutrition Examination Survey (NHANES) 1999–2000, a nationally representative sample conducted about the same time period. NHANES showed a similar median chloroform exposure (1.1 μg m^{-3}) as RIOPA (Jia et al., 2008b; CDC, 2012). However, NHANES had slightly higher concentrations at the upper percentiles than RIOPA, e.g., 16% of individuals had estimated cancer risk exceeding 10^{-4}, and the maximum exposure was 53.9 μg m^{-3}. VOC exposures in RIOPA and NHANES can differ for multiple reasons, including the use of different sampling strategies (convenience sampling in three cities for RIOPA versus national probability-based sampling for NHANES), demographics (RIOPA participants were older and more likely to be females), occupations (RIOPA participants had fewer VOC-related occupations), and employment (fewer employed in RIOPA).

The second VOC considered, 1,4-DCB, is considered a possible carcinogen (causing renal and liver tumors) based on animal studies (IARC, 2012). 1,4-DCB is a component of many commercial products, including moth repellents, deodorants, insecticides and resins (Chin et al., 2012), and inhalation of vapors that have sublimated from these products is a primary exposure route (ATSDR, 2006). Like chloroform, 1,4-DCB exposure is due to predominantly indoor sources. In RIOPA, indoor and personal concentrations (median = 1.40 and 1.88 μg m^{-3}, respectively) were similar; and outdoor levels were typically at the half MDL (0.46 μg m^{-3}). (MDLs in RIOPA were 0.43 and 0.91 μg m^{-3}.) The URF for 1,4-DCB in air is 1.1×10^{-5} (μg m^{-3})$^{-1}$ (OEHHA, 2005), and 23% of personal samples in RIOPA exceeded a 10^{-4} risk. The maximum 1,4-DCB level in RIOPA (2153 μg m^{-3}) corresponds to high risk, 2.4×10^{-2}. Again, the median 1,4-DCB exposure (1.7 μg m^{-3}) in NHANES was close to that in the RIOPA study (Jia et al., 2008b). However, the top percentiles of 1,4-DCB exposure found in NHANES exceeded those in RIOPA, e.g., 30% of NHANES observations exceeded the risk level of 10^{-4}, and the maximum was 2236 μg m^{-3}.

Like 1,4-DCB, styrene is considered as a possible carcinogen (lymphatic, hematopoietic and pulmonary tumors) based on limited human and animal evidence (IARC, 2012). Styrene is widely used in industry as a raw material for plastic and rubber products, and is a component of cigarette smoke, vehicle exhaust and other combustion processes, and thus exposure occurs via inhalation from styrene-contained products in many settings, near certain industrial facilities, in traffic and near smokers (ATSDR, 2010). Because most data fell below the MDL (0.34 and 0.84 μg m^{-3}), personal, indoor and outdoor measurements in RIOPA had the same median (0.42 μg m^{-3}); 75th percentile values were 1.10, 1.07 and 0.42 μg m^{-3}, respectively. The URF for styrene in air is 2.0×10^{-6} (μg m^{-3})$^{-1}$(Caldwell et al., 1998), and 0.2% of personal samples in RIOPA exceeded the risk level of 10^{-4}. The maximum styrene concentration (59.5 μg m^{-3}) gives a risk of 1.2×10^{-4}. RIOPA excluded smokers and smoking households, and participants were predominantly women (75%) (Weisel et al., 2005b), thus, styrene levels may be lower than national norms. Personal styrene exposure was not collected in NHANES 1999–2000.

Table 1 contrasts goodness-of-fit statistics for the three density estimation methods (normal, finite mixture of normals, DPM of normals) and the three VOCs from RIOPA (chloroform, 1,4-DCB and styrene). The table helps to identify situations where a particular method may be advantageous, as discussed below.

3.2. Single normal and GEV distributions

For chloroform, which is roughly lognormally distributed except that 17% of the data is under the MDL, the single normal distribution model fits about as well as the finite mixture of normals and DPM of normals (described below) on the basis of MSE and MAE values, and gives a 21% probability of being below the MDL, similar to that observed (Table 1). However, for 1,4-DCB and styrene, which have more data under the MDL as well as heavy tails, the fit of the single normal distribution model is inferior compared to those of the mixture models. For example, the predicted probability of being below MDL is 28% and 56% for 1,4-DCB and styrene, respectively, compared to 34% and 66% observed, and 33% and 64% estimated by the mixture models. The single normal distribution overestimated the mean of these VOCs since it underestimated the non-detection frequency.

Using the top 5% and 10% of concentrations as extrema, the RIOPA VOC data previously has been fitted to GEV and Gumbel distributions (Su et al., 2012). However, distributional characteristics such as multi-modality and left censoring are not captured by such analyses. For example, the finite mixture of normals discussed next show that at least two (chloroform) to four (1,4-DCB) components are needed to reflect the multiple modes, skewness and extreme values in the observed distributions.

3.3. Mixture of normals

Fitted density plots (and component clusters) are shown in Figs. 1B, 2B and 3B for chloroform, 1,4-DCB and styrene, respectively. The fitted parameters (weight λ_k, location μ_k and dispersion σ_k^2) of each cluster k for the mixture of normals are given in Table 2. The optimal Ks (based on the AICc) were 2, 4 and 3 for chloroform, 1,4-DCB and styrene, respectively. These choices of K clearly reflected the multi-modality and right-skewness of the VOC data, and the resulting mixture of normals closely fitted the observed distributions. For example, Fig. 2B represents the four clusters that fitted the 1,4-DCB data: the first (red) cluster captured the left censoring due to the MDL, the second and third (green and blue) clusters reflected the majority of the data and the skewness, and the fourth (blue) cluster modeled the heavy tail.

Table 1
Goodness-of-fit statistics of each density estimation method for chloroform, 1,4-DCB and styrene sample data from the RIOPA study.

VOCs	(Estimated) Proportion below MDL				MSE			MAE		
	Observed	Normal	MN	DPMN	Normal	MN	DPMN	Normal	MN	DPMN
Chloroform	0.17	0.21	0.23	0.23	0.07	0.07	0.08	7.18	6.89	6.95
1,4-DCB	0.34	0.28	0.33	0.33	31.81	0.08	0.04	167.05	7.00	5.30
Styrene	0.66	0.56	0.64	0.64	32.61	0.07	0.04	160.47	6.10	4.27

MSE: mean squared error; MAE: mean absolute error; (MSE and MAE are multiplied by a scalar of 1000 to reflect the significant figure.) MN: mixture of normals; DPMN: Dirichlet process mixture of normals.

Table 2
Fitted weight, location and dispersion parameters under the finite mixture of normals for chloroform, 1,4-DCB and styrene sample data from the RIOPA study.

	Chloroform			1,4-DCB			Styrene		
	Weight	Mean	SD	Weight	Mean	SD	Weight	Mean	SD
$K = 2$	**AICc = 1774**			AICc = 2403			AICc = 1735		
Cluster 1	**0.11**	**−1.78**	**1.31**	0.16	−1.05	0.96	0.40	−1.12	1.86
Cluster 2	**0.89**	**0.19**	**1.06**	0.84	1.35	2.23	0.60	−0.40	0.62
$K = 3$	AICc = 1778			AICc = 2330			**AICc = 1716**		
Cluster 1	0.12	−1.78	1.23	0.12	−1.05	1.58	**0.41**	**−1.12**	**1.31**
Cluster 2	0.60	0.08	0.90	0.63	0.31	1.14	**0.51**	**−0.35**	**0.54**
Cluster 3	0.28	0.55	1.20	0.25	3.84	1.93	**0.08**	**1.82**	**1.01**
$K = 4$	AICc = 1781			**AICc = 2328**			AICc = 1714		
Cluster 1	0.11	−1.78	1.27	**0.14**	**−1.05**	**1.54**	0.39	−1.12	1.33
Cluster 2	0.07	−0.52	0.25	**0.60**	**0.27**	**1.08**	0.49	−0.37	0.60
Cluster 3	0.05	0.61	0.15	**0.23**	**3.29**	**1.55**	0.04	−0.29	0.08
Cluster 4	0.78	0.24	1.09	**0.04**	**6.64**	**0.67**	0.07	1.90	0.97
$K = 5$	AICc = 1785			AICc = 2329			AICc = 1722		
Cluster 1	0.11	−1.78	1.26	0.14	−1.05	1.52	0.33	−1.12	1.32
Cluster 2	0.17	−0.39	0.43	0.05	−0.24	0.16	0.05	−1.51	1.28
Cluster 3	0.10	0.60	0.21	0.62	0.48	1.21	0.04	−0.29	0.08
Cluster 4	0.58	0.22	1.21	0.04	6.66	0.66	0.51	−0.37	0.60
Cluster 5	0.04	1.31	0.12	0.16	3.86	1.27	0.08	1.86	0.99

Bold: the smallest AICc; SD: standard deviation.

3.4. DPM of normals

Fitted densities using DPM of normals for the three VOCs are shown in Figs. 1C, 2C and 3C. This method clearly captures the censoring, right-skewness, and potential multi-modality of the exposure data. In terms of MSE and MAE, the DPM approach attained slightly lower values than the finite mixture of normals (Table 1).

Panel D on Figs. 1–3 shows results of the sensitivity analysis with the four different gamma distributions used as priors for precision parameter α. As noted before, K stochastically increases with α as $E(K|\alpha,n) \approx \alpha \log(1 + n/\alpha)$ for moderately large n (Antoniak, 1974). The four prior distributions were informative and formed up to 20 clusters that reflected more specific subject matter information. Estimated densities obtained using the four priors nearly overlapped and showed very similar MSE and MAE for each of the VOCs, although the corresponding posterior distribution of the number of clusters K varied (Table 3). The posterior mean of K under all prior settings of α (Table 3) slightly exceeded the K selected using the AICc (Table 2). The higher K in the DPM is due to the prior information of α, and does not introduce any additional complexity or more model parameters. The initial prior variance of α critically influences the extent of smoothing (Escobar and West, 1995). Given K distinct values among the elements of θ, a larger variance leads to increased dispersion among the K group means, which increases the likelihood of multiple modes and decreased smoothness in the resulting predictive distribution (Escobar and West, 1995). The general goals in selecting α and K to partition the data is to avoid over-smoothing and also excessive jaggedness. The prior distributions of α regarding the number of clusters K reflect a subjective assessment that balances these competing goals. Prior distributions might also reflect normative and objective representations of parameter values, although there is no consensus on the "best" way to elicit a subjective prior (Dey and Liu, 2007).

No convergence issues using the DPM method were encountered, and density estimation results were robust given the moderate sample size ($n = 544$). Another advantage of the DPM method is that a constraint to ensure a cluster below MDL is not required since the sampling scheme (4) is data driven. As shown in Eq. (4), the DPM can handle values under the MDL that are represented as a point mass, because a newly sampled value has equal probability $1/(n - 1 + \alpha)$ to be drawn from the observed set of values.

Table 3
Posterior distribution of the number of clusters K based on various prior settings of α as a sensitivity analysis.

	Posterior distribution of K								
	Chlorofrom			1,4-DCB			Styrene		
Prior Settings	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
Setting 1	2.8	2	1.4	32.8	34	20.2	10.9	5	10.8
Setting 2	3.9	3	2.4	5.6	5	2.5	4.6	4	2.8
Setting 3	4.1	4	2.2	7.1	7	3.4	7.9	7	4.4
Setting 4	10.5	9	6.0	15.3	14	6.5	13.1	12	6.0

SD: standard deviation.
Setting 1: $\alpha \sim$Gamma(0.3,0.4); Setting 2: $\alpha \sim$Gamma(1.2,2.5).
Setting 3: $\alpha \sim$Gamma(2.0,1.5); Setting 4: $\alpha \sim$Gamma(5.0,2.0).

3.5. Simulations

Simulation results, summarized in Table 4, show similar patterns for the MSE and MAE criteria. Both finite mixture and DPM of normals provided much better fits than a single normal distribution, except that the former two methods are only slightly better under distribution 1 with $P_0 = 0.15$. For both distributions, as the fraction P_0 of data below the MDL increased, there is evidence of increasing trend of lack of fit for a single normal distribution, while the finite mixture and DPM of normals fitted considerable better and without such trend. The DPM of normals shows advantage of robustness regarding P_0. It fits equally well, or even better, as P_0 increased. For distribution 1, the finite mixture of normals provided a slightly better fit than the DPM of normals, but this trend can be offset since the prior variance of α can be decreased to promote smoothness. In this regard, DPM is much more flexible than the finite mixture of normal. Here, we have used $\alpha \sim$ Gamma(1.2,2.5) which favors 1–5 clusters given our sample size (as the prior information of K). For distribution 2 which is right skewed and with a heavy tail, the DPM of normals provided a much better fit than finite mixture of normals under all settings.

4. Discussion

Finite mixture of distributions has been used to address problems of classification, density estimation and pattern recognition

Table 4
Summary of goodness-of-fit statistics of each density estimation method in the simulation study.

	Proportion below MDL	MSE			MAE		
		Normal	MN	DPMN	Normal	MN	DPMN
Distribution 1	0.15	0.09	0.03	0.08	7.65	4.64	7.11
	0.30	0.19	0.04	0.08	11.19	4.80	7.29
	0.50	0.43	0.05	0.05	16.77	5.26	5.69
Distribution 2	0.15	1.55	0.10	0.02	32.58	8.19	3.57
	0.30	2.53	0.10	0.02	43.69	8.59	3.29
	0.50	2.62	0.12	0.02	46.52	8.22	3.28

MSE: mean squared error; MAE: mean absolute error; (MSE and MAE are multiplied by a scalar of 1000 to reflect the significant figure.) MN: mixture of normals; DPMN: Dirichlet process mixture of normals.
Distribution 1: Normal(0,2^2); Distribution 2: 1/2 Gamma(3,1.5) + 1/2 Uniform(−3,8). Prior distribution on α is Gamma(1.2,2.5).

problems across a wide range of areas (Titterington et al., 1985; McLachlan and Basford, 1988; McLachlan and Peel, 2000). However, applications in the environmental literature have been limited. For human exposure and risk assessment, mixtures of lognormal distributions have been used to model concentrations of radon 222 in drinking water, and the model yielded a high-fidelity fit that was not achievable with any single parametric distribution (Burmaster and Wilson, 2000). In another risk assessment application, a mixture dose–response model was used to derive the upper confidence limits on risk and benchmark doses (Razzaghi and Kodell, 2000). A mixture of distributions of true zero exposures and lognormal distributions with left censoring was used to account for non-detects (Taylor et al., 2001). In a medical intervention that examined biomarkers of aflatoxin, finite mixture of distributions were selected using bootstrap- and cross-validation-based information criterion to portray bimodal biomarker distributions that had a significant fraction of measurements below MDL, and that also varied due to differences in exposure, metabolism, intervention effect and other factors (Chu et al., 2005). To quantify variability and uncertainty in emission inventories, the use of a mixture of lognormal distributions showed a better fit and more efficient estimates of uncertainty to NO_x emission data than the use of a single distribution (Zheng and Frey, 2004).

This paper is one of the first demonstrations of mixture distribution models for environmental exposure data. No application of the DPM method specifically for pollutant concentration and exposure data was identified. Our analysis compared both finite mixture of normals and DPM of normals methods to parametric full distribution models and extreme value models, and included a sensitivity analysis of smoothing parameters. None of this information is available in the literature.

Our intent was to develop and characterize the performance of mixture models for fitting environmental exposures. The methods can be applied to other types of data, e.g., airborne concentrations, biomarkers, and pollutants other than VOCs. In exposure and risk applications, such models can improve the accuracy and realism in estimating cumulative exposure, cumulative risk, population attributable risk (PAR), and uncertainty (of exposure, risk or PAR) based on Monte Carlo simulations or other methods. For example, it would be straightforward and potentially informative to evaluate the differences in risk or numbers of individuals affected at a specific risk level using the cumulative distributions presented in the paper.

Finite mixture of distribution models continued to receive attention from both theoretical and practical points of view. In environmental applications, such as the description and analysis of air pollutant concentration and exposure data, the key advantage of this class of parametric mixture of distributions over a single parametric distribution model is their ability to simultaneously portray the multimodal nature of exposure data, the heavy tails, and observations around MDLs. These models also have advantages over extreme value distributions (e.g., Gumbel or GEV) since the entire distribution is fitted (not just the tail), and it is not necessary to select a cut-off to define extrema. Information on the full distributions of exposure levels can be used to estimate health risks and uncertainty estimates across a population (Su et al., 2012) and facilitate probabilistic analyses. Even if the goal is only to predict extreme values, DPM methods can provide closer fits to any empirical distribution than GEV models, although GEV models often, but not always, perform well in such applications, as demonstrated by Su et al. (2012) for NHANES data. Further, the use of a cluster to represent observations below the MDL is appealing for datasets where a fair fraction of data is believed to be at or near this level. Finally, the parameter estimates corresponding to this lowest cluster may be heuristically interpretable as the parameters corresponding to MDL and its uncertainty. Importantly, we have found that a constrained MLE was required in the presence of data censoring due to the MDL, otherwise the estimated clusters may not contain the mode below the MDL.

Further work may be needed to develop consistent rules and guidance for finite mixture of distributions that address the number of components K, the selection of performance criteria, the effect of the estimation algorithm, use of constraints in the presence of data censoring, and the minimum sample size needed for analysis (Mendell et al., 1991). Such decisions can depend on the nature of the original data and the purpose of the analysis, e.g., focus on extrema or the entire distribution. We assessed performance in terms of the accuracy of point predictions above the MDL, but also compared the proportion of data points predicted to be below the MDL, thus reflecting datasets with censored data. Other techniques to select appropriate models include resampling information criteria (McLachlan, 1987; Chu et al., 2005), likelihood methods (McLachlan, 1987; Chen et al., 2001), and Bayesian approaches that treat K as a parameter (Richardson and Green, 1997). Chu et al. (2005) also suggests that the D-test via the L^2 distance between competing models (Charnigo and Sun, 2004), which has a closed form expression, is advantageous for selecting K as compared to likelihood statistics, which are nontrivial functions of both the parameter estimators and the full dataset.

We note that estimating the "true" numbers of clusters is not a necessary goal for practical applications of mixture of distributions. However, marginalizing over the distribution of K is critical for capturing the uncertainty in the density estimates. Rather than using some pre-specified number of clusters, which is always a concern of the traditional mixture distribution problem, the DPM of normals provides a semi-parametric Bayesian alternative to nonparametric histograms and provides greater flexibility and precision in modeling the underlying characteristics of the sample data. This was clearly demonstrated in simulation results, where the performance of the DPM models was not substantially altered (and sometimes even improved) as the fraction of data below the MDL increased.

The nonparametric DPM of normal distributions assume that observed data randomly arise from sub-distributions with certain probabilities as the finite mixture of distribution models. (Again, sub-populations that an individual observation belongs are not identified.) Compared to the finite mixture models, DPM distributions have advantages in providing a formal assessment of uncertainty for all model parameters, including the number of components K, through generated draws from the posterior distribution. With a suitable Dirichlet process prior structure (Escobar and West, 1995), these models produce predictive distributions qualitatively similar to kernel techniques, and they allow for

differing degrees of smoothing by the choice on priors for precision parameter α. The density estimation results were robust given a moderate sample size ($n = 544$) without any convergence issues noted.

Both types of mixture models explored in this study are well suited to VOC data containing a large fraction of censored data due to MDLs, fat tails, and multiple modes. They offer clear advantages over parametric full distribution models and extreme value models, and also appear appropriate for many other types of environmental data, such as concentrations or doses of persistent and/or emerging compounds and biomarkers. The use of mixture models has the potential to improve the accuracy and realism of models used in a variety of exposure and risk applications, and further environmental applications are warranted.

5. Conclusions

Compared to the finite mixture of normals, DPM of normals has advantages by characterizing uncertainty around the number of components, and by providing a formal assessment of uncertainty for all model parameters through the posterior distribution. The method adapts to a spectrum of departures from standard model assumptions and provides robust estimates of the exposure density even under censoring due to MDL.

Acknowledgments

Research described in this article was conducted under contract to the Health Effects Institute (HEI), an organization jointly funded by the United States Environmental Protection Agency (EPA) (Assistance Award No. R-82811201) and certain motor vehicle and engine manufacturers. The contents of this article do not necessarily reflect the views of HEI, or its sponsors, nor do they necessarily reflect the views and policies of the EPA or motor vehicle and engine manufacturers. The research of Dr. Bhramar Mukherjee was supported by grant NIH ES 20811. We appreciate the input of the HEI reviewers. Additional support was provided by grant P30ES017885 from the National Institute of Environmental Health Sciences, National Institutes of Health entitled “Lifestage Exposure and Adult Disease.”

Disclaimer

The authors declare that they have no actual or potential competing financial interests.

Code

The R code used to generate results in this paper is described in the Supplementary material.

Appendix A. Supplementary material

Supplementary material related to this article can be found at http://dx.doi.org/10.1016/j.atmosenv.2013.05.004.

References

Antoniak, C.E., 1974. Mixtures of Dirichlet processes with applications to nonparametric problems. Annals of Statistics 2, 1152–1174.
Antweiler, R., Taylor, H., 2008. Evaluation of statistical treatments of left-censored environmental data using coincident uncensored data sets: I. Summary statistics. Environmental Science & Technology 42, 3732–3738.
ATSDR, 1997. Toxicological Profile for Chloroform. Agency for Toxic Substances and Disease Registry, Atlanta, GA. Available from: http://www.atsdr.cdc.gov/ToxProfiles/tp.asp?id=53&tid=16.
ATSDR, 2006. Toxicological Profile for Dichlorobenzenes. Agency for Toxic Substances and Disease Registry, Atlanta, GA. Available from: http://www.atsdr.cdc.gov/toxprofiles/tp.asp?id=704&tid=126.
ATSDR, 2010. Toxicological Profile for Styrene. Agency for Toxic Substances and Disease Registry, Atlanta, GA. Available from: http://www.atsdr.cdc.gov/PHS/PHS.asp?id=419&tid=74.
Batterman, S., Jia, C., Godwin, C., Hatzivasilis, G., 2006. A dominant source of VOC exposure: attached garages. Epidemiology 17, S350.
Batterman, S., Su, F.C., Jia, C., Naidoo, R.N., Robins, T., Naik, I., 2011. Manganese and lead in children’s blood and airborne particulate matter in Durban, South Africa. Science of the Total Environment 409, 1058–1068.
Benaglia, T., Chauveau, D., Hunter, D.R., Young, D.S., 2009. mixtools: an R package for analyzing mixture models. Journal of Statistical Software 32.
Biernacki, C., Celeux, C., Govaert, G., 2000. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 719–725.
Blackwell, D., MacQueen, J.B., 1973. Ferguson distributions via Polya Urn schemes. Annals of Statistics 1, 353–355.
Burmaster, D.E., Wilson, A.M., 2000. Fitting second-order finite mixture models to data with many censored values using maximum likelihood estimation. Risk Analysis: An Official Publication of the Society for Risk Analysis 20, 261–272.
Caldwell, J.C., Woodruff, T.J., Morello-Frosch, R., Axelrad, D.A., 1998. Application of health information to hazardous air pollutants modeled in EPA’s cumulative exposure project. Toxicology and Industrial Health 14, 429–454.
CDC, 2012. NHANES 1999–2000 Laboratory Files. Centers for Disease Control and Prevention, Atlanta, GA. Available from: http://www.cdc.gov/nchs/nhanes/nhanes1999-2000/lab99_00.htm.
Charnigo, R., Sun, J., 2004. Testing homogeneity in a mixture distribution via the L2 distance between competing models. Journal of the American Statistical Association 99, 488–498.
Chen, H., Chen, J., Kalbfleisch, J.D., 2001. A modified likelihood ratio test for homogeneity in finite mixture models. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 63, 19–29.
Chin, J.Y., Godwin, C., Jia, C., Robins, T., Lewis, T., Parker, E., Max, P., Batterman, S., 2012. Concentrations and risks of p-dichlorobenzene in indoor and outdoor air. Indoor Air. http://dx.doi.org/10.1111/j.1600-0668.2012.00796.x.
Chu, H., Kensler, T.W., Muñoz, A., 2005. Assessing the effect of interventions in the context of mixture distributions with detection limits. Statistics in Medicine 24, 2053–2067.
D’Souza, J.C., Jia, C., Mukherjee, B., Batterman, S., 2009. Ethnicity, housing and personal factors as determinants of VOC exposures. Atmospheric Environment 43, 2884–2892.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) 39, 1–38.
Dey, D., Liu, J., 2007. A quantitative study of quantile based direct prior elicitation from expert opinion. Bayesian Analysis 2, 137–166.
Escobar, M.D., 1994. Estimating normal means with a dirichlet process prior. Journal of the American Statistical Association 89, 268–277.
Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90, 577–588.
Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1, 209–230.
Ferguson, T.S., 1983. Bayesian density estimation by mixtures of normal distributions. In: Rizvi, M., Rustagi, J., Siegmund, D. (Eds.), Recent Advances in Statistics. Academic Press, New York, US, pp. 287–302.
Fraley, C., Raftery, A.E., 1998. How many clusters? Which clustering method? Answers via model-based cluster analysis. The Computer Journal 41, 578–588.
Hammonds, J.S., Hoffman, F.O., Bartell, S.M., 1994. An Introductory Guide to Uncertainty Analysis in Environmental and Health Risk Assessment, ES/ER/TM-35. US Environmental Protection Agency, Washington, DC.
Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297–307.
IARC, 2012. Agents Classified by the IARC Monographs. World Health Organization, International Agency for Research on Cancer, Lyon, France. Available from: http://monographs.iarc.fr/ENG/Classification/index.php.
Jara, A., 2007. Applied Bayesian non- and semi-parametric inference using DPpackage. R News 7, 17–26.
Jara, A., Hanson, T., Quintana, F.A., Müller, P., Rosner, G.L., 2011. DPpackage: Bayesian semi- and nonparametric modeling in R. Journal of Statistical Software 40, 1–30.
Jia, C., Batterman, S., 2010. A critical review of naphthalene sources and exposures relevant to indoor and outdoor air. International Journal of Environmental Research and Public Health 7, 2903–2939.
Jia, C., Batterman, S., Godwin, C., 2008a. VOCs in industrial, urban and suburban neighborhoods – Part 2: Factors affecting indoor and outdoor concentrations. Atmospheric Environment 42, 2101–2116.
Jia, C., D’Souza, J., Batterman, S., 2008b. Distributions of personal VOC exposures: a population-based analysis. Environment International 34, 922–931.
Kim, H., Bernstein, J., 2009. Air pollution and allergic disease. Current Allergy and Asthma Reports 9, 128–133.
Krishnamoorthy, K., Mallick, A., Mathew, T., 2009. Model-based imputation approach for data analysis in the presence of non-detects. Annals of Occupational Hygiene 53, 249–263.

Ling, Z., Guo, H., Cheng, H., Yu, Y., 2011. Sources of ambient volatile organic compounds and their contributions to photochemical ozone formation at a site in the Pearl River Delta, southern China. Environmental Pollution 159, 2310–2319.
Lubin, J., Colt, J., Camann, D., Davis, S., Cerhan, J., Severson, R., Bernstein, L., Hartge, P., 2004. Epidemiologic evaluation of measurement data in the presence of detection limits. Environmental Health Perspectives 112, 1691–1696.
McLachlan, G.J., 1987. On bootstrapping the likelihood ratio test statistics for the number of components in a normal mixture. Journal of the Royal Statistical Society. Series C (Applied Statistics) 36, 318–324.
McLachlan, G.J., Basford, K.E., 1988. Mixture Models: Inference and Applications to Clustering. Dekker, M, New York.
McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley & Sons, New York, US.
McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley and Sons, New York, US.
MDE, 2010. Top Ten Sources of Volatile Organic Compound in the Baltimore Area 2002. Maryland Department of the Environment. Available from: http://www.mde.state.md.us/Air/air_information/TopTenVOC.asp.
Mendell, N.R., Thode Jr., H.C., Finch, S.J., 1991. The likelihood ratio test for the two-component normal mixture problems: power and sample size analysis. Biometrics 47, 1143–1148.
Meng, X.L., Pedlow, S., 1992. EM: a bibliographic review with missing articles. In: The Proceedings of the Statistical Computing Section of the American Statistical Association, pp. 24–27.
Mueller, P., Quintana, F., 2004. Nonparametric Bayesian data analysis. Statistical Science 19, 95–110.
OEHHA, 2005. Air Toxics Hot Spots Program Risk Assessment Guidelines Part II: Technical Support Document for Describing Available Cancer Potency Factors. California Environmental Protection Agency, Office of Environmental Health Hazard Assessment, Air Toxicology and Epidemiology Section, Sacramento, CA.
Rappaport, S.M., Kupper, L.L., 2004. Variability of environmental exposures to volatile organic compounds. Journal of Exposure Analysis and Environmental Epidemiology 14, 92–107.
Razzaghi, M., Kodell, R.L., 2000. Risk assessment for quantitative responses using a mixture model. Biometrics 56, 519–527.
Redner, R., Walker, H., 1984. Mixture densities, maximum likelihood and the EM algorithm. SIAM Review 26, 195–239.
Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components. Journal of the Royal Statistical Society. Series B (Methodological) 59, 731–792.
Su, F.-C., Jia, C., Batterman, S., 2012. Extreme value analyses of VOC exposures and risks: a comparison of RIOPA and NHANES datasets. Atmospheric Environment 62, 97–106.
Taylor, D.J., Kupper, L.L., Rappaport, S.M., Lyles, R.H., 2001. A mixture model for occupational exposure mean testing with a limit of detection. Biometrics 57, 681–688.
Titterington, D.M., Smith, A.F.M., Makov, U.E., 1985. Statistical Analysis of Finite Mixture Distributions. John Wiley and Sons, Chichester, UK.
USEPA, 2012a. Integrated Risk Information System (IRIS). US Environmental Protection Agency, Washington, DC. Available from: http://www.epa.gov/IRIS/index.html.
USEPA, 2012b. An Introduction to Indoor Air Quality: Volatile Organic Compounds (VOCs). US Environmental Protection Agency, Washington, DC. Available from: http://www.epa.gov/iaq/voc.html.
Weisel, C.P., Zhang, J., Turpin, B.J., Morandi, M.T., Colome, S., Stock, T.H., Spektor, D.M., Korn, L., Winer, A., Alimokhtari, S., Kwon, J., Mohan, K., Harrington, R., Giovanetti, R., Cui, W., Afshar, M., Maberti, S., Shendell, D., 2005a. Relationship of Indoor, Outdoor and Personal Air (RIOPA) study: study design, methods and quality assurance/control results. Journal of Exposure Analysis and Environmental Epidemiology 15, 123–137.
Weisel, C.P., Zhang, J., Turpin, B.J., Morandi, M.T., Colome, S., Stock, T.H., Spektor, D.M., Korn, L., Winer, A.M., Kwon, J., Meng, Q.Y., Zhang, L., Harrington, R., Liu, W., Reff, A., Lee, J.H., Alimokhtari, S., Mohan, K., Shendell, D., Jones, J., Farrar, L., Maberti, S., Fan, T., 2005b. Relationships of Indoor, Outdoor, and Personal Air (RIOPA). Part I. Collection Methods and Descriptive Analyses. Health Effects Institute Research Report 1–127.
Weschler, C.J., 2011. Chemistry in indoor environments: 20 years of research. Indoor Air 21, 205–218.
Zheng, J., Frey, H.C., 2004. Quantification of variability and uncertainty using mixture distributions: evaluation of sample size, mixing weights, and separation between components. Risk Analysis 24, 553–571.