The Annals of Applied Statistics
2020, Vol. 14, No. 4, 1945–1963
https://doi.org/10.1214/20-AOAS1364

INTEGRATIVE STATISTICAL METHODS FOR EXPOSURE MIXTURES AND HEALTH

BY BRIAN J. REICH[1], YAWEN GUAN[2], DENIS FOURCHES[3], JOSHUA L. WARREN[4], STEFANIE E. SARNAT[5] AND HOWARD H. CHANG[6]

[1]*Department of Statistics, North Carolina State University, bjreich@ncsu.edu*
[2]*Department of Statistics, University of Nebraska, yguan12@unl.edu*
[3]*Department of Chemistry, North Carolina State University, dfourch@ncsu.edu*
[4]*Department of Biostatistics, Yale University, joshua.warren@yale.edu*
[5]*Department of Environmental Health, Emory University, sebelt@emory.edu*
[6]*Department of Biostatistics and Bioinformatics, Emory University, howard.chang@emory.edu*

Humans are concurrently exposed to chemically, structurally and toxicologically diverse chemicals. A critical challenge for environmental epidemiology is to quantify the risk of adverse health outcomes resulting from exposures to such chemical mixtures and to identify which mixture constituents may be driving etiologic associations. A variety of statistical methods have been proposed to address these critical research questions. However, they generally rely solely on measured exposure and health data available within a specific study. Advancements in understanding of the role of mixtures on human health impacts may be better achieved through the utilization of external data and knowledge from multiple disciplines with innovative statistical tools. In this paper we develop new methods for health analyses that incorporate auxiliary information about the chemicals in a mixture, such as physicochemical, structural and/or toxicological data. We expect that the constituents identified using auxiliary information will be more biologically meaningful than those identified by methods that solely utilize observed correlations between measured exposure. We develop flexible Bayesian models by specifying prior distributions for the exposures and their effects that include auxiliary information and examine this idea over a spectrum of analyses from regression to factor analysis. The methods are applied to study the effects of volatile organic compounds on emergency room visits in Atlanta. We find that including cheminformatic information about the exposure variables improves prediction and provides a more interpretable model for emergency room visits for respiratory diseases.

1. Introduction. Environmental chemical exposures, such as pesticides, industrial contaminants and air pollution, have major public health consequences. There is a vast body of epidemiologic literature that has identified significant associations between ambient air pollution and numerous adverse health outcomes, including all-cause mortality (Atkinson et al. (2015), Chen et al. (2017)), cardiovascular events (Shah et al. (2015)), respiratory events (Zheng et al. (2015)) and birth outcomes (Stieb et al. (2012), Vrijheid et al. (2011)). The majority of these past studies used single exposure approaches to assess health risks, controlling for other confounders. However, humans are simultaneously exposed to complex mixtures of chemicals from multiple sources. Many components of these mixtures are highly correlated due to their common sources, chemical properties, spatial variation and meteorological drivers. Moreover, chemicals within mixtures have the potential to interact with each other in complex ways that may result in additive, synergistic or competing detrimental effects on

Received September 2019; revised June 2020.
Key words and phrases. Cheminformatics, collinearity, factor analysis, principal components, stochastic search, variable selection.

human health. Therefore, the combined effect of exposure to a multipollutant mixture may differ greatly than the sum of individual observed effects (Billionnet et al. (2012)).

As recommended by the United States (U.S.) National Research Council, there is a need to move from a single-exposure to a multiexposure approach in order to gain an improved understanding of the public health burden of chemical mixtures (National Research Council and others (2004a, 2004b)). Hence, there has been a recent paradigm shift, particularly in air pollution epidemiology, to study health risks associated with exposure to multiple chemicals simultaneously. Knowledge gained from estimating health effects of mixtures will aid in performing more comprehensive risk assessments, designing regulatory policies to effectively minimize health burdens and developing compliance and monitoring strategies for multiple pollutants (Dominici et al. (2010)).

Various statistical techniques have been developed for estimating joint effects of multiple exposures. These techniques broadly fall into two categories, often with specific scientific goals: (1) variable selection for estimating health effects of individual pollutants and (2) dimension reduction for identifying interpretable latent processes (e.g., common sources). Variable selection using stepwise algorithms (Sinisi and van der Laan (2004)) has been used in multipollutant settings to identify pollutants associated with adverse health outcomes after controlling for confounders (Mortimer et al. (2008)). Under high-dimensionality, regularization methods that penalize model complexity via additional constraints on regression coefficients can further improve estimation stability (MacLehose et al. (2015)). However, simulation studies have shown that health effect estimates can be inflated and have overly optimistic confidence intervals (Dominici et al. (2008), Sun et al. (2013)). Recent work has applied Bayesian variable selection to identify exposures with adverse and potentially nonlinear health effects (Antonelli et al. (2020), Bobb et al. (2015), Fang et al. (2019), Sabanés Bové, Held and Kauermann (2015), Wei et al. (2020)).

As the scientific focus has shifted away from health effects of individual chemicals toward the impact of the mixture, dimension reduction techniques, such as principal component analysis and factor analysis, have become more widely used in environmental health studies. These approaches utilize observed correlations between pollutants' dose profiles to produce a smaller number of constructed exposure variables that may be more interpretable than individual pollutants. Examples include profile regression (Moliter et al. (2010)), selforganizing maps (Pearce et al. (2016)) and weighted quantile sum (Carrico et al. (2015)).

We propose major expansions of methods in these two categories through the common theme of integrating auxiliary information on pollutants' chemical and toxicological properties. Under a Bayesian hierarchical framework, auxiliary information enters through prior distributions in a flexible and modular fashion. We hypothesize that this approach will improve estimation performance and provide interpretable findings by leveraging information on chemical structures and toxicological properties. This follows from previous epidemiologic investigations where pollutants are first grouped a priori and analyzed as a group using different methods (Suh et al. (2011), Ye et al. (2017)). We propose approaches that are motivated by recent advances in bioinformatics and computational biology, where analytic tools are being developed that have also increasingly considered leveraging auxiliary information, such as genomic annotations, regulation networks (Li and Li (2010)), and protein-protein interactions (Li et al. (2016)). Specifically, we will develop state-of-the-art Bayesian methods to simultaneously perform variable selection and induce shrinkage of estimates based on interpollutant similarity in chemical/toxicological properties. More importantly, our model-based framework also allows for an assessment of how these chemical/toxicological properties are associated with adverse health effects, allowing findings to be generalized to pollutants not examined in a study. Finally, a Bayesian approach allows straightforward and transparent uncertainty quantification, as opposed to methods that involve multiple screening stages.

The proposed methods are applied to a time series analysis of emergency department visits in Atlanta to estimate the health effects associated with mixtures of volatile organic compounds (VOC) that are not routinely measured at the ambient level. We aim to provide population-based epidemiologic evidence that may help elucidate the biological mechanism of air pollutant toxicity. VOC constituents can serve as better markers for emission sources; therefore, better control strategies may be developed through the identification of specific harmful pollutants. While our specific modeling approaches are driven by the scientific questions of this motivating epidemiologic study, the proposed modeling framework is applicable to supplement existing methods for other applications.

2. Emergency department visits and air pollution in Atlanta.

2.1. *Emergency department visits data.* Since 1998, investigators at Emory University have been collecting patient-level emergency department (ED) visit records for the Atlanta metropolitan area. We analyze the daily number of ED visits from acute care hospitals in Atlanta between January 1, 1998, and December 31, 2008. ED visits for asthma and wheeze, all respiratory diseases and all cardiovascular diseases were identified by their International Classification of Diseases 9th Revision (ICD-9) diagnostic codes and daily counts of each response were aggregated from individual-level records by matching a patient's ZIP code to a ZIP Code Tabulation Area (ZCTA) within the city. Over this time frame the total number (daily mean, standard deviation) of ED visits is approximately 232,000 (61, 24) for asthma and wheeze, 250,000 (66, 18) for cardiovascular diseases and one million for respiratory diseases (274, 98). By regressing city-wide counts onto city-wide ambient concentrations, we do not account for individual variation in exposure. The Atlanta ED visit database is the largest of its kind in the U.S. for assessing air pollution-related morbidity. Prior epidemiologic results from Atlanta have shown cardiorespiratory effects of primary (e.g., carbon monoxide, nitrogen dioxide, elemental carbon), secondary (e.g., ozone) and mixed (e.g., fine particulate matter) origin pollutants (Strickland et al. (2010), Ye et al. (2018)).

2.2. *Air pollution data.* The population-based health studies described in Section 2.1 took advantage of the extensive air quality measurements collected at the Jefferson Street (JST) monitoring site in the SouthEastern Aerosol Research and Characterization (SEARCH) network. Situated in central Atlanta with few large emission sources nearby, JST is considered representative of the urban environment and has generated a long historical record of unique daily speciated pollution data. The dataset includes daily concentrations of 44 identified VOC constituents. VOC sampling details are given by Hansen et al. (2006), and concentrations are reported in part per billion carbon (ppb-C). Complete lists of VOC constituents and their summary statistics are given in Ye et al. (2017) which also presents a traditional epidemiology analysis of these VOC measurements and Atlanta ED visits using single- and multipollutant approaches. The multipollutant approaches relied on defining seven groups of VOC a priori based on chemical structure (e.g., *n*-alkanes, cycloalkanes, alkenes and aromatics). Whereas, in this analysis we utilize quantitative measures to capture similarity in chemical structures between VOCs. Figure 1 (upper-left triangle) plots the sample correlations between the exposure variables; there are many strong positive correlations between the exposures. For this analysis we only included 44 VOC constituents with concentrations above the limit of detection (0.1 ppb-C) on at least 90% of days. The daily VOC measurements were taken at a monitoring location specifically selected to reflect the daily "background" pollution level in Atlanta. Hence, one limitation of the study is that we assume this daily concentration is representative of averaged individual-level exposures.

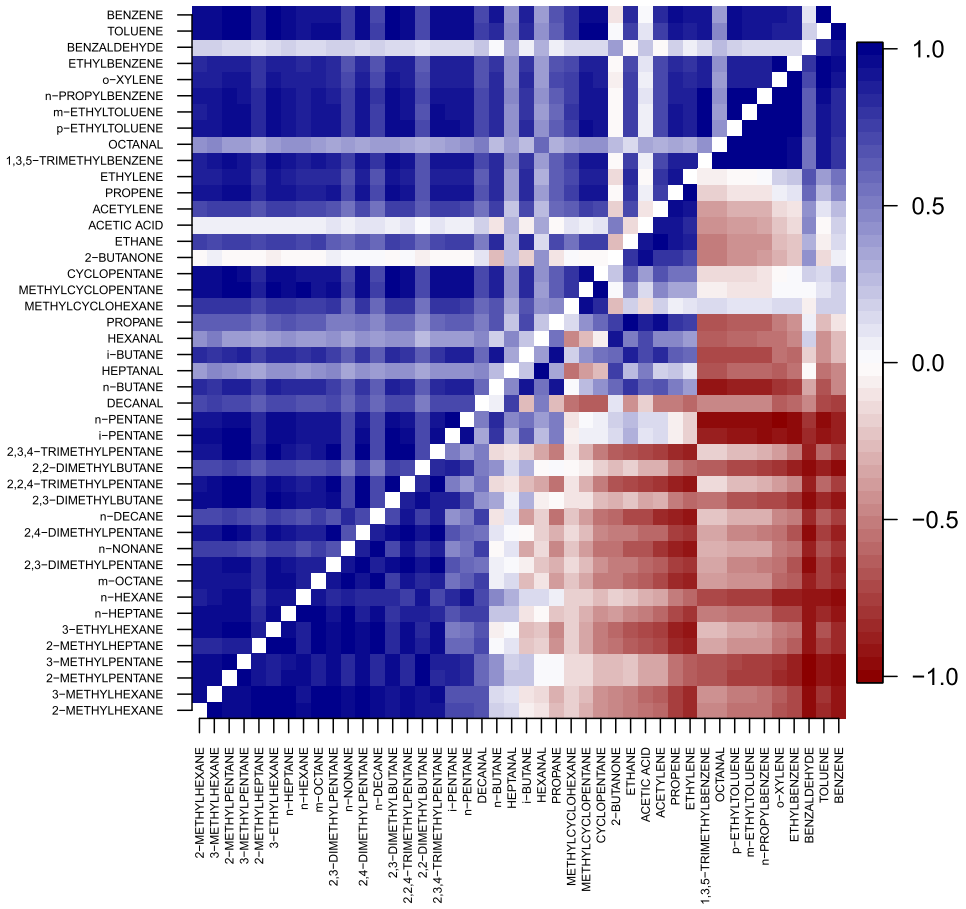

FIG. 1. Sample correlations of daily volatile organic compound air pollutant in Atlanta, 1998 to 2008. *The upper-left triangle plots the sample correlation between the exposure variables* (**X**) *in the dataset*; *the lower-right triangle plots the correlation of VOC-associated auxiliary data* (**Z**).

2.3. *Auxiliary data*. We use the in-silico methods in Fourches, Muratov and Tropsha (2010) to characterize molecular structure. Using the R package `cdk` (Guha (2007)), we extract $q = 74$ numerical descriptors of each exposure variable's molecule structure. Examples of the descriptors include the number of aliphatic carbocycles for the molecule and the fraction of carbon atoms that are SP3 hybridized. Let Z_{jl} denote the (standardized) value of descriptor l for exposure variable j. These descriptors can be used to measure the pairwise structural similarity between toxicants. For example, the correlation of Z_{jl} and Z_{kl} across the $l = 1, \ldots, q$ features is a measure of structural similarity between exposure variables j and k. Figure 1 (bottom-right triangle) plots these correlations. Since these auxiliary variables are functions of chemical properties alone and have no relation with the observed concentrations, their correlation structure is vastly different than the correlation of the observed concentrations of the exposure variables. The correlation of the concentrations varies from study to study, whereas the auxiliary variables represent inherent properties of the chemicals, therefore groupings based on the auxiliary variables are more stable and arguably more biologically meaningful than groupings based on the concentrations.

3. Enhanced Bayesian variable selection by exploiting auxiliary information. For observation $i = 1, \ldots, n$, let Y_i be the response, $\mathbf{w}_i = (w_{i1}, \ldots, w_{iW})^T$ be a vector of potential confounders and $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^T$ be the vector of exposure variables. We supplement these data with further information about the p exposures. Let $\mathbf{Z}_j = (Z_{j1}, \ldots, Z_{jq})^T$ be

the vector of q additional variables relating to exposure j. In our motivating examples these include molecular structure properties, but the methods below easily generalize to other forms of auxiliary information including results of in-vitro experiments. Gathering data across observations, let $\mathbf{X}$ be the $n \times p$ matrix of exposure variables and $\mathbf{Z}$ be the $q \times p$ matrix of auxiliary data. For simplicity, we assume all confounding, exposure and auxiliary variables have been standardized to have mean zero and variance one.

The confounding and exposure variables are linked to the response as $g[\mathrm{E}(Y_i)] = \mathbf{w}_i^T \boldsymbol{\alpha} + \sum_{j=1}^{p} X_{ij}\beta_j$ where g is the link function, $\boldsymbol{\alpha}$ are the regression coefficients associated with the confounding variables and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ are parameters that describe the impact of exposure on the response. For example, in our motivating case study, Y_i represents number of ED visits for day i, X_{ij} represents the jth VOC pollutant and $\boldsymbol{\alpha}$ include meteorology and seasonal trends. Bayesian variable selection seeks to identify a subset of exposures associated with health effects which is equivalent to the subset of coefficients $\beta_1, \ldots, \beta_p$ that are nonzero. Uncertainty about the subset of important variables is encoded in the prior distribution which is commonly (e.g., George and McCulloch (1997), O'Hara and Sillanpää (2009)) taken to be the two-component mixture: $\mathrm{Prob}(\beta_j = 0) = \pi_j$ and given that $\beta_j \neq 0$ the continuous component is $\beta_j | \beta_j \neq 0 \sim \mathrm{Normal}(\mu_j, \sigma^2)$, independent across j. Since the subset of variables that are included in the model is treated as random in the Bayesian analysis, this is referred to as Stochastic Search Variable Selection (SSVS).

The auxiliary data $\mathbf{Z}_j$ are used to define the prior distribution for the effects of the exposure variables. The most direct way to incorporate the auxiliary variables into the Bayesian variable selection prior is via the prior model probabilities,

$$\mathrm{logit}(\pi_j) = \gamma_{01} + \sum_{l=1}^{q} Z_{jl}\gamma_{l1}, \tag{3.1}$$

where γ_{l1} determines the effect of auxiliary variable l on the exposure variable's inclusion probability. We also use the auxiliary variables in the prior mean of the nonzero coefficients

$$\mu_j = \gamma_{02} + \sum_{l=1}^{q} Z_{jl}\gamma_{l2}, \tag{3.2}$$

where γ_{l2} determines the effect of auxiliary variable l on the exposure variable's effect, given it is included in the model.

The Bayesian model is completed by specifying priors for the model parameters. We assume priors

$$\gamma_{l1} | \gamma_{l2} \sim \mathrm{Normal}\big(b\gamma_{l2}, \tau_1^2\big) \quad \text{and} \quad \gamma_{l2} \sim \mathrm{Normal}\big(0, \tau_2^2\big) \tag{3.3}$$

for $l \geq 1$. The parameter b controls the dependence between the effect of an auxiliary variable on the inclusion probability and the effect size; if $b > 0$, then it is more likely that the same set of auxiliary variables determine both which exposures are associated with the health outcome and the magnitude of their effect size. The remaining parameters have uninformative priors $\boldsymbol{\alpha} \sim \mathrm{Normal}(0, c_1^2\mathbf{I})$, $b, \gamma_{01}, \gamma_{02} \sim \mathrm{Normal}(0, c_1^2)$ and $\sigma, \tau_1, \tau_2 \sim \mathrm{HC}(c_2)$, that is, the half-Cauchy prior with scale parameter c_2 (Gelman (2006)). The logistic regression intercept γ_{01} controls the prior expected number of variables included in the model. We have given this parameter a prior mean of zero to center the inclusion probabilities on 0.5, but its prior mean could be reduced to control-model size if the number of exposures is large (e.g., Kwon et al. (2011)). We also consider fixing $\tau_1 = 0$ so that $\gamma_{1i} = b\gamma_{i2}$ and thus auxiliary variables enter both the inclusion probability and the effect size model via the single index $\sum_{l=1}^{q} Z_{jl}\gamma_{l1}$. This shared model has fewer parameters to estimate which is advantageous if indeed inclusion probability and effect magnitude are driven by the same factors.

4. Enhanced principal components regression by exploiting auxiliary information. Collinearity is a major challenge in estimating the effect of an exposure mixture. Bayesian variable selection in multiple regression tends to select one representative exposure from a correlated group which is misleading if all members of the group have significant effects (Ghosh and Ghattas (2015)). Common dimension reduction tools, such as factor analysis (FA) and principal component analysis (PCA), produce linear combinations of exposures that explain their covariance. However, the sample covariance between the exposures is not necessarily indicative of shared health effect; rather it may be an artifact of the creation, transport and fate of pollutants which can vary dramatically across studies. Therefore, we should not expect FA or PCA to return biologically-meaningful combinations. We propose new multivariate methods guided by auxiliary information to overcome this limitation.

Standard principal components regression (PCR) reduces the dimension of the exposure matrix by constructing representative linear combinations of the original exposures. The linear combinations are based on the $p \times p$ sample covariance matrix of $\mathbf{X}$ (which has been standardized), S_X. Let Γ_x be the $p \times d_x$ matrix comprised of the first d_x eigenvectors of S_X. In the regression model the original exposure vector $\mathbf{X}_i$ is replaced with the d_x linear combinations $\mathbf{X}_i^* = \mathbf{X}_i^T \Gamma_x = (X_{i1}^*, \ldots, X_{id_x}^*)^T$, and the model becomes

$$g\big[\mathrm{E}(Y_i)\big] = \mathbf{w}_i^T \boldsymbol{\alpha} + \sum_{l=1}^{d_x} X_{il}^* \gamma_{xl}, \tag{4.1}$$

where $\boldsymbol{\gamma}_x = (\gamma_{x1}, \ldots, \gamma_{xd_x})^T$ are the principal-component effects. With all other covariates held fixed, an increase of one in original exposure j leads to an increase in the linear predictor equal to the jth element of $\boldsymbol{\beta} = \Gamma_x \boldsymbol{\gamma}_x$.

An alternative PCR approach that encourages biologically-meaningful combinations of the exposures to emerge is to replace the weight matrix Γ_x derived from the sample covariance of $\mathbf{X}$ with weights based on the auxiliary data $\mathbf{Z}$, denoted as the $p \times d_z$ matrix Γ_z. Although $\mathbf{Z}$ is not a random variable, we use its covariance as a similarity measure to define groups of exposures with similar chemical structure. Let S_z be the $p \times p$ sample covariance matrix of $\mathbf{Z}$ (which has been standardized) and Γ_z be comprised of the first d_z eigenvectors of S_z. An alternative approach is to take Γ_z to be the eigenvectors of the similarity matrix defined via similarity metrics such as Tanimoto (i.e., Jaccard) similarity. If the number of auxiliary variables is small, then an option is to simply set $\Gamma_z = \mathbf{Z}$.

The eigenvectors Γ_z should lead to linear combinations $\tilde{\mathbf{X}}_i = \mathbf{X}_i^T \Gamma_z = (\tilde{X}_{i1}, \ldots, \tilde{X}_{id_x})^T$ that group together biologically-similar exposures. For example, say $\Gamma_z = \mathbf{Z}$ and that the jth auxiliary variable is binary. Then, element $\tilde{X}_{ij}$ is the sum of the exposure variables on day i that possess auxiliary feature j and thus represents the total exposure to a class of biologically-similar chemicals. If Γ_z are eigenvectors, then the constructed covariates $\tilde{X}_{ij}$ are interpreted as the value of the jth linear combination of the exposure variables for observation i. Using these constructed covariates gives the model

$$g\big[\mathrm{E}(Y_i)\big] = \mathbf{w}_i^T \boldsymbol{\alpha} + \sum_{k=1}^{d_z} \tilde{X}_{ik} \gamma_{zk}, \tag{4.2}$$

where $\boldsymbol{\gamma}_z = (\gamma_{z1}, \ldots, \gamma_{zd_z})^T$ are the PC effects. With all other covariates held fixed, an increase of one in original exposure j leads to an increase in the linear predictor equal to the jth element of $\boldsymbol{\beta} = \Gamma_z \boldsymbol{\gamma}_z$.

In most cases it will not be clear a priori whether dimension reduction based on the sample covariance or auxiliary data is preferred, and so we fit the combined model

$$g\big[\mathrm{E}(Y_i)\big] = \mathbf{w}_i^T \boldsymbol{\alpha} + \sum_{l=1}^{d_x} X_{il}^* \gamma_{xl} + \sum_{k=1}^{d_z} \tilde{X}_{ik} \gamma_{zk}. \tag{4.3}$$

The regression coefficients have priors $\gamma_{xl} \sim \text{Normal}(0, \sigma^2)$ and $\gamma_{zk} \sim \text{Normal}(0, \tau^2)$, so that the prior variances σ^2 and τ^2 determine the relative influence of the two dimension reduction approaches. In addition to studying the variance parameters and regression coefficients $\boldsymbol{\gamma}_x$ and $\boldsymbol{\gamma}_z$, we examine the posterior of the effects of the original exposures,

$$\boldsymbol{\beta} = \Gamma_x \boldsymbol{\gamma}_x + \Gamma_z \boldsymbol{\gamma}_z.$$

We use priors $\boldsymbol{\alpha} \sim \text{Normal}(0, c_1^2 \mathbf{I})$ and $\sigma, \tau \sim \text{HC}(c_2)$; however, this model is a standard generalized linear model and could easily be fit using maximum likelihood estimation.

5. Enhanced factor analysis by exploiting auxiliary information. In PCR the linear combinations of the exposure variables that appear in the response model are determined without input from the health response. The supervised factor analysis model described in this section allows for both the health and exposure data to determine the relevant linear combinations of the exposure variables.

The latent factor model is

$$\begin{aligned} g\big[\mathrm{E}(Y_i|\boldsymbol{\alpha}, \boldsymbol{\theta}_i, \boldsymbol{\delta})\big] &= \mathbf{w}_i^T \boldsymbol{\alpha} + \boldsymbol{\theta}_i^T \boldsymbol{\delta}, \\ \mathbf{X}_i|\mathbf{A}, \mathbf{D}, \boldsymbol{\theta}_i &\sim \text{Normal}(\mathbf{A}\boldsymbol{\theta}_i, \mathbf{D}), \end{aligned} \tag{5.1}$$

where $\eta_i = \boldsymbol{\theta}_i^T \boldsymbol{\delta}$ is the cumulative effect of the exposure variables, $\boldsymbol{\theta}_i \overset{\text{iid}}{\sim} \text{Normal}(0, \mathbf{I}_d)$ are the latent factors, $\mathbf{A}$ is the $p \times d$ ($d < p$) factor loading matrix with (j, l) element A_{jl} and $\mathbf{D}$ is diagonal with diagonal elements $\tau_1^2, \ldots, \tau_p^2$. In this model, $\boldsymbol{\theta}_i$ is the driving latent factor for observation i (e.g., emissions from d sources) that affects both Y_i and $\mathbf{X}_i$. Given $\boldsymbol{\theta}_i$, $\boldsymbol{\delta}$ describes the health effects and $\mathbf{A}$ relates the latent factors to the observed exposure variables. The models for health response and exposure variables are fit simultaneously in a hierarchical Bayesian approach, and so both Y_i and $\mathbf{X}_i$ provide information about $\boldsymbol{\theta}_i$.

The primary interest is not to estimate the $\boldsymbol{\theta}_i$ but rather to use the latent random effect model for dimension reduction and to pool information across responses. For example, marginally over $\boldsymbol{\theta}_i$, the covariation between the exposures is $\text{Cov}(\mathbf{X}_i) = \mathbf{A}\mathbf{A}^T + \mathbf{D}$, and the conditional mean is $\mathrm{E}(\eta_i|\mathbf{X}_i) = \mathbf{X}_i^T \boldsymbol{\beta}$ where

$$\boldsymbol{\beta} = \mathbf{D}^{-1}\mathbf{A}\big(\mathbf{A}^T\mathbf{D}^{-1}\mathbf{A} + I_d\big)^{-1}\boldsymbol{\delta}. \tag{5.2}$$

Therefore, the posterior distribution of $\boldsymbol{\beta}$ can be used for inference on individual exposure effects. The expression of $\boldsymbol{\beta}$ is a function of parameters in both the health and exposure models, confirming that both aspects of the model contribute to effect estimation.

To encourage biologically-meaningful latent factors, we specify a prior for $\mathbf{A}$ that uses auxiliary information. We center the prior for the columns of $\mathbf{A}$ on the eigenvectors of the covariance of $\mathbf{Z}$. The spectral decomposition of the $p \times p$ covariance matrix of the Z_{jl} is denoted as Γ_{jl} for the value of eigenvector l corresponding to exposure j (alternative choices for $\boldsymbol{\Gamma}$ are discussed in Section 4). For exposure j, let $\boldsymbol{\Gamma}_j = (\Gamma_{j1}, \ldots, \Gamma_{jd})^T$ be the d leading eigenvectors. The prior for $\mathbf{A}$ is then

$$A_{jl} \overset{\text{indep}}{\sim} \text{Normal}\big(b\Gamma_{jl}, \sigma^2\big). \tag{5.3}$$

Since each column of $\mathbf{A}$ is assigned a different prior mean, this serves to identify $\mathbf{A}$ as long as $b > 0$. For identification, the number of latent factors d must be less than both p and q.

Two extreme cases illustrate the behavior of this prior. If $b = 0$, then we obtain the usual Bayesian factor analysis model without auxiliary information. In the other extreme, if $b = 1$ and $\sigma = 0$, then $\mathbf{A} = \boldsymbol{\Gamma}$, and we use the principal components of $\mathbf{Z}$ to construct combinations of exposures to relate to health response. We argue that dimension reduction based on covariation of $\mathbf{Z}$, which is driven by biology, is more appropriate than dimension reduction based

on covariation in $\mathbf{X}$ which is driven by pollutant sources and other external factors such as meteorological conditions.

In the absence of prior knowledge, b and σ are treated as unknown parameters to be estimated, and thus the data determine how to balance the response, exposure and auxiliary data to estimate $\mathbf{A}$. The priors are $\boldsymbol{\alpha}, \boldsymbol{\delta} \sim \text{Normal}(0, c_1^2\mathbf{I})$, $b \sim \text{Normal}(0, c_1^2)$ restricted to $b > 0$ and $\sigma, \tau_j \sim \text{HC}(c_2)$.

6. Simulation study. In this simulation study we evaluate the benefits of incorporating auxiliary data in the analysis and compare SSVS, factor analysis and PCR in terms precision for estimating exposure effects and prior sensitivity. Since the three models given in Sections 3, 4 and 5 make very different assumptions and produce different forms of inference, it is difficult to compare them directly in a comprehensive simulation study. Therefore, we conduct separate simulation studies for the sparse SSVS model (Section 3), nonsparse PCR (Section 4) and FA (Section 5) methods.

6.1. *Linear regression simulation.* We generate the $p = 40$ exposure variables X_{ij} as Gaussian with mean zero and autoregressive covariance $\text{Cov}(X_{ij}, X_{ik}) = \rho^{|j-k|}$, independent over i. The auxiliary data are generated as Gaussian with mean zero and $\text{Cov}(Z_{jl}, Z_{jk}) = 0.5^{|l-k|}$, independent over j. For $i = 1, \ldots, n = 500$, we simulate the response Y_i as Poisson with mean $\mu_i = 10\exp(\sum_{j=1}^{p} X_{ij}\beta_j)$. The simulations vary based on the number of auxiliary variables q, the autocorrelation ρ and the true exposure effect, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, which is either:

1. Sparse: $\beta_j = 0.1(Z_{j4} + Z_{j5})_+$;
2. Nonsparse: $\beta_j = 0.05(\sum_{l=1}^{4} Z_{jl}(1 - l/5))_+$,

where $(x)_+ = \max\{0, x\}$. This set-up ensures that $\beta_j \geq 0$ for all j, as we expect increased exposure to increase health risk. Also, the relationship between the auxiliary data and the true effects is nonlinear testing robustness to our assumptions. For the sparse scenario, on average 50% of the β_j are zero compared to only 18% for the nonsparse case where most of the exposure variables have a small harmful effect. For each combination of $q \in \{5, 20\}$, $\rho \in \{0.5, 0.9\}$ and true $\boldsymbol{\beta}$ we generate 100 datasets and apply each method below to each dataset. Large q makes the effect of the auxiliary variables more difficult to estimate, and large ρ increases collinearity and thus difficulty in separating effects of correlated exposure variables.

For each dataset we fit several methods with Poisson likelihood and priors that are special cases of the SSVS model in Section 3. The models are given in Table 1. For all models we use

TABLE 1

Model descriptions: *We consider the following special cases of the Stochastic Search Variable Selection* (*SSVS*) *model for Multiple Linear Regression* (*MLR*) *given in Section* 3. *The models vary depending on their inclusion probabilities,* π_j, *whether they include the auxiliary data* $\mathbf{Z}$, *and the effects of the auxiliary data on the the inclusion probabilities* (*via* γ_{l1}) *and the effect sizes* (*via* γ_{l2})

Model	Description
MLR-NoZ	All variables included ($\pi_j = 1$) and no auxiliary data ($\mathbf{Z} = 0$).
MLR-Z	All variables included ($\pi_j = 1$) and $\mathbf{Z}$ included the effect-size prior.
SSVS-NoZ	SSVS but no auxiliary data ($\mathbf{Z} = 0$).
SSVS-Z	SSVS with $\mathbf{Z}$ in the inclusion ($\gamma_{l1} \neq 0$) but not effect-size ($\gamma_{l2} = 0$) prior.
Full	The full model in Section 3 with $\tau_1 > 0$.
Shared	The full model in Section 3 but with $\gamma_{l1} = b\gamma_{l2}$.

TABLE 2

SSVS simulation results: *Mean squared error* (*times* 1000) *for the regression coefficients* $\boldsymbol{\beta}$ *for the models in Table 1. Coverage percentage of* 90% *intervals for* $\boldsymbol{\beta}$ *is given in the subscript. Data generation varies based on the average proportion of true regression coefficients that equal zero* (*"Sparsity"*), *the number of auxiliary variables* (q) *and the autocorrelation between exposure variables* (ρ). *Mean squared error is averaged over exposure variables and dataset, and the final column gives the largest Monte Carlo standard error in each row to gauge statistical significance between methods. The final row gives the CPU time* (*seconds*) *to generate* 12,000 *iterations for the first simulation setting*

			MLR		SSVS				Max
Sparsity	q	ρ	NoZ	Z	NoZ	Z	Full	Shared	SE
0.50	5	0.5	0.13_{90}	0.12_{90}	0.06_{92}	0.05_{93}	0.05_{94}	0.05_{94}	0.00
0.50	5	0.9	0.37_{89}	0.35_{89}	0.19_{93}	0.16_{94}	0.42_{95}	0.38_{95}	0.11
0.50	20	0.5	0.12_{90}	0.12_{90}	0.06_{93}	0.06_{93}	0.06_{93}	0.06_{93}	0.00
0.50	20	0.9	0.38_{88}	0.38_{89}	0.22_{93}	0.19_{94}	0.24_{94}	0.22_{94}	0.07
0.18	5	0.5	0.14_{90}	0.14_{90}	0.12_{89}	0.12_{86}	0.10_{88}	0.10_{87}	0.00
0.18	5	0.9	0.40_{90}	0.38_{91}	0.37_{90}	0.32_{85}	0.31_{87}	0.32_{85}	0.04
0.18	20	0.5	0.14_{90}	0.14_{90}	0.12_{90}	0.12_{87}	0.12_{87}	0.12_{87}	0.00
0.18	20	0.9	0.39_{90}	0.42_{89}	0.36_{89}	0.34_{86}	0.34_{85}	0.33_{85}	0.02
CPU			54	56	178	178	178	168	

uninformative priors $c_1 = 10$ and $c_2 = 1$. We compute the posterior mean of $\boldsymbol{\beta}$ using MCMC and compute mean squared error (MSE) of $\boldsymbol{\beta}$ averaged over exposure variable and simulated dataset. Computational details are given in the Appendix.

The results are given in Table 2 and Figure 2. As expected, the SSVS model with auxiliary data has the largest reduction in MSE compared to the MLR model without the auxiliary data when the true coefficient vector is sparse. In the sparse cases the SSVS-Z model that uses auxiliary data only in the inclusion probabilities gives the best performance. In cases with

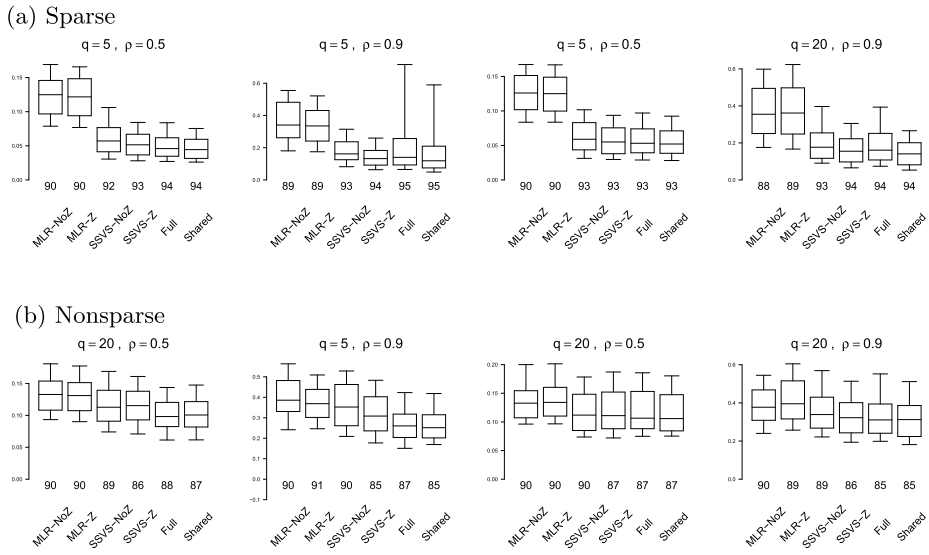

FIG. 2. SSVS simulation results: *Mean squared error* (*times* 1000) *for the regression coefficients* $\boldsymbol{\beta}$ *for the models in Table 1. Coverage percentage of* 90% *intervals for* $\boldsymbol{\beta}$ *is given along the x-axis. Data generation varies based on the average proportion of true regression coefficients that equal zero* (*"Sparse" or "Nonsparse"*), *the number of auxiliary variables* (q) *and the autocorrelation between exposure variables* (ρ). *Mean squared error is averaged over exposure variables and plotted across datasets.*

high autocorrelation ($\rho = 0.9$), the more complicated full and shared models perform poorly. Inspection of the results (Figure 2) shows that their poor performance is driven largely by a few datasets that give large errors, and thus the simpler SSVS-Z model is more stable in this most challenging case. In the less sparse cases the full and shared models have the smallest MSE.

To test for sensitivity to the hyperpriors, we fit the full model to the first and eighth simulation scenarios with all four combinations of hyperpriors $c_1, c_2 \in \{0.5, 10\}$. Across these four scenarios, the MSE (times 1000) varied from 0.050 to 0.054 across the four hyperprior settings, the coverage of 90% intervals varied from 93% to 94% for the first scenario, the MSE (times 1000) varied from 0.35 to 0.38 and the coverage of 90% intervals varied from 83% to 85% for the eighth scenario. Therefore, there is some sensitivity to hyperparameter selection in the final scenario with high autocorrelation and strong sparsity.

6.2. *Factor analysis and principal components regression simulation.* We generate the $p = 40$ exposure variables X_{ij} as Gaussian with mean zero and variance one. The exposures are independent over i and have block-diagonal correlation across exposures with within-block correlation ρ_x. The first of five blocks includes exposures $\{1, 6, \ldots, 36\}$, the second block includes exposures $\{2, 7, \ldots, 37\}$, etc. To give a distinct covariance structure from the exposure data, the $q = 20$ auxiliary variables Z_{jl} are Gaussian with mean zero, variance one, and autoregressive correlation $\mathrm{Cor}(Z_{jl}, Z_{kl}) = \rho_z^{|j-k|}$, independent over l. For $i = 1, \ldots, n = 500$, we simulate the response Y_i as Poisson with mean $\mu_i = 10\exp(\sum_{j=1}^{p} X_{ij}\beta_j)$. The true exposure effect is set to $\boldsymbol{\beta} = 0.1[(1-r)\Gamma_x \mathbf{1} + r\Gamma_z \mathbf{1}]$, where Γ_x and Γ_z are the first three eigenvectors of S_x and S_z, respectively. If $r = 0$, then the true effects are aligned with the covariance of $\mathbf{X}$; if $r = 1$, then the true effects are determined by the auxiliary data. The simulations vary based on the correlation coefficients ρ_x and ρ_z, and the true exposure effect via r. For each combination of $\rho_x, \rho_z \in \{0.5, 0.9\}$ and $r \in \{0.1, 0.9\}$, we generate 100 datasets and apply each method below to each dataset.

For each dataset we fit several methods with Poisson likelihood. The first model is standard multiple linear regression ("MLR") with $\beta_j \sim \mathrm{Normal}(0, \sigma^2)$. We also fit the FA model in Section 5 with ("FA-Z") and without ("FA-NoZ," i.e., with $b = 0$) the auxiliary data. We also fit the three PCR methods given in Section 4: PCR using only Γ_x ("PCR-X") as in (4.1), PCR using only Γ_z ("PCR-Z") as in (4.2) and the full model that uses both Γ_x and Γ_z ("PCR-XZ") as in (4.3). For all models we use uninformative priors $c_1 = 10$ and $c_2 = 1$ and select the number of dimension of the FA/PCR models (d_x and/or d_z) to explain 90% of the variation as measured by cumulative eigenvalues (of S_x and/or S_z). Computational details are given in the Appendix. We compare estimators based on MSE of the posterior mean of $\boldsymbol{\beta}$ averaged over covariate and simulated dataset.

The results are given in Table 3. For the scenarios with $r = 0.1$, the true exposure effects are largely determined by the eigenvectors of $\mathbf{X}$, and thus the auxiliary data are not influential. In these cases both FA-NoZ and PCR-X that do not use the auxiliary data provide a large reduction in MSE compared to MLR. In these cases FA is preferred to PCR when $\mathbf{X}$ has low autocorrelation ($\rho_x = 0.5$) and vice versa. The methods that include auxiliary data in these cases have higher MSE because the auxiliary data are not related to the true $\boldsymbol{\beta}$. PCR-Z has the largest MSE because it incorrectly assumes the true coefficients lie in a low-dimensional space determined by the auxiliary data. In the scenarios with $r = 0.9$ and thus the true exposure effects are largely determined by the auxiliary data, including auxiliary data in the FA model, does not provide a reduction in MSE. The PCR models with auxiliary data do reduce MSE compared with both MLR and PCR without the auxiliary data.

In summary, of the methods considered in this simulation the best vehicle to incorporate auxiliary data in the analysis appears to be the PCR-XZ model. The FA methods use both the

TABLE 3

FA and PCR simulation results: *Mean squared error* (*times* 1000) *for the regression coefficients* $\boldsymbol{\beta}$ *for the multiple linear regression* (*"MLR"*) *model without auxiliary data, factor analysis* (*"FA"*) *models with and without the auxiliary data* $\mathbf{Z}$ *and principal components regression* (*"PCR"*) *based on eigenvectors of* $\mathbf{X}$, $\mathbf{Z}$ *and both* $\mathbf{X}$ *and* $\mathbf{Z}$. *Coverage percentage of* 90% *intervals for* $\boldsymbol{\beta}$ *is given in the subscript. Data generation varies based on the proportion of the true regression coefficients that are derived from the* $\mathbf{X}$*'s eigenvectors* (r) *and the autocorrelation in* $\mathbf{X}$ (ρ_X) *and* $\mathbf{Z}$ (ρ_Z). *Mean squared error is averaged over exposure variables and dataset, and the final column gives the largest Monte Carlo standard error in each row. The final row gives the CPU time* (*seconds*) *to generate* 12,000 *iterations for the first simulation setting*

r	ρ_x	ρ_z	MLR	FA-NoZ	FA-Z	PCR-X	PCR-Z	PCR-XZ	Max SE
0.1	0.5	0.5	0.29_{92}	0.03_{100}	0.04_{100}	0.10_{91}	1.11_{35}	0.20_{92}	0.03
0.1	0.5	0.9	0.29_{92}	0.02_{99}	0.02_{98}	0.10_{91}	0.94_{28}	0.16_{91}	0.03
0.1	0.9	0.5	0.77_{95}	0.17_{99}	0.18_{99}	0.01_{63}	3.40_{41}	0.44_{91}	0.16
0.1	0.9	0.9	0.75_{96}	0.04_{98}	0.04_{97}	0.01_{66}	6.12_{22}	0.22_{91}	0.26
0.9	0.5	0.5	0.31_{92}	0.32_{83}	0.30_{86}	0.40_{60}	0.12_{85}	0.21_{91}	0.01
0.9	0.5	0.9	0.29_{92}	0.34_{73}	0.35_{72}	0.35_{64}	0.07_{83}	0.15_{92}	0.01
0.9	0.9	0.5	0.80_{95}	0.50_{94}	0.52_{94}	0.52_{10}	0.38_{88}	0.49_{91}	0.02
0.9	0.9	0.9	0.78_{96}	0.42_{74}	0.43_{73}	0.43_{10}	0.24_{83}	0.27_{89}	0.02
CPU			60	287	287	34	24	48	

covariance of $\mathbf{X}$ and the regression relationship with Y to estimate the latent factor matrix, but for data with low signal-to-noise ratio, such as air pollution studies, the covariance of $\mathbf{X}$ overwhelms the regression on $\mathbf{Y}$ and latent factor estimation is mostly driven by the covariance of $\mathbf{X}$ regardless of whether auxiliary data is included in the prior. The PCR-Z model is optimal when the true regression coefficients are largely determined by the auxiliary data but is worse than MLR when this assumption is violated. PCR-XZ is robust to this form of model misspecification and consistently outperforms MLR.

Finally, to test for sensitivity to the hyperpriors we fit the FA-Z and PCR-XZ models to the first and eighth simulation scenarios with hyperpriors $c_1 = c_2 = c \in \{0.5, 10\}$. For FA-Z, in the first scenario the MSE (times 1000) is 0.03 (coverage 100%) for $c = 0.5$ and 0.04 (100%) for $c = 10$; for the eighth scenario the MSE is 0.42 (74%) for $c = 0.5$ and 0.43 (73%) for $c = 10$. For PCR-XZ, in the first scenario the MSE is 0.20 (coverage 92%) for both $c = 0.5$ and $c = 10$; for the eighth scenario the MSE is 0.27 (89%) for both $c = 0.5$ and $c = 10$. Therefore, the PCR-XZ methods appears to be the least sensitive to hyperprior selection.

7. Analysis of the Atlanta ED visits data. We estimate short-term associations between ED visits and mixtures of 44 ambient VOC pollutants in Atlanta during 1998 to 2008. We expand the analysis previously conducted by Ye et al. (2017) by incorporating a suite of 74 quantitative auxiliary descriptors on chemical structure instead of assigning pollutants to distinct chemical structure groups.

7.1. *Model comparisons.* To compare methods, we use the same models, priors and computational details as in the simulation study except that we do not fit the factor analysis models because of their poor performance in the simulation study. We also fit these models with a negative binomial likelihood to account for overdispersion but found no marked improvement over the Poisson likelihood, and so these results are excluded. The potential confounder variables $\mathbf{w}_i$ include: natural cubic splines of calendar date with monthly knots, cubic functions of same-day maximum temperature, dew-point temperature, indicators of day-of-week, holidays, seasons, season-day-of-week interactions, season-temperature interactions and indicators for hospital participation periods. There are $n = 3216$ observations (days), $p = 44$

TABLE 4

Cross-validation results for the Atlanta emergency department visit data: *Root mean squared error* (*"RMSE"*) *and test set deviance* (*"Dev"*; *reported as the reduction from the MLR-NoZ model*) *for asthma and wheezing, respiratory and cardiovascular diseases for linear regression models in Table 1 and principal components regression* (*PCR*) *based on the observed exposures* $\mathbf{X}$, *the auxiliary data* $\mathbf{Z}$ *or both*

	Asthma		Respiratory		Cardiovascular	
Model	RMSE	Dev	RMSE	Dev	RMSE	Dev
MLR-NoZ	11.3	0	138.3	0	9.0	0
MLR-Z	10.9	−97	106.8	−3341	9.0	0
SSVS-NoZ	11.6	55	111.7	−2838	9.0	−6
SSVS-Z	11.0	−88	125.0	−1468	9.0	−9
FULL	11.0	−93	125.1	−1425	9.0	−4
SHARED	10.8	−116	106.7	−3342	9.0	−6
PCR-X	12.4	177	113.5	−2653	9.0	−11
PCR-Z	10.9	−104	99.5	−4111	9.0	−6
PCR-XZ	11.5	58	118.7	−2163	9.0	−11

exposure variables and $q = 74$ auxiliary variables. Following Ye et al. (2017), we use the three-day moving average of lags 0, 1 and 2 for asthma and respiratory emergency department visits and the lag 0 concentration for cardiovascular visits. For the PCR models we use $d_x = d_z = 5$ eigenvectors which explains 90% of the variability of S_x and S_z. Because of the large number of auxiliary variables and their collinearity, we use these five eigenvectors of S_z as the auxiliary data $\mathbf{Z}$ for the SSVS models.

We compare models using five-fold cross-validation with days randomly assigned across folds. The metrics for comparison are root mean square prediction error and the test-set deviance, that is, twice the negative log likelihood of the test set data given the posterior mean of the model parameters, summed over all observations and cross-validation folds. Table 4 gives the difference in deviance between each model and the baseline multiple regression model that excludes auxiliary data; negative values indicate better fit.

With few exceptions the advanced methods improve fit compared to the standard Poisson regression model. The SSVS model with shared structure in the inclusion probability and effect size priors has the smallest RMSE and deviance for asthma and wheeze. For the respiratory response the PCR model based on the eigen decomposition of the auxiliary data has the best performance. Perhaps due to the small number of events, none of the exposures are identified as harmful for the cardiovascular response, and the results are fairly similar for all models. Therefore, we do not discuss this health outcome below.

7.2. *Results*. The posterior of the exposure effects β_j for asthma and wheeze ED visits are plotted in Figures 3 and 4 (and the PCR parameter estimates are in Table 5). None of the inclusion probabilities are close to zero or one, likely due to collinearity (Ghosh and Ghattas (2015)). As in the standard epidemiological analysis of Ye et al. (2017), the estimated relative risks of the individual exposures are no more than $\exp(0.01) \approx 1.01$, and so studying the total effect requires considering the entire mixture. Compared to the standard SSVS model without auxiliary data, the shared model has smaller inclusion probability (i.e., $\text{Prob}(\beta_j \neq 0|\mathbf{Y})$) than the SSVS standard model for most of the variables with small estimated effects size but similar inclusion probability variables with larger estimated effect size. Therefore, including the auxiliary data gives a smaller model without attenuating the estimates effects of the variables with high inclusion probability. The posterior 95% interval for b, the parameter that controls the relationship between the model for inclusion probability and effect size in (3.3), is $(-6.0, 6.1)$, and, since this parameter is centered on zero, it appears the auxiliary data

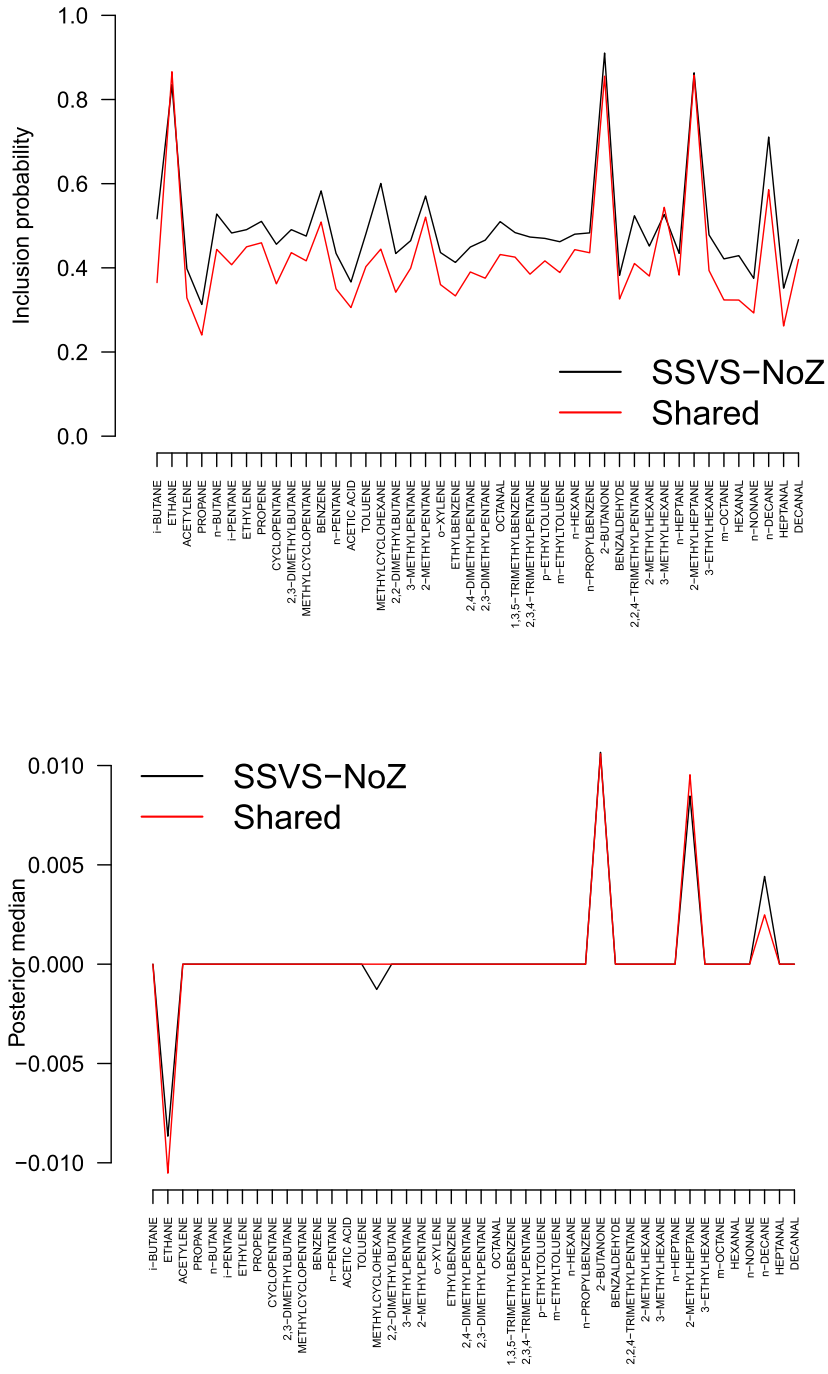

FIG. 3. SSVS results for asthma and wheeze. *Plotted are the posterior inclusion probabilities* (*i.e.*, *the posterior probability that* $\beta_j \neq 0$) *and posterior medians for* β_j *for the SSVS models without auxiliary data* (*"SSVS-NoZ"*) *and the SSVS-shared model. The exposures are ordered by the posterior mean of the shared auxiliary factor,* $\mu_j = \sum_{l=1}^{q} Z_{jl}\gamma_{l2}$, *for the shared model.*

affects the expected effect size more than the inclusion probability. The strongest positive association with 2-Butanone was also identified in Ye et al. (2017). However, in this analysis all 44 VOC pollutants were considered jointly, while Ye et al. (2017) examined associations with 2-Butanone in a multipollutant model controlling for different small sets of pollutants. Even though the shared model is the best fitting model, the PCR-Z model fits nearly as well.

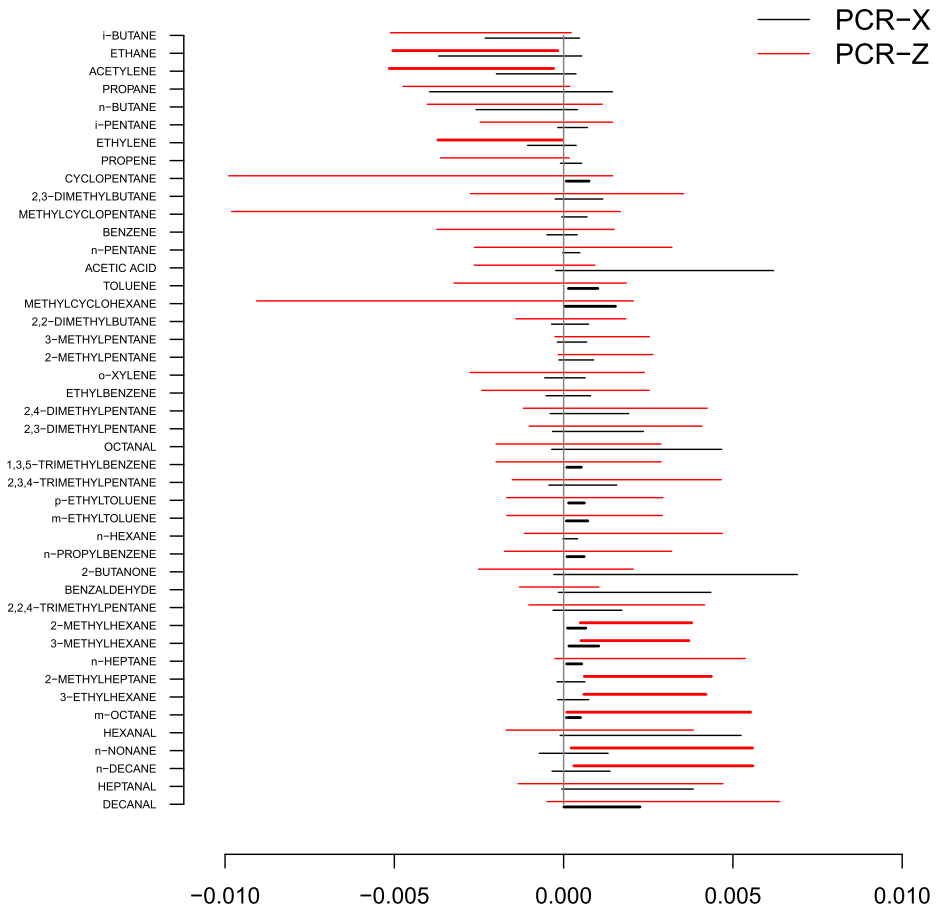

FIG. 4. PCR results for asthma and wheeze. *Posterior 95% intervals for PCR regression using eigenvectors of the correlation of* $\mathbf{X}$ *versus* $\mathbf{Z}$. *Thick lines indicate intervals that exclude zero. The exposures are ordered by the posterior mean of the shared auxiliary factor,* $\mu_j = \sum_{l=1}^{q} Z_{jl}\gamma_{l2}$, *for the shared model.*

Comparing Figures 3 and 4 shows that the PCR-Z model identifies more harmful exposures because it smooths effects across exposures with similar chemical structure. For example, both models flag 2-Methylheptane and n-Decane, but PCR-Z also identifies chemically similar exposures such as 3-Ethylhexane and n-Nonane as harmful. In Ye et al. (2017), asthma ED visits were also found to be associated with these groups of alkanes.

The PCR method based on the auxiliary data is the best fitting model for the all respiratory ED visits. Figure 5 compares the posteriors for PCR-X and PCR-Z. Both methods identify a large group of exposures at the bottom of Figure 5 as being associated with increase risk of respiratory problems. This group includes various alkane hydrocarbons and are primarily emitted from traffic or other combustion sources. Because this group of exposures is also correlated (top left of Figure 1) with many other exposures, the PCR-X model also flags other exposures such as toluene. However, since toluene is an aromatic hydrocarbon and is chemically dissimilar (bottom right of Figure 1) to the alkanes, it is not deemed harmful by the PCR-Z model.

Thus far we have examined the results via the posterior of the regression coefficients of individual exposures. An advantage of incorporating the auxilary data into the analysis is that is can reveal the chemical properties that are common to harmful exposures. For both asthma and all respiratory ED visits, only the first eigenvector in the PCR-Z model has a

TABLE 5

Parameter estimates for the PCR-XZ model: *Posterior median* (*equal-tailed* 95% *interval*) *for the effect of PC's of* $\mathbf{X}$ (γ_{xl}) *and* $\mathbf{Z}$ (γ_{zk}) *and their prior variances* $\mathrm{Var}(\gamma_{xl}) = \sigma^2$ *and* $\mathrm{Var}(\gamma_{zk}) = \tau^2$. *All values are multiplied by* 1000

	PCR-XZ		PCR-X	
Parameter	Median	95% interval	Median	95% interval
γ_{x1}	0.01	$(-0.35, 0.43)$	-0.14	$(-0.21, -0.08)$
γ_{x2}	0.08	$(-1.23, 2.24)$	0.58	$(-0.24, 1.53)$
γ_{x3}	-0.78	$(-5.84, 0.36)$	-2.18	$(-4.72, 0.00)$
γ_{x4}	0.21	$(-0.98, 4.54)$	0.50	$(-1.13, 2.76)$
γ_{x5}	0.20	$(-1.02, 4.51)$	0.18	$(-1.56, 2.21)$
γ_{z1}	-1.39	$(-5.76, 2.64)$		
γ_{z2}	1.16	$(-2.44, 4.32)$		
γ_{z3}	0.60	$(-2.51, 4.71)$		
γ_{z4}	-6.56	$(-11.69, -1.47)$		
γ_{z5}	-2.74	$(-10.25, 1.69)$		
σ	0.90	$(0.01, 6.10)$	1.45	$(0.18, 4.79)$
τ	4.68	$(1.37, 13.75)$		

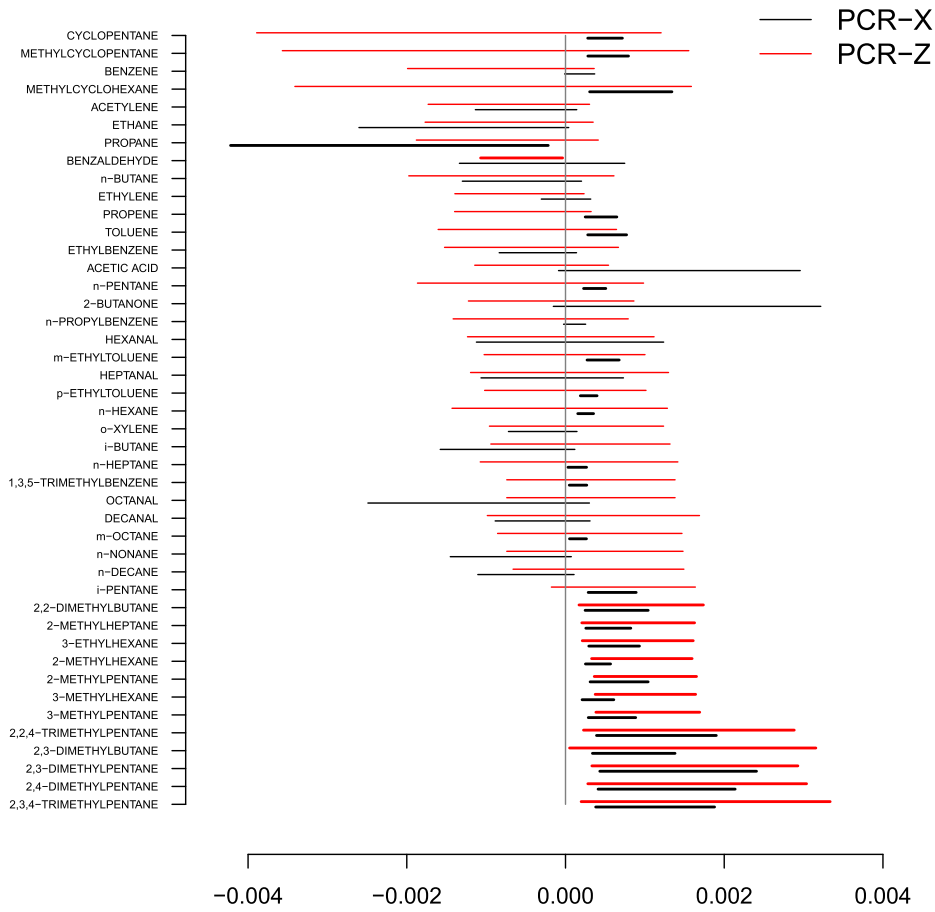

FIG. 5. PCR results for respiratory diseases. *Posterior* 95% *intervals for PCR regression using eigenvectors of the correlation of* $\mathbf{X}$ *versus* $\mathbf{Z}$. *Thick lines indicate intervals that exclude zero. The exposures are ordered by the posterior mean* $\boldsymbol{\beta}$ *under the PCR-Z model.*

credible interval that excludes zero, that is, the posterior probability that γ_{z1} is negative is larger than 0.975. To summarize which chemical properties contribute the most to the first eigenvector Γ_{z1}, we compute $\mathbf{Z}\Gamma_{z1}$, which is proportional to the correlation between each property, and Γ_{z1}. Since Γ_{z1} has a negative relationship with the response, properties with negative (positive) correlation with Γ_{z1} were deemed harmful (protective). The properties (`rcdk` name) with the most negative correlation are the fraction of carbon atoms that are SP3 hybridized (`FractionCSP3`) and the Hall–Kier alpha coefficient (`HallKierAlpha`). The properties with the strongest positive correlation are the number of rings (`NumRings`) and the sum of log surface area for atoms with contribution between 20–25% (`slogp-VSA6`).

8. Conclusions. In this paper we propose a suite of statistical methods that incorporate auxiliary data to estimate the health effects of a mixture of exposure variables. Cheminformatics (Dragon (2019), Guha (2007), Landrum (2019)) and in-vitro (Filer et al. (2016)) characteristics are widely available in the toxicology literature, and therefore methods to incorporate this information can have a broad impact. We found, via simulation studies, that supplementing regression models with auxiliary data improves the precision of estimating the effects of individual exposure variables, and supplementing principal components regression with auxiliary data improves the precision of estimating the effects of linear combinations of individual exposure variables. Including auxiliary data in the factor analysis model, at least using the structure imposed in this paper, did not improve the statistical analysis. When applied to study emergency department visits for respiratory diseases, the principal components regression model with auxiliary data gave better fit that other methods and resulted in a biologically-plausible result.

While stochastic search variable selection is also effective when the signal is very sparse, we found that the principal components regression model that includes principal components of both concentration and auxiliary data (PCR-XZ) was overall the top-performing model. This is the simplest auxiliary-data method we considered, and our simulation study showed that it is fast, insensitive to hyperprior selection, had nominal coverage in all scenarios and performed well in the presence of high correlation between exposures which is common in mixture analyses.

Although the proposed methods are designed to handle a large number of exposure variables, they are not designed to handle a large number of auxiliary variables. In Section 7 we used principal components of the auxiliary data to reduce the dimension, but future work might incorporate stochastic search variable selection to reduce its dimension. Also, our models thus far use only a single source of auxiliary data, namely, cheminformatics variables, but future applications might include data from additional sources, such as in-vitro analysis, and the statistical methodology will need to be enhanced to allow for separate effects for the different types of auxiliary data. Another limitation is that we consider only linear effects. It should be possible to use auxiliary information in the inclusion probability of Bayesian variable selection methods for nonlinear regressions (e.g., Antonelli et al. (2020), Bobb et al. (2015), Fang et al. (2019), Sabanés Bové, Held and Kauermann (2015), Wei et al. (2020)) as in our linear SSVS model, but incorporating auxiliary data in the model for the nonlinear effect is not straightforward.

APPENDIX: MCMC DETAILS

The methods proposed in Sections 3, 4 and 5 could easily be coded in all-purpose Bayesian software, including `OpenBUGS`, `JAGS` or `NIMBLE`. For faster computation we coded the MCMC sampler in `R`. The updates are all standard Gibbs or Metropolis steps, and the code is included in the Supplemental Material (Reich et al. (2020)). For the simulation study we

used 10,000 MCMC samples after discarding the first 2000 as burn-in. For the data analysis in Section 7, the MCMC burn-in is increased to 10,000, and the total number of iterations is increased to 25,000.

Acknowledgments. This work was supported by the National Institutes of Health (ES025128, ES027892, ES031651 and ES028526), the National Science Foundation (DMS-1638521), the Electric Power Research Institute (EPRI, 10002467) and the U.S. Environmental Protection Agency (USEPA, RD834799). The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication. The air pollutant data were supported by the Southeastern Aerosol Research and Characterization (SEARCH) Network, the Electric Power Research Institute and the Atmospheric Research Associates.

SUPPLEMENTARY MATERIAL

R code (DOI: 10.1214/20-AOAS1364SUPP; .zip). We provide R code to implement the methods proposed in the manuscript.

REFERENCES

ANTONELLI, J., MAZUMDAR, M., BELLINGER, D., CHRISTIANI, D., WRIGHT, R. and COULL, B. (2020). Estimating the health effects of environmental mixtures using Bayesian semiparametric regression and sparsity inducing priors. *Ann. Appl. Stat.* **14** 257–275. MR4085093 https://doi.org/10.1214/19-AOAS1307

ATKINSON, R. W., MILLS, I. C., WALTON, H. A. and ANDERSON, H. R. (2015). Fine particle components and health—a systematic review and meta-analysis of epidemiological time series studies of daily mortality and hospital admissions. *J. Expo. Sci. Environ. Epidemiol.* **25** 208–214. https://doi.org/10.1038/jes.2014.63

BILLIONNET, C., SHERRILL, D., ANNESI-MAESANO, I. and GERIE STUDY (2012). Estimating the health effects of exposure to multi-pollutant mixture. *Ann. Epidemiol.* **22** 126–141.

BOBB, J. F., VALERI, L., CLAUS HENN, B., CHRISTIANI, D. C., WRIGHT, R. O., MAZUMDAR, M., GODLESKI, J. J. and COULL, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* **16** 493–508. MR3365442 https://doi.org/10.1093/biostatistics/kxu058

CARRICO, C., GENNINGS, C., WHEELER, D. C. and FACTOR-LITVAK, P. (2015). Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *J. Agric. Biol. Environ. Stat.* **20** 100–120. MR3334469 https://doi.org/10.1007/s13253-014-0180-3

CHEN, R., YIN, P., MENG, X., LIU, C., WANG, L., XU, X., ROSS, J. A., TSE, L. A., ZHAO, Z. et al. (2017). Fine particulate air pollution and daily mortality. A nationwide analysis in 272 Chinese cities. *Am. J. Respir. Crit. Care Med.* **196** 73–81. https://doi.org/10.1164/rccm.201609-1862OC

NATIONAL RESEARCH COUNCIL AND OTHERS (2004a). *Air Quality Management in the United States*. National Academies Press, Washington, DC.

NATIONAL RESEARCH COUNCIL AND OTHERS (2004b). *Research Priorities for Airborne Particulate Matter: IV. Continuing Research Progress* **4**. National Academies Press, Washington, DC.

DOMINICI, F., WANG, C., CRAINICEANU, C. and PARMIGIANI, G. (2008). Model selection and health effect estimation in environmental epidemiology. *Epidemiology* **19** 558–560.

DOMINICI, F., PENG, R. D., BARR, C. D. and BELL, M. L. (2010). Protecting human health from air pollution: Shifting from a single-pollutant to a multi-pollutant approach. *Epidemiology* **21** 187.

DRAGON (2019). Dragon. Available at http://www.talete.mi.it/products/dragon_description.htm.

FANG, X., FANG, B., WANG, C., XIA, T., BOTTAI, M., FANG, F. and CAO, Y. (2019). Comparison of frequentist and Bayesian generalized additive models for assessing the association between daily exposure to fine particles and respiratory mortality: A simulation study. *Int. J. Environ. Res. Public Health* **16** 746.

FILER, D. L., KOTHIYA, P., SETZER, R. W., JUDSON, R. S. and MARTIN, M. T. (2016). tcpl: The ToxCast pipeline for high-throughput screening data. *Bioinformatics* **33** 618–620.

FOURCHES, D., MURATOV, E. and TROPSHA, A. (2010). Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* **50** 1189–1204.

GELMAN, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* **1** 515–533. MR2221284 https://doi.org/10.1214/06-BA117A

GEORGE, E. I. and MCCULLOCH, R. E. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–373.

GHOSH, J. and GHATTAS, A. E. (2015). Bayesian variable selection under collinearity. *Amer. Statist.* **69** 165–173. MR3391636 https://doi.org/10.1080/00031305.2015.1031827

GUHA, R. (2007). Chemical informatics functionality in R. *J. Stat. Softw.* **18**.

HANSEN, D. A., EDGERTON, E., HARTSELL, B., JANSEN, J., BURGE, H., KOUTRAKIS, P., ROGERS, C., SUH, H., CHOW, J. et al. (2006). Air quality measurements for the aerosol research and inhalation epidemiology study. *J. Air Waste Manage. Assoc.* **56** 1445–1458.

KWON, D., LANDI, M. T., VANNUCCI, M., ISSAQ, H. J., PRIETO, D. and PFEIFFER, R. M. (2011). An efficient stochastic search for Bayesian variable selection with high-dimensional correlated predictors. *Comput. Statist. Data Anal.* **55** 2807–2818. MR2811867 https://doi.org/10.1016/j.csda.2011.04.019

LANDRUM, G. (2019). RDKit: Open-source cheminformatics. Available at http://www.rdkit.org.

LI, C. and LI, H. (2010). Variable selection and regression analysis for graph-structured covariates with an application to genomics. *Ann. Appl. Stat.* **4** 1498–1516. MR2758338 https://doi.org/10.1214/10-AOAS332

LI, Z., LIU, Z., ZHONG, W., HUANG, M., WU, N., XIE, Y., DAI, Z. and ZOU, X. (2016). Large-scale identification of human protein function using topological features of interaction network. *Sci. Rep.* **6** 37179.

MACLEHOSE, R. F., DUNSON, D. B., HERRING, A. H. and HOPPIN, J. A. (2015). Bayesian methods for highly correlated exposure data. *Biostatistics* **16** 493–508.

MOLITER, J., PAPATHOMAS, M., JERRETT, M. and RICHARDSON, S. (2010). Bayesian profile regression with an application to the National Survey of Children's Health. *Biostatistics* **11** 484–498.

MORTIMER, K., NEUGEBAUER, R., LURMANN, F., ALCORN, S., BALMES, J. and TAGER, I. (2008). Air pollution and pulmonary function in asthmatic children: Effects of prenatal and lifetime exposures. *Epidemiology* **19** 550–557.

O'HARA, R. B. and SILLANPÄÄ, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Anal.* **4** 85–117. MR2486240 https://doi.org/10.1214/09-BA403

PEARCE, J. L., WALLER, L. A., SARNAT, S. E., CHANG, H. H., KLEIN, M., MULHOLLAND, J. A. and TOLBERT, P. E. (2016). Characterizing the spatial distribution of multiple pollutants and populations at risk in Atlanta, Georgia. *Spat. Spatio-Tempor. Epidemiol.* **18** 13–23.

REICH, B. J., GUAN, Y., FOURCHES, D., WARREN, J. L., SARNET, S. E. and CHANG, H. H. (2020). Supplement to "Integrative statistical methods for exposure mixtures and health." https://doi.org/10.1214/20-AOAS1364SUPP

SABANÉS BOVÉ, D., HELD, L. and KAUERMANN, G. (2015). Objective Bayesian model selection in generalized additive models with penalized splines. *J. Comput. Graph. Statist.* **24** 394–415. MR3357387 https://doi.org/10.1080/10618600.2014.912136

SHAH, A. S. V., LEE, K. K., MCALLISTER, D. A., HUNTER, A., NAIR, H., WHITELEY, W., LANGRISH, J. P., NEWBY, D. E. and MILLS, N. L. (2015). Short term exposure to air pollution and stroke: Systematic review and meta-analysis. *BMJ* **350** h1295. https://doi.org/10.1136/bmj.h1295

SINISI, S. E. and VAN DER LAAN, M. J. (2004). Deletion/substitution/addition algorithm in learning with applications in genomics. *Stat. Appl. Genet. Mol. Biol.* **3** 18. MR2101467 https://doi.org/10.2202/1544-6115.1069

STIEB, D. M., CHEN, L., ESHOUL, M. and JUDEK, S. (2012). Ambient air pollution, birth weight and preterm birth: A systematic review and meta-analysis. *Environ. Res.* **117** 100–111.

STRICKLAND, M. J., DARROW, L. A., KLEIN, M., FLANDERS, W. D., SARNAT, J. A., WALLER, L. A., SARNAT, S. E., MULHOLLAND, J. A. and TOLBERT, P. E. (2010). Short-term associations between ambient air pollutants and pediatric asthma emergency department visits. *Am. J. Respir. Crit. Care Med.* **2010** 307–316.

SUH, H. H., ZANOBETTI, A., SCHWARTZ, J. and COULL, B. A. (2011). Chemical properties of air pollutants and cause-specific hospital admissions among the elderly in Atlanta, Georgia. *Environ. Health Perspect.* **119** 1421–1428.

SUN, Z., TAO, Y., LI, S., FERGUSON, K. K., MEEKER, J. D., PARK, S. K., BATTERMAN, S. A. and MUKHERJEE, B. (2013). Statistical strategies for constructing health risk models with multiple pollutants and their interactions: Possible choices and comparisons. *Environ. Health* **12** 85.

VRIJHEID, M., MARTINEZ, D., MANZANARES, S., DADVAND, P., SCHEMBARI, A., RANKIN, J. and NIEUWENHUIJSEN, M. (2011). Ambient air pollution and risk of congenital anomalies: A systematic review and meta-analysis. *Environ. Health Perspect.* **119** 598–606. https://doi.org/10.1289/ehp.1002946

WEI, R., REICH, B. J., HOPPIN, J. A. and GHOSAL, S. (2020). Sparse Bayesian additive nonparametric regression with application to health effects of pesticides mixtures. *Statist. Sinica* **30** 55–79.

YE, D., KLEIN, M., CHANG, H. H., SARNAT, J. A., MULHOLLAND, J. A., EDGERTON, E. S., WINQUIST, A., TOLBERT, P. E. and SARNAT, S. E. (2017). Estimating acute cardiorespiratory effects of ambient volatile organic compounds. *Epidemiology* **28** 197–206. https://doi.org/10.1097/EDE.0000000000000607

YE, D., KLEIN, M., MULHOLLAND, J. A., RUSSELL, A. G., WEBER, R., EDGERTON, E. S., CHANG, H. H., SARNAT, J. A., TOLBERT, P. E. et al. (2018). Estimating acute cardiovascular effects of ambient PM. *Environ. Health Perspect.* **126** 027007. https://doi.org/10.1289/EHP2182

ZHENG, X., DING, H., JIANG, L., CHEN, S., ZHENG, J., QIU, M., ZHOU, Y., CHEN, Q. and GUAN, W. (2015). Association between air pollutants and asthma emergency room visits and hospital admissions in time series studies: A systematic review and meta-analysis. *PLoS ONE* **10** e0138146. https://doi.org/10.1371/journal.pone.0138146