NIEHS Statistical Approaches for Assessing Health Effects of Environmental Chemical Mixtures in Epidemiology Studies Workshop: Presentation Abstracts

Table of Contents

1. Combining Visualization Methods with Structural Equation Models for the Analysis Of Chemical Mixture Data 1
2. A Methodology for Building Bayesian Networks to Evaluate the Health Effects of Environmental Chemical Mixtures 5
3. A Statistical Framework for Assessing Association Between Health Effects and Combined Exposures 7
4. Bayesian Kernel Machine Regression for Estimating the Health Effects of Multi-Pollutant Mixtures .. 10
5. Cheminformatics Approaches to Analyze the Effects of Environmental Chemical Mixtures 14
6. A Two Step Variable Selection Procedure Using Prioritization of Interactions Followed by LASSO 17
7. Integrating Toxicology and Mechanistic Evidence Into Complex Mixtures Analysis Using a Flexible Bayesian Approach 21
8. Two-Step Shrinkage-Based Regression Strategy for Assessing Health Effects of Chemical Mixtures in Environmental Epidemiology 24
9. A Two Stage Approach to Analysis of Health Effects of Environmental Chemical Mixtures: Informed Sparse Principal Component Analysis Followed by Segmented Regression 30
10. Direct Assessment of Public Health Impacts of Exposure Mixtures: A Bayesian G-Formula Approach 34
11. Do Your Exposures Need Supervision? 38
12. Principal Component Analysis: An Application for Understanding Health Effects of Environmental Chemical Mixture Exposures 42
13. Interpretation Without Causation: A Data Analysis at the Intersection of Statistics and Epidemiology 47
14. Examining Associations Between Multi Pollutant Exposure Profiles and Health Outcomes via Bayesian Profile Regression 50
15. Analysis of Simulated Data Sets using Conformal Predictions 55

16. Analysis of Chemical Mixture Simulated Data Using Regularized Regression Models 58
17. Building Models to Assess the Effects of Chemical Mixtures by Estimating Similar Modes of Action .. 62
18. Application of Principal Component Analysis and Stepwise Regression to Identify the Exposure Variables Associated with Health Outcome and to Determine Dose-Response Relationship..... 66
19. Identifying the Relative Importance of Multiple Correlated Exposures in Predicting a Continuous Outcome Using the Random Forest Ensemble Learning Method..... 70
20. Improving Prediction Models by Adding Interaction Terms Using a Feasible Solution Algorithm 75
21. Factor Mixture Models for Assessing Health Effects of Environmental Chemical Mixtures: An Application Using Simulated Data Sets 76
22. Dimension Reduction for Chemical Exposure Risk Assessment..... 80
23. Set-based Interaction Tests for High-Dimensional Environmental Exposome Data 84
24. Analyzing Mixtures in Epidemiology Data by Smoothing in Exposure Space..... 88
25. Variable Selection and Multivariate Adaptive Spline Assessments to Investigate Effects of Chemical Mixtures in a Prospective Cohort Study of Mother-Child Pairs 91
26. Bayesian Non-Parametric Regression for Multi-Pollutant Mixtures..... 96
27. Modeling Environmental Chemical Mixtures with Weighted Quantile Sum Regression..... 99
28. Assessing Health Associations with Environmental Chemical Mixtures using LASSO and its Generalizations..... 105
29. Assessing the Impact of Environmental Mixtures on Children's Neurodevelopment 111
30. Analysis of the First Simulated Dataset using Nonlinear and Weighted Quantile Sum (WQS) Regression 120
31. Bayesian Methods for Assessing Health Effects of Chemical Mixtures 124
32. Assessing Health Effects of Environmental Chemical Mixtures Using Stepwise Multiple Linear Regression 125
33. Traditional Epidemiological Approaches to Analyze Chemical Mixtures and Human Health..... 127

1. Combining Visualization Methods with Structural Equation Models for the Analysis Of Chemical Mixture Data

Presenting Author: Sophia Banton

Organization: Emory University

Contributing Authors: Sophia A. Banton[1, 2], Ruiyan Luo[2], and Shuzhao Li[1]
[1]*Department of Medicine, Emory University, Atlanta, Georgia*
[2]*School of Public Health, Georgia State University, Atlanta Georgia*

Abstract:

Introduction: Epidemiological studies frequently contain numerous exposure variables that function as predictors and confounders in the statistical analyses that describe them. Given the large number of variables of interest, data reduction methods that preserve the data quality are of utmost importance. It is often necessary in environmental studies to extract a small selection of variables from a larger measurement set for use in hypothesis generation and/or testing. We thus propose combining visualization methods with structural equation models for the analysis of these chemical mixture data.

Methods: We used the two simulated data sets and the real world dataset representing chemical exposures and outcomes from this workshop. Each data set represented a prospective cohort, a cross-sectional study, and a prospective cohort, respectively. The predictors for the first simulated data set and the real world data set were log transformed to achieve an approximately multivariate normal distribution. The second data set was multivariate normal and no transformations were necessary. Following transformation, principal component analysis (PCA) was used to observe the spread of the independent variables alone and then in the presence of the outcome (Y). Next, Spearman correlations (Spearman's $\rho > 0.6$, $p < 0.01$) were used to generate matrixes and correlations to visualize the relationships between and among variables. Hierarchical clustering was used to evaluate the influence of covariates as statistical confounders and visualized with heat maps, where categorical covariates were treated as class labels. Finally, the combination of the patterns detected above and the dendrograms constructed in parallel with the heat maps were used to generate a structural equation model (SEM) of observed variables (path analysis) to explain the relationships among the variables. The SEM network consists of a system of simultaneous linear regression equations that explain the relationships between the variables in each data set. All statistical analyses were conducted using R. All figures were generated using R with the exception of the PCA plots that were generated using Python.

Results: With the assistance of data visualization, a structural equation model was derived for each data set. The patterns observed for each graphical representation of the data were useful in the construction of the statistical model.

For the first data set, the PCA plot of all variables (Figure 1A), alongside the heat map (Figure 1C), demonstrated that covariate Z was confounding with Y and multiple independent variables. Spearman correlation in the form of a matrix (Figure 1B) and correlation plot (Figure 1D) in combination with

hierarchical clustering revealed that there were two major clusters of variables, both in the absence and presence of the outcome Y. The influences of the variables were supported by the SEM network ($p < 0.0001$) (Figure 1E), in which the edges represent the magnitude of the path coefficients. As shown by the SEM system, the variables that contribute most to the outcome Y are X1, X4, X5, and X7 while controlling for the confounding effects of Z. The variables X2 and X3 are highly correlated with X1 and therefore indirectly affect the outcome Y. A similar pattern was observed for variable X6, which is highly correlated with X5.

For the second data set, the PCA plot of all variables (Figure 2A), alongside the heat map (Figure 2C) showed no confounding influence of any covariates (Z1 yellow and Z2 pink). Rather, covariate Z1 was shown to be uncorrelated with any of the variables analyzed and was subsequently removed from the analysis. The Spearman correlation matrix (Figure 2B) and network (Figure 2D) revealed that the second covariate Z2 was weakly correlated with a number of independent variables, and the third covariate (Z3) was only associated with the outcome variable. These observations were also supported by the SEM system ($p < 0.0001$) (Figure 2E). As shown by the path model, the variables that contribute most to the outcome Y are X8, X14, and Z3. The presence of X8 and X14 in the fitted model is interchangeable. Several variables (X4, X11, and X12, X4) directly influence X14, and this is significant since X14 links the outcome Y to remaining variables in the model. In the absence of variable identifications, the inference can be made that biomarkers X14 and X8 are on the causal path to Y, if Y is indeed a true outcome and not a covariate itself.

For the real world data set, the PCA plot (Figure 3A) showed that the four PBDE chemicals (dark red) and the remaining chemicals (PCBs, DDE, chlordanes, and HCB) formed two distinct clusters. The Spearman correlation matrix (Figure 3B) and network (Figure 3C) revealed that the two clusters were connected to the outcome MDI through the maternal age at delivery predictor. The strong association of age with the chemicals and covariates is demonstrated by the heat map in which the subjects are classified by age group (Figure 3D). These observations were also supported by the SEM network ($p < 0.0001$) (Figure 3E). As shown by the path model, the variables that contribute directly to the outcome MDI are the covariates for maternal education and maternal race. These variables are in turn associated with maternal smoking status and maternal age at delivery, respectively. Maternal age is directly related to the exposures PBDE47, pcb74, and pcb194, and each of these chemicals is a hub for the remaining chemical exposures in the data set.

Conclusions: Our analyses indicate that the combination of data visualization and structural equation models can be useful for the assessment of chemical exposures in epidemiology. While Spearman correlation was used in the step of variable selection, future research will examine the utility of alternatives, including partial correlation and mutual information.

Figure 1

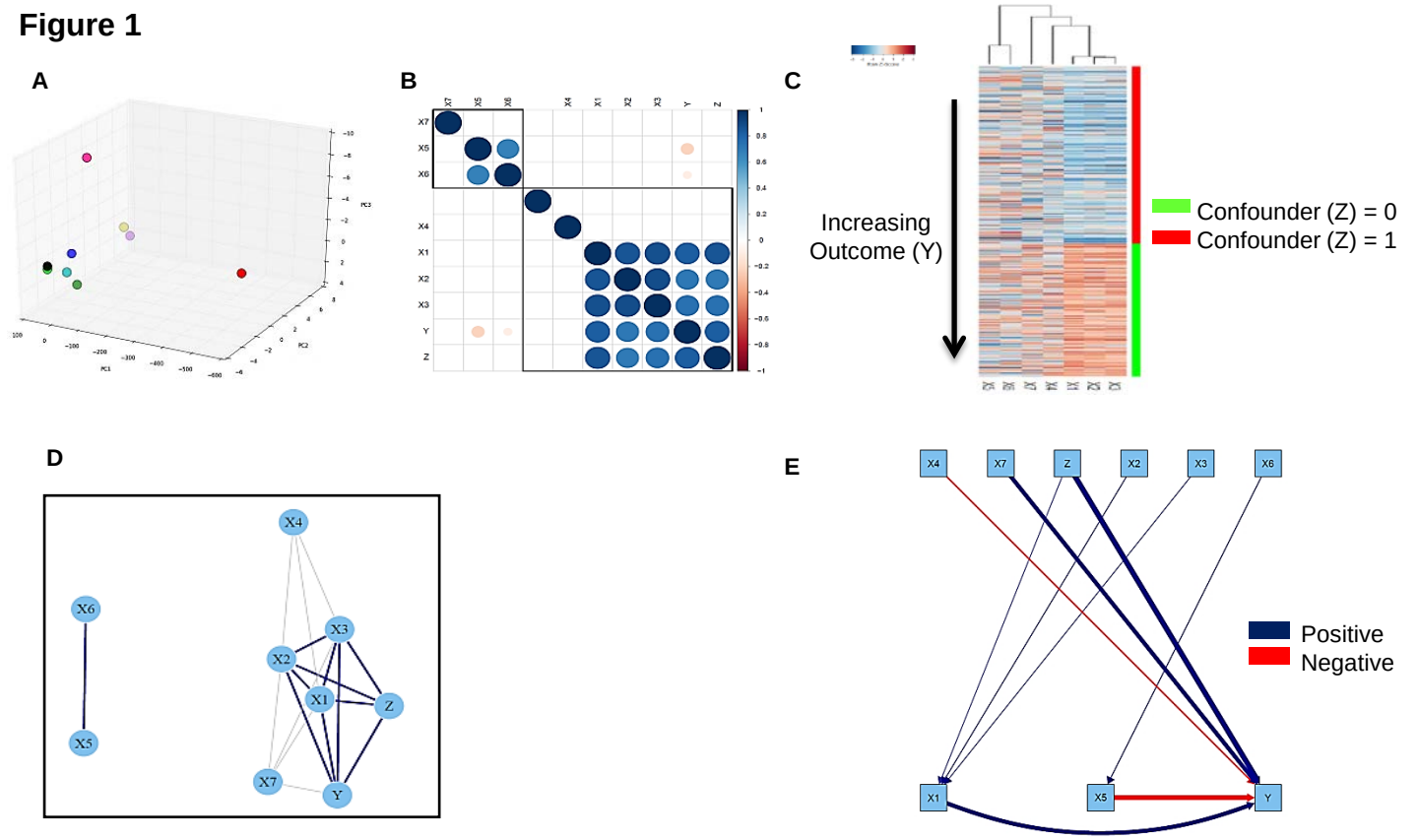


Figure 2

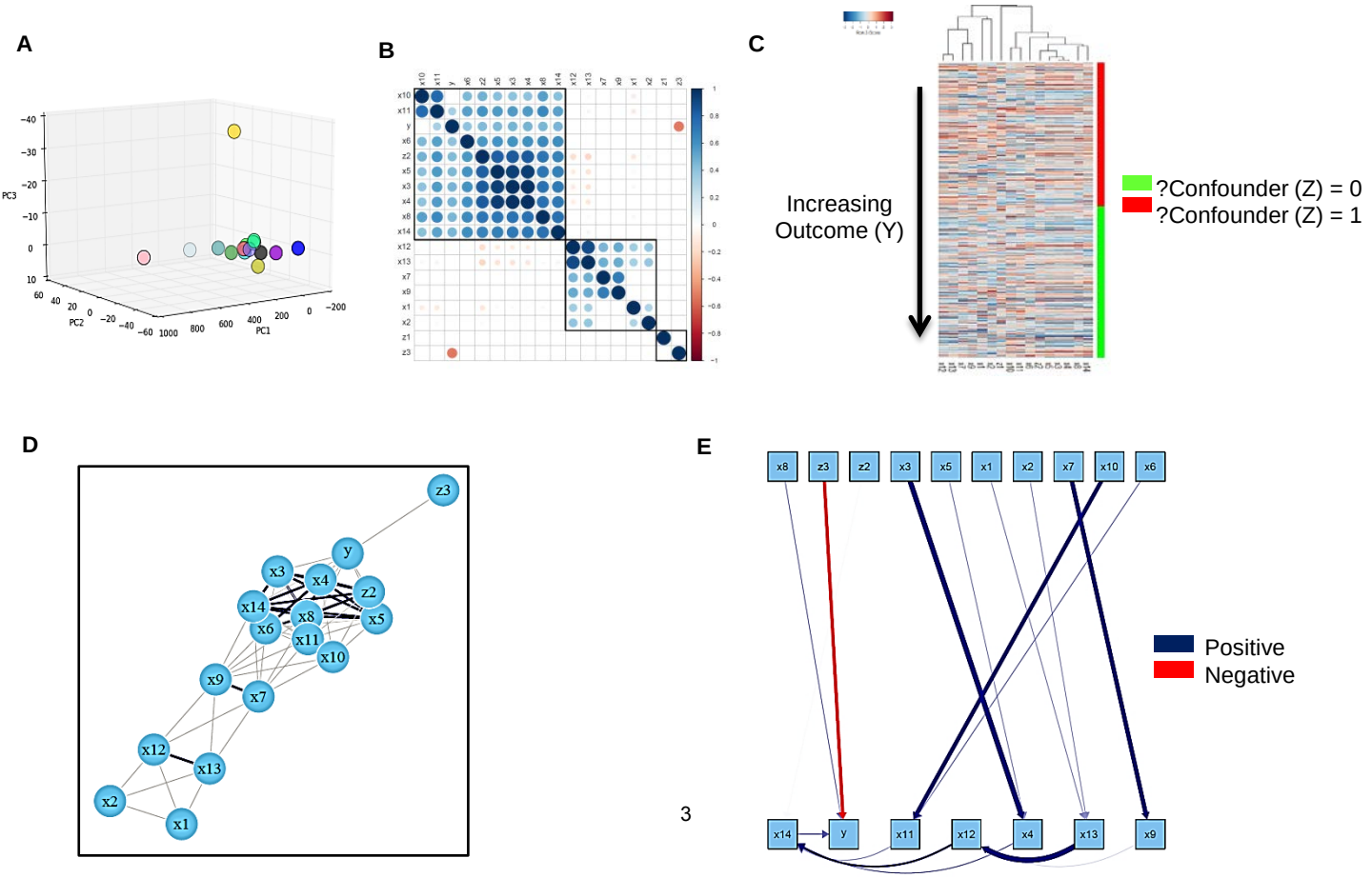
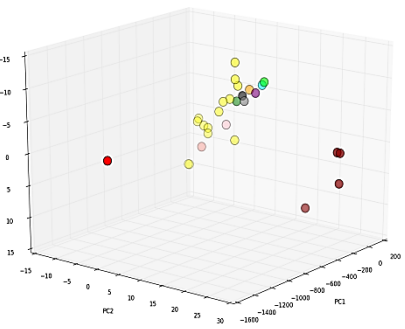
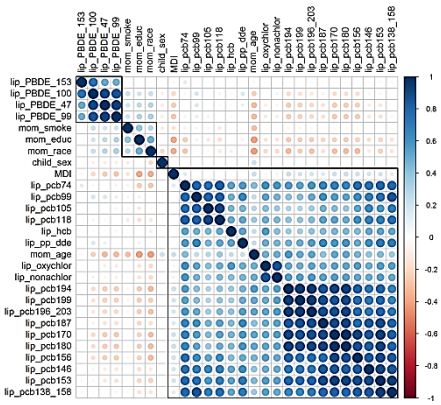


Figure 3

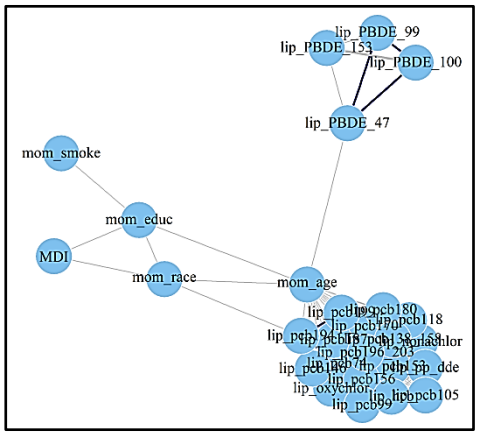
A



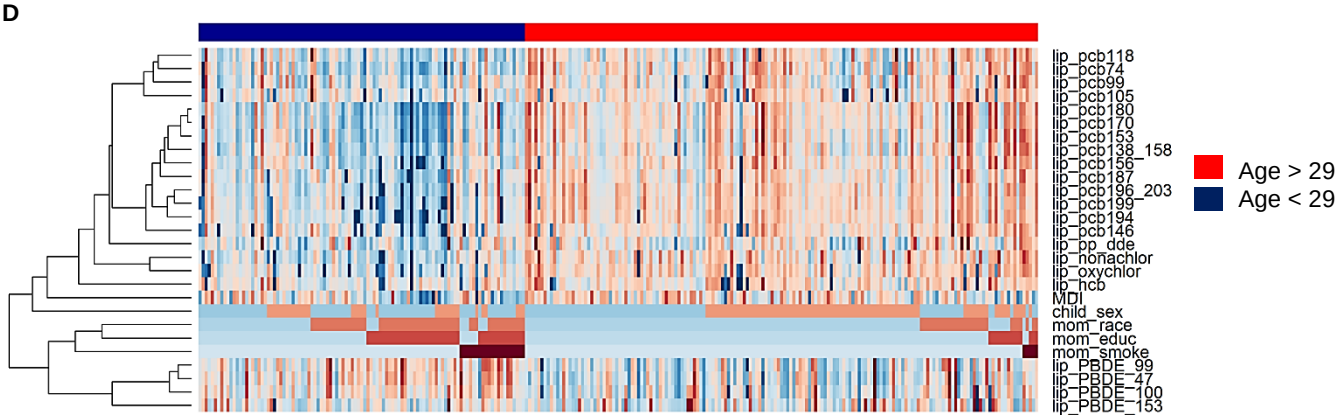
B



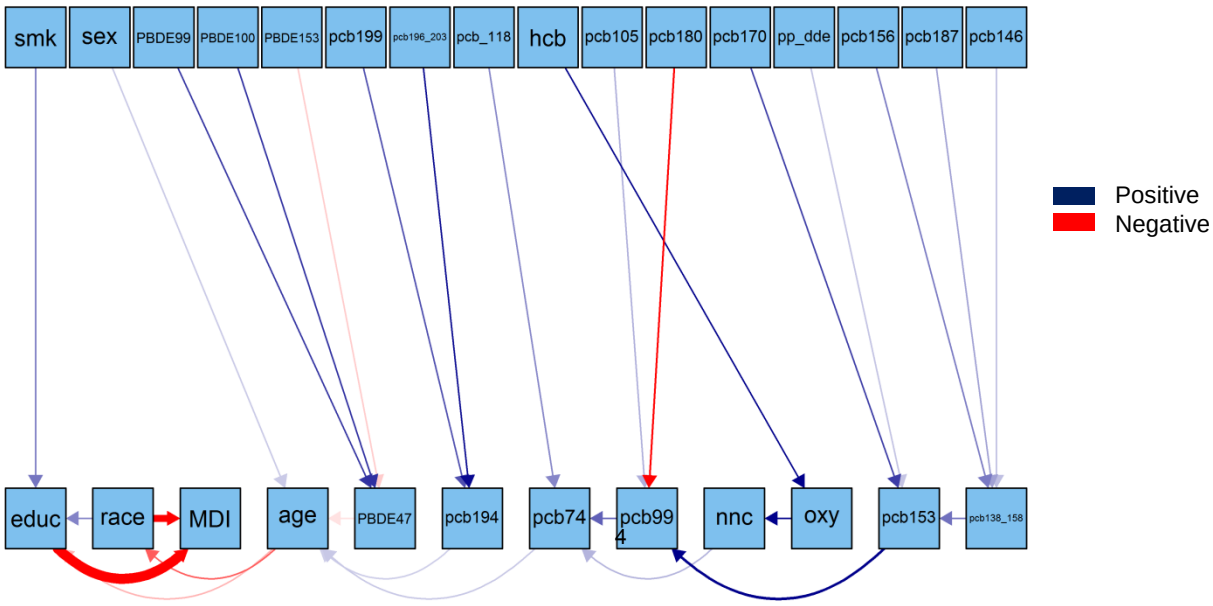
C



D



E



2. A Methodology for Building Bayesian Networks to Evaluate the Health Effects of Environmental Chemical Mixtures

Presenting Author: Sarah Kreidler

Organization: Neptune and Company

Contributing Authors: Sarah M. Kreidler, Ph.D., D.P.T., Tom Stockton, Ph.D., Paul K. Black, Ph.D., and Stephen M. Beaulieu, MSPH

Abstract:

Predicting adverse human health effects due to exposure to chemical mixtures continues to be a challenging problem. Epidemiological studies are expensive and time consuming, and the toxicology of complex chemical mixtures (e.g., synergistic and antagonistic effects) is incompletely understood. Nevertheless, NIEHS and other federal agencies have a responsibility to manage public health risks. There is a need for better predictive modeling approaches that can support these efforts – specifically, to elucidate the relationships among chemical exposures and population characteristics. We propose a methodology for building Bayesian networks to evaluate the health effects of chemical mixtures exposure. The proposed methodology is designed to 1) identify the causal structure, when unknown, among exposures and outcomes through a combination of exploratory data analysis (EDA) and machine learning, 2) estimate individual and joint effects of exposures on health outcomes, and 3) evaluate the sensitivity of the predicted effects to the choice of causal structure.

We represent the causal structure as a directed acyclic graph (DAG). To estimate the DAG, we create a maximal, starting graph based on *a priori* knowledge. For example, if a variable Z is a known confounder for outcome Y and exposure X, we draw edges from Z to X, and from Z to Y. If no *a priori* information is available, undirected edges are added between all variables. Once the initial structure is created, we add and subtract edges based on EDA. In the EDA step, we assume that causal relationships cannot exist if no correlation is present. Therefore, edges are removed if no significant association is observed between the exposure and the outcome. When confounders are present, edges are removed between an exposure and the outcome if no linear association is observed when controlling for the confounders. If undirected edges remain after EDA, a hill-climbing algorithm is applied to determine the final DAG. Once the causal structure is identified, the joint distribution is factored based on the causal structure. Parameters are then estimated via maximum likelihood.

We applied the methodology to three data sets containing real or simulated observational data for chemical mixtures exposures. The data sets represented two different types of study data, with the first representing a prospective cohort epidemiologic study, the second representing a cross-sectional study, and the third being real data from a prospective study of child mental development after maternal exposure to dioxin-like compounds. For each data set, the Bayes net was built using a training set consisting of 80% of the data. The predictive accuracy of the network was assessed on the remaining data.

In the first data set, the exposures were known to occur prior to the outcome, there was a single potential confounder, and no variables were colliders or intermediate variables. The exposure variables were highly skewed and, therefore, were log transformed prior to analysis. After reducing the DAG, the parents of the outcome, Y, were Z, log(X1), log(X2), log(X3), log(X5), log(X6), log(X7). On the training set, the conditional distribution of Y was assumed Gaussian, with an estimated mean of 19.53 + 11.67Z + 3.58log(X1) + 0.75log(X2) – 0.48log(X3) – 4.11log(X5) + 0.21log(X6) + 3.86log(X7) and a standard deviation of 3.05. On the test set, the Bayes net had a mean absolute prediction error of 2.64 for Y.

Simulated data set 2 contained cross-sectional data with fourteen exposure biomarkers, two continuous potential confounders, and one potential binary confounder. No additional information was known regarding the causal structure. Both EDA and hill climbing were used to obtain the final DAG. In the reduced DAG, the parents of the outcome, y, were z3, x1, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, and x14. On the training set, the conditional distribution of y was assumed Gaussian, with an estimated mean of 2.98 - 0.61z3 + 0.02x1 - 0.04x3 + 0.11x4 + 0.01x5 + 0.05x6 - 0.03x7 + 0.04x8 + 0.04x9 + 0.05x10 + 0.1x11 + 0.11x12 - 0.04x13 + 0.07x14 and a standard deviation of 0.47. On the test set, the Bayes net had a mean absolute prediction error of 0.37 for y.

The dioxin exposure data set included maternal education, age, race, smoking status, and sex of the child as covariates. Exposures included 21 different dioxin-like compounds. The outcome was the Mental Development Index (MDI) of the Bayley Scale of Infant Development-II (BSID-II). EDA and hill climbing were used to simplify the DAG. In the final DAG, the parents of the outcome, MDI, were maternal education, age, race, and smoking status, serum PCB194, and serum PCB199. On the training set, the conditional distribution of MDI was assumed Gaussian. The estimated means and standard deviations were dependent on the values of the categorical covariates. On the test set, the Bayes net had a mean absolute prediction error of 10.06 for MDI.

The results from the proposed methodology are promising. Low prediction error was achieved for all data sets, even when limited *a priori* knowledge of the causal structure was available. The methodology is advantageous because it 1) accommodates a variety of complex chemical mixtures data, 2) combines scientific expertise and data-driven machine learning to build the causal structure, 3) supports sensitivity analysis by allowing comparison of predictive accuracy for different starting DAG structures, and 4) allows for new data to be incorporated as they become available, providing a solution that can adapt to advances in the state of knowledge.

Potential improvements to the method include refinements to causal structure learning via methods such as gradient boosting and handling of non-Gaussian outcomes, adoption of a more rigorous cross-validation strategy, and techniques to estimate subsets of the full joint distribution when remaining exposure variables are unknown. In addition, it would be useful to build Web-based tools to support collaborative exploration of different DAG structures by multiple researchers.

We believe that the use of Bayesian networks that combine expert *and* data-driven knowledge represents a practical and robust approach to the analysis of chemical mixtures data, one that ultimately will lead to significant improvements in our ability to characterize potential risks to public health.

3. A Statistical Framework for Assessing Association Between Health Effects and Combined Exposures

Presenting Author: Shuo Chen

Organization: University of Maryland, College Park

Contributing Authors: Shuo Chen, Jing Zhang, Chengsheng Jiang, and Don Milton

Abstract:

The association analysis between exposures of environmental chemical mixtures and health effects is challenging because numerous exposures could be highly correlated and have non-linear and joint impact on health. Regression analysis including shrinkage regression techniques (e.g. LASSO and elastic nets) as widely used tool is often limited for this purpose since there have been few attempts to automatically detect the suitable non-linear transformation (e.g. polynomials) of exposure and interactions between multiple exposures. To fill the gap, we develop a novel statistical procedure consisting four steps to address this challenging issue. Firstly, we conduct the screening step to limit our scope within the exposures that are correlated to health effects with loose a threshold (e.g. p-value <0.1). We next leverage likelihood principle as criteria to determine the suitable univariate transformation each of supra-thresholded continuous exposures (e.g. tuning the exponential term from 0.1 up to cubic terms and natural cubic splines). Moreover, we examine whether the linear combination of highly correlated exposures improves the model fit than individual exposures. In the third step, we apply a least angle regression (LARS) procedure to introduce exposures (or linearly combined/joint exposures) one by one, and when introducing a new exposure or confounding variable we examine the interaction with the pre-entered covariates. Lastly, we conduct model verification by using cross-validation and boot-strapping techniques and prune the final model for robustness and reproducibility. We implement and apply this algorithm to both simulated data sets to seek the optimal regression model that reveal the associations between health effects. As a result, the optimal model for data set 1 is : X12 is the sum of X1 and X2. The estimated parameters and corresponding p-values is showed in Table 1. All exposures are significantly associated with health outcomes with most p-values <0.001 (other than p-value of). X12 is positively associated with the health effects, while X4 has a negative effect. X5 and X7 have non-linear trend associations. The confounding factor Z also modifies the effect of associations of X12 and X5. The adjusted R2 for the selected model is 94% and MSE is 2.659 on 491 degrees of freedom. When treating the exposures as multivariate normally distributed variables, we observe that several exposures X1, X2, X3 are highly correlated, and similarly for X5 and X6.For synthetic data 2, we identify a model as: The estimated parameters and corresponding p-values is showed in Table 1. All exposures are significantly associated with health outcomes, with estimated parameters and corresponding p-values demonstrated in Table 2. All exposures have positive effects, and z3 also modifies the associations. The adjusted R2 for the selected model is 53% and MSE is 0.45 on 492 degrees of freedom. Different from data set 1, most variance (roughly 80%) are explained by confounding factors When treating the exposures as multivariate normally distributed variables, we observe that several exposures X3, X4, X5 are highly correlated, and similarly for X12 and X13.

Figure 1: Figure 1: Flowchart of the algorithm

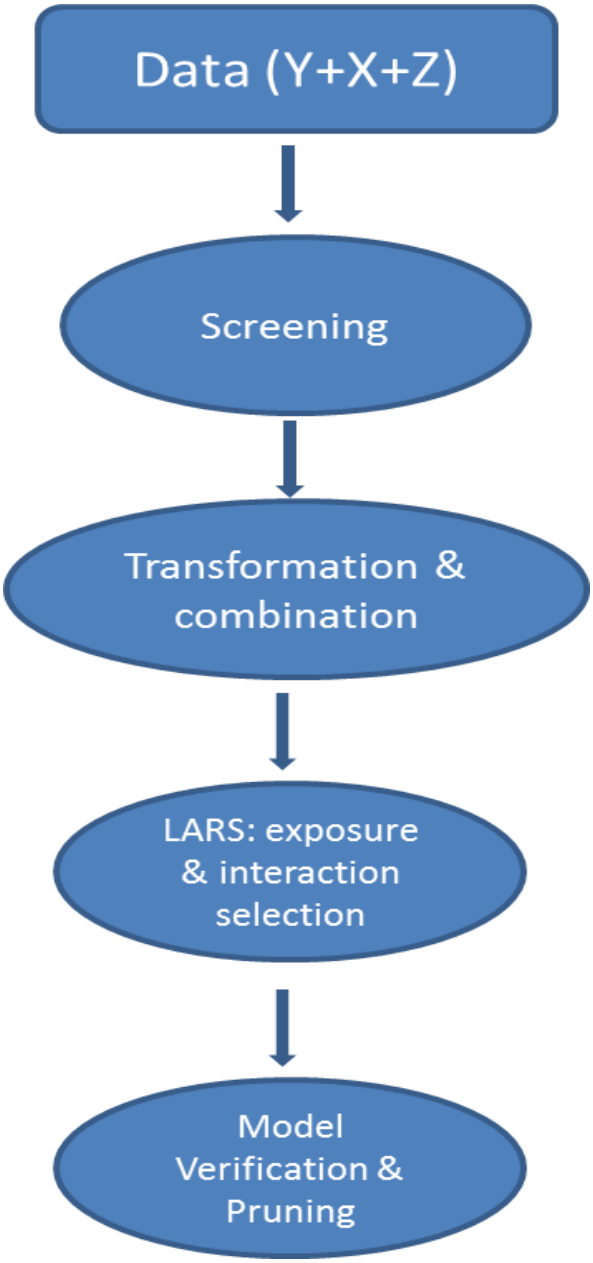


Table 1: Regression results for data set 1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15.1829	0.5745	26.43	0.0000
Z	15.1113	0.9025	16.74	0.0000
X12	3.8273	0.3315	11.54	0.0000
X4	-1.0121	0.1526	-6.63	0.0000
I(X5^0.5)	-8.1576	0.3819	-21.36	0.0000
X7	6.9422	0.3686	18.84	0.0000
I(X7^2)	-1.0474	0.0907	-11.55	0.0000
Z:X12	-1.2395	0.3596	-3.45	0.0006
Z:I(X5^0.5)	-1.4362	0.5625	-2.55	0.0110

$$Y = \beta_0 + \beta_1 Z + \beta_2 X_{12} + \beta_3 X_{12} \times Z + \beta_4 X_4 + \beta_5 \sqrt{X_5} + \beta_6 \sqrt{X_5} \times Z + \beta_7 X_7 + \beta_8 X_7^2 + \varepsilon$$

Table 2: Regression results for data set 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.0527	0.1683	18.14	0.0000
z2	0.0089	0.0012	7.40	0.0000
z3	-0.2271	0.0786	-2.89	0.0040
x6	0.1190	0.0315	3.78	0.0002
x12	0.5705	0.0875	6.52	0.0000
x11	0.0969	0.0353	2.74	0.0063
z3:x6	-0.0952	0.0387	-2.46	0.0141
z3:x12	-0.6523	0.1191	-5.48	0.0000

$$y = \beta_0 + \beta_1 z_2 + \beta_2 z_3 + \beta_3 x_6 + \beta_4 x_6 \times z_3 + \beta_5 x_{12} + \beta_6 x_{12} \times z_3 + \beta_7 x_{11} + \varepsilon$$

4. Bayesian Kernel Machine Regression for Estimating the Health Effects of Multi-Pollutant Mixtures

Presenting Author: Birgit Claus Henn

Organization: Boston University

Contributing Authors: Jennifer Bobb, Linda Valeri, Birgit Claus Henn, and Brent Coull

Abstract:

We applied Bayesian kernel machine regression (BKMR) to the two simulated datasets. This approach, described in detail in Bobb and Valeri (2014), estimates the multivariate exposure-response function. It also estimates the *posterior inclusion probability* for each of the mixture components, which quantifies (on a scale from 0 to 1) how important the pollutant is in predicting the health outcome.

Data Set #1:

We first z-scored each of the pollutants (x_m: $m = 1, \ldots, 7$) so that they would be on the same scale. We then fit the model $E(Y_i) = f(x_{i1}, x_{i2} \ldots, x_{i7}) + \gamma z_i$, adjusting for the confounder. The posterior inclusion probabilities were high (>0.95) for pollutants 1, 2, 4, 5, and 7 and low (<0.05) for pollutants 3 and 6, suggesting that pollutants 1, 2, 4, 5, and 7 contribute to the health outcome while pollutants 3 and 6 do not. Higher exposure levels were associated with higher levels of the outcome for some of the pollutants (x_1, x_2, x_7) and with lower levels of the outcome for others (x_4, x_5), for the other pollutants at their median value.

BKMR estimates the exposure-response function f in a very flexible way. Therefore, any cross-section of f can be plotted as desired. See **Figures 1**, **2**, and **3** as examples. Our results suggested both nonlinear and interactive effects of the mixture components. The coefficient of determination (R^2) was 0.95. The residual standard deviation (posterior mean) was 2.46 (95% credible interval: 2.31—2.62).

We define two exposures to interact if the joint exposure-response function $f(x_m, x_{m'})$ cannot be expressed as $f(x_m) + f(x_{m'})$. We investigated possible two-way interactions by making plots, such as those in **Figure 2**. Visual inspection of these plots indicated that the shape of the exposure response function of one pollutant x_m did not vary qualitatively depending on the level of the second pollutant $x_{m'}$. To investigate whether these potential two-way interactions were statistically significant, we fit parametric models based on the BKMR fit, both with and without interaction terms. In particular, we fit a regression model with natural cubic splines with 3 degrees of freedom for pollutants 1, 2, 4, 5, and 7 and linear terms for pollutants 3 and 6 and for the potential confounder. We conducted an analysis of variance (ANOVA) for the addition of interaction terms. This analysis indicated statistically significant interactions of pollutants 1 and 2 ($P = 0.03$), 1 and 7 ($P < 0.01$), 5 and 7 ($P < 0.01$), 2 and 7 ($P < 0.01$), and marginally significant interaction of pollutants 2 and 5 ($P = 0.10$). **Figure 3** investigates potential three-way interaction by plotting the joint exposure response function of pollutants 1 and 7, with pollutant 5, fixed at its 10th, 50th, and 90th percentiles. The corresponding ANOVA test based on a parametric model did not indicate statistically significant three-way interaction for these pollutants. Similarly, we did not find statistically significant three-way interaction among pollutants 1, 2, and 7.

Data Set #2:

We first z-scored the pollutants (x_m: $m = 1, \ldots, 14$) so that they would be on the same scale. We then fit the model $E(Y_i) = f(x_{i1}, x_{i2} \ldots, x_{i14}) + \boldsymbol{\gamma z}$, adjusting for the three potential confounders ($\boldsymbol{z}$). The highest posterior inclusion probabilities were 0.63 for pollutant 12 and 0.58 for pollutant 6. Plots of the estimated exposure-response function are in **Figures 4** and **5**. From the exposure-response function, we also estimated various quantities to summarize the health effect of the mixture (**Figure 6**): the total effect (**6A**), pollutant-specific effects (**6B**), and the successive contributions of each pollutant to the total effect (**6C**).

With the high degree of correlation among several pollutants, it is hard to definitively say which pollutants contribute to the outcome. The most likely candidates are pollutants 1, 6, and 12. The least likely candidates are 3, 4, 5, and 7 **(Figure 6B)**. For most pollutants, higher exposures were associated with higher levels of the outcome (**Figure 4**), though these associations were not statistically significant (**Figure 6B**). This lack of statistical significance could be due to colinearity as all pollutants were statistically significant in single-pollutant models (results not shown). Overall, higher levels of exposure to the mixture were associated with higher levels of the health outcome (**Figure 6A**). The more pollutants an individual is exposed to at high levels, the greater the risk tended to be (**Figure 6C**), though this increasing risk with increasing numbers of pollutants tapered off after approximately 6 pollutants (1, 12, 6, 10, 14, 11).

We investigated possible two-way interactions by plotting $f(x_m, x_{m'})$ for each pair of pollutants (m, m'). **Figure 5** shows plots that were most suggestive of interaction. To investigate whether these suggestive interactions were statistically significant, we fit parametric models based on the BKMR fit, both with an without interaction terms. In particular, we fit a regression model with linear and quadratic terms for pollutants 2, 12, and 13 and linear terms for the other pollutants and for the potential confounders. We conducted an analysis of variance (ANOVA) for the addition of interaction terms. This analysis indicated statistically significant interactions of pollutants 12 and 13 ($P < 0.01$), 2 and 13 ($P < 0.01$), 2 and 12 ($P < 0.01$), and marginally significant interactions of pollutants 9 and 13 ($P = 0.07$). **Figure 6B** also supports this result, indicating that the pollutant-specific health effect estimates for pollutants 2, 12, and 13 varied depending on the levels of the other pollutants.

The coefficient of determination (R^2) was 0.56. The residual standard deviation (posterior mean) was 0.45 (95% credible interval: 0.42—0.48).

Figures for NIEHS Workshop

Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures

Jennifer Bobb, Linda Valeri, Birgit Claus Henn, and Brent Coull

Data Set #1: Chemical Mixture Simulated Data

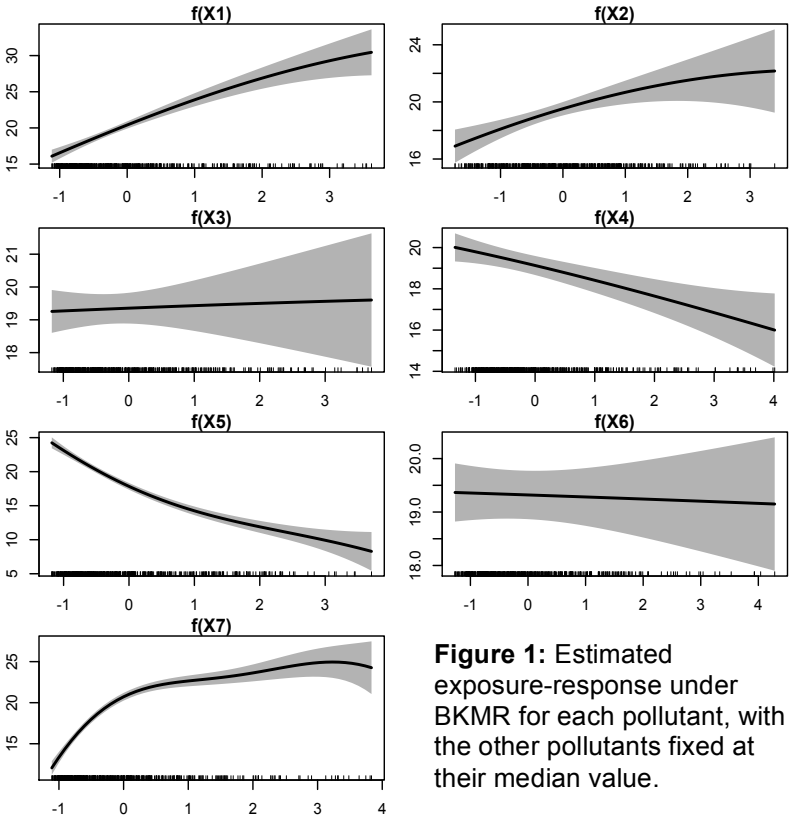


Figure 1: Estimated exposure-response under BKMR for each pollutant, with the other pollutants fixed at their median value.

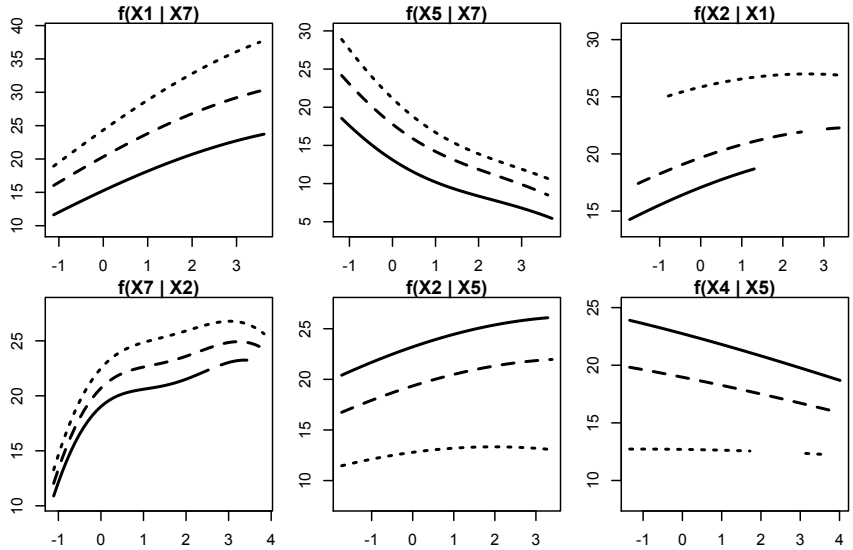


Figure 2: Estimated exposure-response function under BKMR for select pairs of pollutants. Shows cross-sections $f(x_m \mid x_{m'})$ of the estimated exposure response function for a given pollutant x_m at different levels of a second pollutant $x_{m'}$, where the remaining pollutants are fixed at their median value. The levels of $x_{m'}$ are its 10th (solid line), 50th (dashed line), and 90th (dotted line) percentile.

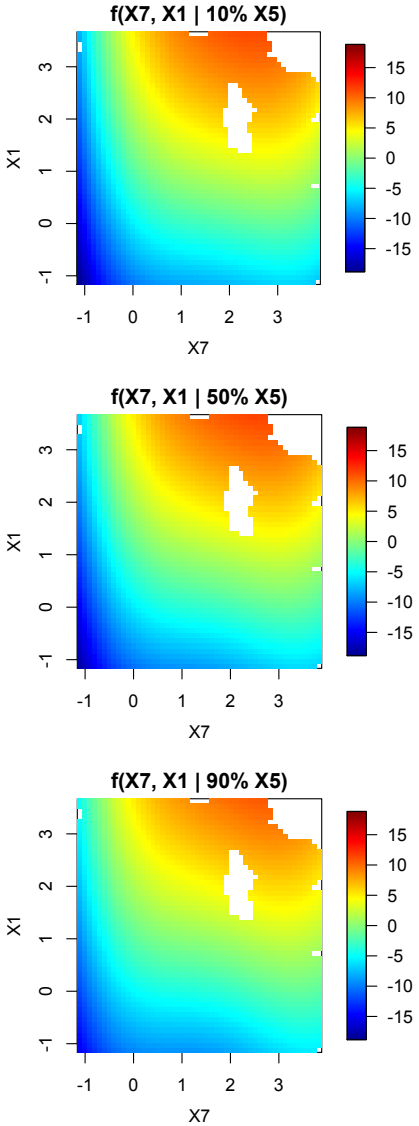


Figure 3: Estimated exposure-response under BKMR for pollutants 1 and 7, with pollutant 5 fixed at its 10th (top panel), 50th (middle panel), and 90th (bottom panel) percentiles, and for the remaining pollutants fixed at their median value. Estimated exposure-response surface plotted for points within 0.5 units from an observed data point.

Data Set #2: Mixture Simulated Data with an Environmentally Relevant Correlation Pattern

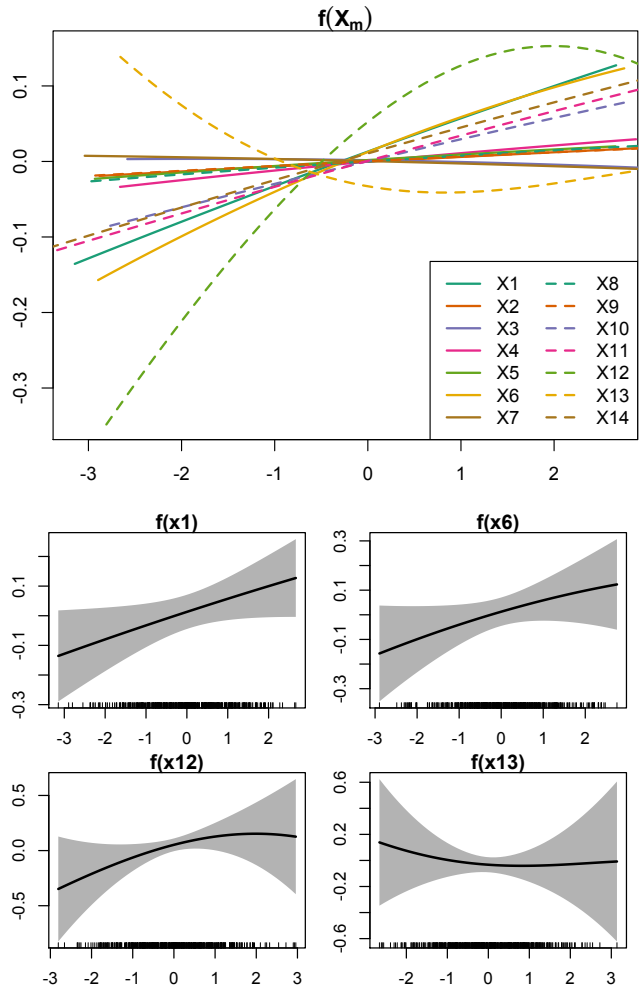


Figure 4: Estimated exposure-response under BKMR for each pollutant, with the other pollutants fixed at their median value (top panel). Four of the functions with pointwise 95% credible intervals are highlighted (bottom panel).

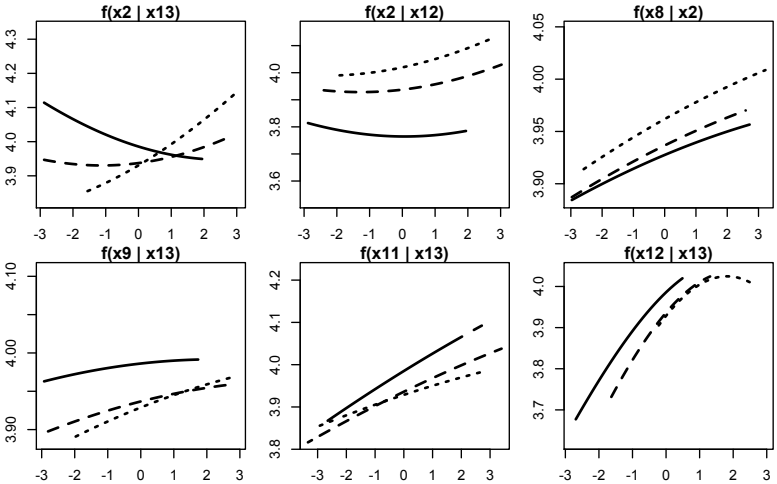


Figure 5: Estimated exposure-response function under BKMR for select pairs of pollutants. Shows $f(x_m \mid x_{m'})$, where $x_{m'}$ is fixed at its 10th percentile (solid line), 50th percentile (dashed line), and 90th percentile (dotted line). The other pollutants are fixed at their median value.

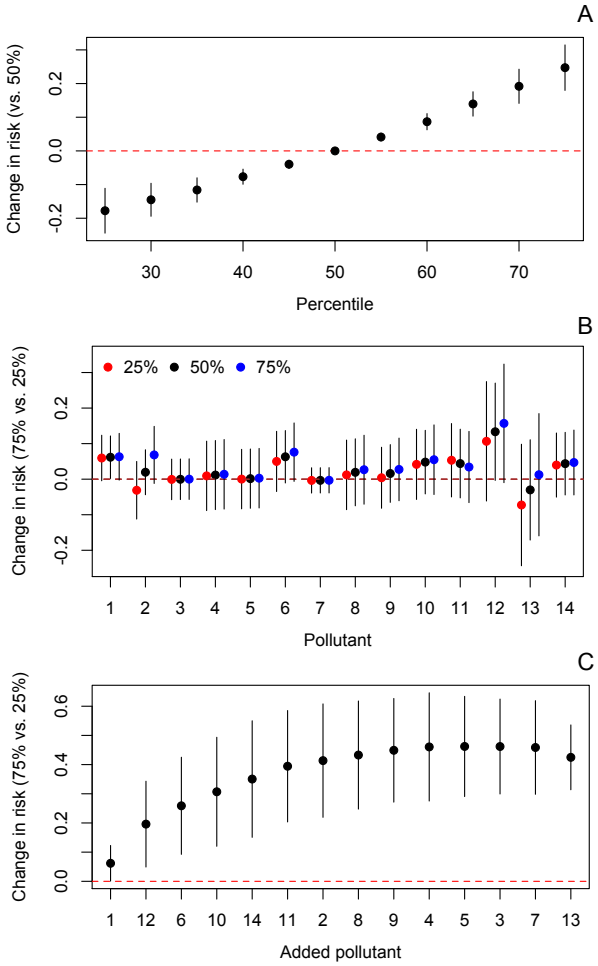


Figure 6: Estimated change in risk (95% credible intervals) when (**A**) all of the pollutants are at a particular percentile (25th to 75th) of their distribution to when all of the pollutants are at their median value; (**B**) comparing a pollutant at the 75th percentile to the 25th percentile, for all other pollutants fixed at either their 25th (red points), 50th (black points), or 75th (blue points) percentile; and (**C**) comparing a subset of pollutants at their 75th percentile to their 25th percentile, for the others fixed at their median value. Pollutants are successively added to the subset from left to right (e.g., left point denotes pollutant X1, second point from left denotes X1 plus X12, etc).

5. Cheminformatics Approaches to Analyze the Effects of Environmental Chemical Mixtures

Presenting Author: Denis Fourches

Organization: North Carolina State University

Contributing Authors: Denis Fourches*[,1] and Ryan Lougee[2]

[1] Department of Chemistry, Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, USA.
[2] Department of Toxicology, North Carolina State University, Raleigh, North Carolina, USA.

Abstract:

Individuals in modern societies are exposed to multiple and diverse chemicals from various sources, such as food, medicines, cosmetics, and environmental pollutants. As expensive and time-consuming toxicological studies are traditionally done one chemical at a time, developing cheminformatics methods to analyze and reliably forecast the overall biological outcomes of exposure to chemical mixtures is of high importance. Herein, we are using different cheminformatics techniques that take as inputs a series of variables characterizing the toxicants and/or the nature of the exposure in order to forecast the overall health outcomes.

Regarding **Simulated Dataset 1**, we analyzed the contributions of each exposure using multi-linear regression (MLR) representing Y (continuous modeling) as a function of the exposure variables. As shown in *Table 1*, certain variables, such as x1 (R^2=0.78 correlation with Y; MLR regression coefficient equal to +2.9), x5, and x7, are very important for the overall prediction performances of the MLR model (R^2 =0.92, MSE = 9.6, *Figure 1*).

Regarding **Simulated Dataset 2**, the obtained MLR models were significantly less predictive (R^2 ~0.52, MSE = 0.6, *Figure 2*). Very weak linear contributions indicated the need to use non-linear techniques, such as Support Vector Machines and Associative Neural Networks. The SVM model for three-classes obtained very reasonable prediction performances.

Regarding the **NIEHS REAL WORLD dataset**, we analyzed the whole set of 270 mother-child pairs, the associated Mental Development Index (MDI) scores, and the distributions of the 22 exposure chemicals. First, after data preprocessing steps, confounder-adjusted associations between the 22 chemicals and MDI scores were computed using a two-stage hierarchical analysis. Most of the measured exposures were associated with negligible absolute differences in MDI scores. Second, the molecular structures of the chemicals were taken into account to compute 2D fragment descriptors and generate CBRA[1] radial plots (*Figure 3*). Interestingly, structurally similar compounds shared similar exposure distribution in the studied population of children. Third, ADDAGRA[2] visualization plots confirmed the outlying distributions for DDE and PBDE47, the high similarity of exposure within the PCBs and the PBDEs, as well as the very weak associations with MDI scores.

We will also discuss the rationale and benefits of characterizing toxicants with molecular descriptors computed from their chemical structures to better analyze such types of mixture exposure data.

References

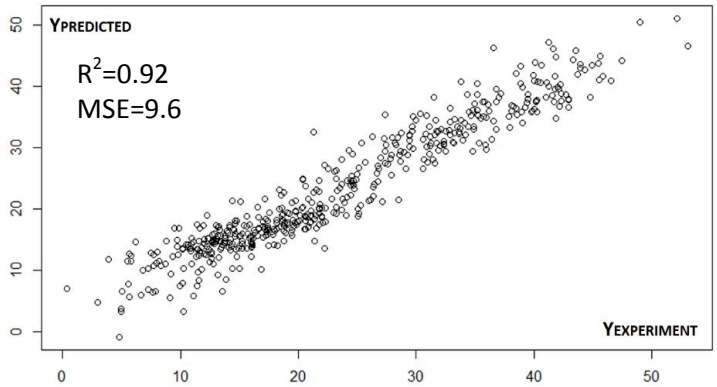
1. Low Y, Sedykh A, Fourches D, Golbraikh A, Whelan M, Rusyn I, Tropsha A. 2013. Integrative chemical-biological read-across approach for chemical hazard classification. Chem. Res. Toxicol. 26(8):1199-1208.
2. Fourches D, Tropsha A. 2013. Using graph indices for the analysis and comparison of chemical datasets. Molecular Informatics. 32(9-10):827–842.

Simulated Dataset 1

Table 1. Exposure coefficients found in the best MLR model (R^2=0.92, MSE = 9.6) and the different folds according to a 5-fold cross-validation procedure.

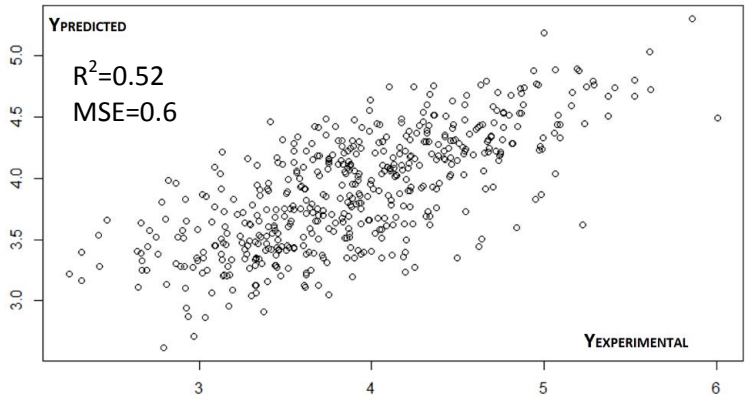
	Full Set	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	p value
Intercept	14.36	15.15	14.35	14.15	13.97	14.15	
X1	2.91	2.74	2.91	3.24	2.85	2.88	< 2e-16
X2	3.19	2.70	2.95	3.21	3.75	3.42	4.06E-07
X3	-0.01	0.27	0.15	-0.30	-0.06	-0.13	0.981
X4	-0.98	-0.911	-1.07	-0.93	-1.16	-0.88	6.31E-08
X5	-3.55	-3.60	-3.44	-3.55	-3.57	-3.54	< 2e-16
X6	-0.14	-0.39	-0.18	-0.01	-0.12	-0.01	0.497
X7	2.93	2.75	2.91	3.00	2.99	3.00	< 2e-16
Z	11.41	11.71	11.36	11.29	11.45	11.27	< 2e-16

Fig 1. Predicted versus experimental outcomes using one individual MLR model



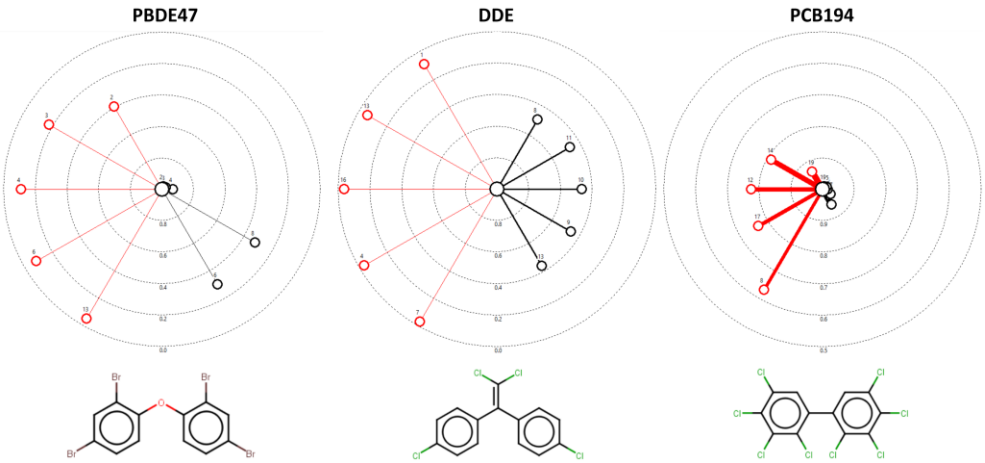
Simulated Dataset 2

Fig 2. Predicted versus experimental outcomes using one individual MLR model



NIEHS Real World Dataset

Fig 3. CBRA radial plots for PBDE47, DDE, and PCB194. Structural neighbors are represented in *black*, whereas neighbors in the Children Exposure space are in colored in *red*.



6. A Two Step Variable Selection Procedure Using Prioritization of Interactions Followed by LASSO

Presenting Author: Jiang Gui

Organization: Dartmouth College

Contributing Authors: Jiang Gui and Margaret R. Karagas

Abstract:

A two step variable selection procedure using prioritization of interactions followed by LASSO.

We propose a LASSO penalized regression approach to select relevant predictors and interactions to build a predictive model for the outcome of interest. The interactions mentioned in this abstract are all statistically significant – that is defined as departure from additivity in linear regression model. First, for each predictor x, we used cubic smoothing spline to smooth the outcome and x and plotted them to visualize their functional relation. If a non-linear relationship was identified, we tried possible transformations, such as log, square root, to remove the nonlinearity. We then selected candidate interactions using two methods:

1. For each pair of predictors, we calculated the correlation between the cubic spline smoothed outcome and the product of the two predictors. Then we calculated the same correlation using the two predictors separately. We ranked the pairs based on the marginal difference between the first correlation the maximum of the second set of correlations.

2. We used linear regression to calculate and rank the adjusted p-value for interaction effects for any pair of two predictors after adjusting covariates.

Top models from both methods and other univariate predictors were entered into a LASSO penalized regression model. Cross-validation was used to select the turning parameter. One of the strengths of LASSO penalized method is that it selects at most one variable from a group of highly clustered predictors. This was a useful remedy for this set of simulation data as some of the exposure variables were correlated.

For dataset 1, we found that square root transformation improved the linear relationship between the outcome and all predictors. However, we did not find any statistically significant interactions. We fitted a LASSO penalized regression on the outcome and square root transformed predictors and nuisance variable Z. Using cross-validation, we were able to identify a predictive model using X1, X2, X4, X5, X7, and Z. We refit a linear regression model using the 5 transformed predictors and nuisance variable Z to obtain a coefficient of determination (R square) of 0.9385 (Table 1).

For dataset 2, we did not find any transformation that improved the linear relationship between outcome and predictors. In search for candidate interactions, from the first method, we found that the product of X2 and X12, X2 and X13 improved the univariate correlation by a margin of 0.202 and 0.203

respectively (i.e., became linear). Notably, the 2-way correlations, both with X2, were in the opposite direction, suggesting a three-way interaction among X2, X12, and X13 (Figure 1). Using the second method, we identified a significant interaction between X3 and X10 (Table 2). We then fit a LASSO penalized regression model using the three way interaction of X2, X12, X13 and two way interaction of X3 and X10, the remaining predictors and covariates. Cross-validation was applied to select the turning parameter and the final model that included Z2 , Z3, X1, X6, X14, two way interactions of X3 and X10 and three way interactions of X2, X12, X13 (Table 3).

We applied this method to the pregnancy and birth cohort study. First, we used scatter plot to check whether the variables were normally distributed. We found that most predictors were right-skewed. We used log transformation to stabilize predictor's variance. We obtained scatter plot on transformed data and found that all variables follow normal distribution. Based on correlation difference between interaction and main effect, we identified five pairs of exposures. We ran a LASSO penalized regression model using the 5 two-way interactions, the remaining predictors and covariates. The final model included lip_PBDE_99, lip_PBDE_153, lip_pcb105, lip_pcb146, lip_pcb187, lip_pcb194, and an interaction between lip_oxychlor and lip_pcb187. We refit a regular linear regression model to get the unbiased estimate (Table 4).

Table 1: Final model for data set 1.

Predictors	Coefficients	P-value
Intercept	11.2091	< 2e-16
Square root (X1)	8.3892	< 2e-16
Square root (X2)	3.7957	0.000222
Square root (X4)	-2.5159	7.11e-11
Square root (X5)	-8.7818	< 2e-16
Square root (X7)	7.5241	< 2e-16
Z	10.5147	< 2e-16

Figure 1: Cubic spline smoothed plot indicating a three-way interaction.

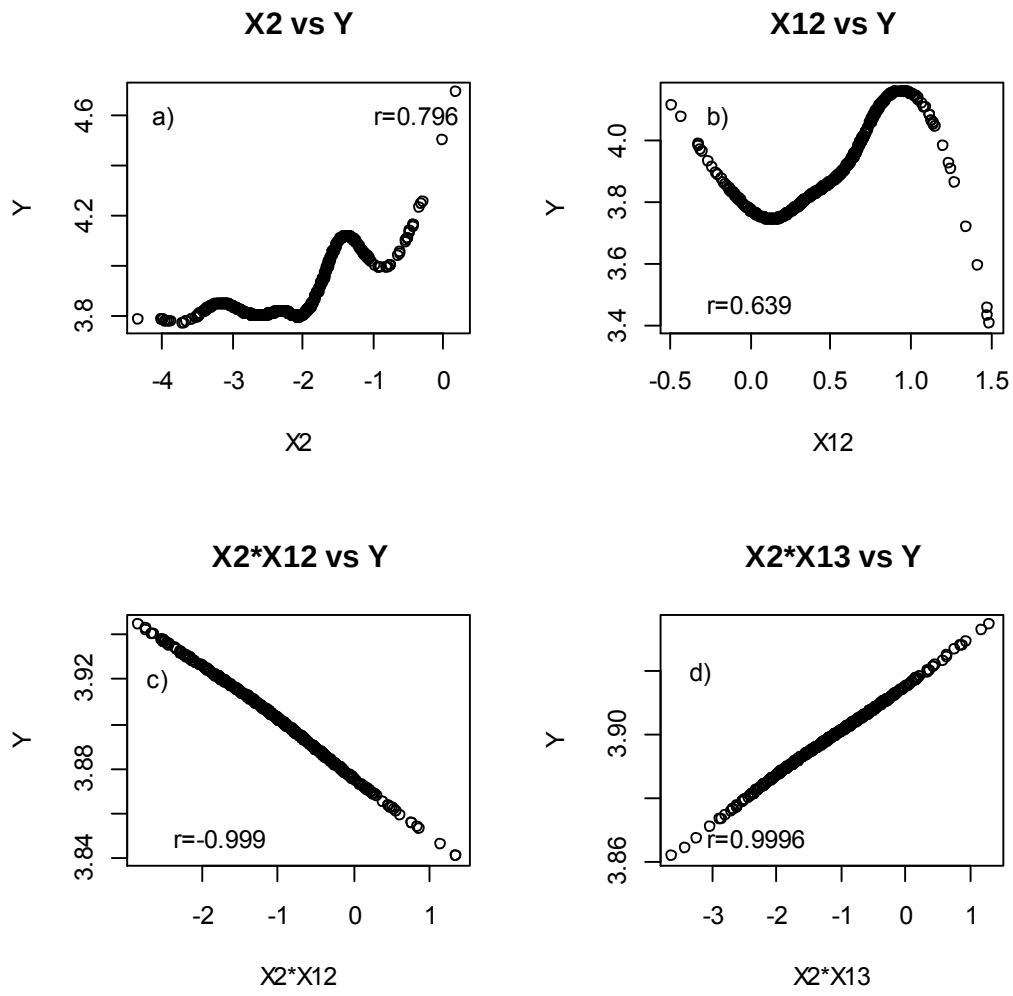


Figure 1 a) is the cubic spline smoothed scatter plot of X2 vs Y. b) is the cubic spline smoothed scatter plot of X12 vs Y. c) is the cubic spline smoothed scatter plot of X2*X12 vs Y. Here X2*X12 denote the inner product of X2 and X12. d) is the spline smoothed scatter plot of X2*X13 vs Y.

Table 2. Top adjusted p-value for all pair-wise interactions.

Interaction model	Adjusted p-value
X2, X13	0.0030
X3, X10	0.0031
X4, X10	0.0033
X8, X10	0.0055
X5, X10	0.0076

Table 3: Final model for data set 2.

Predictors	Coefficients	P-value
Intercept	3.198581	< 2e-16
X1	0.069188	0.03263
X6	0.055202	0.04932
X3	-0.141024	0.02661
X10	0.016664	0.67759
X2	-0.186233	0.00628
X12	0.347766	0.48824
X13	1.086567	0.02339
X14	0.073957	0.71953
Z1	0.004789	0.8427
Z2	0.006857	0.00012
Z3	-0.605989	< 2e-16
X3*X10	0.0523931	0.0033
X2*X12	0.029869	0.88911
X2*X13	0.541121	0.0081
X12*X13	-0.739856	0.04878
X2*X12*X13	-0.339453	0.04612

Table 4: Final model for real data.

Predictors	Coefficients	P-value
Intercept	98.5941	< 2e-16
Log(lip_PBDE_99)	-2.0353	0.00117
Log(lip_PBDE_153)	0.6264	0.29476
Log(lip_pcb105)	1.5771	0.04980
Log(lip_pcb146)	1.0138	0.28819
Log(lip_pcb187)	-3.8250	0.00715
Log(lip_pcb194)	4.4924	3.57e-06
Log(lip_oxychlor)	-0.4947	0.55563
Log(lip_oxychlor)*Log(lip_pcb187)	-0.4123	0.51372
zchild_sex	0.4586	0.70301
zmom_educ	-2.6630	0.10391
zmom_age	-3.6674	0.00752
zmom_race	1.4590	0.29696
zmom_smoke	0.1482	0.94285

7. Integrating Toxicology and Mechanistic Evidence Into Complex Mixtures Analysis Using a Flexible Bayesian Approach

Presenting Author: Ghassan Hamra

Organization: Drexel University School of Public Health

Contributing Authors: Ghassan Hamra, Richard MacLehose, David Richardson, Stephen Bertke, and Robert Daniels

Abstract:

Summary: When studying disease risk due to a group of exposures, researchers will often take two approaches: sum exposures based on weights derived from toxicology research or consider potentially correlated exposures individually. However, we are rarely interested in the effect of any individual exposure on disease risk, since exposures always occur together. Weights, problematically, do not translate directly from toxicology to observational epidemiology research. When incorrect weights are applied, risk estimates will be incorrect because the summed exposures are effectively misclassified.

Methods: I propose a Bayesian approach to allow flexibility in the estimation of weighted sums of complex mixtures. In the contexts of datasets 1 and 2, this model is of the following form:

$$Y = \alpha + \beta\left(\sum_{i=1}^{I} \omega_i x_i + r\right) + \sum_{j=1}^{J} \varphi_j z_j$$

where Y is a linear outcome of interest (such as a biomarker of inflammation), z_j are potential confounders, r is a reference exposure (often TCDD in toxicology), and x_i are exposures for which a weight is estimated. This approach directly parameterizes weights that can otherwise be thought of as the ratio of two risk coefficients for two exposures of interest. I apply a second stage prior to the weights that truncates their estimation at zero; this ensures that implausible (negative) weights are not estimated. I further apply a hierarchical, group level mean to the weights to stabilize their estimation in the presence of statistical instability.

Results: The table summarizes results for these methods applied to datasets 1 and 2. The first columns show the mean and SD for the risk coefficient and weights for a model with no prior applied. Estimated weights for dataset 1 include negative values, which are not plausible (i.e., an exposure cannot be weighted to be negative when summing with other exposures). When the truncation is applied, weights in both datasets gain substantial statistical precision. When a shared group mean is added to the truncation, weights in dataset 2 are stabilized toward it. While there is stabilization of some weights from dataset 1, results are less susceptible to shrinkage toward a group mean. This suggests that there is more support for a difference in weights for dataset 1. In both datasets, when the truncation and shared group mean are applied, risk estimates for the complex mixture are increased. The final columns show

results for estimation of weights when only a subset of the exposures is considered. Weights appear somewhat susceptible to changes in the other exposures included in the model.

Conclusions: The proposed approach does not rely on a single fixed weight value for components of the pollutant mixture. This approach is contingent upon selection of a reference exposure, which is best determined directly from toxicological evidence. In the current example, I chose reference weights based on the exposure whose linear risk coefficient was smallest in a model including all the exposures. This approach is flexible, in that reliable evidence from toxicology can be directly applied to inform estimation of the weights from observational data. *I will present results where a mixture prior is applied to estimation of the summed risk using the real world dataset at the NIEHS mixtures workshop July 13-14, 2015.*

		no prior		truncation		truncation+shrinkage		truncation +subset of exposures	
Dataset 1		mean	SD	mean	SD	mean	SD	mean	SD
β		0.0344	0.0147	0.0244	0.0106	0.0655	0.0788	0.0591	0.0780
Weights									
	x1	96.5	33.6	149.3	51.3	91.2	59.1	103.0	58.2
	x2	101.1	36.9	54.0	35.3	35.2	31.5	50.0	36.5
	x4	-32.2	12.1	6.6	6.3	4.2	5.0	3.6	4.2
	x5	-117.0	38.8	0.6	0.7	0.4	0.5	n/a	n/a
	x6	-4.9	7.5	1.2	1.3	0.7	1.0	n/a	n/a
	x7	97.0	32.4	127.5	43.3	78.8	50.8	n/a	n/a
	group mean	n/a	n/a	n/a	n/a	19.6	21.5	n/a	n/a
Dataset 2									
β		0.00054	0.00017	0.00044	0.00013	0.00376	0.01036	0.00055	0.00021
Weights									
	x1	96.4	57.1	100.4	55.7	61.5	58.2	97.0	52.9
	x2	46.2	53.4	63.9	44.8	48.2	49.9	62.1	43.5
	x3	11.3	75.2	47.3	38.9	39.3	44.7	43.4	37.1
	x4	40.0	77.0	52.4	41.9	41.5	46.0	49.8	40.3
	x5	27.8	60.0	45.9	37.1	38.4	43.6	42.6	35.5
	x7	26.9	73.1	65.3	50.0	48.6	51.7	84.0	57.0
	x8	67.1	63.4	78.5	52.7	53.2	54.0	107.5	57.7
	x9	62.6	72.3	79.6	55.6	54.1	54.9	110.6	61.3
	x10	87.8	64.4	99.9	58.2	61.3	58.8	n/a	n/a
	x11	82.2	68.8	93.0	58.4	59.1	58.0	n/a	n/a
	x12	84.1	86.6	95.3	64.1	58.2	58.0	n/a	n/a
	x13	42.2	87.1	80.9	58.9	53.9	55.3	n/a	n/a
	x14	93.5	69.9	94.8	60.0	58.4	57.3	n/a	n/a
	group mean	n/a	n/a	n/a	n/a	40.9	43.2	n/a	n/a

*x3 and x6 are the reference exposures in datasets 1 and 2, respectively.

1. Hamra GB, MacLehose R, Richardson D, Bertke S, Daniels RD. Modelling complex mixtures in epidemiologic analysis: additive versus relative measures for differential effectiveness. *Occupational and environmental medicine.* 2014;71(2):141-146.

8. Two-Step Shrinkage-Based Regression Strategy for Assessing Health Effects of Chemical Mixtures in Environmental Epidemiology

Presenting Author: Xindi Hu

Organization: Harvard University

Contributing Authors: Xindi Hu

Abstract:

In epidemiological studies, an additive model with only main effects cannot sufficiently explain the relationship between chemical mixtures and the outcome. This abstract reports improved results from including variable interactions to capture the nonlinearity in the model, while keeping model parsimony by leveraging methods reported in recent literature. Here interactions are defined as "effect of one exposure on a health outcome depends on the level of other exposures."

Highlights:

Dataset 1:

- Exposures X_1, X_2, X_5 and X_7 contribute to the outcome, while others do not.
- There are interactions between X_1 and X_5, X_4 and X_5, X_1 and X_7.
- The joint effect of exposure to the mixture is more than additive.
- Joint dose-response function with standardized coefficients is $\log(Y)=0.142*\log(X_1)+0.010*\log(X_2)-0.128*\log(X_5)+0.132*\log(X_7)+0.188*Z+0.017*\log(X_1)*\log(X_5)+0.003*\log(X_4)*\log(X_5)-0.003*\log(X_1)*\log(X_7)-0.015*\log(X_5)*\log(X_5)-0.008*\log(X_7)*\log(X_7)$
- Mean square error = 0.045

Dataset 2:

- Exposures x_3, x_6, x_8, x_{10}, x_{12}, and x_{14} contribute to the outcome, while others do not.
- There are interactions between exposure and confounders, but no interactions between two exposures.
- Joint dose-response function with standardized coefficients is $y=0.0352*x_3+0.031*x_6+0.0247*x_8+0.0272*x_{10}+0.0401*x_{12}+0.0335*x_{14}+0.0983*z_2-0.2034*z_3+0.023*x_{10}*z_2-0.0401*z_3*x_{12}-0.0147*z_3*x_6$.
- Mean square error = 0.120

Real world dataset:

- Lipid normalized PBDE-99, PCB-105, PCB-153, PCB-156, PCB-187, PCB-199, mom_educ and mom_smoke contribute to the outcome, while others do not.

- There are interactions between chemical exposures and between chemical exposure and maternal factors.
- Mean square error = 0.195

Introduction:

LASSO (Tibshirani 1996) and the other shrinkage-based methods (ridge, elastic-net) have gained popularity for their ability to stabilize coefficient estimates and produce sparse models. Bien et al. (Bien et al. 2013) improved LASSO by adding convex constraints to estimate interactions between independent variables. In practice, however, hierarchical LASSO can be sensitive to the assumptions of strong or weak hierarchy when data has high-dimensionality. Sun and colleagues (Sun et al. 2013) proposed employing classification and regression tree (CART) as the initial step to screen important variables. Simulation studies have found that this can reduce the bias of estimates of non-zero coefficients. Here these two methods are combined to form a two-step strategy: tree-based screening coupled with hierarchical LASSO. This strategy is also compared to some single-step approaches in simulated datasets to shed light on choosing appropriate analytical methods given different data structures.

Methods:

The model takes the form $Y = \beta_0 + \sum_j \beta_j X_j + \frac{1}{2}\sum_{j \neq k} \Theta_{jk} X_j X_k + \varepsilon, \quad \beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, \Theta \in \mathbb{R}^{p \times p}, \Theta = \Theta^T, \Theta_{jj} = 0$. As proposed by Bien et al., strong hierarchical LASSO estimates the coefficients by solving the optimization problem, $argmin_{\mu,\beta,\Theta}\ \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \mu - x_i^T\beta - \frac{1}{2}x_i^T\Theta_{x_i}\right)^2 + \lambda 1^T(\beta^+ + \beta^-) + \frac{\lambda}{2}\|\Theta\|_1\ \ subject\ to\ \ \Theta = \Theta^T, \left\|\Theta_j\right\|_1 \leq \left(\beta_j^+ + \beta_j^-\right), \beta_j^+ \geq 0, \beta_j^- \geq 0$. The constraint implies that if the interaction term has a coefficient of non-zero, then the main effect will also be estimated as non-zero. The side effect is that a strong interaction term will force the model estimates on main effects to be large, increasing the biases. Therefore, some authors argued for weak hierarchy (Liu et al. 2014), which can be easily achieved by removing the symmetry constraint.

CART is a tree-structured nonparametric method with little assumptions on data structure. Its algorithm recursively partitions observed data until the tree explains most of the outcome variability. The tree is then pruned to remove less important nodes and to generate a list of important variables.

Before the initial tree-based screening, the sample datasets are first examined visually for distribution normality. Dataset 1 is log-transformed while dataset 2 is not. Then, continuous variables are standardized to have a mean of 0 and a standard deviation of 0.5, while binary variables are intact, as suggested by (Gelman 2008). I compare results from multiple methods using the single-step and two-step strategies. In the single-step strategy, the standardized dataset are fit using ridge, LASSO, elastic-net (EN), strong hierarchical LASSO (SHL), and weak hierarchical LASSO (WHL). In the two-step strategy, CART is first built to select the "important variables" and then they are used as the input for the five aforementioned methods. All analyses are conducted in R 3.1.3, with rpart and hierNet package. Tuning

parameters (i.e. lambda and alpha) are selected based on the value that minimizes the mean squared error from a 10-fold cross-validation.

Results:

SHL and WHL perform well in capturing variable interactions while fitting a relatively parsimonious model at the same time. Figure 1 visualizes the main effects and interactions for datasets under the strong and weak hierarchical assumptions. For dataset2, when the dimensionality is higher, the final product of SHL and WHL can have too many regressors to be easily interpreted. A comparison of one-step versus two-step strategy is shown in Table 1. For dataset 1, the benefit of CART is outweighed by the side effects: the number of main effects is reduced from 5 to 3, the number of interactions is reduced from 2(or 3) to 1, however the MSE goes up by 26%. On the other hand, CART does benefit model building for dataset 2. Nine candidate variables are selected from the original 17 for the subsequent shrinkage methods. In the final model, the number of main effects decreases from 12 to 8, and the number of interaction terms is reduced from 6(or 4) to 3. More importantly, the MSE of two-step model almost stays the same, and the difference between SHL and WHL is also minimized. As shown in Figure 2, this generates a more interpretable model for dataset 2, with 8 main effects and 3 pairwise interaction terms.

For the real world dataset, 15 out of the 28 variables were selected by the CART screening. The main effects and interaction terms estimated by SHL and WHL are visualized in Figure 3. The MSE are 0.206 and 0.195 respectively. Unlike the simulated datasets, the SHL and WHL generated quite different model results. Under the WHL, the MDI score is negatively associated with PBDE 99, PCB 187, mom's education and mom's smoking status; and positively associated with PCB105, PCB153, PCB 156, and PCB 199.

Conclusions:

The results presented here demonstrate that shrinkage-based regression methods, coupled with an initial tree-based screening algorithm when necessary, can expand the linear regression framework to include non-additive interactions and produce succinct, interpretable models. In both datasets, only subsets of exposures are found to contribute to the variability in the outcome. Interactions between exposures and between exposures and confounders are also found. Based on MSE from 10-fold cross-validation, the models that best explained both datasets are selected. Two-step strategy is only beneficial when the dataset has high dimensionality. When the proposed strategy is applied to the real world dataset, the strategy is sensitive to the strong and weak hierarchical assumption. This could be due to the low signal-to-noise ratio or strong non-linearity in the real world dataset. The performance on the real world dataset is less satisfying than the simulated datasets, thus the current methods need to be further improved before putting into practical use.

References:

1. Bien J, Taylor J, Tibshirani R. 2013. A LASSO for hierarchical interactions. The Annals of Statistics 41:1111-1141.

2. Gelman A. 2008. Scaling regression inputs by dividing by two standard deviations. Statistics in medicine 27:2865-2873.
3. Liu Y, Wang J, Ye J. An efficient algorithm for weak hierarchical LASSO. In: Proceedings of the Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, ACM, 283-292.
4. Sun Z, Tao Y, Li S, Ferguson K, Meeker J, Park S, et al. 2013. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: Possible choices and comparisons. Environmental Health 12:85.
5. Tibshirani R. 1996. Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society Series B (Methodological):267-288.

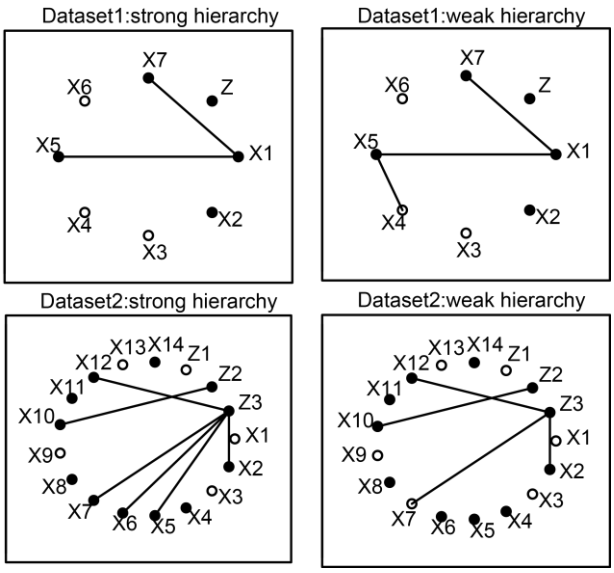


Fig.1. Wheel plot showing the main effects and interactions for strong and weak hierarchical lasso. Filled nodes represent nonzero main effects, edges represent nonzero interactions.

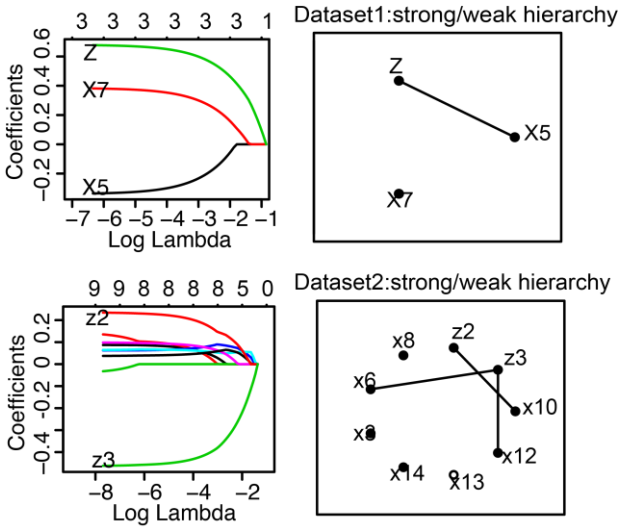


Fig.2. Main effects and interactions for simulated datasets generated from the two-step strategy. For both datasets, strong and weak hierarchical lasso generate the same interaction sparsity.

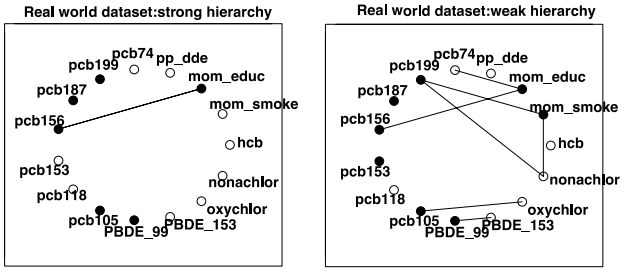


Fig.3. Main effects and interactions in the real-world dataset analyzed by the two-step strategy. Strong hierarchy and weak hierarchy yielded different model estimates.

Table 1. Compare the results of five shrinkage methods: one-step v.s. two-step strategy

	(A) One-step strategy			(B) Two-step strategy			
Methods	Main effects	Interactions	MSE[a]	Main effects	Interactions	MSE	Δ_{MSE}
Dataset 1							
Ridge	All	/	0.061	X5, X7, Z	/	0.070	15%
Lasso	X1, X5, X7, Z	/	0.059	X5, X7, Z	/	0.068	15%
EN[b]	X1, X2, X3, X5, X7,Z	/	0.057	X5, X7, Z	/	0.069	22%
SHL[c]	X1, X2, X5, X7,Z	X1*X5, X1*X7	0.045	X5, X7, Z	X5*Z	0.056	26%
WHL[d]	X1, X2, X5, X7,Z	X1*X5, X1*X7, X4*X5	0.045	X5, X7, Z	X5*Z	0.056	26%
Dataset 2							
Ridge	All	/	0.134	x10, x12, x13, x14, x3, x6, x8, z2, z3	/	0.132	-1%
Lasso	x4, x6, x8, x10, x11, x14, z2, z3	/	0.132	x10, x12, x14, x3, x6, x8, z2, z3	/	0.131	-1%
EN	x4, x6, x8, x10, x11, x14, z2, z3	/	0.133	x10, x14, x3, x6, x8, z2, z3	/	0.134	0%
SHL	x2, x4, x5, x6, x7, x8, x10, x11, x12, x14, z2, z3	x10*z2, x12*z3, x7*z3, x6*z3,x5*z3, x2*z3	0.116	x10, x12, x14, x3, x6, x8, z2, z3	x10*z2, x6*z3, x12*z3	0.117	1%
WHL	x2, x4, x5, x6, x8, x10, x11, x12, x14, z2, z3	x10*z2, x12*z3, x7*z3, x2*z3	0.118	x10, x12, x14, x3, x6, x8, z2, z3	x10*z2, x6*z3, x12*z3	0.120	1%

a. Mean square error of the model selected from 10-fold cross validation; b. Elastic-net; c. Strong hierarchical lasso; d. Weak hierarchical lasso

Table 2. Main effects and interactions for real-world datasets generated from the two-step strategy.

	Main effects	Nonachlor	Oxychlor	PBDE 153	PCB 153	PCB 156	PCB 187	PCB 199	PCB 74	Mom _educ	Mom _smoke
PBDE 99	-0.035	0	0	-0.014	0	0	0	0	0	0	0
PCB 105	0.022	0	-0.011	0	0	0	0	0	0	0	0
PCB 153	0.002	0	0	0	-0.002	0	0	0	0	0	0
PCB 156	0.007	0	0	0	0	-0.007	0	0	0	0.007	0
PCB 187	-0.011	0	0	0	0	0	0.011	0	0	0	0
PCB 199	0.027	-0.010	0	0	0	0	0	0	0	0	0.003
Mom_ educ	-0.100	0	0	0	0	0.007	0	0	0.031	-0.024	0
Mom_ smoke	-0.008	0.004	0	0	0	0	0	0.003	0	0	0

9. A Two Stage Approach to Analysis of Health Effects of Environmental Chemical Mixtures: Informed Sparse Principal Component Analysis Followed by Segmented Regression

Presenting Author: Roman Jandarov

Organization: University of Cincinnati

Contributing Authors: Roman A. Jandarov, Liang Niu, and Susan M. Pinney

Abstract:

Analysis of health effects of exposure to real-world environmental chemical mixtures poses various challenging problems to researchers. These problems are often related to dimensionality of the potential exposures of interest, the complex correlation structure in the exposures, high uncertainty in the measurements of the exposures (e.g. high number of measurements close to limit of detection (LOD) of the exposure), possible non-linear and interacting relationship between the exposures and the health endpoints that may depend on the magnitude of the exposure mixtures, the presence of continuous and categorical confounders, and difficulty of interpreting the result of statistical models. In an attempt to resolve these issues, we propose a two-stage approach that can be applied to the analysis of health effects associated with environmental chemical mixtures. In the first stage of our approach, we propose to reduce the dimensionality of the exposure variables using a novel informed sparse principal components analysis (PCA). In the second stage of the approach, we propose to analyze the effects of these lower dimensional exposure variables (principal scores) using a segmented linear regression analysis.

In general, PCA allows extracting a small number of important variables from a higher dimensional set of exposures that explain most of the variability in the data. These important variables are called PCA scores. In PCA, the PCA scores are calculated by projecting the original exposures onto the vectors called PCA loadings. Introducing sparcity to PCA scores makes the loading vectors sparse and sparcity helps to increase interpretability of the scores. In traditional sparse PCA, sparsity of the loadings is controlled by a parameterized penalty function that adds a penalty to the loadings in the optimization problem using the absolute values of the loadings. This penalty function is not informed by prior information about how/if the original exposures are reliable. In our informed sparse principal component analysis, in contrast to the traditional sparse PCA, we propose to penalize the PCA loadings not just by their absolute values, but by also adding weights to the algorithm to inform the penalty function about which exposures are more reliable than the others. These weights can be obtained by experts' knowledge on how exposures are measured and which measurements tend to be more prone to measurement error. For example, exposure weights can be calculated as a proportion of measurements close to the LOD for each exposure variable, or those where the coefficient of variation for the quality control samples is large. By incorporating information about reliability of the exposures, the informed sparse PCA therefore can potentially eliminate variables that do not contribute to explaining the total variability of

the data and unreliable variables; and construct PCA scores that are sparse, interpretable and more reliable.

There are a huge number of methods to model a potential non-linear relationship between the outcome and the predictors. While some of these methods may result in very accurate predictions of the outcome (e.g. random forests, models with splines, deep artificial neural networks, etc.), interpreting the effects of the predictors on the outcome in these methods is always very challenging. Keeping in mind that interpretability of the estimated effects from the health models is crucial in the analysis of exposures to environmental chemical mixtures; we propose to analyze the effects of PCA scores obtained from informed sparse PCA using a simple segmented linear regression. In segmented linear regression with confounders, we obtain estimates for the slopes by allowing for possible multiple breakpoints in the relationship between principal scores and the health endpoint. The result of the model is easy to interpret in order to characterize the effects of chemical exposures. In this model, we do not include interaction terms between the principal scores since it is expected that correlations between these variables is low because of the first stage of the approach.

Therefore, we note that in the first stage of our two-stage approach, by utilizing informed sparse PCA, we attempt to resolve the issues with dimensionality of the exposures, the complex correlation structure, and high uncertainty in the measurements of the exposures. In the second stage of the approach, we model the possible non-linear relationship between the exposures and the health endpoints by also accounting for confounders using a very interpretable segmented linear model.

In our draft, we plan to demonstrate the utility of the two-stage approach in simulated data and in an application to two data sets made available by the NIEHS workshop.

In this document, we summarize our preliminary application of the two stage analysis approach to simulated datasets available at the NIEHS workshop's website. We omit the details of the optimization problem and algorithms that are necessary to conduct informed sparse principal component analysis (IsPCA) in the first stage. A more detailed version of this document (with answers to summary questions listed on the website) can be found at https://www.dropbox.com/s/0bkl6g89tcmc9o7/applicationIsPCAsummary.pdf?dl=0

Analysis of Data Set 1: In the first stage of our approach, we apply informed principal component analysis to data on exposure variables X1 - X7. For informed sparse PCA, we assume that all weights for the exposure variables are equal based on the assumption that there is not mis-measurement in these data.

Table 1 presents the loadings for the first four principal components obtained from IsPCA. Using this table, it is clear that the first principal component score (PC1) is a mixture of the exposures X1, X2 and X3; the second principal component score (PC2) is primarily a mixture of the exposures X5 and X6; the third principal component score (PC3) is a mostly related to X4; and the fourth principal component score (PC4) has its mass on X7.

In the second stage of the approach, we want to study the health effect of principal scores calculated in the first step. As described in the abstract, we use segmented linear regression to model the relationship between Y and PC1, PC2, PC4 and PC4 given the confounding variable Z:

Y~PC1+PC2+PC3+PC4+Z + possible interactions between Z and PCs

In our final model, we included interactions PC1*Z and PC2*Z. For variables PC1-PC4, we assumed the possibility of break points to capture the potential non-linear relationship between Y and these variables.

Estimates of the slopes with corresponding confidence intervals for the final model are given in Table 2. Based on these slopes and the estimated break points, Figure 1 shows the effects of principal scores on Y. Table 2 and Figure 1 can easily be used to understand the ranges where principal scores and Y are significantly associated: for example, we see that PC1 is associated with Y after the breakpoint, PC2 is associated negatively with Y only in the middle range between the breakpoints, PC3 is associated with Y between the breakpoints, and PC4 is associated with Y after the first breakpoint. In the final model, Z and interaction of Z with PC2 (PC2*Z) were also significantly associated with Y (with p.values<0.001). The adjusted R^2 for this model was around 0.91.

Analysis of Data Set 2: From the boxplots of the variables, we see that different exposures X1-X14 have different ranges and values. Since information on LODs for these variables are not given in this exercise, based on our domain expert, we calculated weights for the exposure variables based on their ranges. These weights were 0.13, 0.09, 0.06, 0.07, 0.06, 0.07, 0.20, 0.08, 0.27, 0.10, 0.11, 0.80, 0.40, and 0.16. Using these weights, we can order the exposures by assumed reliability as follows: X5, X3, X4, X6, X8, X2, X10, X11, X1, X14, X7, X9, X13, and X12. Here, for example, we are assuming that X12 and X13 are of the lowest quality. In general, we can construct weights using prior information on LODs of the exposures and/or other information.

In the first stage of our analysis, we apply IsPCA to exposure data using the weights from above. The resulting loading vectors for the first 5 principal components are given in Table 3. We note here that these loadings are different from what we expect to get from the traditional sparse PCA since traditional PCA assumes equal weights for all variables. In Table 3, we see that the least reliable variables X12 and X13 ended up being eliminated by the algorithm. From Table 3, we can interpret PC1 as a mixture of X3, X4 and X5, PC2 as a mixture of X6 and X8, PC3 as a mixture of X8, X10 and X11, PC4 as a mixture of X3,X4, X5, X8 and X14 and PC5 as a mixture of X1 and X2.

In the second stage, we used segmented regression model to investigate the effects of PC1-PC5 on Y given the confounder variables Z1, Z2, and Z3:

Y~PC1+PC2+PC3+PC4+PC5+Zs + possible interactions between Zs and PCs

Again, we assumed the possible non-linearity between the scores and Y. In our final model, the results showed that only PC1 had a broken-line relationship with Y (as seen in Figure 2), while other principal scores didn't show any evidence of non-linearity. The association between PC1 was not significant in any of the segments. At the 0.05 level, the following main effects of principal scores were statistically significant: PC3 (estimate of the slope: 0.068514, p.value = 0.01481), PC5 (estimate of the slope: 0.176889, p.value < 0.0001). Additionally, confounders Z3 (estimate of the slope: -0.82947, p.value = 0.0001), Z2 (estimate of the slope: - 0.004907, p.value = 0 0.00664) and interaction term PC5*Z3 (estimate of the slope: -0.2082, p.value < 0.0001) were also statistically significantly associated with Y. The adjusted R^2 for this model was around 0.51.

Table 1: IsPCA loadings for first 4 components

	PC1	PC2	PC3	PC4
X1	0.56	0	0.15	0.26
X2	0.57	0	0.19	0.24
X3	0.6	0	0.17	0.26
X4	0	-0.1	0.95	-0
X5	0	0.71	0	0
X6	0	0.71	0	0
X7	0	0	0	0.9

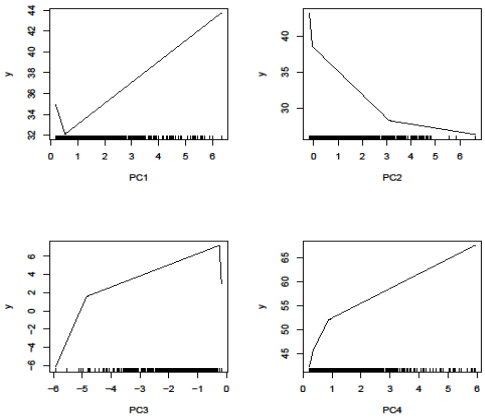


Figure 1: Effects of principal scores

Table 2: Estimated slopes and confidence intervals

$PC1				$PC3			
	Est.	CI(95%).l	CI(95%).u		Est.	CI(95%).l	CI(95%).u
slope1	-8.24	-22.27	5.80	slope1	7.22	-4.95	19.38
slope2	2.01	0.94	3.08	slope2	1.22	0.78	1.65
				slope3	-54.06	-233.70	125.60
$PC2							
	Est.	CI(95%).l	CI(95%).u	$PC4			
slope1	-37.86	-161.90	86.14		Est.	CI(95%).l	CI(95%).u
slope2	-3.25	-3.71	-2.78	slope1	25.99	-23.39	75.37
slope3	-0.54	-1.80	0.71	slope2	11.92	7.50	16.34
				slope3	3.06	2.68	3.44

Table 3: IsPCA loadings for first 5 components

	PC1	PC2	PC3	PC4	PC5
X1	0	0	0	0	0.6
X2	0	0	0	0	0.8
X3	0.58	0	0	0.44	0
X4	0.52	0.01	0	0.47	0
X5	0.62	0	0	0.4	0
X6	0	0.86	0	0.05	0
X7	0	0	0	0	0
X8	0.09	0.51	0.11	0.38	0
X9	0	0	0	0	0
X10	0	0	0.77	0	0
X11	0	0	0.63	0.06	0
X12	0	0	0	0	0
X13	0	0	0	0	0
X14	0	0	0	0.53	0

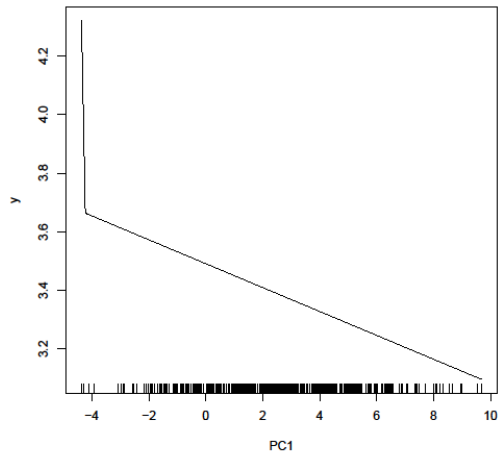


Figure 2: Effects of principal scores 1

10. Direct Assessment of Public Health Impacts of Exposure Mixtures: A Bayesian G-Formula Approach

Presenting Author: Alexander Keil

Organization: University of North Carolina, Chapel Hill

Contributing Authors: Alexander P. Keil, Jessie P. Buckley, and Jessie K. Edwards

Abstract:

One goal of assessing complex mixtures is to identify interventions or policy recommendations to improve public health. This approach fits naturally in environmental epidemiology as much of our work seeks to provide evidence for regulation of hazardous agents. Furthermore, the correlated nature of environmental exposures has important implications for what and how to intervene. Here, we present an approach to estimate the health effects of interventions or policy changes related to exposures occurring in mixtures.

We outline a Bayesian approach to the parametric g-formula to estimate the effects of interventions on a complex exposure mixture. The quantity of interest is the population mean of *Y*, referred to as E(*Y*), under various independent and joint interventions on exposures in the simulated datasets. We use a linear model for *Y* including all *X*s, *Z*s, and their two-way interactions, and coefficients are shrunk towards the null value using a Bayesian LASSO. Using this model, we estimated the posterior predictive distributions of *Y*, which correspond to the population distributions of *Y* under each intervention. We implemented our approach in STAN v2.6.0.

In the Table, we summarize a set of policy-relevant questions regarding independent and joint interventions. We focus on two types of interventions to reduce exposure: 1) an intervention to reduce only the highest exposures and 2) an intervention that reduces all exposures. We implement intervention 1 by capping exposure at the 75th percentile (Figures 1a and 2a) and intervention 2 by reducing all exposures by half (Figures 1b and 2b). We estimate the effects of individual interventions on each exposure as well as joint interventions on all exposures or sets of exposures with high correlations that may indicate a common source (dataset 1: *X*1, *X*2, *X*3; dataset 2: *X*3, *X*4, *X*5). Each effect, referred to as Diff(*Y*), is defined as the difference between E(*Y*) under the intervention and E(*Y*) under no intervention. To compare the dose-response functions for key interventions, we plot E(*Y*) over increasingly extreme downward shifts of the exposure distribution (Figure 1c and 2c). We assume that exposures are reported as measured in dataset 1 and that exposures were natural-log transformed in dataset 2.

In dataset 1, E(*Y*) under no intervention was 23.3. An intervention to shift the distribution of *X*7 had the strongest effect on *Y* of any single *X* intervention (Diff(*Y*) if the distribution of *X*7 was halved: -2.5; 95% CI: -2.7, -2.3; Figure 1a,b). The effect of a joint intervention to halve the distributions of *X*1, *X*2, and *X*3 reduced E(*Y*) to 19.8 (Diff(*Y*): -3.5; 95% CI: -4.1, -2.9); indicating that effects of *X1, X2*, and *X*3 were approximately additive. Similarly, as exposure distributions were incrementally reduced, a joint

intervention on $X1$, $X2$, and $X3$ reduced E(Y) more than decreasing all Xs (Figure 1c). In dataset 2, E(Y) under no intervention was 3.9. An intervention to shift the distribution of $X13$ had the strongest effect on Y of any single X intervention, though the estimate was imprecise (Diff(Y) if the distribution of $X13$ was halved: -0.18; 95% CI: -0.62, 0.27; Figure 2a,b). Under a joint intervention to halve the distributions of $X3$, $X4$, and $X5$, E(Y) was 3.9 (Diff(Y): -0.02; 95% CI: -0.08, 0.04), suggesting that an intervention on a source of $X3$, $X4$, and $X5$ may not be effective in reducing Y. Intervening only on $X4$ reduced E(Y) more than a joint intervention on all Xs (Figure 2c).

These results suggest that interventions should carefully target specific exposures or exposure combinations. Our approach offers a method for determining *which* exposures should be targets of interventions. For example, in dataset 2, we estimated that an intervention on a hypothetical source of three highly correlated exposures may not be as effective as intervening on a specific single component of the mixture. Furthermore, shifting the distribution of X was generally more effective than capping X at the 75th percentile in these datasets, which has useful policy implications for determining *how* to intervene. Both of our intervention types correspond to potential realistic scenarios. For example, in occupational settings some workers might be monitored for exposure and pulled out of exposed jobs when reaching a certain exposure cap. Similarly, in environmental settings, use of sulfur dioxide scrubbers may proportionately reduce sulfur dioxide exposures in all coal-fired power plants in which they are implemented. In real life applications, prior knowledge regarding exposures, their sources, and feasible exposure reduction approaches could be synthesized to test additional targeted interventions.

Our approach dovetails with other Bayesian methods for complex exposure mixtures. Consequently, it could be used to compare the magnitudes of main-effect and product term parameters, as well as joint dose-response functions, rather than utilizing the posterior predictive distribution. Similarly, precision in our estimates may be improved by using stronger priors, which may be informed by prior research or results from hierarchical models. Our novel extension of the more traditional Bayesian approach, in which we estimate the effects of hypothetical interventions, may strengthen traditional risk assessment as it allows for direct comparison of policy alternatives and can accommodate cost benefit analyses. In a real world setting, we can use prior knowledge on costs and feasibility to identify potential interventions or policy alternatives to regulate exposures. We can then compare health outcomes under each scenario to inform decisions about public health policy.

Study question	Intervention type		Workshop question	Figure
	Cap[a]	Shift[b]		
Independent effects				
How much can we reduce *Y* by limiting exposure to each *X*?	Cap each *X* at κ	Divide each *X* by δ	1, 2	1/3
What is the incremental reduction in *Y* associated with lower exposure limits on *X*?	Cap each *X* at a range of κ	Divide each *X* by a range of δ	1, 2	2/4
Joint effects				
How much can we reduce *Y* by limiting exposure limits to all *X*?	Cap all *X* at κ	Divide all *X* by δ	3, 4	2/4
What is the incremental reduction in *Y* associated with lower exposure limits on all *X*?	Cap all *X* at a range of κ	Divide all *X* by a range of δ	3, 5	2/4
How much can we reduce *Y* by limiting exposure to the *p X* that are most strongly associated with *Y*?	Cap *p X* at κ	Divide *p X* by δ	3	Not shown
Non-intervention quantities[c]				
What is the conditional effect of each *X* on *Y*?	n/a	n/a	1, 2	Not
What is the magnitude of each *n*-way interaction term?	n/a	n/a	3, 5	Shown

Table. Examples of interventions to assess independent and joint effects of exposures (X) on outcome (Y)

[a] If *X*>κ, set *X* to κ, where κ is a quantile of the exposure distribution, an actual or potential standard, or other meaningful level.

[b] Shift the entire exposure distribution of *X* (e.g., divide each observed *X* value by δ).

[c] Posterior distributions of coefficients estimated in the outcome model.

Figure 1. Effects of single and joint interventions on the mean outcome (E(*Y*)) in dataset 1. The effect, Diff(*Y*) is defined as the difference between E(*Y*) under the intervention and E(*Y*) under no intervention. The left panels plot the change in E(*Y*) following interventions to A) cap exposures at the 75th percentile or B) reduce all exposures by half. Panel C plots E(*Y*) over increasingly stringent reductions in exposures.

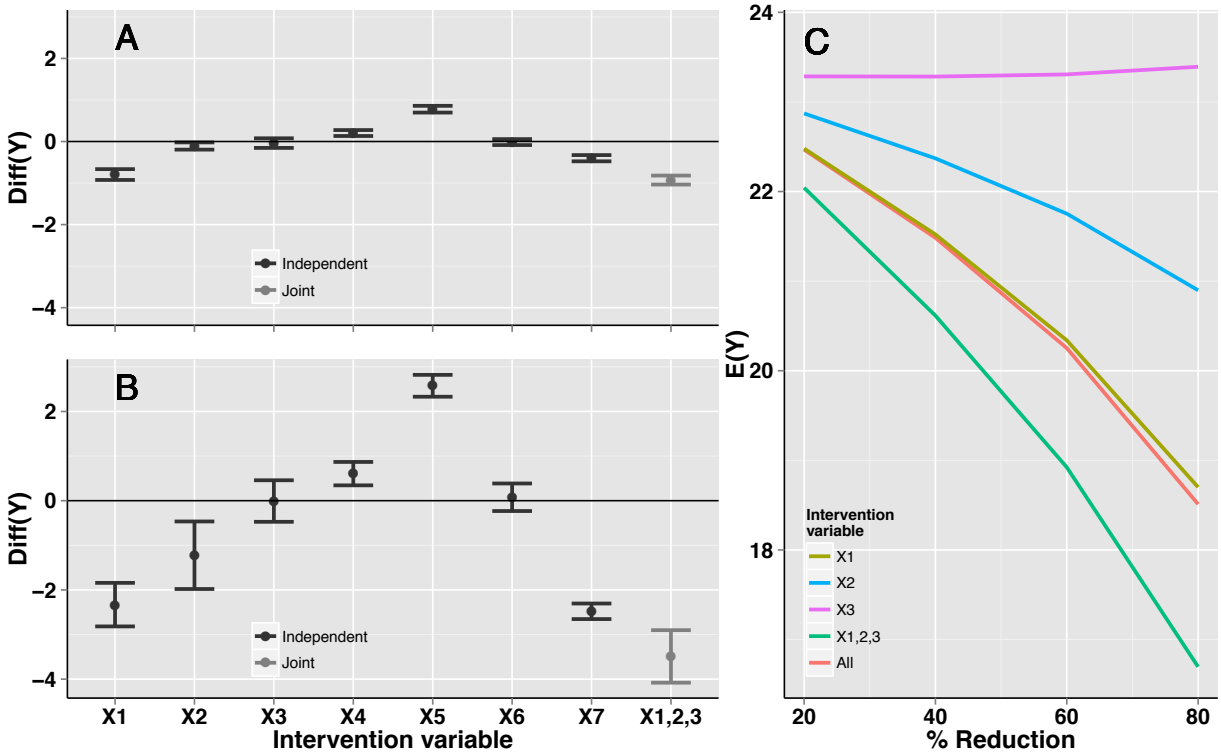
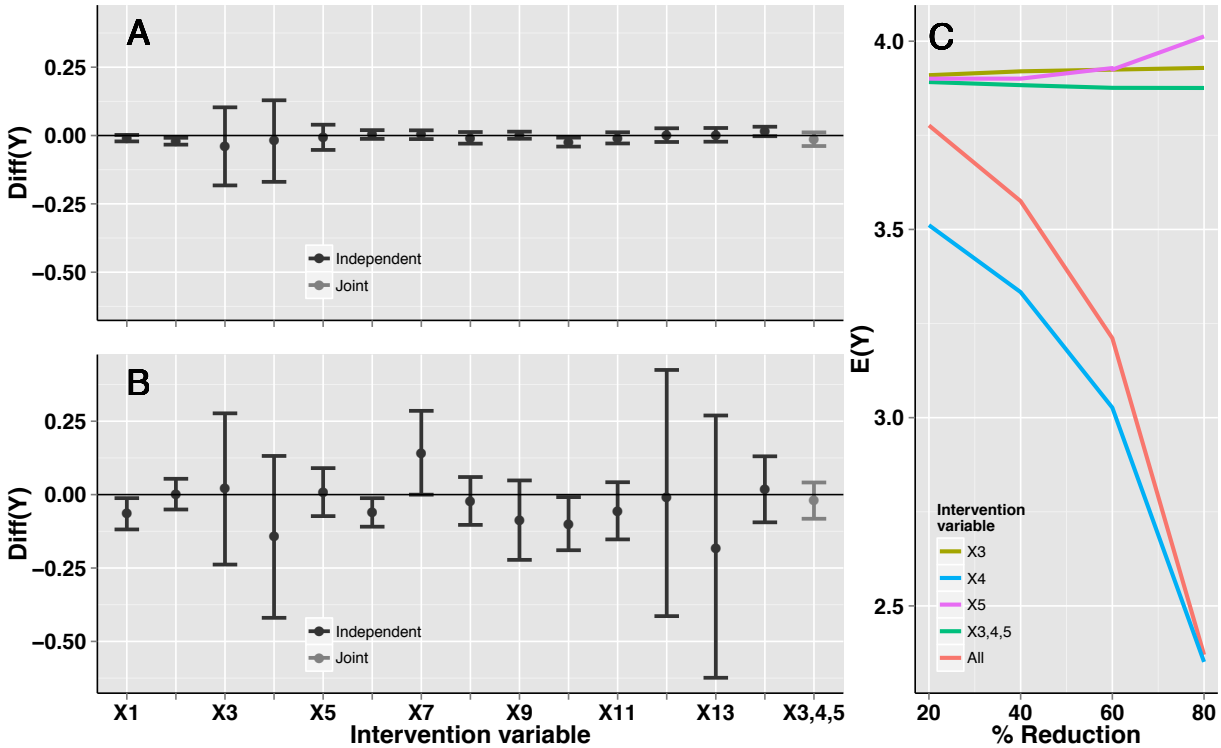


Figure 2. Effects of single and joint interventions on the mean outcome (E(*Y*)) in dataset 2. All panels correspond to the same interventions as given in Figure 1.



11. Do Your Exposures Need Supervision?

Presenting Author: Jenna Krall

Organization: Emory University

Contributing Authors: Jenna R. Krall, Howard H. Chang, Katherine M. Gass, W. Michael Caudle, and Matthew J. Strickland

Abstract:

Background: Determining the environmental exposures most associated with a health outcome is challenging when exposures are highly correlated. Statistical and epidemiologic methods, with various strengths and limitations, have been introduced to estimate such associations. Here we applied two methods, with different advantages and disadvantages, to help reach a consensus about the presence and/or absence of associations. Specifically, we used both an unsupervised approach, which groups exposures irrespective of the outcome, and a supervised approach, which uses the outcome to determine the most important exposures. By comparing the two sets of results, we can estimate associations between groups of exposures and the outcome and also determine which exposures within the groups might be most important.

Methods: Our unsupervised approach used principal component analysis (PCA) to reduce multicollinearity and to estimate joint associations between combined exposures and the outcome. We selected principal components (PCs) that explained most of the variability in the exposure data and applied varimax rotation to improve interpretability. We estimated associations between the PC scores, which are combinations of original exposures, and the outcome using linear regression with and without adjustment for confounding. The PCA approach is well suited for highly correlated exposures, but does not separate out the most important exposures within a group, and the regression coefficients for the new exposures can be difficult to interpret. We compared the PCA approach to using the exposures directly in both unadjusted linear models and linear models adjusting for other exposures and potential confounders.

We also determined the exposures most associated with the outcome using classification and regression trees (C&RT). C&RT is a supervised form of hierarchical clustering where the data are successively split into dichotomous groups, such that each resulting group contains increasingly similar responses for the outcome. Each binary split is determined by the value of the exposure that best explains the outcome (based on analysis of variance) for that partition of the data. After the full (saturated) tree is grown, a pruning rule (based on cross-validation) is applied to select the final tree, which balances the tradeoff between parsimony and minimizing cross-validation error. To control for confounding, we applied C&RT to the residuals obtained by regressing out the effects of the confounders from both the outcome and exposure variables.

Results: For the first simulated dataset, four PCs explained most (88.9%) of the variability in the exposure data. The varimax-rotated loadings had good separation (Figure 1a), with rotated PC 1 (rotPC1) primarily representing X1-X3, rotPC2 primarily representing X5-X6, rotPC3 primarily representing X7, and rotPC4 primarily representing X4. The regression results for simulated dataset 1 are shown in Figure 1b. Using the PCA approach, we found that the groups corresponding to rotPC1-rotPC3 were significantly associated with the outcome in both unadjusted and adjusted models. While we did not find that the rotPC associated with X4 was associated with Y, X4 was negatively associated with the outcome in an adjusted model and the association was statistically significant. We also applied C&RT to simulated dataset 1 to identify the exposures most associated with Y. Using a pruned tree on the exposures and outcome (on the residuals (r) with the effect of Z regressed out), we identified X1, X5, and X7 as the most important exposures. These three exposures correspond to exposures grouped in rotPC1-rotPC3. Figure 2 shows the pruned regression tree as well as the range of exposure residual values for four selected terminal nodes of the tree.

We identified six PCs in simulated dataset 2, which explained 88.7% of the variability in the exposure data. The rotPCs generally categorized exposures as: rotPC1: X3-X5, X8, X14, rotPC2: X12-X13, rotPC3: X10-X11, rotPC4: X6-X7, X9, rotPC5: X2, rotPC6: X1 (Figure 3a). We found that rotPC1-rotPC4 and rotPC6 were significantly positively associated with the outcome, after controlling for all the rotPCs and confounders (Figure 3b). The adjusted association for rotPC5 was positive, but not statistically significant. Using C&RT, we did not find that any exposures were predictive of the outcome. Once we regressed out the effect of the confounders, the exposures explained very little of the outcome.

Because the exposures in the real-world dataset were right-skewed, we logged the exposure data for both the PCA approach and regression analysis. In the PCA analysis, we identified four PCs, which explained 78.9% of the variability in the exposures (Figure 4a). The second rotPC consisted primarily of PCBs 74, 99, 105, 118, and 138/158 and some PP-DDE, and rotPC1 consisted primarily of the remaining PCBs and most of the remaining PP-DDE. All of the PBDEs primarily included in rotPC3 and rotPC4 consisted mostly of Hexachlorobenzene, Trans-Nonachlor, and Oxychlordane. In this dataset, we found substantial evidence of confounding. After controlling for confounders, none of the rotPCs were significantly associated with Mental Development Index (MDI) (Figure 4b). C&RT did not find evidence that any one exposure was predictive of MDI.

Conclusion: We found that when exposures are all highly correlated, PCA can be used to reduce multicollinearity in linear models and determine those groups of exposures most associated with an outcome. C&RT does well in identifying important exposures when confounding is not severe, however when a confounder is very correlated with several exposures and the outcome, it can be difficult to distinguish the effect of those exposures from the effect of the confounder. We demonstrated that pairing an unsupervised approach, PCA, and a supervised approach, C&RT, can lead to better intuition about which exposures are most associated with an outcome. A lack of consistency across these methods highlights the uncertainty in the results and demonstrates how conclusions can vary according to the statistical model chosen for analysis.

Figure 1: Results for simulated dataset 1 using the PCA approach

(a) Varimax-rotated PC loadings (b) Estimated regression coefficients by rotPC group

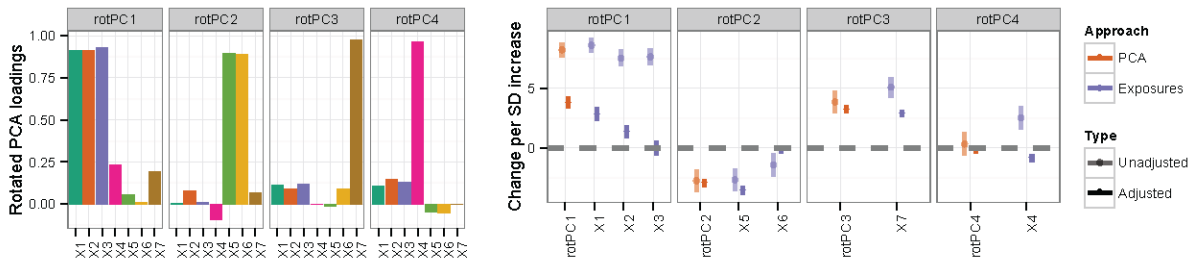


Figure 2: Pruned regression tree using C&RT for simulated dataset 1. For four selected terminal nodes, we show the distribution of residuals for the exposure variables that define that terminal node.

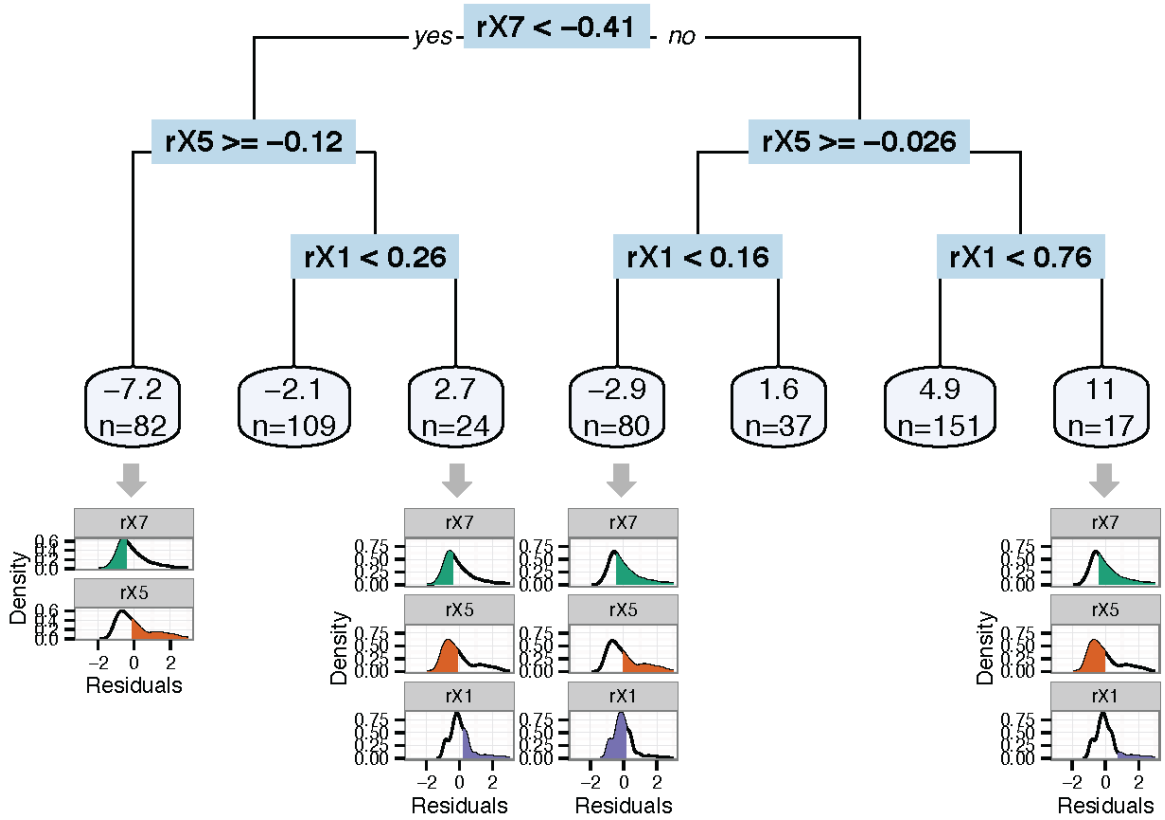
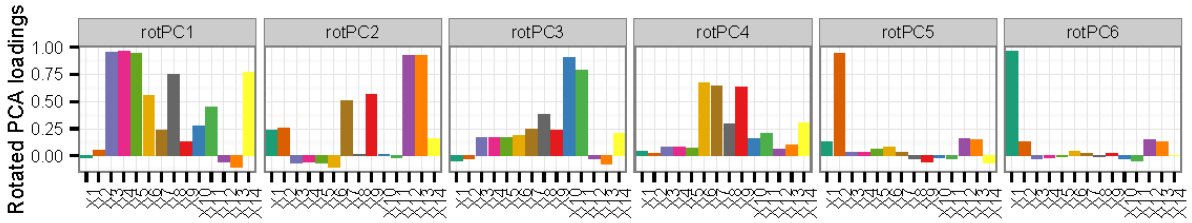


Figure 3: Results for simulated dataset 2 using the PCA approach

(a) Varimax-rotated PC loadings



(b) Estimated regression coefficients by rotPC group

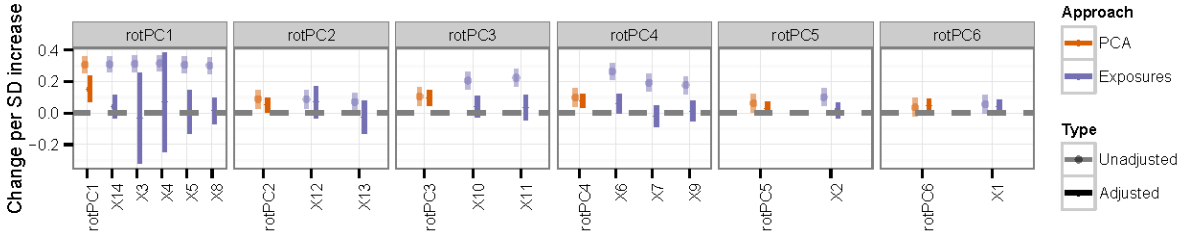
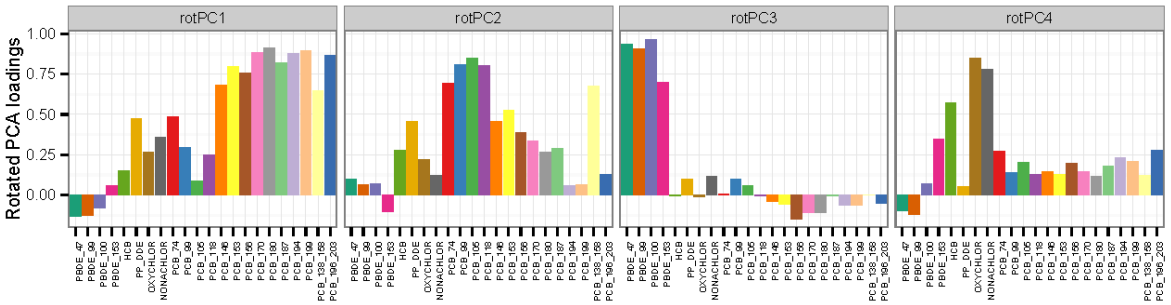
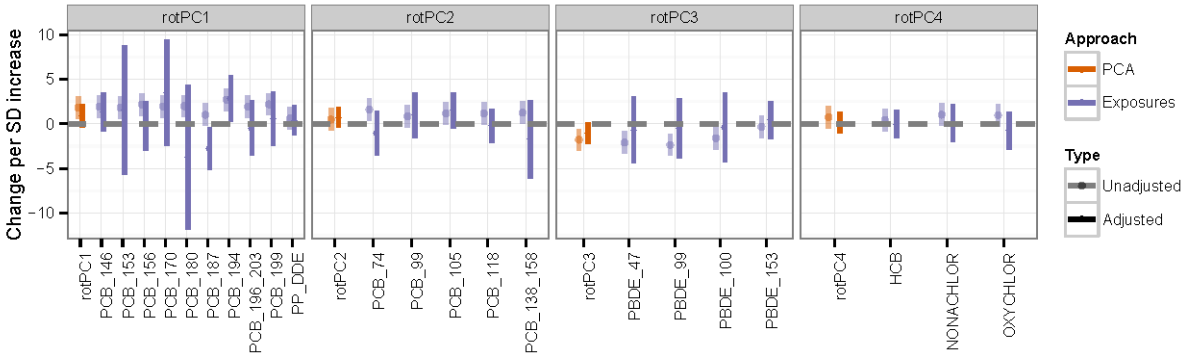


Figure 4: Results for the real-world dataset using the PCA approach

(a) Varimax-rotated PC loadings



(b) Estimated regression coefficients by rotPC group



12. Principal Component Analysis: An Application for Understanding Health Effects of Environmental Chemical Mixture Exposures

Presenting Author: Cristina Murray-Krezan

Organization: University of New Mexico

Contributing Authors: Johnnye Lewis, Curtis Miller, and Cristina Murray-Krezan

Abstract:

Data Set 1

We approached the analysis of Data Set 1 as a **dose-response function**, where the dependent variable Y was considered to be some outcome measure of exposure, X1-X7 were assumed to be measurements of individual chemical exposure, and Z was a potentially confounding dichotomous variable.

Statistical Analysis: All continuous independent variables (X1-X7) were highly skewed and the dependent variable Y was somewhat skewed. The logs of X1-X7 were approximately normally distributed and were used for all analyses. Many variables were highly correlated to each other, in particular X1, X2, and X3 (all $r \geq 0.86$) and X5 and X6 ($r=0.70$), indicating potential joint-exposure effects. We implemented two regression approaches, both using principal component analysis (PCA). We used model selection techniques based on AIC to guide selection of significant interactions. After reducing the interactions, we determined whether any main effects not contributing to an interaction effect could be eliminated.

Method 1: In effort to reduce the data, we performed one PCA on all seven log-transformed X variables to yield the principal component (PC) scores PC1-PC7. We selected PC1-PC4 since they accounted for >90% of the common variance in the independent variables. The first three PCs were dominated by unique sets of variables (PC1 by X1, X2, and X3; PC2 by X5, X6; and PC3 and PC4 by both X4 and X7). A regression model was fitted to Y with independent variables Z, PC1, PC2, PC3, PC4, and the interactions between Z and the PCs.

Method 2: We performed PCA on X1-X3 and separately on X5-X6, using PC11, PC12, PC13, PC21, and PC22 for linear regression analysis rather than the original variables to reduce the effects of multicollinearity. We linearly regressed Y onto the following variables: Z, PC11, PC12, PC13, log X4, PC21, PC22, log X7; all second order interactions with Z; interactions between log X4, log X7, and the PCs; and select third order interactions.

Results: We found both methods had similar results. Figures 1a and 1b show how well our models fit the observed data and report fit statistics for model comparison. Note that the final model for Method 2 contained 14 independent variables compared to 8 for Method 1. We present Method 2 results since the interpretation of the PCs is more straightforward given that each is comprised only of the correlated variables. The final regression model is reported in Equation 1. The residuals from the model were normally distributed and when plotted against the predicted $\hat{Y}$, showed random distribution and had constant variance.

PC11 interacts with covariate Z indicating that Y decreases as PC11 increases in the presence of Z ($P<0.0001$). Z also interacts with PC12 and PC21 with a higher Y as PC12 and PC21 increase in the presence of Z. The interaction between Z and PC12 results in the largest change in Y out of all the regressors. Log X7 interacts significantly with Log X4, PC11 (negative effects on Y), and PC21 (positive effect on Y). Log X4 interacts significantly with PC21 (negative effect). PC22 is significant as a main effect only and is negatively related to Y with the second largest influence on Y of all the regressors. PC31 is not significant in this model.

Data Set 2

We developed a **dose-response function** for Data Set 2, where the dependent variable Y was considered to be some outcome measure of exposure, X1-X14 were assumed to be measurements of individual chemical exposure, Z1 and Z2 were continuous potentially confounding variables, and Z3 was a dichotomous potentially confounding variable.

Statistical Analysis: All variables were normally distributed. Correlation analysis yielded high correlations among many variables, especially between (X3, X4, X5, X8), (X7, X9), (X10, X11), and (X12, X13), all with $r>0.70$, indicating joint-exposure effects. Analysis was similar to those described for Data Set 1, including model selection.

Method 1: A single PCA was performed on the 14 X variables. PC1-PC7 were retained as independent variables for the regression since they accounted for >90% of the variability. PC1 was dominated by X3-X6, X8, X11, and X14; PC2 by X1, X2, X7, X9, X11, and X13; PC3 and PC4 by X1, X2, X10, X11; PC5 by X1, X2; PC6 by X6; and PC7 by X6, X7, and X14. We regressed Y onto Z1-Z3, PC1-PC7, and all second-order interactions between the Z covariates and X variables.

Method 2: PCA consisted of four PCAs on the sets of correlated variables yielding PC11-PC14, PC21-PC22, PC31-PC32, and PC41-PC42. We regressed Y onto Z1-Z3, PC11-PC14, PC21-PC22, PC31-PC32, PC41-PC42, X1, X2, X6, X14, all second order interactions between the Zs, first-order PCs, and remaining Xs; and select three-way interactions.

Results: We found both methods had very similar results. Figures 2a and 2b show how well our models fit the observed data and report fit statistics for model comparison. Note that the final models for both methods had the same number of independent variables (12). The PCs from Method 1 were dominated by many variables making this model especially difficult to interpret. Instead, we present the results from Method 2, given by the dose-response model in Equation 2. The residuals were normally distributed and had constant variance.

We found that covariate Z1 was not significant. Z2 interacted significantly with PC31 (positive effect on Y), and X14 (negative effect); Z3 interacted significantly with X2, PC11, and PC41 (all negative effects on Y in the presence of Z3). In fact, Z3 PC41 had at least 3 times the effect on Y that any other variable had indicating that those with a positive Z3 and higher values of X12 and X13 had significantly decreased Y values.

Discussion:

Principal component regression is often used to isolate variance contributed by highly correlated variables so that multicollinearity is minimized in the regression model since PCs obtained from the same analysis are orthogonal and hence uncorrelated with each other. Our two methods were similar and fit the observed data reasonably well. More work needs to be done to interpret the contributions of the PCs and how they account for the variance in Y.

Data Set 1

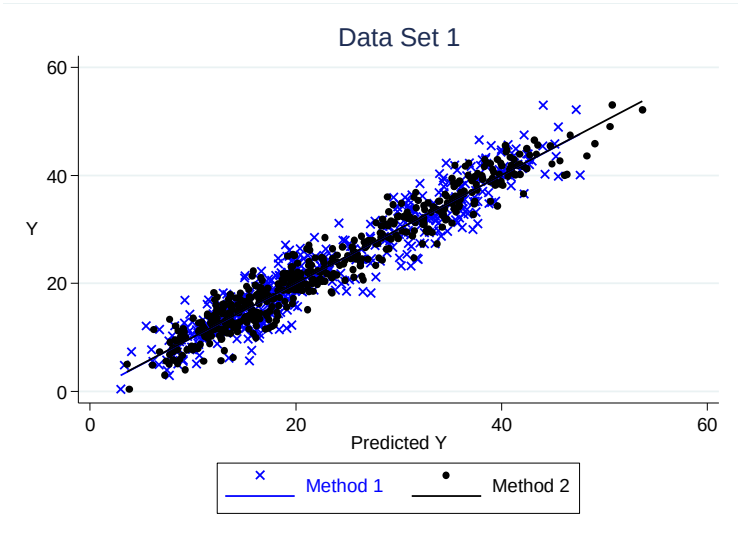


Figure 1(a). Observed Y vs. Predicted Y from the two methods with linear regression lines for *Y* regressed on $\hat{Y}$.
Method 1: R^2 = 0.911, F=628.79, RMSE =3.26
Method 2: R^2 = 0.941, F=554.20, RMSE = 2.64

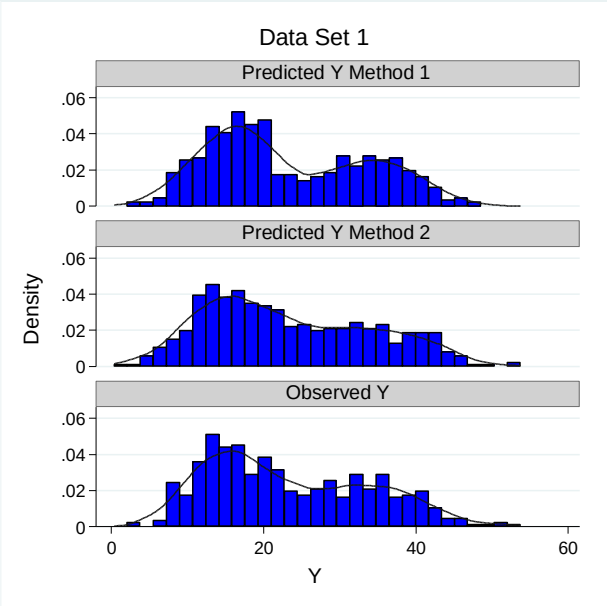


Figure 1(b). Distributions of $\hat{Y}$ from Methods 1 and 2, and distribution of observed *Y*.
Kolmogorov-Smirnov test for equal distributions (Methods 1 and 2): p = 0.9022

Dose-Response Model

$$
\begin{aligned}
Y = {} & 18.77 + 9.28 * Z - 2.31 * PC_{11} + 0.98 * PC_{12} - 1.14 * \log X_4 + 3.08 * PC_{21} - 2.54 * PC_{22} + 3.61 * \log X_7 \\
& -2.67 * Z \times PC_{11} + 3.26 * Z \times PC_{12} + 0.54 * Z \times PC_{21} \qquad \textbf{(Eq. 1)} \\
& -0.63 * \times PC_{11} \times \log X_7 - 0.90 * PC_{21} \times \log X_4 + 0.51 * PC_{21} \times \log X_7 - 0.68 * \log X_4 \times \log X_7
\end{aligned}
$$

Data Set 2

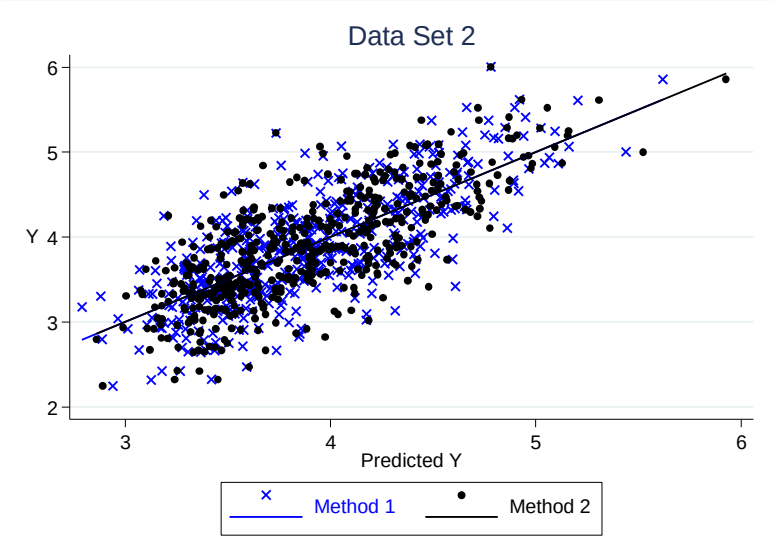


Figure 2(a). Observed Y vs. Predicted Y from the two methods with linear regression lines for *Y* regressed on $\hat{Y}$.
Method 1: R^2 = 0.558, F=628.66, RMSE = 0.434
Method 2: R^2 = 0.558, F=629.30, RMSE = 0.434

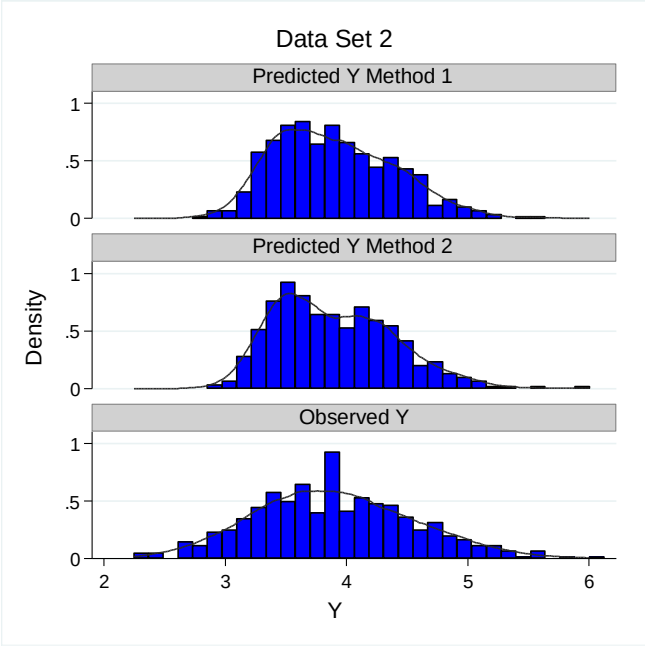


Figure 2(b). Distributions of $\hat{Y}$ from Methods 1 and 2, and distribution of observed *Y*.
Kolmogorov-Smirnov test for equal distributions (Methods 1 and 2): p > 0.99

Dose-Response Model

$$Y = 3.99 + 0.0086 * Z_2 - 0.88 * Z_3 + 0.097 * X_2 + 0.065 * PC_{11} - 0.0059 * PC_{31} + 0.33 * PC_{41} + 0.17 * X_{14}$$
$$+0.0034 * Z_2 \times PC_{31} - 0.0028 * Z_2 \times X_{14} - 0.14 * Z_3 \times X_2 - 0.055 * Z_3 \times PC_{11} - 0.42 * Z_3 \times PC_{41} \quad \textbf{(Eq. 2)}$$

13. Interpretation Without Causation: A Data Analysis at the Intersection of Statistics and Epidemiology

Presenting Author: Emily Mitchell

Organization: National Institute of Child Health and Human Development, National Institutes of Health

Contributing Authors: Emily M. Mitchell, Neil J. Perkins, Germaine M. Buck Louis, and Enrique F. Schisterman

Abstract:

Background

Assessing health effects of environmental chemical mixtures is a common challenge in epidemiological studies and is particularly relevant for risk assessment. Generally, prediction modeling is more straightforward than etiologic investigation, which requires a nuanced blending of epidemiologic and statistical methods. Statistical techniques estimate exposure effects under the assumption of a correct model. Causal diagrams explicitly defining the structure of exposure effects, even highly correlated exposures, is a much better tool than traditional forward or backward variable selection for etiologic investigation. *A priori* knowledge here is limited to the stated exposures of interest and potential confounders for both datasets. In the absence of *a priori* knowledge of the underlying relations, we ascertain these relations through observed associations.

Data Set 1

This dataset contains 7 exposure variables (X_1 - X_7), a single binary confounder (Z) and a continuous outcome (Y). Initial investigation into the contribution of Z on Y suggested that Y was likely a mixture of normal distributions, stratified by levels of Z (Figure 1). Each of the exposures was log-transformed for normality prior to analysis. While log-transformation on skewed independent predictors is not always necessary, additional investigation suggested that the relationship between the exposures and the outcome was linear on the log-scale of the exposures, thus justifying the transformation (Figure 2). Nearly all of the exposures (except X_4) appeared to contribute to the distribution of Y. The marginal association between several of the exposures on Y was different among the levels of Z (Figure 2), indicating a statistical interaction and effect modification between the exposure and Z.

Several groups of exposures were highly correlated, suggesting a joint mode of exposure in environmental exposures or a common ancestor. For instance, pairwise correlations between log(X1), log(X2), and log(X3) exceeded 0.85. Without a specific etiological question, we performed principal components analysis (PCA) on these variables for inclusion in a regression model. This method can be viewed as a variation on the additivity assumption of chemical mixtures. Additional analysis revealed a potential interaction between log(X5) and log(X6), due to strong association seen between the weighted sum as well as the weighted difference of these variables on the outcome.

The final model relating the joint distribution of the exposure to the outcome was defined as a linear regression model with parameter estimates given in Table 1. The contributions of log(X1), log(X2), and log(X3) were included as PC1, a weighted sum based on the first principal component. Interactions between PC1 and Z, and between log(X5) and log(X6), were also included. The AIC for our proposed model was 2471, whereas the AIC under a model containing all exposures (log-transformed) and Z was 2542, indicating that our proposed model provides a better fit to the data.

Data Set 2

A multiple linear regression relating all individual exposures (X) and the confounders (Z) to the outcome revealed few significant associations, likely due to highly-correlated exposures. Without etiologic knowledge we assume Z_1, Z_2, and Z_3 are potential confounders of the exposures and exposures may have a common source due to high correlations (Figures 3 & 4). All variables potentially contribute to the outcome.

Separate linear regression models for each X, adjusting for Z_1 and Z_2, indicated strong associations between the individual exposures and the outcome. The risk in interpreting these estimates and their accompanying statistical tests, however, lies in the possibility of a common, unmeasured source of exposure resulting from confounding. In order to avoid this pitfall while simultaneously reducing dimensionality of the multiple linear regression model, separate regression models were fit for each exposure X, adjusting for a subset of principal components on the remaining variables. Since no prior knowledge was given on the relationship between the exposures, PC's were chosen by a new selection criterion, where the criterion for adding a PC to the model was a 10% or greater change in the exposure parameter of interest. Z_1 and Z_2 were included in all models, and models were stratified by Z_3 to permit assessment of potential effect modification. Regression coefficients, standard errors, and corresponding p-values are provided in Table 2 for each of the 14 exposure-specified PC adjusted models. PC adjustment indicated several exposures potentially associated with the outcome as well as evidence of interaction with Z_3 across most of the chemicals. It remains clear from these analyses that structural knowledge is indispensable for dimensionality reduction.

Discussion

The high correlation between some of the exposures could conceivably be the same exposure subject to measurement error, measured multiple times. *A priori* background knowledge of the biology, chemical structure, and modes of action could help navigate potential causal structures to be assessed. While beyond the scope of this particular analysis, methods to test the causal assumptions could be applied. The current analyses represent only a preliminary approach based solely on statistical investigation. While these statistical methods may improve detection of statistical significance by reducing multi-collinearity, the result unfortunately is purely numerical. Rather, accurate and thoughtful conceptualization of specific study questions could help guide further analysis into etiological underpinnings given the available data. In epidemiology, correctly specifying the causal structure is paramount and should be prioritized over improved model fit. The overall goal is to gain efficiency on the parameter of interest, while appropriately accounting for a thoughtful set of confounders. Thus it is

imperative to identify and characterize the exposure of interest, and build the corresponding causal DAG, before proceeding with statistical analysis, especially in scenarios with highly correlated mixtures. If the causal structure is not correctly specified, all the results may be biased. By combining *a priori* knowledge of potential causal associations with appropriate analytical techniques, vast advances towards providing pertinent knowledge concerning etiological effects of mixtures of exposures could be achieved.

14. Examining Associations Between Multi Pollutant Exposure Profiles and Health Outcomes via Bayesian Profile Regression

Presenting Author: John Molitor

Organization: Oregon State University

Contributing Authors: John Molitor, Eric Coker, and Silvia Liverani

Abstract:

Background: Our approach to modeling the health effects of chemical mixtures is implemented via a semi-parametric Bayesian approach we denote as profile regression. Bayesian profile regression is an adaptable, dimension reduction technique, capable of clustering multiple covariate data, thus overcoming unstable estimation problems associated with multicolinearity that we see with conventional multivariate linear regression. Furthermore, our approach enables a simplified examination of the joint health effects of correlated variables to aid in identifying risk drivers, which may otherwise prove computationally intensive or challenging in terms of interpretation of the joint effects when using conventional regression approaches. This Bayesian profile regression approach is able to partition and cluster covariate data into exposure profiles with similar risk without relying on the hard clustering techniques that have been well described in the literature. Avoidance of hard clustering is advantageous in that uncertainty is appropriately handled during the clustering process, and therefore the presumptive parameterization that can overly influence the clustering results and risk estimates downstream is likely to be avoided. Estimation of exposure profiles of interest are obtained via averaging over all clustering obtained via the iterative estimation process. For interpretative reasons, a single "best" clustering can be obtained, though uncertainty (C.I.'s, standard errors) related to this clustering is also assessed via averaging over all the clustering obtained via the estimation process, meaning that consistent clustering results in lower standard errors. Importantly, our approach is readily implemented using the PReMiuM package in R, thus making our approach widely accessible to researchers in a variety of settings.

Method: Our approach for the purposes of analyzing these two simulated datasets involved a multi-stage process that exploits various features of the PReMiuM package. In order to establish our clustering variables we first implemented the variable selection feature of the PReMiuM package. In this first stage, covariates were switched on and off iteratively so as to identify the most probable combination of covariates that drive the clustering risk profiles. Covariates with relatively high median probabilities of being selected as candidate clustering variables were included as clustering variables in the second stage of the Bayesian profile regression. Covariates with relatively low median probabilities of being selected as a candidate clustering variable were excluded in the second stage as clustering variables and are rather included as fixed effects in the Bayesian profile regression model. Covariates, "Z" were included in all analyses. Once the second stage profile regression was completed we were able to examine, via graphical output, the exposure profile clusters most likely to contribute to the outcome, as well as which

fixed effects that were likely to contribute to the outcome (Figures 1, 3). Finally, we examined the potential for interaction via predictions of Y for various pre-specified exposure profiles.

Results:

Dataset 1

Covariate data for dataset 1 were non-normally distributed and were thus discretized into three categories using the "cut" function in R. All analyses used Z as a confounder, separate from the clustering. (See dose-response equation accompanying pdf document.) Variables X1, X2, X3, and X5 were selected from the initial-stage profile regression variable selection procedure, while X4, X6, X7, and Z were analyzed as fixed effects in the second stage model. (We encourage readers to examine Figure 1 for graphical display of typical exposure mixtures associated with various levels of “risk” – levels of Y.) We found that two of the five exposure clusters (clusters 4 and 5) that were identified from the profile regression were associated with large values of Y (Figure 1), with cluster 5 exhibiting the largest Y value. Within cluster 4, high levels of X1 tended to drive the increased levels of Y, whereas within cluster 5, higher levels of X1, X2, and X3 tended to drive the observed high levels of Y, and these variables are highly correlated with each other. Given that clusters 1 and 2 were not associated with high values of Y and that elevated levels of X5 tended to predominate within these two clusters, it was apparent that increased levels of X5 were associated with decreased levels of Y. Cluster 3 was predominantly characterized by low levels for each of the clustering variables and was not associated with a high value for Y. Regarding the fixed effects in our profile regression model, variable X7 and the confounding variable Z each significantly contributed to an increase in Y, while the association between X6 and Y was marginal, and variable X4 did not contribute.

The R package allows for prediction of Y via predefined exposure profiles, even if some of the exposures are left unspecified. We utilized this feature to examine interaction between X1 and X5 by predicting Y via “pseudo-profiles”, consisting of (X1, X2, X3, X5). X1 and X5 were allowed to vary across category levels, while X2 and X3 were treated as missing. Figure 2 shows results of this analysis suggesting presence of interaction, meaning the effect of one exposure on Y depends on levels of other exposure(s).

Dataset 2

The biomarker covariate data were also non-normal and were discretized into three categories using the "cut" function in R. Variables X3 through X14 were selected from the initial-stage profile regression variable selection procedure, while X1, X2, Z1, Z2, and Z3 were analyzed as fixed effects in the second stage model. See Figure 3 for characterization of exposure mixtures and associated risks. We found that two of the six exposure profiles clusters (clusters 5 and 6) that were identified from the profile regression contributed to high levels of Y. Cluster 5 was associated with high levels of X6-X14, while cluster 6 (also high risk) was associated with high levels of X3-X8, X10, X11, and X14 . Note that several exposures display similar association patterns (e.g. X3, X4, X5, and X12, X13). PReMiuM allows for computation of root mean square error (here 0.43, 0.84 for standard regression).

With more information, we could produce GIS maps displaying spatial clustering of individual profiles, and compare risks associated with substantively driven predefined exposure mixtures, all of which could be used to inform exposure-reduction policy decisions.

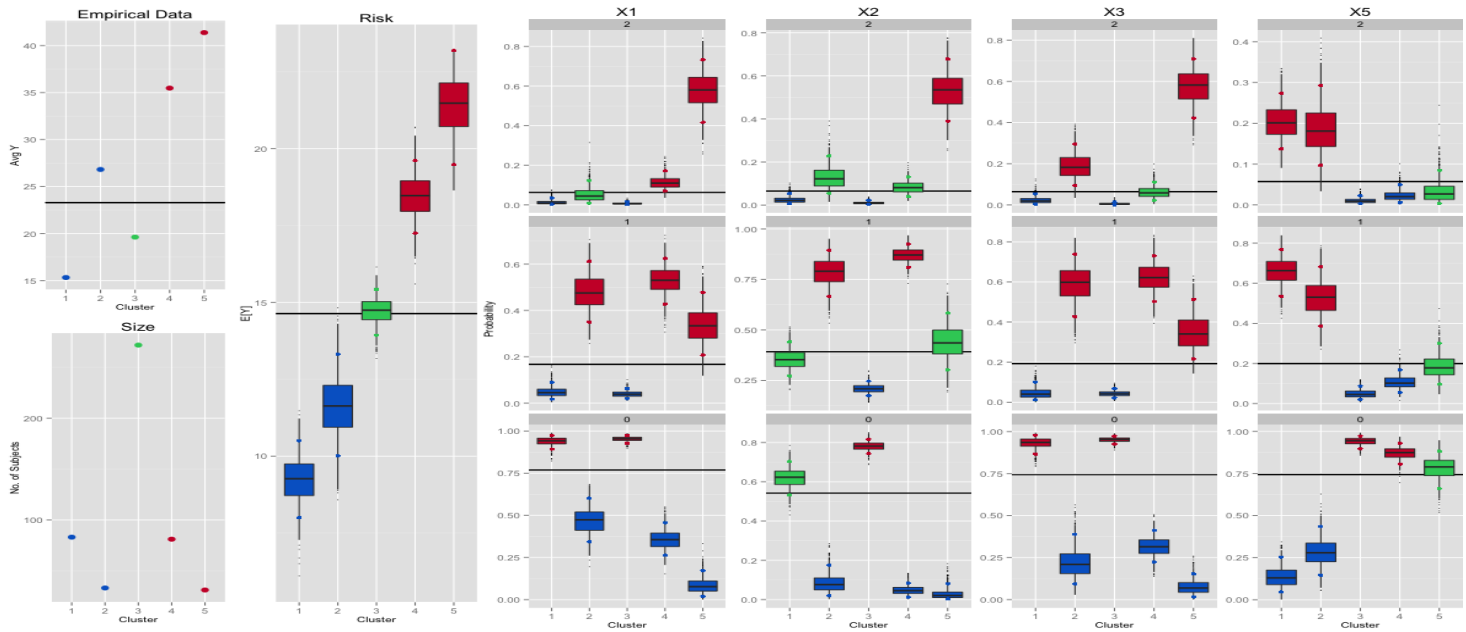


Figure 1: Optimal clustering for dataset 1. For each covariate, estimates of probabilities of membership to each of three categories (0,1,2) are provided for each cluster. For cluster 1 (low risk), X_1, X_2, and X_3 values are more likely to be in category 0, while X_5 values are most likely in category 2. Conversely, for clusters 4 and 5 (high risk), X_1, X_2, and X_3 are more likely in category 2, while X_5 is more likely to be in category 0.

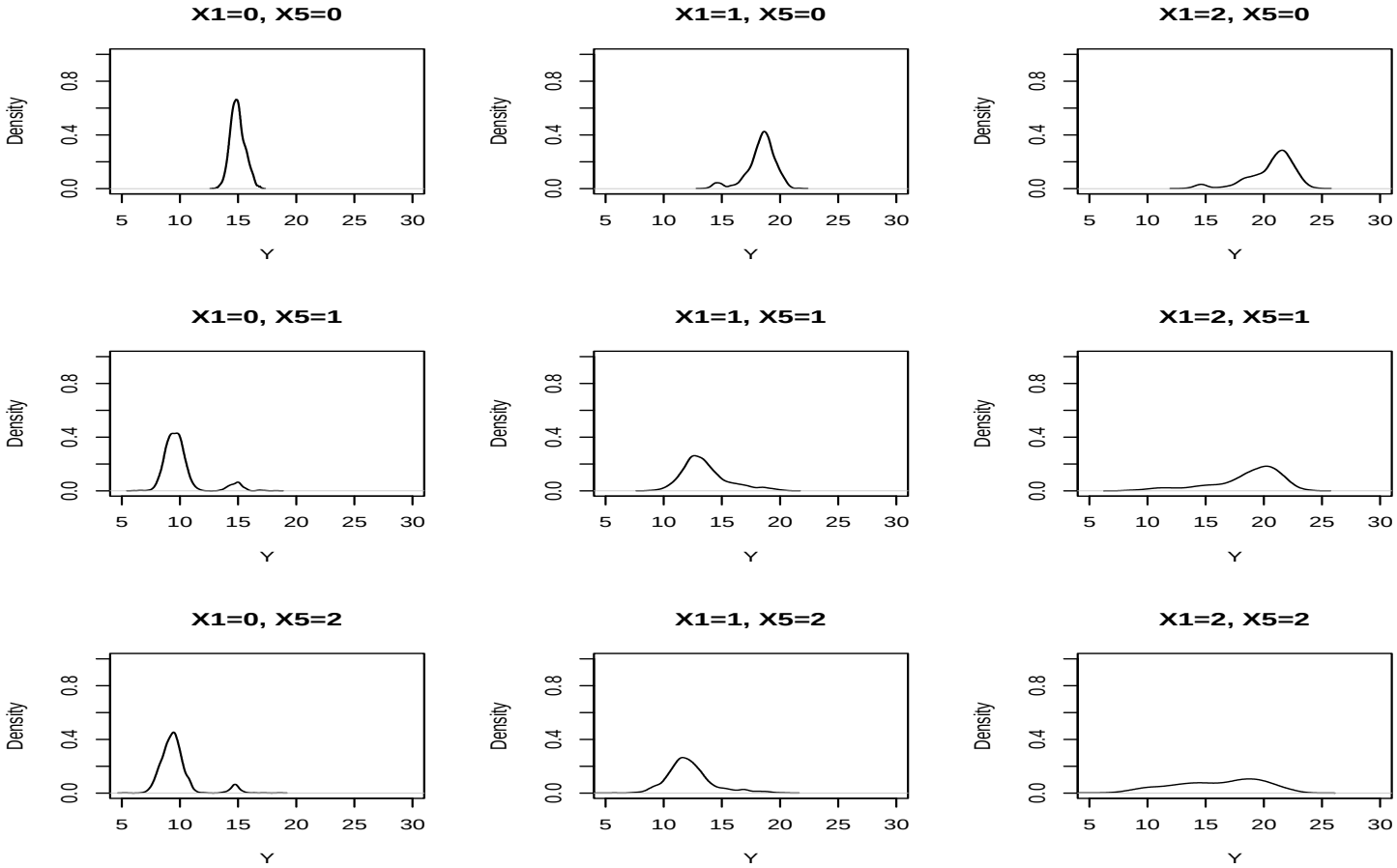


Figure 2: Predicted densities of Y for various pseudo-profiles specified with varying categorical values for X_1 and X_5, with other exposure covariates left unspecified. Each row displays estimated density of Y as X_1 varies across categories 0,1,2 with X_5 fixed to a particular exposure category. Note difference in density estimates as X_1 ranges from 0,1,2 (each row) conditional on different specifications of X_5, suggesting that the effect of increase in X_1 depends on levels of X_5, i.e. interaction.

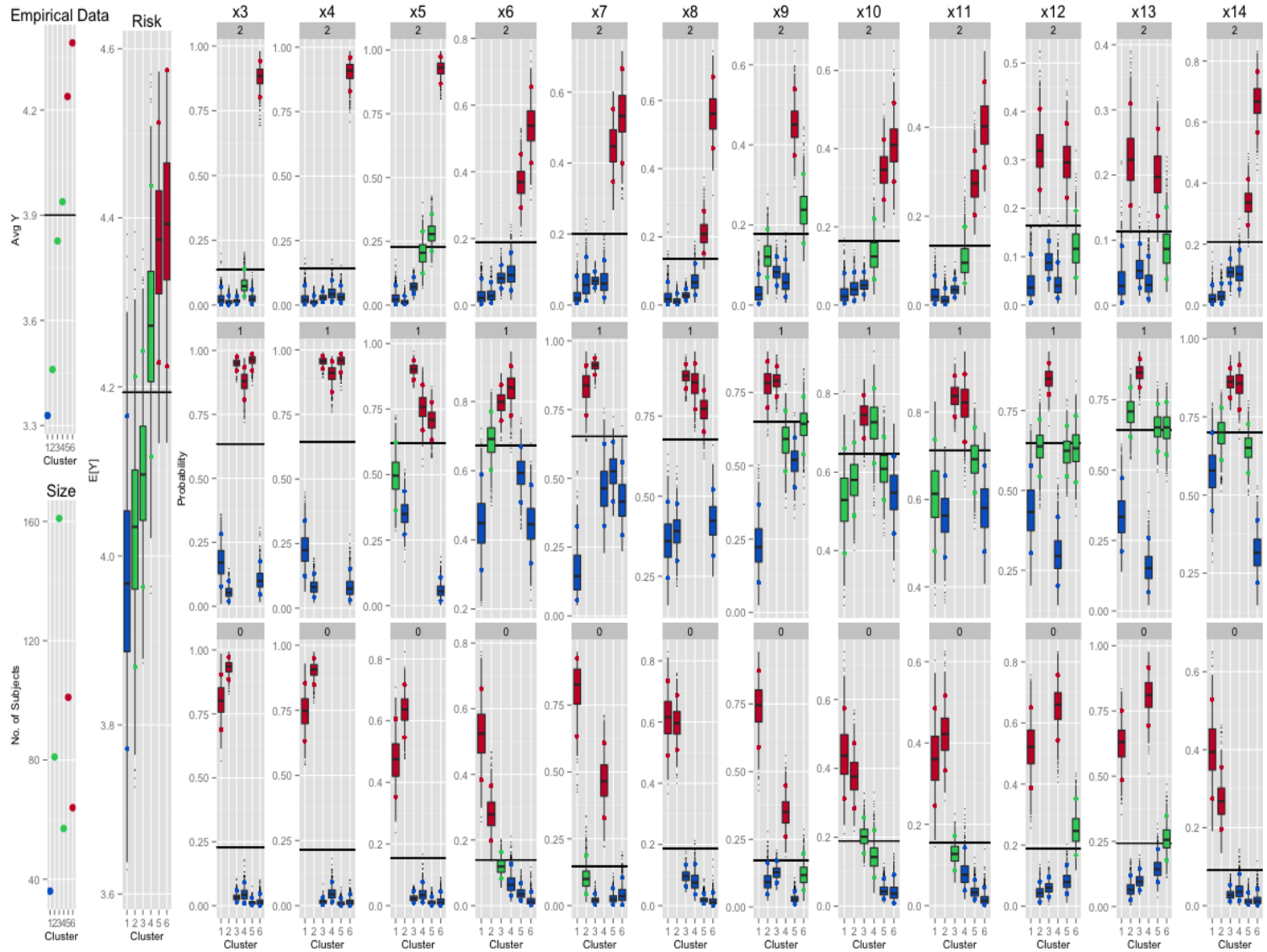


Figure 3: Characterization of typical exposure profiles for dataset 2. For each covariate, estimates of probabilities of membership to each of three categories (0,1,2) are provided for each cluster. Cluster 5 (high risk) was associated with high levels of X_6-X_{14}, while cluster 6 (also high risk) was associated with high levels of X_3-X_8, X_{10}, X_{11} and X_{14}. Note that several exposures display similar association patterns (e.g. X_3, X_4, X_5 and X_{12}, X_{13}).

Joint dose-response function: Individual-level exposures X_{ij} for individual i and exposure j are clustered into groups at each iteration of the iterative estimation process. These clusters (defined at each iteration) are then used as random effects in a standard regression model, with θ_c denoting the "effect" for cluster c on continuous outcome y_i. We further denote confounders as Z_i. We then set up a standard regression equation as,

$$y_i = \theta_{c_i} + \beta Z_i + \epsilon_i$$

where $\epsilon \sim N(0, \sigma^2)$. Here $c_i = c$ denotes an allocation variable indicating membership of individual i to cluster c. Given cluster parameters allocations, c_i, conditional independence between covariates is assumed, and $X_{ij} = c \sim \text{Multinomial}(1, \phi_{c_{ij}})$ where $\phi_{c_{ij}}$ is the vector of probabilities associated with cluster c for each of the possible categories that could be observed for covariate j. Our joint model is then,

$$p(Y_i, X_i|\theta, \beta, \sigma) = \sum_{c=1}^{\infty} \psi_c p(X_i|\phi)\, p(Y_i|\theta, \beta, \sigma).$$

with mixture weights ψ_c modeled according to a "stick breaking" representation according to a Dirichlet process prior. Note that a multivariate normal model is also available for continuous outcomes, though we utilized categorical exposures here.

15. Analysis of Simulated Data Sets using Conformal Predictions

Presenting Author: Ulf Norinder

Organization: Swedish Toxicology Sciences Research Center (Swetox)

Contributing Authors: Ulf Norinder

Abstract:

Conformal prediction (CP) represents a framework that offers an intuitive extension to the application of machine-learning methods for data analysis where focus is on predictions with pre-defined confidence levels. A conformal predictor will make correct predictions on new instances, e.g. mixtures or compounds, corresponding to a user defined confidence level. The set confidence level can be altered depending on the situation at hand, which allows for flexibility and adaption to risks that the user is willing to take. For a formal description and for proofs of the mathematical theorems on which the conformal prediction framework is built, see Vovk et al. 2005. The method has recently been applied to data analysis within the regulatory domain (Norinder et al. 2015).

Method of Analysis: For the analyses of datasets #1 and #2 two algorithms of choice have been utilized within the CP framework, namely, random forests (RF) and partial least squares (PLS). The confidence level for the framework was set to 80 %, (i.e., the errors from the resulting predictions on the external test set should not exceed 20%). The reason for using two different methods is that although RF implicitly handles interactions, (i.e., the synergistic and antagonistic effect (AB) of variables A and B), and may provide more accurate predictions than other algorithms a more detailed analysis of the importance of the variables and particularly interaction effects within the derived model is more difficult to achieve. The PLS method, on the other hand, although requiring explicit additions of interactions effects, offers possibilities for more explicit understanding of the importance of the variables in the model.

The dataset was randomly split 100 times into a training set (80%) and an external test set (20%) and 100 models were constructed for each of the two algorithms (RF and PLS, respectively). The outcome presented is an average prediction for each instance, when part of the external test set, across all 100 generated models.

To evaluate the signal-to-noise ratio in the analyses 4 additional random (white noise) variables were added to each dataset. These were also included in the explicit interaction effect variables added to the PLS method.

After analysis of the overall variable importance across all 100 models for the PLS method based on the generated training sets (not the external test sets), variable selection was performed to include the final variables and interaction effect variables.

Results: The results are presented in Table 1 and Figures 1 –3 (marked yellow in Table 1).

Table 1. Average external test set predictions

Dataset	Method	Variables	R2	adj-R2	RMSE	Spearman ran	CP-validity	Binary
DataSet1	RF		0.916	0.914	3.165	0.950	82.3	90.0 (≥90)
DataSet1	PLS	removed x4,x6. Added x1*Z	0.909	0.908				
DataSet1	PLS	removed x4,x6. Added x1*Z, x2*x3	0.910	0.908				
DataSet1	PLS	removed x4,x6. Added x1*Z, x2*x3, x5*x7	0.913	0.911				
DataSet1	PLS	removed x4,x6. Added x1*Z, x5*x7	0.913	0.912	3.177		80.3	
DataSet2	RF		0.507	0.490	0.460	0.691	80.3	70.2 (≥70)
DataSet2	PLS	removed x2,x7,x9,x13, z1	0.490	0.472	0.467		80.9	
data_05182015	RF		0.160		9.438	0.367	82.2	71.8 (≥70)

CP-validity should be above 80 (%) since the confidence level was set to 80 for regression.

For dataset #1 x1*Z and x5*x7 were added as possible interaction effects.

For dataset #2 it is difficult to identify interaction effects with significant statistical significance.

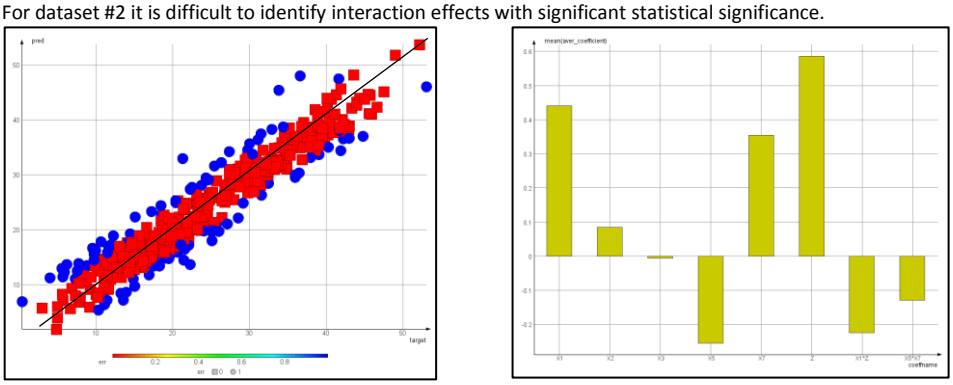


Figure 1. PLS – dataset #1 external predictions vs. experimental (left) and coefficients plot (right). Blue circles are invalid predictions.

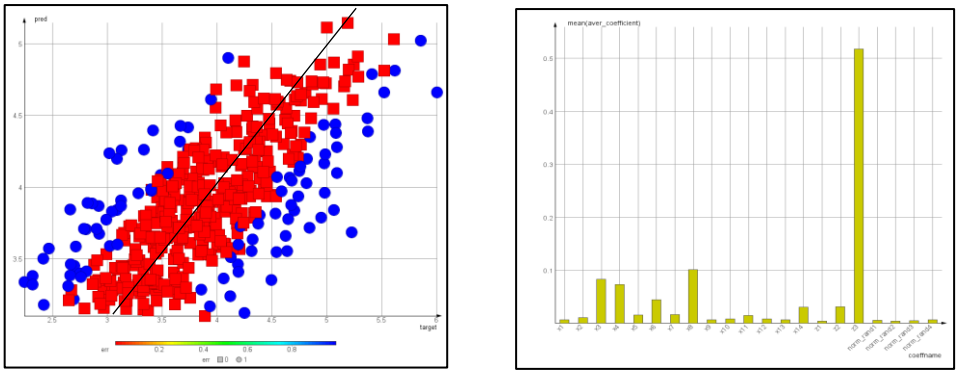


Figure 2. RF – dataset #2 external predictions vs. experimental (left) and coefficients plot (right). Blue circles are invalid predictions.

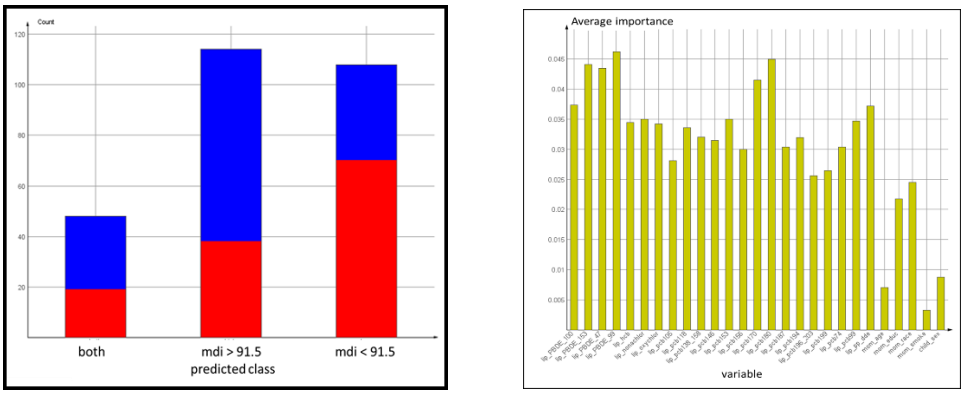


Figure 3. RF – data 05182015 external test set classification (left) and variable importance plot (right). Blue (mdi > 91.5), red (mdi < 91.5)

References

Norinder U., Carlsson L., Boyer S., Eklund M., 2015. Introducing conformal prediction in predictive modeling for regulatory purposes. A transparent and flexible alternative to applicability domain determination. Regulatory toxicology and pharmacology: RTP, 71(2), 279-84.

Vovk V., Gammerman A., Shafer G., 2005. Algorithmic learning in a random world, Springer, New York.

16. Analysis of Chemical Mixture Simulated Data Using Regularized Regression Models

Presenting Author: Sung Kyun Park

Organization: University of Michigan

Contributing Authors: Sung Kyun Park,[1,2] Yin-Hsiu Chen,[3] Weiye Wang,[1] John Meeker,[2] and Bhramar Mukherjee[3]

[1]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA
[2]Department of Environmental Health Sciences, School of Public Health, University of Michigan, Ann Arbor, MI, USA
[3]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA

Abstract:

Introduction: Our research team has previously considered application of various statistical strategies to examine health effects of multiple pollutants and their interactions using modern variable selection and machine learning tools (Sun et al. 2013 *Environ Health,*12:85). We examined the two simulated datasets using these methods.

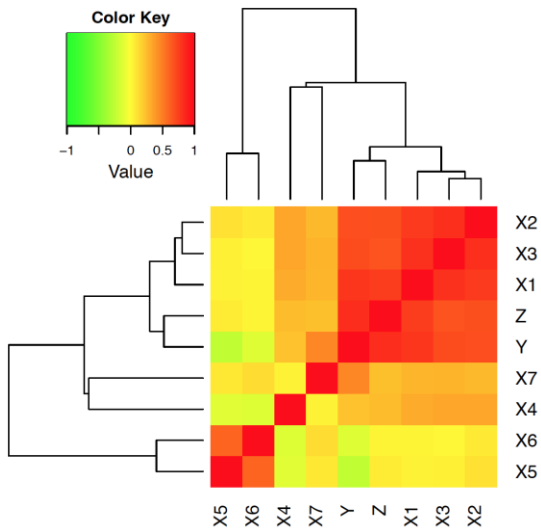
Methods: For both datasets, we conducted univariate analyses of each predictor, pairwise correlations among predictors, and confounder(s)-adjusted marginal associations between each predictor and outcome. Smoothing plots using penalized splines were used to determine potential non-linear relationships. We standardized each predictor (i.e., centered then scaled by its standard deviation, [x – mean(x)]/SD(x)) to make differently distributed predictors comparable and to reduce correlation between possible higher-order terms and pairwise interactions. We evaluated pairwise interactions using cross-product terms between predictors as well as predictors and confounders. These 'statistical' interactions imply departures from additive joint effects. For variable selection methods, we used adaptive least absolute shrinkage and selection operator (LASSO), adaptive elastic-net (E-Net), a combination of LASSO and ridge regression, and LASSO for hierarchical interaction (hierNet). Model goodness-of-fits and predictions were evaluated by adjusted R^2 and out-of-bag (OOB) adjusted R^2 using cross-validation. We also computed the mean squared error (MSE) and the mean squared prediction error (MSPE) to compare the prediction performance.

Results: For ***data 1***, a clear log-linear dose-response association was found with X7, thus X7 was log-transformed and then standardized. Quadratic terms for all other predictors were considered. Among the variable selection methods used, **adaptive LASSO** was chosen as the final model because of highest adjusted R^2 (0.950) and adjusted OOB R^2 (0.947) and lowest MSE (5.670) and MSPE (6.084) (**Table 1**). Although there were no significant marginal associations with X3 and X6, suggestive interactions between X3 and X4 and X5 and X6 were detected, and thus all predictors were included in the final model. We also found significant interactions of X1*X2; X1*log(X7); X4*X5; and X5*log(X7) and a significant quadratic term of X2 ($X2^2$). The final model also included marginally significant terms of X2*X5 and X5*Z. Thus, our final model included a total of 17 predictors (see the estimated coefficients of the final model in **Table 1**).

For ***data 2***, because of high correlations between X3 and X4 (r=0.99), X3 and X5 (r=0.94), and X12 and X13 (r=0.91), X4, X5, and X13 were dropped in the variable selection procedure. Although a marginal association with Z1 was weak, its interactions with other predictors were significant and we found a better model prediction with Z1 in the model, thus we decided to keep Z1. Smoothing plots suggested non-linear associations with X2, X6, X9, and X12, thus quadratic terms were considered in variable selection. Because we found significant interactions between Z3, a binary confounder, and several predictors, we also considered three-way interactions between those interactions with Z3 and other predictors. Among the variable selection methods used, **adaptive E-NET** was chosen as the final model because of highest adjusted R^2 (0.568) and lowest MSE (0.174) and MSPE (0.191) (**Table 1**). This final model suggests that X1, X2, $X2^2$, X3, X6, X8, X9, X10, X11, X12, and X14 along with Z2, are important predictors for Y. Z3 is an important effect modifier. The final model included the following interactions: X1*Z1; X2*X8; X10*Z1; and X10*Z2. Z3 seems to modify the effects of X2; X7; X12; X1*X7; X1*X9; X2*X14; X3*X14; X3*Z2; X7*Z1; $X1^2$; and $X6^2$. Thus, our final model included a total of 28 predictors (see the estimated coefficients of the final model in **Table 1**). Despite a little larger (worse) MSE and MSPE than adaptive E-Net, adaptive LASSO which selected 21 predictors could be an option if parsimoniousness is considered.

Conclusion: There is no general consensus model emerging from different selection methods, as expected. There is evidence of interaction and non-linearity in the dose-response relationships. These statistical models should be supplemented with subject-matter knowledge regarding the pollutants considered and their sources. In addition, these models can be strategically constructed to derive interpretable and policy relevant summary quantities of health risk.

Data 1



Data 2

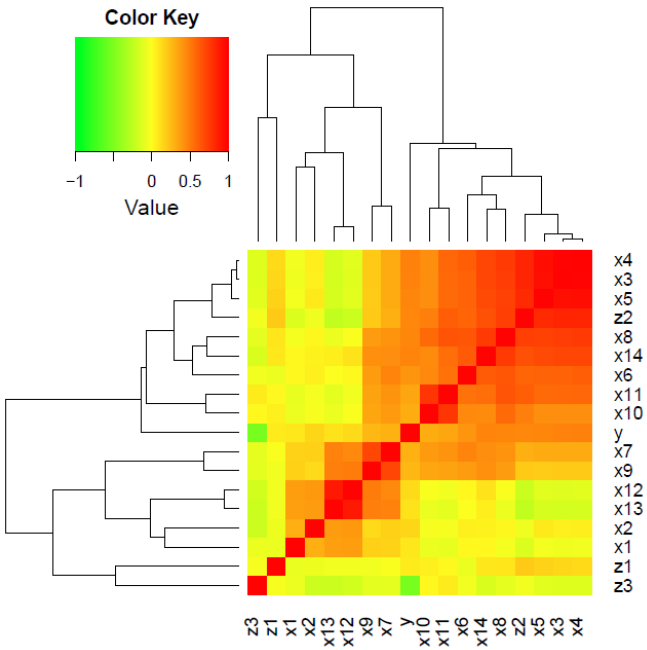


Figure 1. Heat maps of pair-wise correlations for Data 1 and Data 2

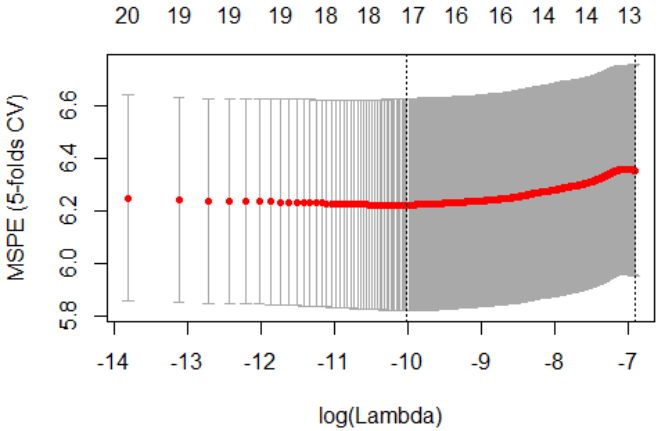


Figure 2. Test MSPE against log(λ) from adaptive LASSO in Data 1. Adaptive E-NET provided a similar result. (λ=8.6e-5).

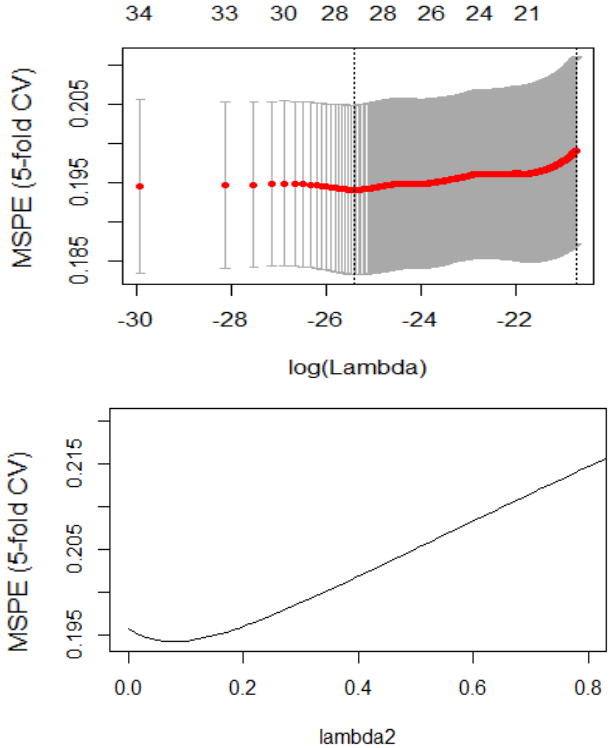


Figure 3. MSPE against λ1 and λ2 from adaptive E-NET in Data 2 ($\lambda 1=9.1\times10^{-12}$, λ2=0.079).

Table 1. Summary of variable selection results for Data 1 and Data 2.

	Data 1				Data 2			
	Adaptive LASSO		Adaptive E-Net	hierNet	Adaptive LASSO	Adaptive E-Net		hierNet
Adjusted R^2	0.950		0.950	0.948	0.564	0.568		0.541
OOB adj R^2	0.947		0.946	0.943	0.525	0.523		0.490
MSE	5.670		5.711	5.806	0.178	0.174		0.188
MSPE	6.084		6.101	6.394	0.194	0.191		0.208
No. Var	17	β (SE)	17	23	21	28	β (SE)	24
Variables selected	X1	3.20(0.24)*	X1	X1	X1	X1	0.05(0.02)*	X1
	X2	1.21(0.22)*	X2	X2	X2	X2	0.05(0.03)	X2
	X3	0.08(0.23)	X3	X3	X6	X3	0.06(0.04)	X3
	X4	-0.63(0.13)*	X4	X4	X8	X6	0.05(0.03)	X6
	X5	-5.15(0.36)*	X5	X5	X10	X8	0.03(0.04)	X8
	X6	-0.17(0.15)	X6	X6	X11	X9	0.02(0.03)	X10
	log(X7)	3.42(0.12)*	log(X7)	log(X7)	X12	X10	0.05(0.04)	X11
	Z	10.85(0.32)*	Z	Z	X14	X11	0.04(0.04)	X12
	$X5^2$	0.43(0.08)*	$X5^2$	$X2^2$	Z2	X12	0.11(0.04)*	X14
	X1*X2	-0.52(0.11)*	X1*X2	$X4^2$	Z3	X14	0.03(0.04)	Z2
	X1* log(X7)	0.62(0.12)*	X1* log(X7)	$X5^2$	X2*X8	Z2	0.13(0.04)*	Z3
	X2*X5	-0.23(0.17)	X2*X5	$X6^2$	X10*Z1	Z3	-0.46(0.04)*	$X2^2$
	X3*X4	-0.25(0.10)*	X3*X4	X1*X2	X10*Z2	X1*Z1	-0.04(0.02)	X1*X9
	X4*X5	0.30(0.12)*	X4*X5	X1* log(X7)	$X2^2$	X2*X8	0.03(0.02)	X2*X8
	X5*X6	0.25(0.13)*	X5*X6	X2*X5	X2*Z3	X10*Z1	-0.05(0.02)*	X2*Z2
	X5* log(X7)	-0.33(0.12)*	X5* log(X7)	X2* log(X7)	X7*Z3	X10*Z2	0.08(0.02)*	X3*Z2
	X5*Z	-0.44(0.31)	X5*Z	X3*X4	X12*Z3	$X2^2$	0.03(0.01)*	X6*Z2
				X4*X5	X1*X7*Z3	X2*Z3	-0.08(0.04)	X7*Z2
				X4*Z	X3*Z2*Z3	X7*Z3	-0.07(0.04)	X8*Z2
				X5*X6	X7*Z1*Z3	X12*Z3	-0.11(0.05)*	$X10^2$
				X5* log(X7)	$X6^2$*Z3	X1*X7*Z3	0.06(0.04)	X10*Z1
				X5*Z		X1*X9*Z3	0.04(0.04)	X10*Z2
				X6*Z		X2*X14*Z3	-0.03(0.03)	X12*Z3
						X3*X14*Z3	-0.03(0.04)	$X14^2$
						X3*Z2*Z3	-0.03(0.04)	
						X7*Z1*Z3	0.06(0.03)*	
						$X1^2$*Z3	-0.04(0.02)	
						$X6^2$*Z3	-0.04(0.02)*	

LASSO, least absolute shrinkage and selection operator; hierNet, LASSO for hierarchical interaction; MSE, mean square error; MSPE, mean square prediction error; OOB, out-of-bag.

In β (SE) column, " * " means statistically significant at $P<0.05$.

17. Building Models to Assess the Effects of Chemical Mixtures by Estimating Similar Modes of Action

Presenting Author: Harrison Quick

Organization: Centers for Disease Control and Prevention

Contributing Authors: Harrison Quick, Julia M. Gohlke, and Tran Huynh

Abstract:

The proposed methodology builds from the aim to separate exposures based on similar modes of action (or "pathways") from physico-chemical properties and previous toxicity studies. In the absence of this information, however, we group covariates by investigating the between-covariate correlations, grouping together covariates which are highly correlated (say, R>0.5). For each group of covariates, we first fit a model consisting of the main effects for each covariate. We then investigate potential interactions between covariates within each (supposed) pathway. Finally, we combine our group-specific models into a single model, consider interactions between pathways, and proceed with standard model selection techniques (e.g., removing non-significant covariates) to obtain a parsimonious model.

This general strategy is easily extended in the case of potential confounding variables. For a binary (or, more generally, categorical) confounder Z, we fit separate conditional models for each level of Z. In addition to allowing for interactions between the covariates and Z, this also allows for a different error variance parameter for each level of Z. In the case of a continuous confounder z', we first use our strategy without accounting for z', obtain a parsimonious model, and then consider interaction terms between our covariates and z'. An alternative would be to discretize z' – say, by using quantiles – and then treat the discretized z' as a categorical confounder Z. We could then allow for quantile-specific regression coefficients. Unfortunately, this could require splitting our data into several small pieces, so such an approach may not always be ideal. Finally, when confronted with numerous potential confounders, we may restrict our attention to those which are highly correlated with either the covariates or the response.

As shown in Figure 1(a), we identified the following groups for Dataset #1: {x1, x2, x3}, {x4}, {x5, x6}, and {x7}. In our group-specific models, we identify {x1}, {x4}, {x5}, and {x7} as being significant covariates. Upon combining our models, however, x4 is no longer significant, perhaps due to its moderate correlation with the covariate x1. For both levels of Z, exposure to x1 and x7 are positively associated with outcome Y, while exposure to x5 is negatively associated with the outcome. We then consider all two-way interactions between x1, x5, and x7. For Z=0, we identify significant interactions between x1 and both x5 and x7 – in each case, as x1 increases, the effect of the other covariate is diminished. For Z=1, however, only the two-way interaction of x1 and x5 is found to be significant; here again, increases in x1 diminish the protective effect of x5. Our joint dose-response functions for Dataset #1 are given in

equations (1) and (2) of the attached supplement, along with residual plots. These models achieve adjusted R-squares of 0.63 and 0.81, respectively.

Our results for Dataset #2 are much more complex. First and foremost, partitioning the 14 covariates into groups was not nearly as clear-cut as for Dataset #1. Nonetheless, as shown in Figure 2(a), we identified the following groups: {x1}, {x2}, {x3, x4, x5, x6, x8, x10, x11, x14}, {x7, x9}, and {x12, x13}. Unlike in Dataset #1, we did not notice a difference in variability between Z=0 and Z=1, thus we fit a conditional model in which each level of Z shares the same variance parameter. Here, we identify x2, x5, x6, x10, x12, and x14 as the exposures which contribute to the outcome and x1, x3, x4, x7, x8, x9, x11, and x13 as those which do not. Similarly, the continuous confounder z2' appears significantly correlated with the response, but z1' is not. While we failed to identify any significant interactions between our covariates and the response – or between the covariates and z2' and the response – there are substantial differences between the models for Z=0 and Z=1. For instance, x2, x5, x6, x10, and x12 are all positively associated with the response when Z=0, but only x14 is positively associated with the response for Z=1, with x12 being negatively associated with response Y. In total, our joint dose-response function for Dataset #2 is provided in Table 1 of the attached pdf, which yields an adjusted R-square of 0.54.

Supplemental materials for "Building models to assess the effects of chemical mixtures by estimating similar modes of action" by Quick, Gohlke, and Huynh

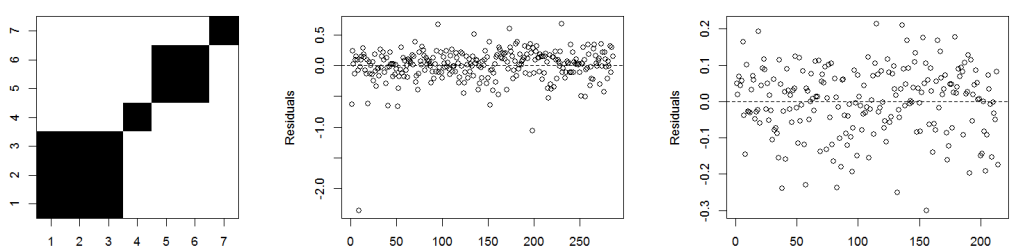
1 Dataset #1

For this analysis, we (natural) log-transformed both Y and X. Our final models were:

$$\ln Y_i \mid Z_i = 0 \sim N\,(2.89 + 0.18 \ln x_{i1} - 0.20 \ln x_{i5} + 0.21 \ln x_{i7} + 0.09 \ln x_{i1} \ln x_{i5} - 0.06 \ln x_{i1} \ln x_{i7}, \sigma_0^2 Æ \quad (1)$$

$$\ln Y_i \mid Z_i = 1 \sim N\,(3.36 + 0.20 \ln x_{i1} - 0.16 \ln x_{i5} + 0.13 \ln x_{i7} + 0.04 \ln x_{i1} \ln x_{i5}, \sigma_1^2 .Ĭ \quad (2)$$

Plots of our covariate groups (i.e., supposed pathways) and residuals can be found in Figure 1.



(a) Grouping (b) $Z_i = 0$, $\sigma_0^2 = 0.071$ (c) $Z_i = 1$, $\sigma_1^2 = 0.008$

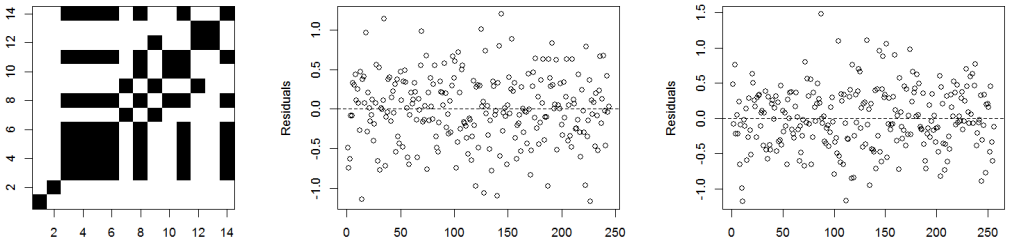
Figure 1: Plots for Datasets #1. Panel 1(a) shows our covariate groups (based on highly correlated covariate pairs, e.g., $|\mathrm{Cor}(x_j, x_{j'})| > 0.5$), and panels 1(b) and 1(c) display residual plots for each level of Z.

2 Dataset #2

Sticking with the notation in the abstract, the binary confounder is denoted as Z while the two continuous confounders are denoted as Z_1 and Z_2. After fitting our conditional models, we found that $\sigma^2_{Z=0} \approx \sigma^2_{Z=1}$, so we refit the model using a shared variance parameter, σ^2. Our final model is given in Table 1. Plots of our covariate groups and residuals can be found in Figure 2.

Covariate	Effect for $Z_i = 0$ Est. (95% CI)	Effect for $Z_i = 1$ Est. (95% CI)
(Intercept)	3.434 (3.107, 3.761)	3.226 (3.102, 3.349)—
x_2	0.084 (0.002, 0.165)	—
x_5	0.088 (0.022, 0.154)	—
x_6	0.094 (0.026, 0.162)	—
x_{10}	0.123 (0.033, 0.213)	-0.148 (-0.322, 0.026)
x_{12}	0.505 (0.316, 0.695)	0.166 (0.057, 0.275)
x_{14}	—	0.007 (0.004, 0.011)
$z^{ĭ}_2$	0.005 (0.001, 0.010)	

Table 1: Estimated regression model for Dataset #2. The estimate variance from this model is $\sigma^2 = 0.19$. Aside from x_{12} for $Z_i = 1$ (which is significant at the 0.1 level), all coefficients are significant at the 0.05 level. Cells marked with "—" denote a covariate which was insignificant for that level of Z_i.



(a) Grouping (b) $Z_i = 0$ (c) $Z_i = 1$

Figure 2: Plots for Dataset #2. Panel 2(a) shows our covariate groups (based on highly correlated covariate pairs, e.g., $|\mathrm{Cor}(x_j, x_{j^{ĭ}})| > 0.5$), and panels 2(b) and 2(c) display residual plots for each level of Z.

18. Application of Principal Component Analysis and Stepwise Regression to Identify the Exposure Variables Associated with Health Outcome and to Determine Dose-Response Relationship

Presenting Author: Sheikh Rahman

Organization: Northeastern University

Contributing Authors: Sheikh M. Rahman[1] and April Z. Gu[1]

[1] Department of Civil and Environmental Engineering, Northeastern University, 400 Snell Engineering Center, 360 Huntington Ave, Boston, MA 02115, USA

Abstract:

Introduction: Most health outcomes are the effect of exposure to pollutants, and in most cases the effects are worsened by the interaction of multiple exposures. The presence of other exposures may alter the effect of a specific exposure on the outcome, resulting in a confounding effect. This study was conducted to explore statistical approaches to find the exposures that are significantly associated with the outcome, possible confounding factors, and dose response function. This study also analyzed the possibilities of exposure interaction effects on the outcome.

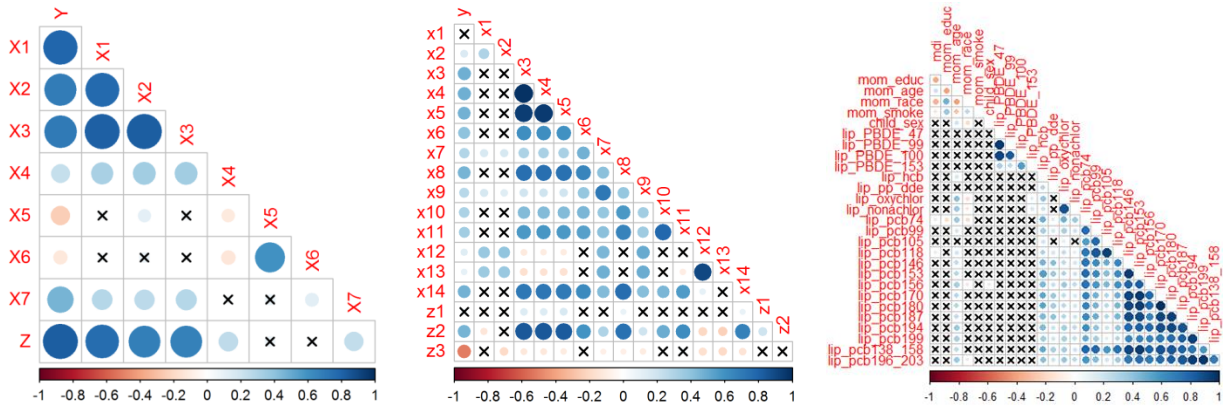
Methodology: The analyses of the present study focused on the two simulated datasets as well as one real world dataset. Dataset_1 consists of 7 exposure variables and 1 possible confounding binary variable. Dataset_2 consists of 14 exposure variables as well as 3 possible confounding variables representing a cross-sectional study. The real world dataset contains mental development index (MDI) as the outcome variable, 22 continuous exposure variables, and 5 possible dichotomous covariate variables.

We have estimated the correlation matrix along with the p-values to find out the co-linearity between the outcome parameter and the exposure variables. Significant correlation with the outcome variable is used to identify which variables can contribute to the outcome. We have further conducted Principal Component Analysis (PCA) on the exposure variables to identify the variables that can capture significant amount of variances in the data. Finally, forward stepwise regressionwas applied to estimate the dose-response function between the exposures and outcome. All the analyses have been conducted separately on the three datasets using statistical software R-3.0.2.

Results and Discussion: From the correlation matrix of dataset_1 presented in Figure 1-a, it can be seen that variables X1, X2, X3, X7, and Z show higher correlation with outcome variable Y. However, X2 and X3 show very high correlation and might be collinear with each other. On the other hand, variables x3, x4, x5, x6, x8, x14, z2, and z3 show higher correlation with outcome y for dataset_2 (Figure 1-b), while correlation of x1 and z1 with y is not significant at 95% confidence level. For the real world dataset only pcb_74 and pcb_118 show significant correlation with mdi, but 4 of the 5 covariate variables show significant correlation.

PCA is used to find a linear projection of high dimensional data into a lower dimensional subspace by maximizing the variances retained and minimizing least square error. Scree plots, which present the eigenvalues of different Principal Components (PCs), are presented in Figure 2 and show sharp bend at third PC. However, since eigenvalue of PC3 is less than 1 for dataset_1, the first two PCs can be selected to represent dataset_1, which can explain approximately 45% and 21% of the variances,respectively. Though for dataset_2 third PC has eigenvalue greater than 1, it can represent less than 10% of the variances and the first two PCs can explain approximately 40% and 19% of the variances. PCA loadings, which represent the correlation of exposure variables with the PCs and can be used to identify the variables accounting for the significant amount of variation in the data, are presented in Figure 2-d, e, and f. For dataset_1, X1, X2, X3, and Z are seen to be mostly correlated with PC1, hence they are the most significant variables responsible for the variation. Furthermore, X5 and X6 are highly correlated with PC2, but X4 and X7 do not show higher correlation with any of the PCs. For dataset_2, variables correlated with PC1 are x3, x4, x5, x6, x8, x10, x11, x14, and z2; while x12 and x13 are highly correlated with PC2. X7 and x9 are correlated with both the PC1 and PC2. For the real world dataset, almost 15 of the variables are highly correlated with PC1 and 44 of them are correlated with PC2.

In order to determine how much different variables can affect the outcome and dose-response function, we have conducted step-wise forward regression analysis. Step-wise forward regression finds out the most significant models by adding variables to the model starting with the intercept until the new model further improves the model selection criteria such as adjusted R-square, t-statistic. Table 1 summarizes two regression models of dataset_1 generated with all exposure variables and all variables except Z. Both the models show high F-statistics and when Z is added to the model, it improves the model accuracy as adjusted R-square increases from 0.808 to 0.917. High F-stat value of 650 indicates that two models with or without Z are significantly different from each other. From the regression model we may conclude that Z is a possible confounding variable for outcome in dataset_1. In addition, outcome variable Y can be estimated as a function of exposure variables X1, X2, X4, X5, X7, and Z with coefficients reported in table 1. We have also found significant models for dataset_2 with high F-stat, and the results are presented in Table 2. For dataset_2, z2 and z3 significantly improve the model as adjusted R-square increases to 0.501 from 0.297. From F-test we have also found that the model with and without z's are significantly different from each other and outcome y of dataset_2 can be modelled as a function of x6, x11, x12, z2, and z3. For the real world dataset, consideration of covariates improves model accuracy since adjusted R-square increased and outcome, mdi, is a function of mom_educ, mom_race, child_sex, lip_pcb118, and lip_PBDE_100 (Table 3). Additionally, we have estimated a regression model with the interaction terms between all the exposure variables to find out the effect of interaction between variables on the outcome and summarized in Table 3. It shows that interaction term X*X2, X5*X7 and Z*X5 is significant at 0.05 significance level and also increases the adjusted R-square of the model slightly. Similar effect of interaction can be also obtained for dataset_2. In summary, the present study applies different statistical approaches to explore the significant exposure variables along with confounding variables associated with health outcome, their dose-response relationship, and possible effects of exposure interaction on outcome.

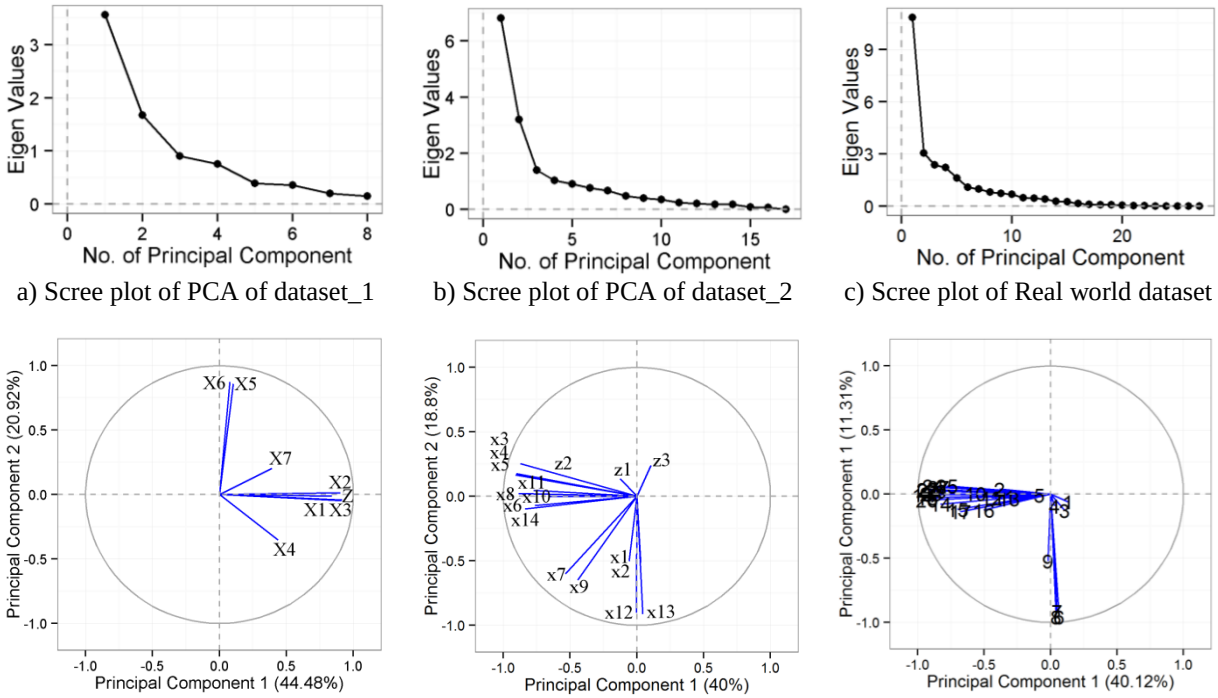


a) Dataset_1. Y is the outcome variable

b) Dataset_2. y is the outcome variable

c) Real world dataset. mdi is the outcome variable

Figure 1. Correlation matrix of the datasets. Circle size is corresponding to the magnitude and the color corresponds to the sign. × sign indicates that, the correlation is not significant on a 0.05 significance level.



a) Scree plot of PCA of dataset_1

b) Scree plot of PCA of dataset_2

c) Scree plot of Real world dataset

d) PCA loadings of variables along PC1 and PC2 for dataset_1

e) PCA loadings of variables along PC1 and PC2 for dataset_2

f) Variables PCA loadings along PC1 and PC2 for real world dataset

Figure 2. Plot of PCA results for the both dataset_1 and dataset_2. Scree plot is the plot of Eigen values of different Principal Components (PCs). PCA loadings of the variables represent the contribution of the variables along PCs.

Table 1. Summary of regression analysis result of dataset_1. F-statistic shown in the table indicates the models are significantly different from the model with no variable.

Parameter	Model without Z			Model with Z		
	Estimate	Standard Error	p-value	Estimate	Standard Error	p-value
Intercept	12.15	0.65	< 2.00E-16	14.29	0.44	< 2.00E-16
X1	6.22	0.35	< 2.00E-16	2.91	0.26	< 2.00E-16
X2	6.27	0.80	2.5E-14	3.18	0.54	6.38E-09
X4	-1.02	0.27	0.000169	-0.97	0.18	6.78E-08
X5	-3.52	0.22	< 2.00E-16	-3.62	0.14	< 2.00E-16
X7	3.02	0.23	< 2.00E-16	2.92	0.15	< 2.00E-16
Z	--	--	--	11.43	0.45	< 2.00E-16
Adjusted R-squared	0.808			0.917		
F-statistic (p-value)	919.7 (< 2.2e-16)			421.1 (2.2e-16)		

Table 2. Summary of regression analysis result of dataset_2. F-statistic shown in the table indicates the models are significantly different from the model with no variable.

Parameter	Model without Z			Model with Z		
	Estimate	Standard Error	p-value	Estimate	Standard Error	p-value
Intercept	3.16	0.15	< 2.00E-16	3.21	0.17	< 2.00E-16
x1	--	--	--	0.06	0.03	0.0529
x2	0.07	0.04	0.0438	--	--	--
x4	0.14	0.03	1.75E-06	--	--	--
x6	0.05	0.03	0.11878	0.06	0.03	0.0239
x10	0.07	0.04	0.05144	--	--	--
x11	--	--	--	0.09	0.04	0.0106
x12	0.22	0.08	0.00689	0.16	0.07	0.0237
x14	0.14	0.06	0.00887	0.08	0.05	0.0963
z2	--	--	--	0.01	0.00	6.61E-08
z3	--	--	--	-0.62	0.04	< 2.00E-16
Adjusted R-squared	0.297			0.501		
F-statistic (p-value)	36.21 (< 2.2e-16)			72.65 (2.2e-16)		

Table 3. Summary of regression analysis result of Real world dataset. F-statistic shown in the table indicates the models are significantly different from the model with no variable.

Parameter	Model without covariates			Model with covariates		
	Estimate	Standard Error	p-value	Estimate	Standard Error	p-value
Intercept	89.8	1.28	< 2.00E-16	95.3	1.15	< 2.00E-16
lip_pcb74	0.71	0.33	0.0306	--	--	--
lip_PBDE_100	-0.07	0.04	0.0874	-0.06	.04	0.109
mom_educ	--	--	--	-6.49	1.55	3.77e-05
mom_race	--	--	--	-5.34	1.38	0.0001
child_sex	--	--	--	-3.28	1.15	0.0046
lip_pcb118	--	--	--	0.22	0.11	0.0525
Adjusted R-squared	0.02			0.205		
F-statistic (p-value)	3.77 (0.02)			14.87 (7.4e-13)		

Table 4. Regression model of dataset_1 with interaction terms. Adjusted R-squared: 0.923

Parameter	Intercept	Z	X5	X7	X1	X2	X4	X1:X2	X5:X7	Z:X5
Estimate	12.22	11.32	-3.15	3.52	4.98	4.66	-1.29	-1.32	-0.51	-0.62
Std. Error	0.62	0.62	0.28	0.25	0.59	0.65	0.24	0.34	0.15	0.29
p-value	<2E-16	<2E-16	<2E-16	<2E-16	3E-16	2E-12	1.7E-07	0.0001	0.0008	0.036

19. Identifying the Relative Importance of Multiple Correlated Exposures in Predicting a Continuous Outcome Using the Random Forest Ensemble Learning Method

Presenting Author: Anne Starling

Organization: Colorado School of Public Health

Contributing Authors: Anne P. Starling, Katerina Kechris, Dana Dabelea, and John L. Adgate

Abstract:

Background: A common problem in environmental epidemiology is the analysis of a set of correlated chemical exposures which may have synergistic or antagonistic relationships in their associations with an outcome of interest. Traditional regression-based methods tend to perform poorly in such scenarios, and methods of testing the significance of statistical interactions between exposures have limited power. We propose to use the recursive partitioning method of random forests, which may be implemented using freely-available R software [1,2,3]. This method has recently found numerous applications in fields including statistical genomics, but the use of this method in environmental epidemiology has thus far been limited [4].

The random forest algorithm has a number of advantages relevant to the problem of multiple correlated exposures: it is non-parametric and can identify non-linear associations, and it allows for interactions (non-additive effects) of combinations of variables in the prediction of a continuous or categorical outcome. The basic principle of a regression tree is a hierarchical approach in which observations are divided at each node of a tree based on values of the predictor above or below a cutoff value, determined as the cutoff leading to the greatest difference between groups in mean outcome value [4]. "Random forests" consist of a large set of decision trees constructed with between-tree variation introduced by random sampling of observations at the root of each tree and of predictors at each binary split. The collection of trees performs better than a single tree and avoids over-fitting, and may be applied to classification or regression problems.

Parameters of the forest which may be specified by the user, include the number of trees and the number of variables sampled at each binary division. Sensitivity analyses can be performed to describe the impact of changing these parameters on the relative importance of the predictors. An internal validation procedure is incorporated into the process because each tree uses only a random sample of the observations, and the unselected "out-of-bag" observations are used for testing the prediction error of the tree.

Methods: We applied this method, using the R package randomForestSRC, to the two simulated datasets, and compared the results to those of multiple linear regression models with no interaction terms. The relative importance of each predictor is defined as the average change (over all trees) in the mean squared error of the model when values of that variable are randomly permuted; in other words, the overall improvement in model fit provided by that variable.

Analyses were conducted in R version 3.1.3.

Results: The relative importance of the predictors in each dataset differed from the significance of the predictors determined by linear regression models. In Data Set #1, the variables with the highest relative importance from the random forest model were X1, X7 and X5, where variable X4 contributed very little information (Table 1). In the linear regression model, variables X1, X7 and X5 were identified as significant predictors of the outcome (type III sum of squares p-value <0.0001) but variable X4 was also significant.

Partial dependence plots may provide insight into the marginal association of each variable with the outcome (Figure 1). The positive association between X3 and Y identified by the random forest model disagrees with the non-significant inverse association identified by the linear regression model (Table 1). The high degree of correlation between X3 and X1 (r=0.88) and between X3 and X2 (r=0.88) may explain this discrepancy.

In Data Set #2, the variables with the greatest relative importance were Z3, X4, X3, and X8 (Table 2). Of slightly less importance were variables Z2 and X6. In the linear regression model, only one exposure, X6, and two confounders (Z3, Z2) had significant p-values for the type III sum of squares. Both random forest and linear regression results indicate that the exposures in Data Set #2 were relatively weak predictors of the outcome, compared to those in Data Set #1. However, the random forest model indicates that X4 and X3 have some importance in predicting the outcome. The percent variance explained by the predictors in the random forest models (Data Set #1, R2=0.9134; Data Set #2, R2=0.4987) were similar to the adjusted R2 of the linear regression models (Data Set #1, R2=0.9167; Data Set #2, R2=0.4951).

The random forest model automatically considers interactions between predictors. Partial dependence plots can also show the joint effects of two predictors on an outcome to explore non- additive effects, as shown for X1 by quintiles of X2 in Data Set #1 (Figure 2).

Conclusions: Use of the random forest ensemble learning method led to different conclusions regarding the importance of certain predictors when compared to linear regression models, possibly due to multicollinearity in the linear regression models. A limitation of the random forest approach is that there is no single regression equation which can be produced and therefore the results of the model are not directly comparable to the results of more familiar regression methods, although predicted values of the outcome can be estimated for a given forest and a given set of exposure values. Depending on the goal of the analysis, variable importance ranks may be useful as presented, or may be used to select or transform variables in a more traditional analytic method. However, there is currently no standard criterion for the level of variable importance which should be retained in the model. The combination of variable importance ranking and partial dependence plots may provide a comprehensive view of the role of multiple, correlated predictors and their independent and joint associations (linear or non-linear) with continuous or categorical outcomes.

References:

1. Breiman, L. 2001. Random forests. Machine Learning, 45:5-32.
2. Liaw, A. and M. Wiener. 2002. Classification and regression by randomForest. R News, 2/3:18-22.
3. Ishwaran H. and Kogalur U.B. 2007. Random survival forests for R. R News, 7(2):25-31.
4. Coull, B.A., G.A. Wellenius, B. Gonzalez-Flecha, E. Diaz, P. Koutrakis, J.J. Godleski. 2011. The toxicological evaluation of realistic emissions of source aerosols study: statistical methods. Inhalation Toxicology 23(S2):31-41.

Identifying the relative importance of multiple correlated exposures in predicting a continuous outcome using the random forest ensemble learning method

Starling AP, K Kechris, D Dabelea, JL Adgate

Table 1. Linear regression results and variable importance from random forest for Data Set #1.

Parameter	Estimate	Standard Error	t value	Pr > |t|	Variable Importance in random forest
X1	2.91	0.29	9.98	<0.0001	66.1
X2	3.19	0.62	5.15	<0.0001	4.03
X3	-0.01	0.31	-0.02	0.98	6.11
X4	-0.98	0.18	-5.49	<0.0001	0.327
X5	-3.55	0.18	-20.1	<0.0001	16.4
X6	-0.14	0.21	-0.68	0.50	1.41
X7	2.93	0.15	19.36	<0.0001	17.6
Z	11.4	0.45	25.4	<0.0001	50.1

Figure 1. Partial dependence plots for each predictor variable in Data Set #1.

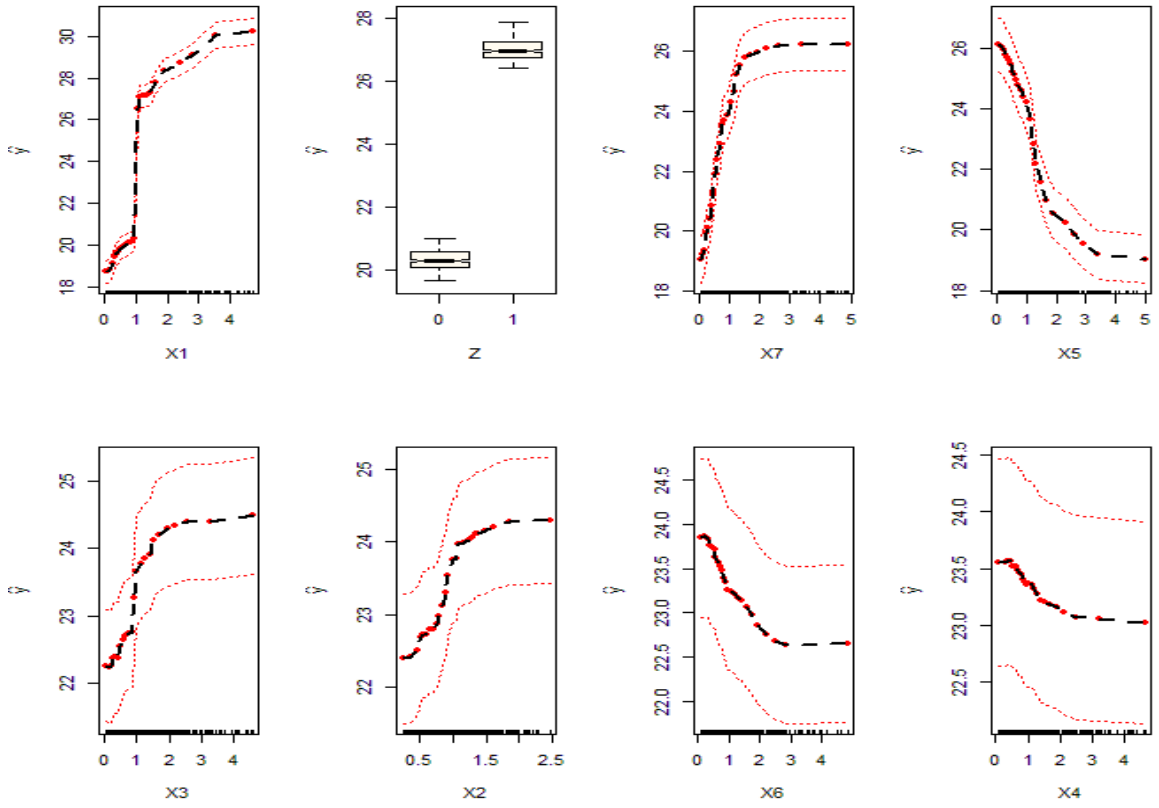
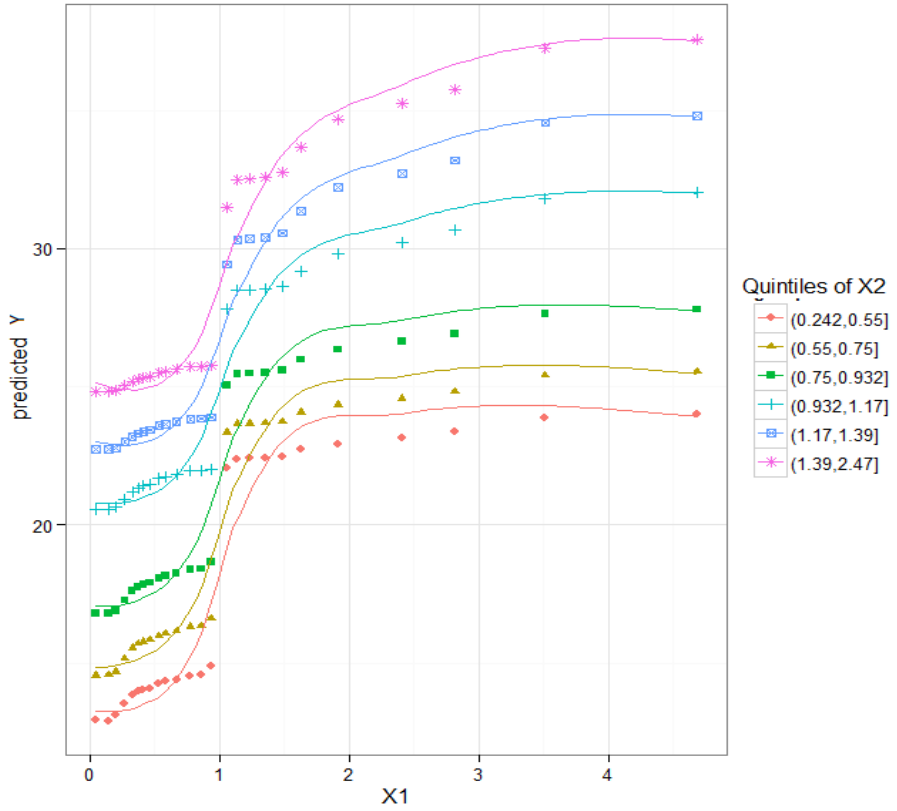


Table 2. Linear regression results and variable importance from random forest for Data Set #2.

Parameter	Estimate	Standard Error	t value	Pr > |t|	Variable Importance in random forest
X1	0.058	0.033	1.75	0.080	0.0018
X2	0.018	0.031	0.59	0.554	0.0051
X3	-0.030	0.103	-0.29	0.774	0.0242
X4	0.053	0.114	0.46	0.644	0.0242
X5	0.004	0.043	0.10	0.923	0.0075
X6	0.060	0.030	1.99	0.047	0.0112
X7	-0.031	0.062	-0.50	0.620	0.0025
X8	0.017	0.041	0.41	0.679	0.0199
X9	0.025	0.059	0.42	0.673	0.0017
X10	0.052	0.047	1.13	0.260	0.0049
X11	0.049	0.052	0.95	0.341	0.0061
X12	0.222	0.150	1.48	0.138	0.0060
X13	-0.083	0.152	-0.54	0.586	0.0053
X14	0.054	0.051	1.05	0.293	0.0093
Z1	0.006	0.014	0.41	0.685	0.0001
Z2	0.006	0.002	3.35	0.001	0.0131
Z3	-0.609	0.044	-13.8	<0.0001	0.1733

Figure 2. Partial dependence plot for variable X1 by quintiles of variable X2 in Data Set #1.



20. Improving Prediction Models by Adding Interaction Terms Using a Feasible Solution Algorithm

Presenting Author: Arnold Stromberg

Organization: University of Kentucky

Contributing Authors: Li Xu, Joshua Lambert, Bernhard Hannig, and Arnold Stromberg

Abstract:

Consider the problem of identifying interactions in existing data sets. This is typically done using machine learning or by using a software-aided selection process (forward, backward, stepwise) followed by looking for specific lower-order interactions between the reduced set of variables. Usually the computational complexity of finding these interactions prevents researchers from looking at all. The algorithm that will be presented provides a set of potentially interesting interactions that exist in the data set of interest. The algorithm can work for any objective function and for practically any model of interest. In the poster we will discuss the findings of the algorithm with the NIEHS data and future uses of the algorithm in other settings.

21. Factor Mixture Models for Assessing Health Effects of Environmental Chemical Mixtures: An Application Using Simulated Data Sets

Presenting Author: Heidi Sucharew

Organization: Cincinnati Children's Hospital Medical Center

Contributing Authors: Heidi Sucharew, Patrick Ryan, Shelley Ehrlich, Erin Haynes, and Monir Hossain

Abstract:

Background: It is well established that fitting common multivariable regression models with highly correlated predictors (exposures) can fail to converge and estimated coefficients may be unstable if convergence is achieved (Hoerl,Technometrics,1970). More recent approaches consider Bayesian hierarchical (MacLehose,Epidemiology,2007) and latent variable models (Sanchez,Biometrics,2012). However, high-dimensional predictor models remain an analytic challenge both in implementation and in coefficient interpretation. We propose to leverage the benefits of factor and latent class models in handling and evaluating complex correlated data structures. Factor models cluster components by defining regression relationships between the observed exposures and the underlying continuous factors with the goal of capturing common content among measured values. Latent class analysis clusters subjects with similar exposures allowing us to identify exposure subtypes. The combination of factor and latent class models are termed factor mixture models (Muthen&Shedden,Biometrics,1999). With this approach we are able to relax the within class independence assumptions of the common latent class analysis, such that observed exposures within a class follow a factor model imposed structure on the covariance matrix and mean vector. Herein, we illustrate the use of factor mixture models to assess health effects of environmental chemical mixtures using two simulated datasets with data collected at a single time point.

Methods: We began the analysis by conducting an exploratory factor analysis of the exposure variables and used the results to specify the pattern of factor loadings in the factor mixture model. Factor loadings were specified to be class invariant, i.e., equal across classes. No restrictions on the means of the observed exposures variables were specified, i.e., intercepts of all indicators of the factors were class specific. Factor mixture models were fitted with an increasing number of classes and the fit of the different models were compared using AIC, BIC, and Vuong-Lo-Mendell-Rubin likelihood ratio test. The probability of belonging to each of the classes was predicted for each subject and was allowed to be predicted by covariates. Mplus 7.1 was used for model fitting and SAS 9.4 to evaluate associations and exposure interactions on outcome using linear regression models.

Results:

Dataset 1: There are 500 subjects with seven continuous exposure measurements (x1-x7), one binary potential confounder (z), one continuous outcome variable (y) and no missing data.

The correlation coefficients indicate that exposure variables x1,x2, and x3 are highly correlated, as well as x5 with x6. Assessing one exposure at a time, x1,x2,x3, and x7 were positively associated with y and

x5 and x6 were negatively associated with y. X4 was not significantly associated with y. The factor model has a two-factor structure, based on eigenvalues of the correlation matrix greater than 1, with x1,x2,x3, x4, x7 loading on factor 1 and x5,x6 loading on factor 2. The model has two cross-loadings; x4 has a small negative loading on factor 2 and x7 has a small positive loading on factor 2. Fit measures indicated that the three-class model provided a better fit over the single-class and two-class models and the four-class model did not improve fit. Factor loadings are shown in Table 1. Z was incorporated as a covariate in the factor mixture model and determined to be an important covariate for factor 1 and predictor of class membership.

The characteristics of the three clusters of subjects are shown in Table 2. Class 1 represents low values of y with z=0 and average factor scores at 0 for both factor 1 and 2. Class 2 represents med/high values of y with z=1 and high factors scores for factor 1 and low factor scores for factor 2. Class 3 represents high values of y with z=1 and high factors scores for both factor 1 and 2.

Through the factor loadings, all 7 exposures (x1-x7) contribute to the outcome and quantified by regression estimates in Table 3 (adjusted R2=0.80). There was evidence of a class by factor 1 interaction on the outcome, such that the impact of the additive effect of x1,x2,x3,x4,x7 on y varied significantly by class ($p<0.01$). The additive effect, or linear combination of x5,x6,x4,x7 (factor 2) was negatively associated with y and did not vary by class (Table 3, Figures 1 and 2).

Data set 2: There are 500 subjects with 14 continuous exposure measurements (x1-x14), three potential confounders (z1-z3), one continuous outcome variable (y), and no missing data.

Assessing one exposure at a time, all 14 exposures (x1-x14) were positively associated with y. The factor model had a three-factor structure with the two-class model indicating best fit. Factor loadings are shown in Table 4. z2 was determined to be an important covariate for factors 1,2, and 3, and z2 and z3 were predictors of class membership. z1 was not significant in the model and subsequently removed.

The characteristics of the two clusters of subjects are shown in Table 5. Class 1 represents low values of y with low factor 2 scores and high factor 3 scores. Class 2 represents high values of y with high factor 2 scores and low factor 3 scores. Through the factor loadings, all 14 exposures (x1-x14) contribute to the outcome and quantified by regression estimates in Table 6 (adjusted R2=0.58). There was evidence of a class by factor 3 interaction on the outcome, such that the impact of the additive effect of x1,x2,x7,x9,x12,x13, x14 on y varied significantly by class ($p<0.01$). For class 2, an increase in factor 3 score was associated with higher y values, whereas factor 3 had less of an impact for subjects in class 1. The additive effect, or linear combination of x3,x4,x5,x14,x8,x6,x2,x11, x7 (factor 1) and of x10,x11,x9,x7,x8,x6,x14,x2 (factor 2) were positively associated with y and did not vary by class (Table 5, Figures 3-5).

Conclusion: We were able to identify exposure subtypes (classes) and factor structures that were associated with the outcome y. A limitation of this approach is that we did not have an alternative dataset to validate the factor structure, although cross-validation techniques could be used, and the interpretation of the factor scores hinge on the rational of the additive exposure construct.

Factor mixture model (Muthen&Shedden,Biometrics,1999):
Let **X** (matrix notation) denote the multivariate observed exposure variables, such that x_i is subject i's value on exposure variable x and **Z** the covariate matrix. The factor model can be expressed as shown in Equation 1. The regression intercepts are denoted as v, the regression slopes or factor loadings as Λ_x, and the regression residuals as ε_i. The effect of **Z** on **X** is captured by the regression coefficient Γ_x. Factor scores are denoted by η. The model assumes that the residuals have zero autocorrelations and are uncorrelated with the factors. The factor scores η vary by covariate Z with regression coefficients Γ_η.

Equation 1: $X_i = v + \Lambda_x \eta_i + \Gamma_x z_i + \varepsilon_i$
$\eta_i = \Gamma_\eta z_i + \zeta_i$

The factor model can be extended to include a latent class variable. Let k = 1,… K represent latent classes

$$c_{ik} = \begin{cases} 1 \; if subject \; i \; belongs \; to \; class \; k \\ 0 \; otherwise \end{cases}$$

Then we attached subscript k to parameters that may vary across classes as shown in Equation 3. Note that **A** contains the intercepts of the factors for each class and is of dimension number of factors x number of classes.

Equation 3: $X_{ik} = v_k + \Lambda_{xk}\eta_{ik} + \Gamma_{xk} z_i + \varepsilon_{ik}$
$\eta_{ik} = \mathrm{A}c_i + \Gamma_{\eta k} z_i + \zeta_{ik}$

The probability of belonging to each of the classes is predicted for each subject using multinomial regression and class membership may be predicted by covariates **Z.** Then, the outcome y is predicted by regression Equation 4 including subject factor scores η_i and assigned class membership c_i.

Equation 4: $Y_i = Bc_i + \Delta\eta_i + Hz_i + \varepsilon_i$

Data set 1:

Table 1. Class-invariant parameters in the partially invariant 3-class model

	Factor loadings Mean (SD)
	Factor 1
X3	1.00 (0)
X2	0.63 (0.10)
X1	0.54 (0.05)
X4	0.55 (0.10)
X7	0.29 (0.14)
	Factor 2
X6	1.00 (0)
X5	1.07 (0.21)
X4	-0.25 (0.07)
X7	0.16 (0.07)

Table 2. Characteristics of the 3-class partially invariant model

	Class 1 – Low (n=288)	Class 2 (n=158)	Class 3 – high (n=54)
Observed data:			
Y	15.6 (5.2)	32.1 (6.2)	38.9 (6.5)
Z=1, n(%)	2 (1%)	158 (100%)	54 (100%)
X1	0.50 (0.26)	1.55 (0.43)	3.36 (0.63)
X2	0.75 (0.30)	1.24 (0.32)	1.60 (0.35)
X3	0.60 (0.44)	1.55 (0.70)	2.74 (0.94)
X4	1.04 (0.76)	1.39 (0.88)	1.69 (0.97)
X5	1.19 (0.99)	1.34 (1.03)	1.18 (0.99)
X6	1.14 (0.89)	1.17 (0.79)	1.22 (0.85)
X7	0.92 (0.82)	1.33 (1.03)	1.65 (1.26)
Model estimated:			
Factor 1	0.01 (0.38)	3.43 (0.42)	3.75 (0.47)
Factor 2	0.00 (0.63)	-4.03 (0.58)	2.22 (0.56)

Data shown as mean (SD) unless noted otherwise.

Table 3. Linear regression evaluating the association between class and factors with Y

	Est (SE)	p-value
Class		<0.01
1	15.5 (0.3)	
2	-0.8 (3.5)	
3	43.1 (5.3)	
Factor 1 class 1	6.5 (0.8)	<0.01
Factor 1 class 2	4.6 (0.9)	<0.01
Factor 1 class 3	1.4 (1.4)	0.32
Factor 2	-4.3 (0.4)	<0.01

Adjusted R^2 = 0.80

Figure 1. Y versus Factor 1 with class indicator

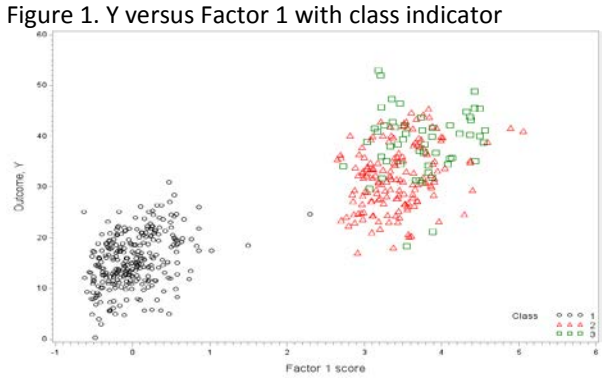
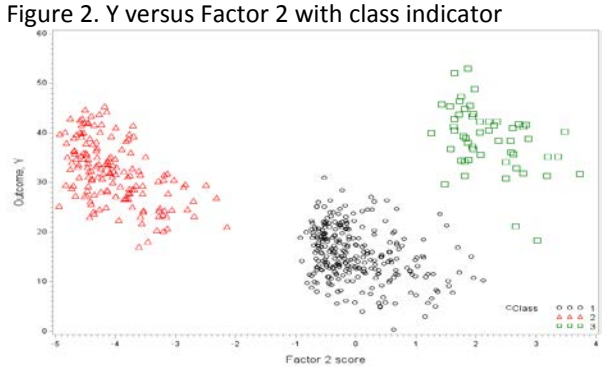


Figure 2. Y versus Factor 2 with class indicator



Data set 2:

Table 4. Class-invariant parameters in the partially invariant 2-class model

	Factor loadings Mean (SD)		Factor loadings Mean (SD)
	Factor 1		Factor 3
X4	1.00 (0)	X13	1.00 (0)
X3	1.01 (0.01)	X12	0.95 (0.02)
X5	1.11 (0.02)	X9	0.93 (0.07)
X14	0.27 (0.02)	X7	0.91 (0.06)
X8	0.34 (0.03)	X2	0.91 (0.10)
X6	0.27 (0.04)	X1	0.80 (0.09)
X2	0.09 (0.03)	X14	0.32 (0.07)
X11	0.12 (0.06)		
X7	0.06 (0.02)		
	Factor 2		
X10	1.00 (0)		
X11	0.96 (0.04)		
X9	0.50 (0.07)		
X7	0.45 (0.07)		
X8	0.76 (0.10)		
X6	0.73 (0.12)		
X14	0.39 (0.08)		
X2	-0.19 (0.08)		

Table 5. Characteristics of the 2-class partially invariant model

	Class 1 – Low (n=281)	Class 2 - High (n=219)
Observed data:		
Y	3.5 (0.4)	4.4 (0.5)
Z1	1.9 (1.6)	2.2 (1.6)
Z2	20.8 (21.8)	40.7 (22.7)
Z3=1, n(%)	228 (81%)	28 (13%)
X1	0.95 (0.70)	1.12 (0.69)
X2	-2.24 (0.75)	-1.96 (0.75)
X3	0.81 (1.37)	1.99 (1.18)
X4	1.82 (1.34)	3.02 (1.16)
X5	-0.04 (1.55)	1.28 (1.39)
X6	0.57 (0.98)	1.31 (0.96)
X7	1.22 (0.52)	1.46 (0.56)
X8	2.33 (0.93)	3.17 (0.88)
X9	1.22 (0.56)	1.45 (0.54)
X10	3.00 (0.74)	3.32 (0.72)
X11	5.04 (0.76)	5.38 (0.75)
X12	0.43 (0.34)	0.54 (0.34)
X13	0.51 (0.34)	0.62 (0.35)
X14	1.03 (0.67)	1.67 (0.61)
Model estimated:		
Factor 1	1.73 (1.33)	1.67 (1.15)
Factor 2	0.41 (0.55)	0.63 (0.55)
Factor 3	0.03 (0.32)	-0.24 (0.33)

Data shown as mean (SD) unless noted otherwise.

Table 6. Linear regression evaluating the association between class and factors with Y

	Est (SE)	p-value
Class		<0.01
1	3.30 (0.04)	
2	4.27 (0.04)	
Factor 1	0.08 (0.02)	<0.01
Factor 2	0.20 (0.04)	<0.01
Factor 3 class 1	-0.04 (0.08)	0.61
Factor 3 class 2	0.24 (0.09)	<0.01
Z3	-0.21 (0.05)	<0.01

Adjusted R^2 = 0.58

Figure 3. Y versus Factor 1 with class indicator

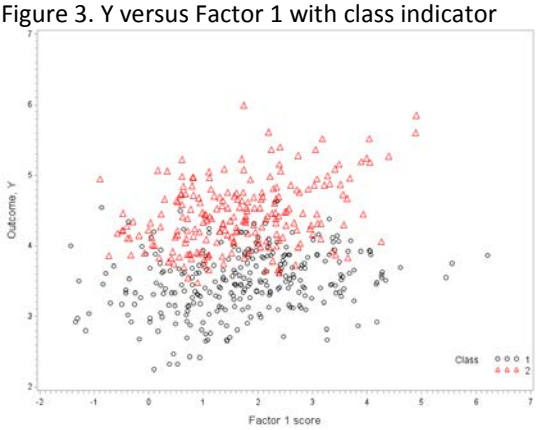


Figure 4. Y versus Factor 2 with class indicator

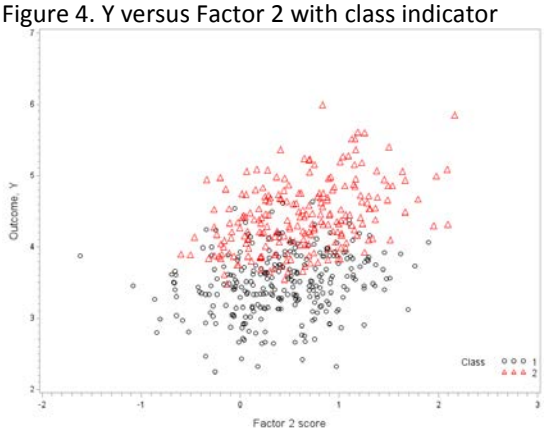
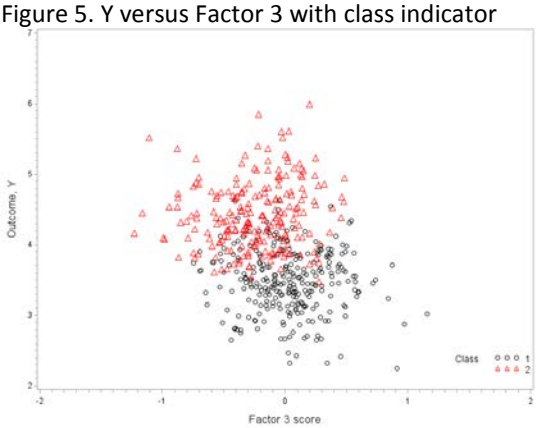


Figure 5. Y versus Factor 3 with class indicator



22. Dimension Reduction for Chemical Exposure Risk Assessment

Presenting Author: Jeffrey Switchenko

Organization: Emory University

Contributing Authors: Jeffrey M. Switchenko, Lance A. Waller, and P. Barry Ryan

Abstract:

Data: The chemical mixture simulated data (Dataset #1) is considered data from a prospective cohort study, where the outcome cannot cause the exposure, and correlations between exposure variables can be thought of as caused by common sources or modes of exposure. The mixture simulated data for Dataset #2 involves an environmentally relevant complex correlation pattern, intended to represent data from a cross-sectional study of 14 biomarkers from biomonitoring data potentially associated with a biomarker of effect. The real world data come from a prospective pregnancy and birth study. Exposures include polychlorinated biphenyl congeners (PCB), polybrominated diphenyl ether (PBDE), and organochlorine pesticides, and the outcome is the Mental Development Index of the Bayley Scale of Infant Development-II.

Motivation for methodology: In selecting statistical methods, we plotted the outcome data variable Y vs. each exposure data variable (X1-X7 for Dataset #1, X1-X14 for Dataset #2, and the 14 PCBs, 4 PBDEs, and 4 organochlorine pesticides for the real world dataset), and stratified these plots by the available covariates: Z for Dataset #1, Z1-Z3 where Z1 and Z2 were dichotomized at their median value for Dataset #2, and for the real world data, child's gender, maternal age at delivery, education, race, and smoking status during pregnancy. Correlation plots were also produced.

For Dataset #1, we noticed that Z completely delineates X1 (Figure 1), and Y vs. X2, X3, and X4 appear to add noise to the original Y-X1 relationship. For Y vs. X5 (Figure 1), Z appears to separate the values of Y, and similar patterns were discovered for Y vs. X6 and X7. What follows is an attempt to understand the relationship between Y and the X's, while both controlling for Z and assessing interaction. For Dataset #2, we noticed that the dataset contained a large number of exposure variables and confounders, which could have an association with each other and the outcome variable. Based on scatterplots, it was clear that Z1 did not affect the relationship between Y and each X, and it was not considered a confounder. Z2 and Z3 appeared to influence some of the relationships, while several X's had nearly zero correlation with Y. In addition, it was clear that X3, X4, and X5 were all strongly correlated with one another ($r>0.94$, $p<0.001$), with X3 and X4 nearly identical. We chose a method which can handle classification of an outcome variable, while minimizing the amount of model error. For the real world data, we noticed the high number of exposure variables, most of which were highly correlated with one another. In addition, several outlying and influential points were discovered, primarily contributed by two individuals. We chose a method to reduce the high data dimensionality, and checked results with and without outlying points.

Methods: For Dataset #1, each exposure data variable was fit in a linear regression model with Z and its interaction with Z. Significant interaction terms were included in a multivariable weighted least squares

(WLS) analysis, where the regression was weighted by squared residuals to limit observed heteroscedasticity. The weighted least squares function is minimized when estimating the model parameters. Non-significant terms were removed in a backward selection process using a removal criteria of 0.05. For Dataset #2, given the number of variables and confounders, we chose to implement a classification and regression tree (CART) analysis to present a snapshot of the relationship of the data variables relevant to the outcome, and visualize the important associations. The "leaves" on the bottom of the tree indicate the estimated value of Y for the combination of true/false variable statements listed above them. For the real world data, we chose a principal components regression (PCR) approach for reducing the number of dimensions in the exposure data. MDI was fit as a function of the exposure principal components with the largest variability within the data along with covariates and interactions between the exposure principal components and covariates. Only significant terms remained in the model.

Results: In Dataset#1, of the 7 initial exposure variables, X1, X2, X4, X5, and X7 contribute to the outcome, while X3 and X6 do not. For X1 and X5, the amount of contribution depends on Z. If Z=0, each unit increase in X1 results in a 5.78 increase in Y, controlling for the other model variables; if Z=1, each unit increase in X1 results in a 2.86 increase in Y. X1 and X5 interact with Z, such that the relationship between X1 or X5 and Y is modified by the presence of Z, controlling for the other exposure variables. The root mean square error is 1.01, and the adjusted R-squared value is 0.914, indicating a very good fit to the data. In Dataset #2, we find complex relationships given the exposure variables X1-X14 and confounders Z1-Z3. Of the 14 initial exposure variables, X3, X6, X8, X10, X12, X13, and X14 contribute to the outcome. X1, X2, X4, X5, X7, X9, and X11 do not contribute. The contribution is based on the regression tree listed above: If Z3=1, X3>0.826, and Z2>49.035, then the estimated value of Y is 3.995. The residual mean deviance is 0.180 and the root mean square error is 0.424. For the real world data, the PCB variables loaded evenly into the first principal component factor (PC1), which was the only significant exposure variable in the regression model. We found a significant interaction between PC1 and mother's age at delivery as well, and the R-squared value was 0.22. This did not differ when removing the outlying points.

Discussion: We have described the relationship between each outcome and each exposure variable, controlling for known covariates, identified strong patterns and relationships between certain exposure variables and an outcome in a classification scheme, and reduced the dimensionality of a large exposure dataset. However, it is possible that a structural equation model approach could provide a more comprehensive approach to identifying appropriate latent exposure variables and relationships with the outcomes, and could provide an alternative approach to analyzing our data.

Figures and Tables:

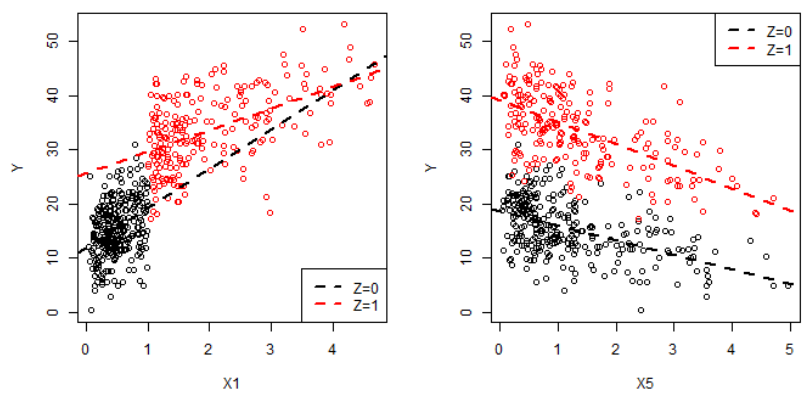
Dataset #1:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_4 + \beta_4 X_5 + \beta_5 X_7 + \beta_6 Z + \beta_7 X_1 Z + \beta_8 X_5 Z$$

Weighted sum of squares $= \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$

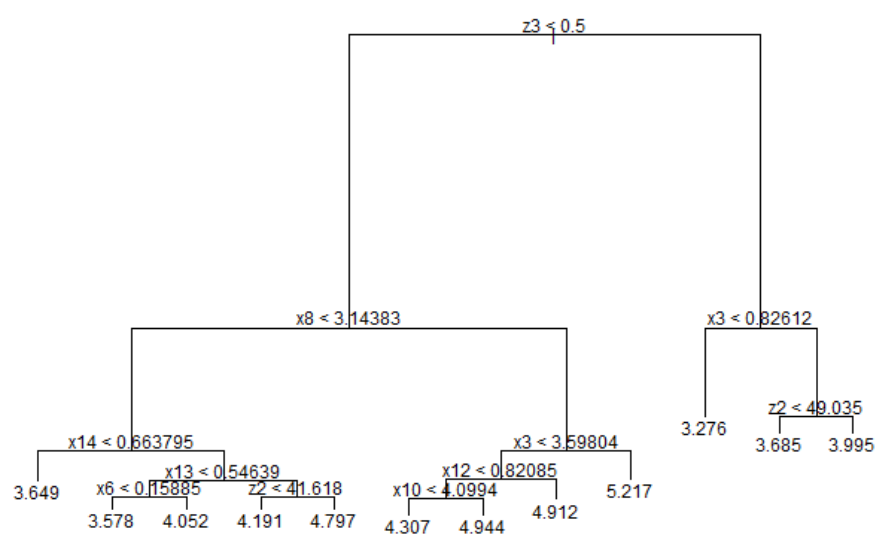
Variable	Estimate	95% confidence interval
X1 – Z=0	5.783	(4.159, 7.408)
X1 – Z=1	2.864	(2.346, 3.383)
X2	2.122	(1.088, 3.156)
X4	-0.947	(-1.297, -0.596)
X5 – Z=0	-3.330	(-3.705, -2.955)
X5 – Z=1	-4.007	(-4.364, -3.650)
X7	3.040	(2.738, 3.342)

Equation 1 (top left): Equation of Y given X's and Z, and weighted sum of squares function to minimize. Table 1 (middle left): Estimates of Y given a 1-unit increase in X, yielded from the WLS analysis. X1 and X5 depend on Z. 95% CIs also provided. Figure 1 (bottom left): Scatterplots of Y vs. X1 and X5, stratifying by Z. Z=0 is denoted in black, and Z=1 is denoted in red. Each strata results in differing slopes, leading to a significant interaction.

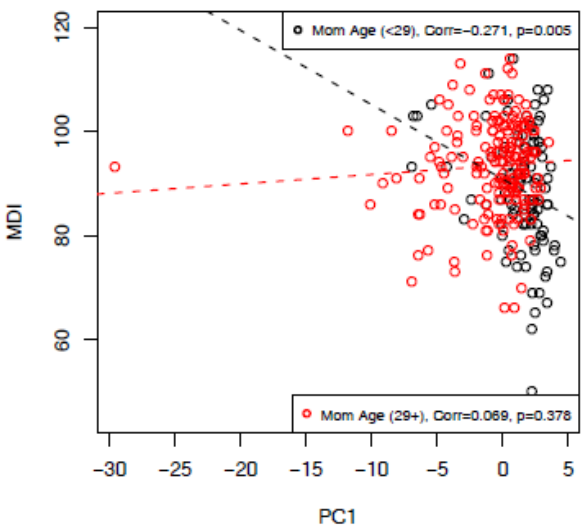


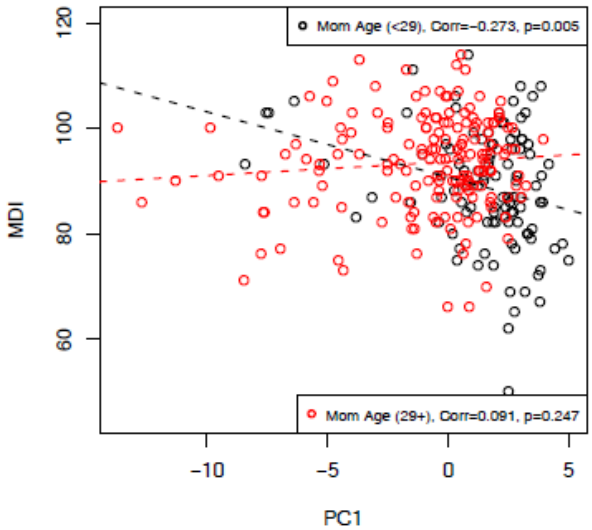
Dataset #2:

Figure 2: Regression tree resulting from a model which included X1-X14 as well as Z1-Z3. Only relevant exposure biomarker variables and confounders are included in the final tree model.



Real world data:





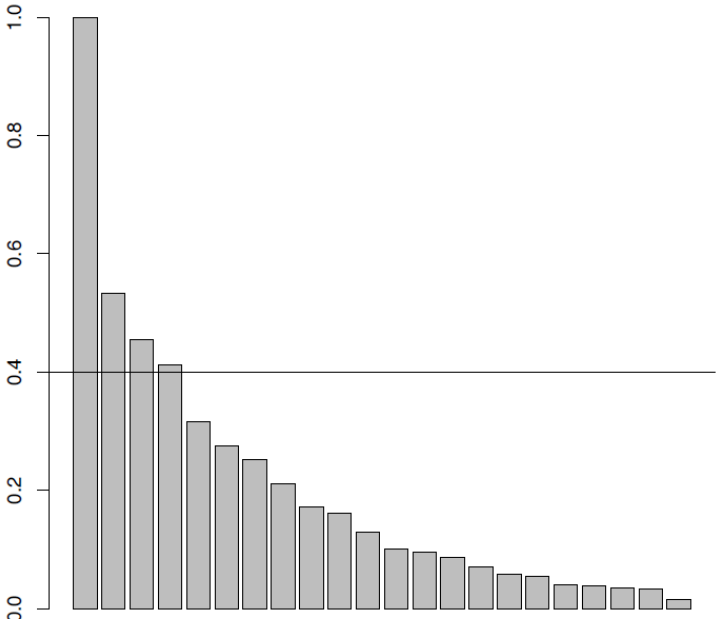


Figure 3: Mental development index vs. principal component with most variation (PC1), stratified by mother's age (with and without outliers)

Figure 4: Standard deviations of 22 principal components from 22 exposure variables, scaled as a proportion of the largest standard deviation. 0.4 chosen as cut-off for PCs containing largest share of variability in exposure data.

23. Set-based Interaction Tests for High-Dimensional Environmental Exposome Data

Presenting Author: Sandra Taylor

Organization: University of California, Davis

Contributing Authors: Sandra L. Taylor, Kyoungmi Kim, and Irva Hertz-Picciotto

Abstract:

Background: Environmental epidemiological studies seek to identify compounds that contribute to health outcomes and quantify the impacts of exposure. Exposure data often consists of large numbers of correlated variables that may have substantial overlapping effects but only small unique effects of each on the outcome. Discerning exposure effects under these circumstances is challenging and further complicated by confounding factors.

Methods: We propose a strategy to identify an optimum subset of exposures that together explains a high proportion of the variability of the outcome, and to quantify the unique effects of exposures singly and jointly using both non-linear and interaction terms. Our approach is outlined in Figure 1. We normalize all continuous variables to mean 0 and variance 1 to quantify the effects in a standardized metric across different exposures. We separate exposures and confounders into two sets based on their correlation (positive or negative) with the outcome. Variables in each correlation set, wherein variables act similarly to either increase or decrease the outcome, are potentially redundant with each other to varying degrees. We therefore conduct step-wise variable selection within each set to identify the exposures that extract maximum information about the outcome and avoid redundancy. We use likelihood ratio tests to sequentially add linear main effect terms to a model, selecting the predictor yielding the largest change in likelihood among significant ($p < 0.01$) predictors. Once main effect linear terms are identified independently in each set, we combine the two sets and add two-way intra- and inter-set interactions to the model using the same process. To capture non-linear effects, we then evaluate higher order polynomial terms of the main effects. Finally, we use the 'change-in-estimate' (>10%) criterion to assess excluded variables as potentially important confounders and select a final model in consideration of redundancy among the variables, while balancing the goals of parsimonious prediction and obtaining least biased estimates for each term. The final model identified through this process is the joint dose-response function and allows estimation of the outcome as a function of the exposure mixture with consideration of confounders and effect modification.

To quantify the contribution of each predictor, we propose two novel metrics. To estimate the relative contribution of each term to the total variability explained by the final model, we calculate the incremental change in the log likelihood ($\Delta termj$) for each term when this term is added to the model containing all other predictors in the final model; we then calculate the percent contribution of each term j as $\%termj = \Delta termj/\sum j \Delta termj$. Because variables can be involved in multiple terms (e.g., interactions), we also calculate the percentage of the total variability that each individual predictor contributes to the final model ($\%varj$) in an analogous manner. We use 1,000 bootstrap samples to

evaluate the stability of the predictors selected in the final model and to generate confidence intervals for %termj and %varj.

Results: We demonstrate our method on two simulated data sets. In Data Set 1, 5 exposures (X1-X4, X7) and a confounder (Z) were positively correlated with the outcome Y; 2 exposures (X5, X6) were negatively correlated. Within each of these sets, X1-X3, and X5 and X6 were strongly correlated ($\rho > 0.65$) and provided redundant information about the outcome. Our approach identified 3 main effects (X1, X5, X7) and that of a confounder Z, including non-linear terms for X5 and X7, and two interactions Z:X1 and X5:X7. X2, X3, X4 and X6 did not appreciably explain variability in Y beyond Z, X1, X5 and X7. With the exception of the Z:X1 interaction, predictors in the final model were selected in greater than 90% of the bootstrap samples. Total R2 for the final model was 0.94; Z accounted for about 30% of the variability explained by the model, followed by X5, X7 and X1 (Table 1). The Z:X1 interaction indicates that Z is an effect modifier of X1 resulting in differential effects of X1 on Y depending on exposure to Z. Y has a complex relationship with X5 and X7 as demonstrated by the higher order and interaction terms for these variables (Figure 2). Importantly, the inter-set interaction of X5 and X7 indicates that X5 has a buffering effect that absorbs the positive impact of X7 on Y when working jointly.

Finally, we considered each of the omitted exposures as potential confounders of exposures retained in the final model. We found both X2 and X4 to confound the association of X1 with Y in the final model. No other omitted exposures had confounding effects and no additional predictors in the final model were affected. However, X2 was highly correlated with X1, but minimally correlated with the outcome when conditioned on X1; thus we opted to omit it as redundant to X1. Ancillary information would be needed to ascertain which, if either, is causal. With regard to X4, based on the a priori project information and given the project objective of parsimonious prediction rather than identifying the most unbiased associations for each of the individual factors, we also omitted X4 from the final model. Applying our method to Data Set 2 was similarly successful in identifying and quantifying the contributions of exposures and confounders to outcome variability.

Conclusion: Our approach was effective at identifying and quantifying the unique and joint contribution of the primary exposures to variance in the outcome, and achieved extremely high explanatory power for Data Set 1. Advantageous aspects of our approach are that it is scalable to high-dimensional exposome data, easily incorporates confounders, flexibly incorporates interactions and higher order terms, is intuitively interpretable, and can be readily implemented using standard statistical software.

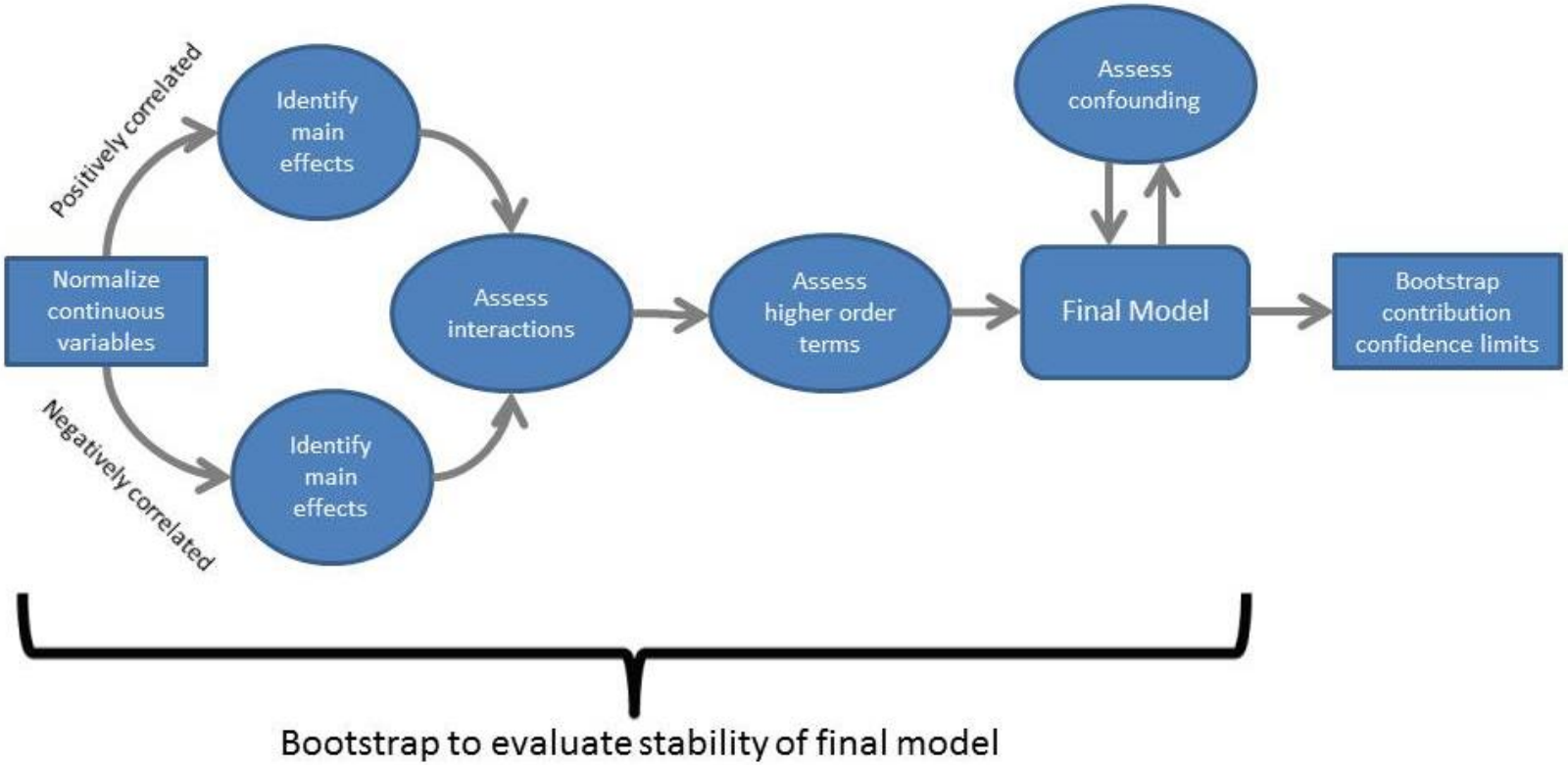


Figure 1. Analytical strategy for variable selection and assessment

Table 1. Results of Variable Selection for Data Set 1: Contribution of each term and variable to the model and predictable variance in the outcome Y

Variable (set mode)	Predictor term in model	Beta Estimate*	Contribution to Predictable variance by each term ($\Delta term_j$)	Relative % Contribution to predictable variance ($\%term_j$)	Total effect of each variable (Δvar_j)	Relative % Total effect of each variable ($\%var_j$)
Z (+)	Z	0.956	157.264	23.9 [19.1, 28.5]	275.04	30.9 [28.0, 33.6]
X1 (+)	X1	0.532	41.134	6.3 [4.1, 8.8]	126.79	14.2 [11.4, 17.0]
	Z:X1	-0.250	8.872	1.3 [0.4, 2.7]	-	-
X7 (+)	X7	0.420	191.856	29.2 [24.6, 33.0]	231.50	26.0 [23.3, 28.8]
	$X7^2$	-0.195	39.551	6.0 [3.9, 8.4]		
	$X7^3$	0.031	11.754	1.8 [0.6, 3.5]		
X5 (-)	X5	-0.386	188.514	28.7 [25.0, 32.5]	256.78	28.8 [26.4, 31.4]
	$X5^2$	0.047	11.996	1.8 [0.7, 3.5]		
	X5:X7	-0.041	6.065	0.9 [0.2, 2.1]	-	-

*A beta estimate for a predictor indicates the expected increase of decrease in the outcome, in SD units, given a one SD increase in predictor with all other predictor held constant.

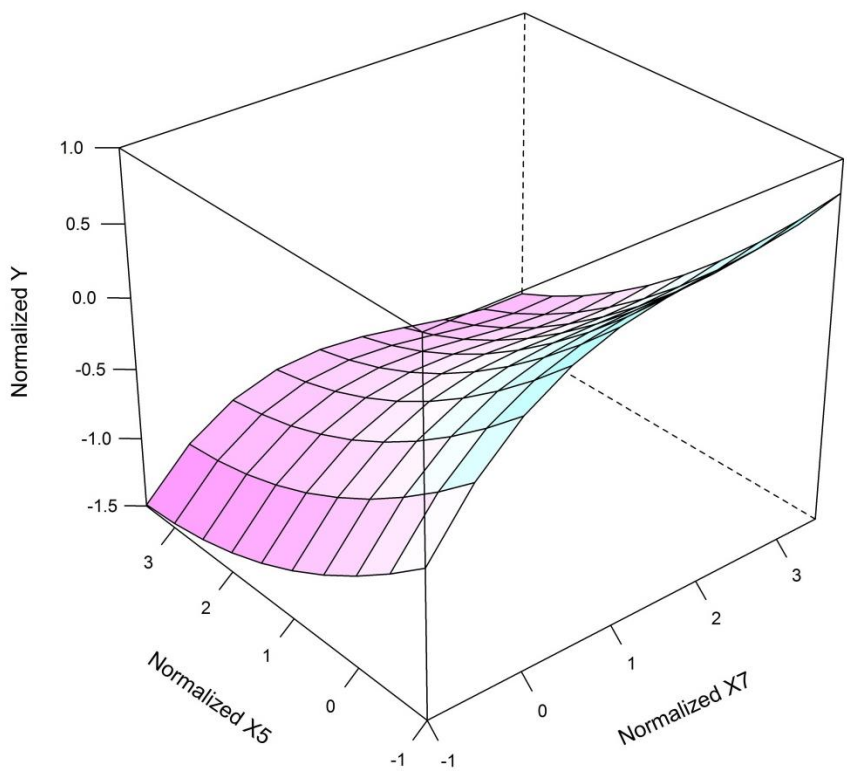


Figure 2. Relationship of Y to X5 and X7.

24. Analyzing Mixtures in Epidemiology Data by Smoothing in Exposure Space

Presenting Author: Veronica Vieira

Organization: Boston University

Contributing Authors: Thomas F. Webster and Veronica M. Vieira

Abstract:

Background: Analysis of the health effects of mixtures is an important topic in both environmental epidemiology and toxicology, although the two fields approach the problem differently.[1] We combine ideas from both[2] to analyze the synthetic data sets posted by NIEHS. Briefly, consider the joint distribution of exposures as defining an exposure space; the outcome represents a surface in this space. Examination of the shape of the levels sets—contours or isoboles as they are called in toxicology—provides information regarding interaction (or not) between exposures. We regard our method primarily as an exploratory data analysis approach that can be used to examine types of "interactions" between exposure variables.

Methods: Suppose that each record of the epidemiologic data set contains a health outcome Y, a set of p exposures X, and covariates Z. Our basic approach uses generalized additive models (gam) to examine the data in exposure space:

$g[Y] = S[X] + \gamma' Z$ (1)

where g[.] is a link function (as in generalized linear models) and S[X] is a smooth function of the exposures. As the outcomes in both synthetic data sets appear to be normally distributed continuous data, we use the identity link. However the general method is also applicable to binary and other types of outcome data. Here we use loess for the smooth function, choosing the degree of smoothing (span) by minimizing the AIC, representing a tradeoff between bias and variance. We have used this approach previously for mapping geographically distributed data in two dimensions.[3] We can deviate from the notation above by including within the smoothing term the non-linear parts of the function (exposure or covariates, as needed), and placing linearly modeled variables outside.

Following smoothing, we take 2 dimensional slices of the multi-dimensional object, mapping the value of the outcome using a color scale and drawing the isoboles. Parallel straight line isoboles (usually of negative slope) imply that the variables can be modeled using toxic equivalent factors (TEFs). For example, the 2 variable joint dose response function f[X1, X2] can then be expressed as f[X1, X2] = h[w1X1 + w2X2] where wi are constants (that need not sum to 1) and h[.] is the dose-response function describing the outcome as a weighted sum of the exposures. The relative potency (TEF) of the compounds can be estimated from the slopes of the isoboles. The shape of the function h[.] can be examined by plotting outcome as a function of w1X1 + w2X2 along one dimensional cross-sections or rays. Other isobole shapes are informative about "interactions," defined in the toxicological sense of the word relative to concentration addition, CA.[1] For example, negatively sloped isobles that bow

downward (positive second derivative) are supra-linear ("synergistic") relative to concentration addition.[4] Positively sloped isoboles (straight or curved) are types of "antagonism" where the compounds are acting in opposite directions. There are other types: e.g., horizontal or vertical isoboles imply a variable does not contribute to the outcome.

As currently implemented in R, our method can only smooth up to 8 dimensions before crashing. We also examined histograms and correlations among variables and used multivariable regression and stepwise regression for initial exploration of the data and for initial variable selection. As a result, we may have failed to pick up some variables.

Results: The two synthetic data sets were quite different as discussed below. We begin with results for data set #2 which has more variables but is simpler in other ways than data set #1.

Dataset #2: Initial analysis of dataset 2 suggested that the exposures were multivariate normal and the outcome was normal. Scatterplots showed no obvious non-linearities. There were high degrees of correlation, >0.90, between some exposure variables (x3-x4-x5 and x12-x13), making it difficult to separate their effects; it also produced high variance inflation factors (VIFs) in multivariable regression. We therefore did an initial step- wise regression (acknowledging that this approach is imperfect), pruning the data set to the following 8 predictors: x1, x4, x6, x10, x12, x14, z2, z3. Inclusion of quadratic terms or binary cross-products did not change results much, but suggested a non-linearity involving z2. Initial smoothing suggested treating z3 (binary covariate) as a linear predictor. We jointly smoothed on the remaining 7 variables (x1, x4, x6, x10, x12, x14, z2), and adjusted for z3 outside the smooth. The results suggest that all of these exposure variables (x1, x6, x10, x12, x14) are TEF with respect to each other—i.e., non-interactive from the toxicologic (CA) point of view—and are positive linear predictors of the outcome (e.g., Fig 1). Potencies relative to x1 were 0.6 for x4, 1 for x6, 1.5 for x10, 2.6 for x12, 1.1 for x14. Removing the binary covariate z2 from the smooth produced close to the same results. The smooths suggest non-linear effects of z2, but it is a nuisance variable. Multivariable regression on these variables (including z3 and z33) produced a model with Rsq=0.51 and the following beta coefficients for the exposures: $\beta1=0.059$ ($p=0.06$), $\beta4=0.039$ ($p=0.2$), $\beta6=0.059$ ($p=0.04$), $\beta10=0.087$ ($p=0.009$), $\beta12=0.16$ ($p=0.03$), $\beta14=0.067$ ($p=0.15$). Omitting z33 produced similar results. Ratios of these beta coefficients are similar to the TEF estimates above.

Dataset #1: Initial analysis of dataset 1 suggested that the exposures were log normal and the outcome was normal; Z was binary. Preliminary stepwise regression identified exposures X1, X2, X4, X5, and X7 as important predictors, while X3 and X6 did not contribute to the outcome. X1 and X2 were highly correlated and inclusion of both produced higher VIFs in multivariable regression. A multivariable regression of log transformed exposures X1, X4, X5, X7, the quadratic terms for X1, X5, and X7, interaction terms for X7 with X1 and X5, and controlling for Z produced the most parsimonious model with an R-squared of 0.95. The multivariable regression of X1, X4, X5, X7 (not transformed) had an R-squared of 0.91. We therefore jointly smoothed exposures X1, X4, X5, X7, with Z modeled parametrically (X2 was omitted because of high correlation with X1). We identified the following interactions: For X1 and X7, the model produced negatively sloped isoboles that bow downward (positive second derivative), suggesting that the interaction between these two exposures is "synergistic." (Fig 2). The isoboles for X5

and both X1 and X7 were positively sloped, with the predictions highest for low values of X5. This suggests that X5 is acting in the opposite direction than X1 and X7 (one type of antagonism). The isoboles for X4 and X5 were negatively sloped, with outcome decreasing as both exposures increased. This indicated that X4 and X5 are acting in the same direction; X4 has an antagonistic interaction with X1 and X7.

Conclusions: Our approach is designed for exploratory data analysis of mixtures in epidemiology. We identified cases in the two data sets where variables appeared to be TEF with respect to each other, antagonistic or synergistic (relative to concentration addition). Further modeling could be used to refine these results. The use of smooths and toxicological concepts provides more information than standard methods such as inclusion of cross- product terms in regressions. Our conclusions on interactions (or lack thereof) are not statistical tests. Our method has limitations with the number of variables it can currently jointly smooth and is not designed as a data reduction tool or to handle very highly correlated data.

References

1. Howard GJ, Webster TF. Environ Health Perspect 2013; 121:1–6.
2. Webster TF, Vieira VM, Korrick SA. Analysis of the Effect of Mixtures: Application of a New Method to a Study of Organochlorines and ADHD. ISEE 2014. (Seattle, 24-28 August 2014).
3. Vieira V, Webster T, Weinberg J, Aschengrau A, Ozonoff D. Environmental Health 2005; 4:11
4. Howard GJ, Webster TF. J. Theor Bio 2009; 259:469–477.

Fig 1. Cross-section isoboles of x1-x6 from Data Set #1 show a TEF pattern

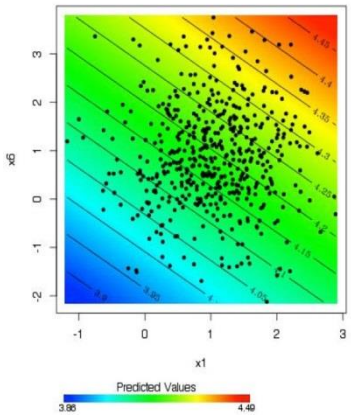
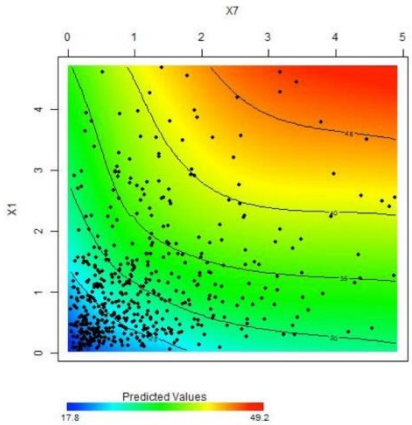


Fig. 2. Cross-section isoboles of X1-X7 from Data Set #2 show synergism



25. Variable Selection and Multivariate Adaptive Spline Assessments to Investigate Effects of Chemical Mixtures in a Prospective Cohort Study of Mother-Child Pairs

Presenting Author: Katrina Waters

Organization: Pacific Northwest National Laboratory

Contributing Authors: Lucas C. Tate, Greg F. Piepel, Aimee E. Holmes, Dennis G. Thomas, and Katrina M. Waters

Abstract:

Statistical Modeling Approaches Considered: Approaches for statistically assessing the health effects of environmental chemical mixtures in an epidemiology study should ideally be chosen based on knowledge of the underlying subject matter and variables in the data. However, this was not possible for the two test data sets (TDSs). It was also not possible to consider the *combined mixture model* [1], which includes the effects of chemicals with a common mechanism (via concentration addition) and chemicals with different mechanisms (via independent joint action).

Several statistical modeling methods were applied to the TDSs, including variable selection regression (VSR), regression trees, random forests, partial least squares, neural networks, support vector machines, and multivariate adaptive regression splines (MARS). No method was clearly superior for either of the TDSs. For each TDS, the VSR and MARS results are summarized, with VSR serving as a baseline. The VSR approach used best-subsets and stepwise regressions, with several diagnostics used to select models. Terms retained in the models had p-values <0.01. The VSR analyses were performed using the Minitab 17 statistical package. MARS is a nonparametric regression approach that identifies important variables and uses splines to model the relationships between the predictor variables and a continuous outcome variable[4]. Up to cubic splines can be used, but linear splines were used for these analyses. MARS analyses were conducted using the *earth* package[5] in the R programming language.

Stratification by the binary covariate was considered for each TDS to assess whether the relationships with the remaining predictor variables were consistent for the two strata. Models were validated using 10-fold cross-validation (CV) to assess over-fitting of the data and guide model selection.

Results for Data Set 1: This data set containing 500 records was described as a prospective cohort study, with a continuous outcome (Y), seven continuous exposure variables (X1-X7), and a binary covariate (Z). Notable pairwise correlations were observed involving exposure variables in the groups (X1, X2, and X3) and (X5, X6). Log transformations of the exposure variables simplified their distributions, but did not result in any measureable gains to model fits or impact important model terms chosen by the VSR and MARS approaches. Stratification on the binary Z was investigated for the VSR and MARS approaches, but there was little difference in the important variables and model fits. Hence, the modeling results for the non-stratified data set without log transformations of X1-X7 are summarized.

The VSR and MARS approaches yielded models involving X1, X2, X4, X5, X7, and Z with X3 and X6 not being selected. The VSR model had R2=0.945 and contained 12 terms (Table 1), including two squared terms and three interaction terms (the latter indicating effects beyond an additive model). Coefficients of interaction terms (Table 1) that are negative (positive) indicate a joint effect that is lower (higher) than would be expected in an additive model. The predicted vs. measured (PvM) plot for the VSR and MARS models are in Figure 1.

The MARS approach produced a first-degree model with 13 terms (Table 2) and R2=0.946. Interactions were considered, but did not contribute to the model fit. The VSR and MARS models had good CV performance. The MARS CV showed potentially similar performance among a family of models (containing from 6–14 terms) that could be explored given information about the variables. A bootstrap method was used to investigate variable importance in the MARS model; the aggregate results agreed with the results from the full data set. The bootstrap results also indicated volatility in variable importance, which is likely due to strong correlations involving certain exposure variables. Individual variable effects based on the fitted MARS model are shown in Figure 2. Caution is needed in interpreting individual effects given some stronger correlations.

Results for Data Set 2: This data set containing 500 records is considered as a cross-sectional study, with a continuous outcome variable (y), 14 continuous exposure biomarkers (x1-x14), two continuous covariates (z1, z2) and a binary covariate (z3). Two groups of biomarkers [(x3, x4, x5) and (x12, x13)] had very strong pairwise correlations, with many other pairwise correlations being relatively strong (including several pairs involving z2). Both the VSR and MARS approaches yielded notably different results when the data were stratified by z3, so modeling results are presented separately for z3=0 and z3=1.

The VSR model results for the two values of z3 are listed in Table 3 with the PvM plot in Figure 3. Although second-degree models were considered, they provided only marginal improvements to model fits. Hence, the models in Table 3 include only first-degree terms.

The MARS models were selected based on CV results (see example in Figure 4) after CV showed that the original MARS models with total R2=0.623 were over-fitting. Results for the CV models are summarized in Tables 4 and 5, with separate R2 values of 0.472 (z3=0) and 0.245 (z3=1) and a total data set R2=0.538. The PvM plot (Figure 5) suggests that values of z3 may represent two different groups in the population, with the outcome variable tending to be larger for z3=0. Variable importance was investigated for each model with the bootstrap method, which revealed very high volatility, particularly among highly correlated variables.

VSR and MARS selected similar biomarkers of importance with a few minor differences (see Tables 3, 4, and 5). For the z3=0 stratum, VSR selected x5 while MARS selected x3; these variables are highly correlated, which may result from a common mode of exposure. For z3=1, VSR selected an additional term x14 as compared to MARS. Neither VSR nor MARS provide evidence of important interactions beyond an additive model.

Summary/Conclusions: TDS1 was modeled effectively using an assortment of methods, although interpreting results must be done with caution because of certain highly correlated exposure variables. TDS2 was more difficult to model, with lower R2 values and over-fitting issues. The MARS models were selected using CV to avoid over-fitting. The MARS methodology provided a very flexible approach without assuming a model form. It also provided important information regarding individual variable effects.

Table 1. Summary of VSR Model Fit for TDS1

Model	Coefficients
Intercept	11.766
X1	4.473
X2	4.450
X4	-0.959
X5	-5.179
X7	7.370
Z	10.836
$(X5)^2$	0.535
$(X7)^2$	-1.179
X1*X2	-1.487
X1*X7	0.464
X5*X7	-0.446
$R^2 = 0.945$	$CVR^2 = 0.942$

Table 2. Summary of MARS Model Fit for TDS1 Limited to Main Effects

Model	Coefficients
(Intercept)	23.0640774
Z	10.5397227
h(1.63282-X1)	-4.1346103
h(X1-1.63282)	2.9925117
h(1.08821-X2)	-3.7072375
h(3.52697-X4)	0.7616012
h(X4-3.52697)	-3.2743707
h(0.52139-X5)	6.9418541
h(X5-0.52139)	-4.5446649
h(X5-1.92941)	2.3178779
h(0.70711-X7)	-10.4152238
h(X7-0.70711)	2.8369383
h(X7-2.16572)	-2.2637288
Importance: Z, X5, X7, X1, X4, X2	
GCV 7.031618	RSS 3173.018
GRSq 0.9404181	RSq 0.9460116

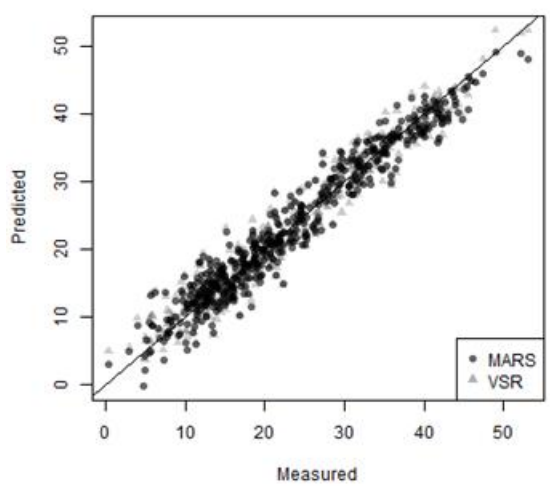


Figure 1. Predicted versus Measured Plot for VSR and MARS Models Fit to TDS1

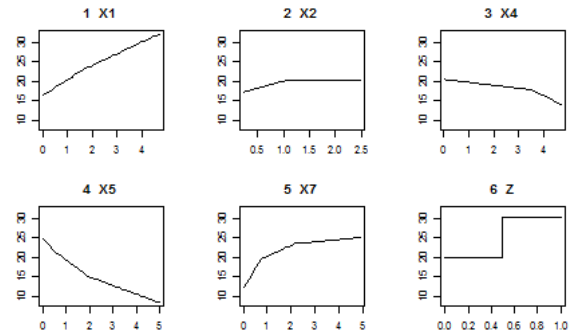


Figure 2.TDS1 MARS Model Variable Effects with Other Variables at Median Values

Table 3. Summary of VSR Model Fit for TDS2

	Coefficients	
Model	z3 = 0	z3 = 1
Intercept	3.5453	3.1698
x5	0.0881	–
x6	0.1165	–
x12	0.5936	–
x14	–	0.1351
z2	0.0067	0.0084
Combined $R^2 = 0.539$	$R^2 = 0.452$ $CVR^2 = 0.412$	$R^2 = 0.276$ $CVR^2 = 0.260$

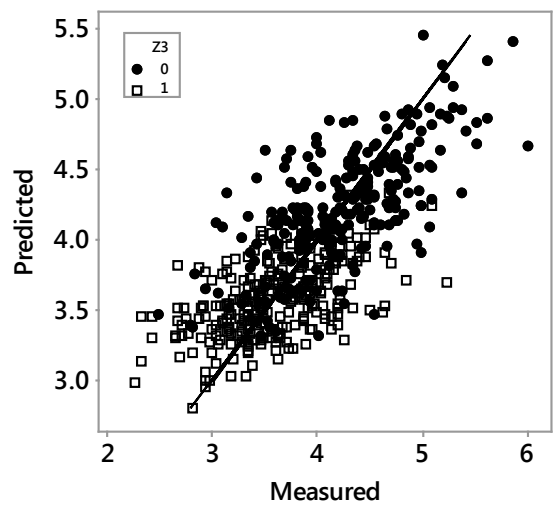


Figure 3. Predicted versus Measured Plot for VSR Models Fit to TDS2 Separately by z3

Table 4. Summary of MARS Model Fit for TDS2 z3=0 Stratum Limited to Main Effects	
Model	Coefficients
(Intercept)	4.5771104
h(x3-3.12469)	0.4248934
h(x6+0.33975)	0.1556095
h(1.02463-x12)	-0.5901474
h(60.234-z2)	-0.0091927
Importance: z2, x12, x6, x3	
GCV 0.2066718	RSS 46.77643
GRSq 0.4358774	RSq 0.4724098

Table 5. Summary of MARS Model Fit for TDS2 z3=1 Stratum Limited to Main Effects	
Model	Coefficients
(Intercept)	3.16634617
h(z2+5.764)	0.01167121
Importance: z2	
GCV 0.2057097	RSS 51.43466
GRSq 0.2331271	RSq 0.2451093

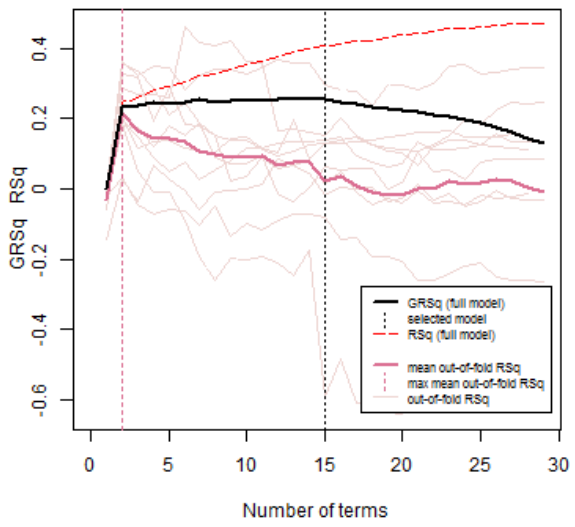


Figure 4. CV of TDS2 MARS Model Fit to z3=1

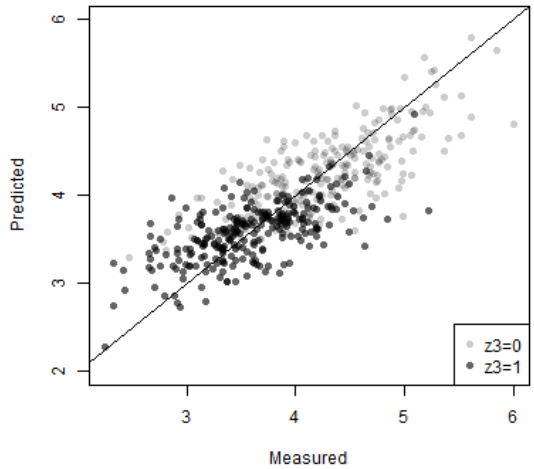


Figure 5. Predicted versus Measured Plot for MARS Model Fit to TDS2 Separately by z3

References

1. Computational Approach to the Toxicity Assessment of Mixtures. 2006. *A Computational Framework for Assessing the Toxicity of Chemical Mixtures*. http://wang.tox.ncsu.edu/model5/
2. Craven P, G Wahba. "Smoothing Noisy Data with Spline Functions." 1979. *Numerische Mathematik.* 31:377-403.
3. Friedman JH, BW Silverman. 1989. "Flexible Parsimonious Smoothing and Additive Modeling." *Technometrics*, 31:3-21.
4. Friedman JH. 1991. "Multivariate Adaptive Regression Splines." *The Annals of Statistics*, 1:1-67.
5. Milborrow S. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. 2015. *earth: Multivariate Adaptive Regression Splines*. R package version 4.2.0. http://CRAN.R-project.org/package=earth

26. Bayesian Non-Parametric Regression for Multi-Pollutant Mixtures

Presenting Author: Ran Wei

Organization: North Carolina State University

Contributing Authors: Ran Wei, Subhashis Ghoshal, Brian Reich, and Jane Hoppin

Abstract:

Many modern epidemiological studies simultaneously investigate the effect of several exposure variables. The statistical challenge is to identify the subset of harmful exposure variables and to estimate the joint effects of multiple exposures. In order to choose an appropriate subset of variables related to the outcome measurements, we propose a Bayesian nonparametric regression model with continuous shrinkage priors for variable selection and model prediction. Our general approach is to decompose the dose-response function as the sum of nonlinear main effects and two-way interaction terms, and apply novel Bayesian variable selection methods to identify important exposures and interactions. The advantage of this approach is that the results are easily interpretable because the signal is allocated to individual exposures and synergistic pairs. The primary challenge when fitting this standard additive model is that the number of parameters explodes even for a moderate number of exposure variables leaving the analysis susceptible to over-fitting. Our approach to overcoming this challenge is to apply a prior that aggressively shrinks many of the terms towards zero, thus mitigating the noise of including unimportant exposures and allowing us to isolate the effects of the important variables. Unlike ad-hoc screening procedures like forward/backward selection, we accomplish variable selection within a single Bayesian hierarchical model. This permits valid statistical inference while properly accounting for all sources of uncertainty, including the uncertainty about the subset of variables to be included. The proposed method is applied to two simulated datasets, and we find evidence in both that including interactions between exposures improves predictive performance.

Model description: We assume the health response Y is normal with mean $f(X_1, \ldots, X_p)$ and variance σ^2, where $X_1, \ldots, X_p$ are the explanatory variables. For notational convenience, we include both the exposure variables of interest and confounders in the p explanatory variables. The joint dose-response function is decomposed as the sum of main-effect and interaction functions, $f(X_1, \ldots, X_p) = C + \sum_{j=1}^{p} f_j(X_j) + \sum_{l<k} f_{lk}(X_l, X_k)$; the main effects include non-linear effects for both confounder and exposures and the second-order terms include both interactions between confounder and exposures, pairs of exposures, and pairs of confounders. After basis expansion, such as $f_j(X) = \sum_{k=1}^{m} B_k(X)\theta_{jk}$, the unknown coefficients are assigned priors $\theta_{jk} \overset{iid}{\sim} N(0, \sigma^2\lambda_j)$, for $k = 1, \ldots, m$. We select continuous shrinkage priors for the variance components, as described below.

For the main effects, we use B-spline basis functions for the B_k with $m = 5$; for interactions we use the outer product of B-spline functions. This results in many terms, and thus variable selection is required. Exposures are completely eliminated only if the entire curve f_j or f_{lk} is zero. The key observation is that this is equivalent to setting the variance parameter λ_j (or λ_{lk} for interactions) to zero. We can conduct model selection by shrinking the values of λ_j through a continuous shrinkage prior. We let $\lambda_j = \lambda_0\varphi_j$ and $\lambda_{lk} = \lambda_0\varphi_{lk}$, where λ_0 controls the overall variance and ϕ_j is the proportion of variance allocated to main effect j, with the proportions adding to one, $\sum_j \phi_j + \sum_{l<k} \phi_{lk} = 1$. The proportions ϕ are then given a Dirichlet prior that encourages most of the proportions to be near zero and a select few to be large. The main effects of confounder variables are not included in the variable selection model, and are given separate variance parameters. Standard Markov chain Monte Carlo techniques are used to implement this approach.

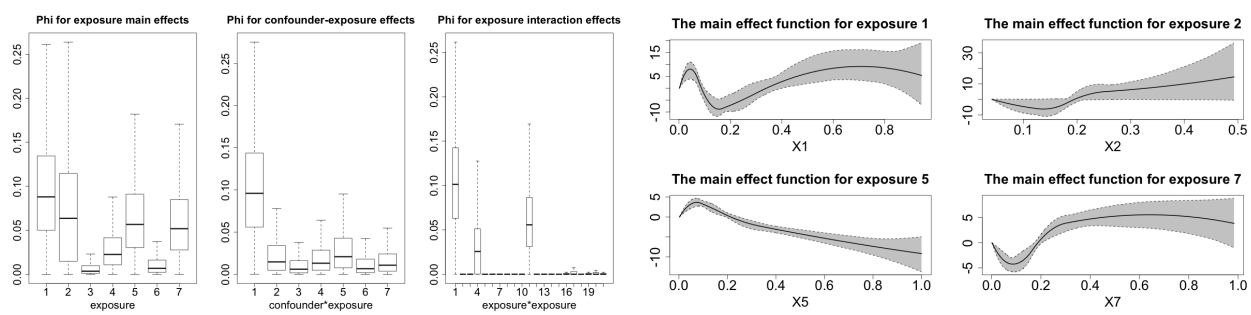


Figure 1: Results for simulated data set 1: Posterior distribution of (from left to right) the variance proportions φ for the main effects, confounder $\times$ exposure interactions, exposure $\times$ exposure interactions, and selected main effect curves $f_j(X)$.

Simulated data set 1. There are seven predictors and one binary confounder variable. Figure 1 plots the posterior of the φ proportions and selected main effect functions. Clearly, the main effects for X_1, X_2, X_5 and X_7 and interactions between $X_1 \times X_2$ and $X_2 \times X_7$ contribute to the outcome. The left middle plot also demonstrates the significant contribution of the binary confounder through interaction with X_1. The main effect functions plotted on the right of Figure 1 demonstrate non-linear dose-response functions.

We compared the main-effect-only model with the full model with interactions using

cross-validation. The mean square error for main-effects model is 20.2 compared to 13.1 for the full model. Also, including the interactions increased predictive R-squared from 82.8% to 88.9%. Therefore, it appears the predictions are accurate and the interactions are important.

Simulated data set 2. In simulated data 2, there are 14 chemical exposures, two continuous confounding variables and a binary confounding variable. Here we include main effects for all exposures and confounders, interactions between the exposures and the binary confounder, and the interactions between each pair of exposures. Figure 2 shows that the important effects do not emerge as clearly as for the first data set. The main effects for X_2, X_4, X_5 and X_{13} and interactions $X_2 \times X_{13}$, $X_{10} \times X_{14}$, $X_{11} \times X_{13}$ and $X_{13} \times X_{14}$ contribute the most to the mean response. Furthermore, the exposures with the strongest interaction with the binary confounder are X_2, X_5 and X_{13}.

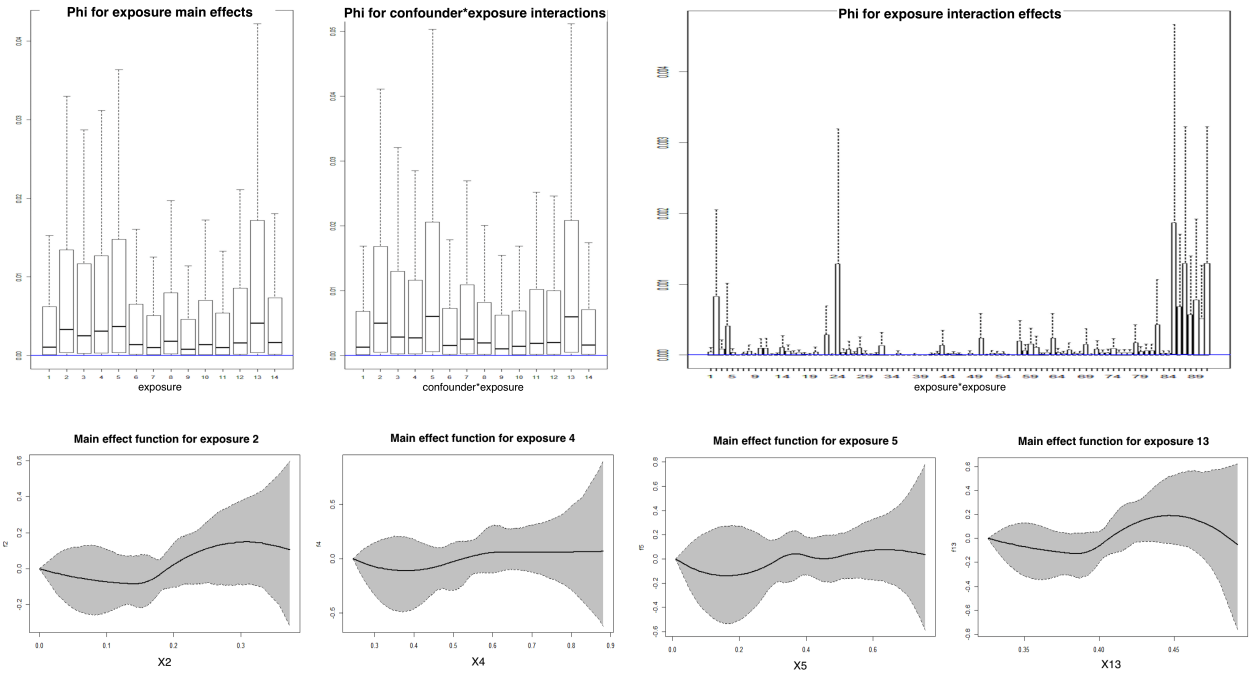


Figure 2: Results for simulated data set 2: Posterior distribution of (from up left to up right) the variance proportions $\boldsymbol{\varphi}$ for the main effects, confounder $\times$ exposure interactions, exposure $\times$ exposure interactions. Selected main effect curves $f_j(X)$ (from bottom left to bottom right).

We also performed cross validation. The mean squared prediction error was 0.35 for the main-effect-only model compared to 0.32 for the full model, and prediction R-squared is increased from 17.2% for the main-effects-only model to 22.9% for the full model. Therefore, predictions are less precise in general for the second data set compared to the first, and the contribution of interactions is less prominent.

27. Modeling Environmental Chemical Mixtures with Weighted Quantile Sum Regression

Presenting Author: David Wheeler

Organization: Virginia Commonwealth University

Contributing Authors: David Wheeler and Jenna Czarnota

Abstract:

Introduction: We present an application of weighted quantile sum (WQS) regression to model the association of a mixture of chemical exposures and a continuous outcome variable in two simulated data sets. Estimation of chemical weights and the resulting WQS index while considering the correlation between compounds allows us to make generalized inference about the mixture effect and identify the individual chemicals ('bad actors') most strongly associated with the outcome.

Methods: The WQS method is constrained to have associations in the same direction for chemical exposures and risk, and is designed for variable selection over prediction. WQS regression estimates a weighted linear index in which the weights are empirically determined through the use of bootstrap sampling. The approach considers data with *c* correlated components scored as ordinal variables into quantiles that are reasonable to combine (i.e., all chemicals) into an index and potentially have a common adverse outcome. The weights are constrained to sum to 1 and be between 0 and 1, thereby reducing dimensionality and addressing issues associated with collinearity. For this analysis, the chemical concentrations were scored into quartiles denoted by q_i, where q_i = 0, 1, 2, or 3 for i = 1 to c. A total of B = 1000 bootstrap samples (of the same size as the total sample, N = 500) were generated from the full dataset and used to estimate the unknown weights, **w**, that maximized the likelihood for b = 1 to B for the following model

$$g(\mu) = \beta_0 + \beta_1\left(\sum_{i=1}^{c} w_i q_i\right) + \mathbf{z}'\phi\Big|_b \qquad [1]$$

subject to the constraints $\sum_{i=1}^{c} w_i\Big|_b = 1$ and $0 \le w_i \le 1$ for i = 1 to c. In the above equation, w_i represents the weight for the ith chemical component q_i and the term $\sum_{i=1}^{c} w_i q_i$ represents a weighted index for the set of c chemicals of interest. Furthermore, **z** denotes a vector of covariates determined prior to estimation of the weights, ϕ are the coefficients for the covariates in **z** , and g(.) is any monotonic and differentiable link function that relates the mean, μ, to the predictor variables in the right hand side of the equation. Because the outcome variables in this analysis are continuous, an identity link was assumed for g.

For each bootstrap sample, the relative strength of the test statistic for β_1, the parameter estimate for the weighted index, was used to estimate the final vector of weights $\overline{w}$, and the weighted quantile score was estimated as $\mathrm{WQS} = \sum_{i=1}^{c} \overline{w}_i q_i$. Finally, the significance of the WQS index was determined using the original data set and the model

$$g(\mu) = \beta_0 + \beta_1 \mathrm{WQS} + \mathbf{z}'\phi, \qquad [2]$$

where β_1 is the parameter associated with a unit (quartile) increase in the weighted sum of exposure quartiles (WQS index).

Results Data Set #1: The data included seven exposure variables (X1-X7) with pairwise Spearmen correlations among the exposures ranging from -0.10 to 0.88. The WQS index was significantly related to the outcome variable ($p < 0.001$), and a one quartile increase in the WQS index was associated with a 5.74 unit (95% CI: 5.07, 6.40) increase in the response (Table 1). The components of X1, X3, and X7 received the highest estimated weights (Table 2) with w1 = 0.323, w3 = 0.147, w7 = 0.482, respectively, and were therefore identified as important exposures (i.e., exposures contributing to the outcome). Components X2 and X4 received smaller weights of w2 = 0.037 and w4 = 0.011 and therefore contributed little to the mixture effect. Components X5 and X6 received negligible weight (< 0.001) and were therefore considered to be unassociated with the outcome. The distribution of the estimated weights for each component is shown in Figure 1.We did not consider interactions between chemicals in this analysis. The adjusted r-square of the WQS model was 0.80 and the root mean square error was 107.95.

Table 1: WQS model parameter estimates for data set #1.

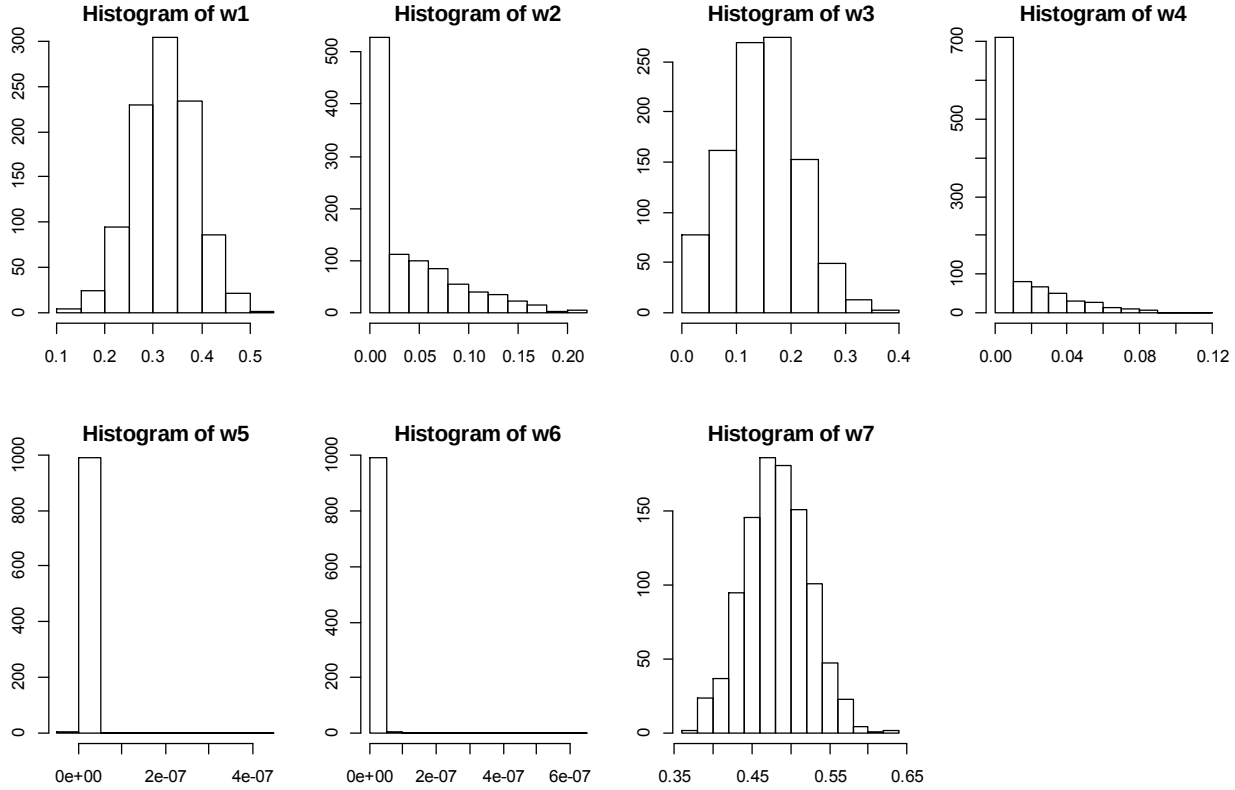
	Estimate	SE	95% CI	p-value
Intercept	9.83	0.44	(8.96, 10.70)	<0.001
Z	11.37	0.60	(10.21, 12.54)	<0.001
WQS	5.74	0.34	(5.07, 6.40)	<0.001

Adjusted R^2 = 0.80; RMSE = 107.95

Table 2: Estimated WQS chemical weights using the relative test statistic for data set #1.

	Estimate
w1	0.323
w2	0.037
w3	0.147
w4	0.011
w5	0.000
w6	0.000
w7	0.482

Figure 1: Histogram of estimated weights for X1-X7 across 1000 bootstrap samples for data set #1.



Results Data Set 2: The data included 14 exposures variables (X1-X14) with pairwise Spearmen correlations among the exposures ranging from -0.14 to 0.99. The WQS index was significant ($p<0.001$), and a one quartile increase in the WQS index was associated with a 0.30 unit (95% CI: 0.23, 0.38) increase in the outcome (Table 3). The most heavily weighted exposures in the index included X1, X6, X8, and X12 with weights w1 = 0.134, w6 = 0.190, w8 = 0.101, and w12 = 0.168, and thus contributed most strongly to the outcome (Table 4). Exposures X4, X10, X11, and X14 each received between 5 and 10% of the total weight and were considered as less important variables that still contributed to the outcome. Finally, exposures X2, X3, X5, X7, X9, and X13 received less than 5% of the total weights, and therefore were considered to not contribute to the outcome. The distribution of the estimated weights for each component is shown in Figure 2. The adjusted r-square of the WQS model was 0.51 and the root mean square error was 10.16.

Table 3: WQS model parameter estimates for data set #2.

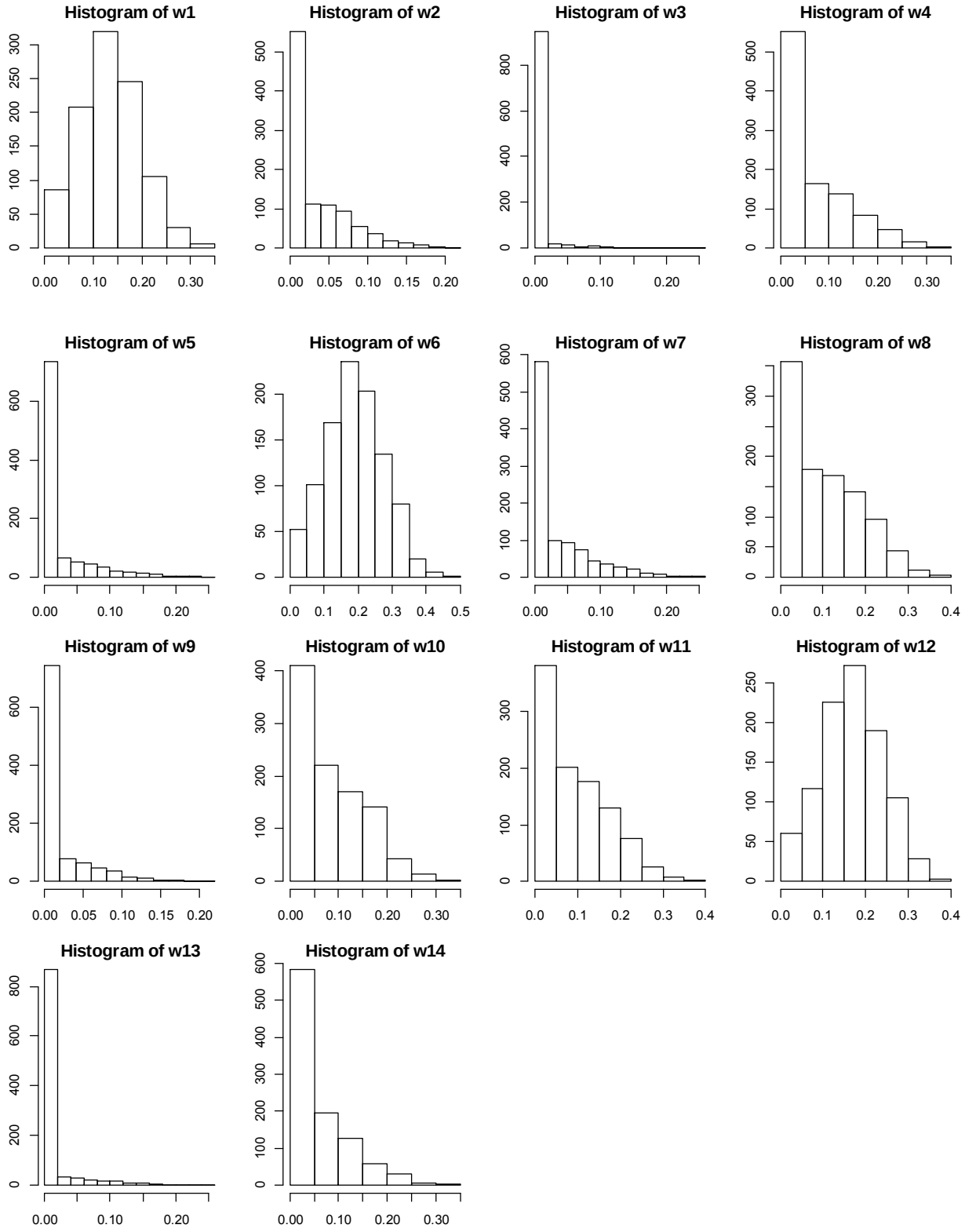
	Estimate	SE	95% CI	p-value
Intercept	3.53	0.07	(3.40, 3.66)	<0.001
z1	0.01	0.01	(-0.02, 0.03)	0.606
z2	0.01	0.00	(0.01, 0.01)	<0.001
z3	-0.60	0.04	(-0.68, -0.52)	<0.001
WQS	0.30	0.04	(0.23, 0.38)	<0.001

Adjusted R^2 = 0.51; RMSE = 10.16

Table 4: Estimated WQS chemical weights using the relative test statistic for data set #2.

	Estimate
w1	0.134
w2	0.032
w3	0.004
w4	0.063
w5	0.022
w6	0.190
w7	0.033
w8	0.101
w9	0.017
w10	0.081
w11	0.089
w12	0.168
w13	0.011
w14	0.054

Figure 2: Histogram of estimated weights for X1-X11 across 1000 bootstrap samples for data set #2.



28. Assessing Health Associations with Environmental Chemical Mixtures using LASSO and its Generalizations

Presenting Author: Changchun Xie

Organization: University of Cincinnati

Contributing Authors: Changchun Xie, Aimin Chen, and Susan M. Pinney

Abstract:

When there are multiple chemical exposures and some of them are highly correlated, there are two challenges in traditional multiple regression: 1) the number of chemical exposures could be greater than the number of samples; 2) multicollinearity. The situation can become more complicated when interactions among exposures and nonlinear effects are involved. Testing interaction between exposures might not be feasible due to their large number. A promising technique called the least absolute shrinkage and selection operator (LASSO) and its generations (least angle regression, elastic net, group LASSO) can be used to handle these challenges. LASSO, proposed by Tibshirani (1996), is a penalized least squares procedure that minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. We have applied LASSO to analyze the two simulated datasets and the real world dataset using the R package "glmnet". It is well known that the p-values from a significant test (for example, chi-squared test or F-test) designed for fixed linear models are not appropriate for adaptively selected models such as those in forward stepwise regression. The p-values from the covariance test (Lockhart et al., 2014), which accounts for the adaptive nature of LASSO modeling are provided using the R package "covTest". We also suggest an approach to test interactions when there are many potential chemical exposures.

Significance testing: Classic theory for significance testing in linear regression operates on two fixed nested models. For example, chi-squared test can be used to compute the drop in residual sum of squares (RSS) from regression on $A \cup \{i\}$ and A, where $A \cup \{i\}$ and A are fixed subsets of {1,2,…,p},

$$R_i = (RSS_A - RSS_{A\cup\{i\}})/\sigma^2$$

and compares it to a χ_1^2 distribution (σ^2 is assumed to be known. If not, F-test can be used). Unfortunately, when $A \cup \{i\}$ and A are not fixed subsets, the use of a χ_1^2 null distribution is not valid anymore. For example, in forward stepwise regression, we enter predictors one at a time, at each step choosing the predictor i that gives the largest drop in RSS. The maximum R_i will clearly be larger than χ_1^2 under the null and Type I error is inflated.

The covariance test was proposed by Lockhart et al. (2014) to account for the adaptive nature of LASSO. Let A be the active set of predictors just before knot λ_k and suppose the predictor i enters at λ_k. Let $\hat{\beta}(\lambda_{k+1})$ be the solution at the next knot, λ_{k+1}, using predictors $A \cup \{i\}$ and $\widetilde{\beta_A}(\lambda_{k+1})$ be the solution using predictors A at λ_{k+1}. The covariance test statistic is defined by

$$T_k = (\langle y, X\hat{\beta}(\lambda_{k+1})\rangle - \langle y, X_A\widetilde{\beta_A}(\lambda_{k+1})\rangle)/\sigma^2$$

Lockhart et al. (2014) show that under the null hypothesis that the current LASSO model contains all truly active variables, $T_k \xrightarrow{d} Exp(1)$. This null hypothesis will change at each step where the set of the active variables changes. The p-value from this test cannot be interpreted in the classic sense based on a fixed null hypothesis since this test is a conditional test.

Results: The following summarizes the findings:

Simulated Data Set #1:

X1, X2, X3 are highly correlated. Log transformation was used for X1, X2, X3 and X7 due to nonlinear relationship with Y, based on scatter plots. We retained the original variable names. Based on 10-fold cross-validation, the optimal LASSO fit did not select X3 and X6. The estimated coefficients are:

(Intercept)	24.292	X5	-3.264
X1	3.330	X6	0
X2	0.287	X7	3.363
X3	0	Z	11.372
X4	-0.103		

We also did our own simulations by generating our own outcome variable Y using X1-X7 and Z in Data Set #1. We found that LASSO can select right variables among X1, X2 and X3, which are highly correlated, while Random Forest (another popular method for high dimensional data) cannot separate X1, X2 and X3 although signals were given to one or two of the three variables. When a signal of interaction without main effects was given, LASSO can select the both involved variables by main effects when the interaction term was not in the model. When the interaction term was added in the model, LASSO detected the interaction and the main effects of the involved variables were gone. Based on this observation, we suggest testing the interactions only on the variables selected by LASSO, making it feasible to test interactions for many potential chemical exposures. We added the two-way interaction terms from all the variables selected above and re-ran LASSO. Based on 10-fold cross-validation, the optimal LASSO fit detected 4 interactions and gave the following estimated coefficients:

(Intercept)	23.725	X1X7	0
X1	1.988	X1z	4.265
X2	1.311	X2X4	0
X3	0	X2X5	0
X4	-0.405	X2X7	0
X5	-3.349	X2z	0
X6	0	X4X5	-0.019
X7	3.060	X4X7	0.028
Z	10.647	X4z	0
X1X2	0	X5X7	0
X1X4	0	X5z	0
X1X5	0	X7z	0.704

The LASSO p-value from covariance test using the LASSO solution path.

	LASSO p-value
Z	0.000
X1	0.000
X1z	0.000
X7	0.000
X5	0.000
X2	0.178
X7z	0.014
X4X5	0.469
X4X7	0.668
X4	0.000

Simulated Data Set #2:

x3, x4, x5, x8, x14 and z2 are highly correlated. x12 and x13 are highly correlated. Based on scatter plots, no nonlinear relationships between independent variables and y were detected. Based on 10-fold cross-validation, the optimal LASSO fit did not select x1, x2, x3, x5, x7, x13 and z1. The estimated coefficients are:

(Intercept)	3.600	x10	0.017
x1	0	x11	0.017
x2	0	x12	0.003
x3	0	x13	0
x4	0.033	x14	0.077
x5	0	z1	0
x6	0.038	z2	0.004
x7	0	z3	-0.502
x8	0.034		
x9	0.002		

Then we added the two-way interaction terms from all the variables selected above and re-ran LASSO. Based on 10-fold cross-validation, the optimal LASSO fit detected 6 interactions and gave the following estimated coefficients:

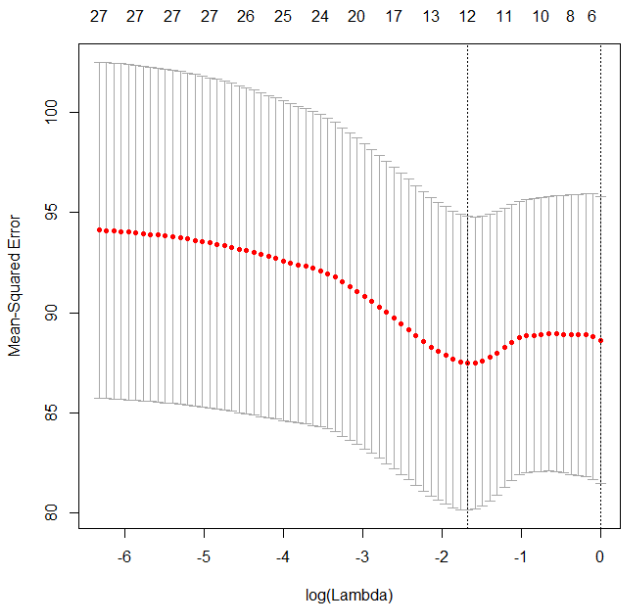
(Intercept)	3.783	x4z2	0
x1	0	x4z3	0
x2	0	x6x8	0.005
x3	0	x6x10	0
x4	0	x6x11	0
x5	0	x6x14	0
x6	0.0118	x6z2	0
x7	0	x6z3	0
x8	0	x8x10	0
x9	0	x8x11	0.002
x10	0	x8x14	0
x11	0	x8z2	0
x12	0	x8z3	0
x13	0	x10x11	0
x14	0.032	x10x14	0.001
z1	0	x10z2	0.001
z2	0	x10z3	0
z3	-0.494	x11x14	0.007
x4x6	0	x11z2	0
x4x8	0	x11z3	0
x4x10	0.014	x14z2	0
x4x11	0	x14z3	0
x4x14	0	xz2z3	0

The LASSO p-value from covariance test using the LASSO solution path (All p-values are rounded to 3 decimal places).

	LASSO p-value
z3	0.019
x4x10	0.000
x10x14	0.304
x10z	0.989
x8x14	0.830
x11x14	0.760
x6x8	0.537
x8	0.956
- x8x14	NA
x14	0.974
x8x11	0.965
-x8	NA
x6	0.707

Real World Data Set:

Log transformation was used for all the exposure variables (we retained the original variable names). lip_PBDE_47, lip_PBDE_99 and lip_PBDE_100 are highly correlated. lip_pcb153, lip_pcb156, lip_pcb170, lip_pcb180, lip_pcb187, lip_pcb194, lip_pcb199, lip_pcb138_158 and lip_pcb196_203 are highly correlated. The 5 covariates are retained in all models. 10-fold cross-validation found that lambda=0.187 gave the cross-validated minimum mean squared error. With this lambda, the LASSO fit selected 7 exposure variables. The estimated coefficients are:



(Intercept)	99.126	lip_pcb74	0
child_sex	-3.179	lip_pcb99	0
mom_educ	-6.136	lip_pcb105	0.824
mom_age	-0.530	lip_pcb118	0
mom_race	-4.671	lip_pcb146	0.985
mom_smoke	0.326	lip_pcb153	0
lip_PBDE_47	-0.331	lip_pcb156	0
lip_PBDE_99	-0.550	lip_pcb170	0
lip_PBDE_100	0	lip_pcb180	0
lip_PBDE_153	0	lip_pcb187	-1.842
lip_hcb	0	lip_pcb194	1.552
lip_pp_dde	0	lip_pcb199	0
lip_oxychlor	-0.131	lip_pcb138_158	0
lip_nonachlor	0	lip_pcb196_203	0

LASSO is based on the size of coefficients to select variables. We may want to retain the most common exposure such as lip_PBDE_47 and lip_PBDE_153 in all models. Re-running LASSO. The estimated coefficients are:

(Intercept)	100.865	lip_pcb74	0
child_sex	-3.238	lip_pcb99	0
mom_educ	-6.197	lip_pcb105	0.894
mom_age	-0.648	lip_pcb118	0
mom_race	-4.599	lip_pcb146	0.973
mom_smoke	0.380	lip_pcb153	0
lip_PBDE_47	-1.216	lip_pcb156	0
lip_PBDE_99	0	lip_pcb170	0
lip_PBDE_100	0	lip_pcb180	0
lip_PBDE_153	0.126	lip_pcb187	-1.862
lip_hcb	0	lip_pcb194	1.549
lip_pp_dde	0	lip_pcb199	0
lip_oxychlor	-0.158	lip_pcb138_158	0
lip_nonachlor	0	lip_pcb196_203	0

However, the LASSO p-values for all the exposure variables are >0.05.

Conclusions and Discussion: We analyzed the two simulated data sets and a real world data set and showed LASSO can be a promising technique to handle multicollinearity and high dimensional data of chemical exposures. Generically, the usual p-values and confidence intervals do not exist for LASSO estimates. Many methods have been proposed using resampling and data splitting (Bühlmann et al., 2011). Recently, a significance test for the predictor variable that enters the current LASSO model along the LASSO solution path has been proposed by Tibshirani's group (Lockhart et al., 2014). This test has a simple and has an exact asymptotic null distribution without the need of resampling or data splitting. However, the p-value from this test cannot be interpreted in the classic sense based on a fixed null hypothesis since the test is a conditional test.

References:

Lockhart, R, Taylor, J., Tibshirani R. J. and Tibshirani, R. (2014). A significance test for the LASSO. *The Annals of Statistics* 42:413-468.

Bühlmann, P. and Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B* 58:267-288.

29. Assessing the Impact of Environmental Mixtures on Children's Neurodevelopment

Presenting Author: Rengyi (Emily) Xu

Organization: University of Pennsylvania, Perelman School of Medicine

Contributing Authors: Rengyi Xu[1], Michelle Ross[1], Mingyao Li[1], Andrew J. Worth[2], and Wei-Ting Hwang[1]
[1]Department of Biostatistics and Epidemiology, [2]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

Abstract:

Objective: To determine which exposure or combination of exposures in the mixture is associated with children's neurodevelopment.

Methods: Exploratory data analysis (EDA) was first performed including descriptive statistics and graphs to examine the distributions of the outcome, children's Mental Development Index (MDI), and individual exposures and relationships between them by computing Pearson's correlation coefficients. Normality of the outcome variable was tested using the Shapiro-Wilk test [1].

We implemented the following two data-driven approaches to analyze the data: (1) stepwise multiple linear regression (MLR) [2], and (2) principle component analysis (PCA) [3]. In Approach #1, a stepwise regression was first performed including only those exposures that were statistically significant at the 0.1 level from univariate linear regression models and the 5 confounders as main effects. Next, a second stepwise regression was performed to examine all possible two-way interactions using only those variables selected in the first step. The final model is the one with the set of variables that results in the smallest Akaike information criterion (AIC) [2,4]. In Approach #2, we first addressed the issue of possible collinearity by applying PCA to the following three groups of exposure variables (i.e., 14 polychlorinated biphenyl (PCB) congeners, 4 polybrominated diphenylether (PBDE) congeners, and 4 organochlorine pesticides) separately.

. We then subjected the top 2 or 3 principle components (PC) to the stepwise variable-selection process using multiple linear regressions. For a given fitted model, model fit statistics including R-squared and root mean squared error (RMSE) were computed. The level of collinearity was assessed by the Variance Inflation Factor (VIF) [2] with VIF>10 indicating the presence of collinearity. Standard regression diagnostic methods [2] were implemented to identify possible violations of model assumptions including non-normality and non-linear effects. Standardized coefficients were computed to compare the relative influence of different explanatory variables. A Box-Cox transformation was also explored to improve the normality of MDI. [5].

Results: The outcome MDI had a mean of 91.6, median of 92, and standard deviation (SD) of 10.3, ranging from 50 to 114. The distributions of chemical exposure variables are different. A scatter-matrix plot of MDI and the exposure variables (Figure 1-3) suggests that MDI has a non-linear relationship with oxychlor, and a positive association with PCB156, PCB196, and PCB199. We also observed a positive

association between many of the exposures in the same category. Figure 4 shows that the distribution of MDI is associated with mother's smoking status, race and education, indicating that these variables should be considered as potential confounders and/or effect modifiers. The histogram of MDI in Figure 4 suggests that its distribution is approximately normal. However, the Shapiro-Wilk test indicated otherwise ($p<0.05$). We explored a Box-Cox transformation of MDI, which suggested a square root transformation would be appropriate. However, given the difficulty in interpreting the model, we decided to use the original scale in our analysis.

Results are summarized in Table 1 and 2. Under Approach #1, PCB74, PCB105, PCB118, PCB194, PCB196_203, PCB199 and PBDE100 were identified as having a p-value less than 0.1 in the univariate regressions. Together with the 5 confounders, the final model included PCB118, PBDE100, mother's education, mother's race, child's gender as main effects, and an interaction between mother's education and PCB118. Using the standardized coefficients, the sizes of the associations with MDI in order of magnitude were: mother's education (std. β=-6.12), mother's race (std. β=-4.93), child's gender (std. β=-3.55), the interaction between mother's education and PCB118 (std. β=3.43), PBDE100 (std. β=-0.90), and PCB118 (std. β=0.72). Diagnostic plots of the residuals suggested no obvious departures from normality or homoscedasticity. Subject 357 (noted as 243 in Figure 5) seems to be an outlier but not an influential point (Figure 5). The final model had a VIF of 1.30, suggesting no excess collinearity among selected predictors. To use the model as a risk assessment tool, we can calculate fitted values and rank them into low, intermediate, and high risk groups, or compute predicted values and prediction intervals (PI) for a new set of data. To help visualize the multi-dimensional nature of the model, the fitted surface of MDI stratified by mother's education as a function of the two exposures PCB118 and PBDE100 while fixing the remaining covariates at their means is presented in Figure 6.

In Approach #2, the first three PCs in the PCB group identified explained 91.8% of the total variance, the first two PCs in the PBDE group explained 96.8% of the total variance and the first two PCs in the organochlorine pesticides group explained 64.2% of the total variance (Figure 7). For PCB, the first PC (PC1) loaded on PCB146, PCB153 and PCB138_158 equally; the second PC (PC2) loaded on PCB99, PCB105 and PCB118 with loadings of 0.40, 0.52 and 0.46 respectively, and the third PC (PC3) loaded on PCB199 and PCB196 with loadings of 0.53 and 0.43 (Figure 8). For PBDE, PC1 loaded on PBDE47 and PBDE100 equally, and the PC2 loaded solely on PBDE153 (Figure 9). PC1 in the pesticides group loaded on oxychlor and nonachlor equally and PC2 loaded on HCB and DDE with loadings of 0.50 and 0.76. (Figure 10) After the forward selection analysis that included the 7 PCs and their two-way interactions (Table 2), the variable that had the strongest association with MDI was mother's education (std. β=-9.13). The only exposure variable identified was PC2 from the PCB group, along with an interaction with mother's education. The AIC, VIF, RMSE and R-squared in this model were very similar with the model in Approach #1.

Conclusions: Appropriate statistical analysis for assessing the exposure-outcome association in a mixture requires multiple steps to explore and model the relationships between variables. There are many statistical models that can be applied. Taking multiple approaches and finding similar results bolster findings. Any proposed models should be validated using independent datasets and/or laboratory experiments. Including knowledge from the scientific content area is also essential in guiding

the selection of statistical methods and can further strengthen the validity and interpretability of the statistical analysis results.

Funding source: NIH grants P30ES013508 and P42ES023720

References:

1. Shapiro, S. S., Wilk, M. B. (1965) An analysis of variance test for normality (complete samples). *Biometrika* 52 (3–4): 591–611.

2. Weisberg, S. (2005) Applied Linear Regression, third edition, Hoboken NJ: John Wiley.

3. Hotelling, H. (1993) Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24: 417-441,498-520.

5. Akaike, H. (1974) A new look at the statistical model identification, *IEEE Transactions on Automatic Control* 19(6):716-723.

6. Box, GEP, Cox, DR. (1964) An analysis of transformations. *Journal of the Royal Statistical Society* 26(2): 211-252.

(Figures and Tables are listed below)

Figure 1: Scatter matrix plot of outcome MDI and PCB exposures variables.

Figure 2: Scatter matrix plot of outcome MDI and PBDE exposures variables.

Figure 3: Scatter matrix plot of outcome MDI and organochlorine pesticides exposures variables.

Figure 4: Histogram of MDI and boxplots of MDI by the confounders.

Figure 5: Residual and other diagnostic plots based on the stepwise regression model.

Figure 6: Fitted surface plot of MDI by exposures (PCB118 and PBDE100) stratified by mother's education while adjusting for the remaining covariates in the final model.

Figure 7: Selection of the top PC based on the scree plot.

Figure 8: Factor loading plot of the top 3 PCs in PCB group.

Figure 9: Factor loading plot of the top 3 PCs in PBDE group.

Figure 10: Factor loading plot of the top 3 PCs in organochlorine pesticides group.

Table 1: Regression analysis results from Approach 1.

Table 1: Regression analysis results from Approach 2.

Figure 1: Scatter matrix plot of outcome MDI and PCB exposures variables.

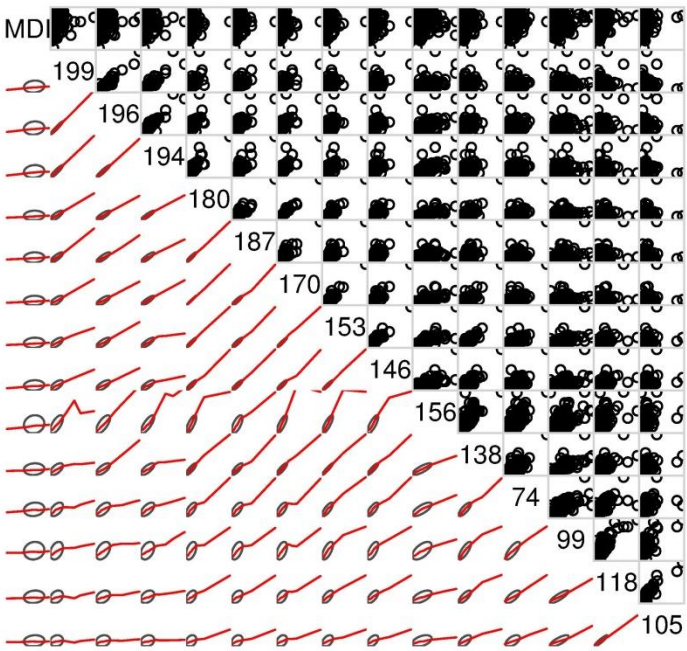


Figure 2: Scatter matrix plot of outcome MDI and PBDE exposures variables.

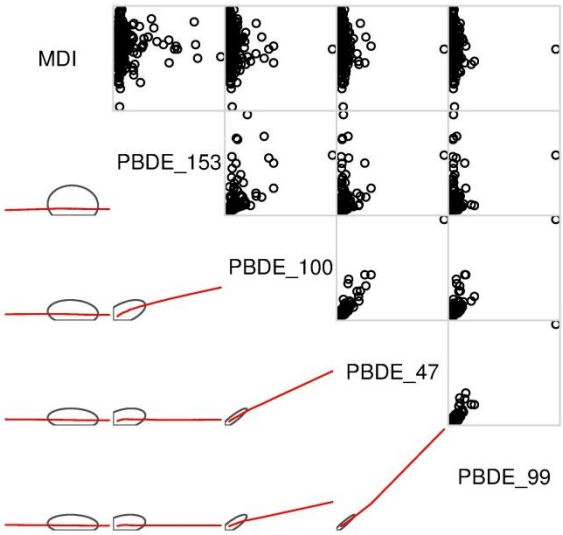


Figure 3: Scatter matrix plot of outcome MDI and organochlorine pesticides exposures variables.

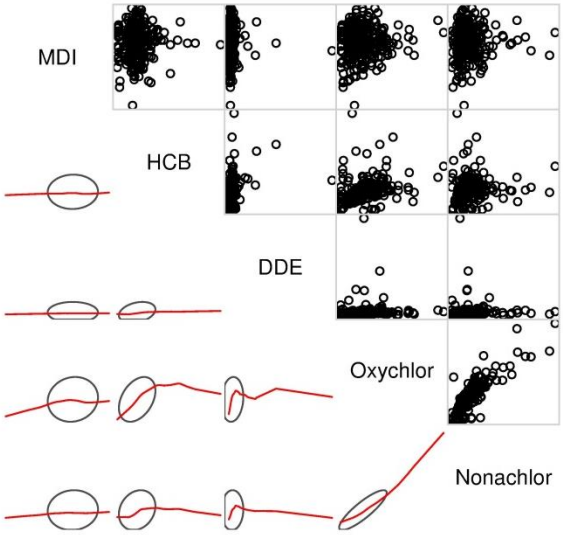


Figure 4: Histogram of MDI and boxplots of MDI by the confounders.

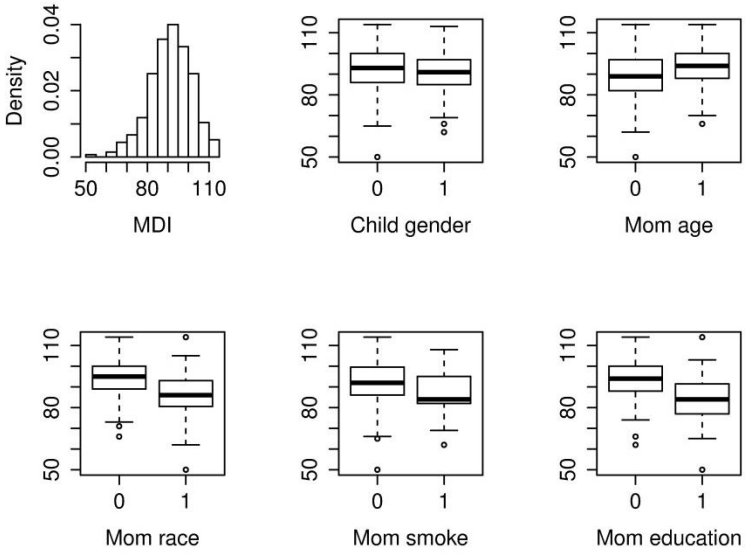


Figure 5: Residual and other diagnostic plots based on the stepwise regression model.

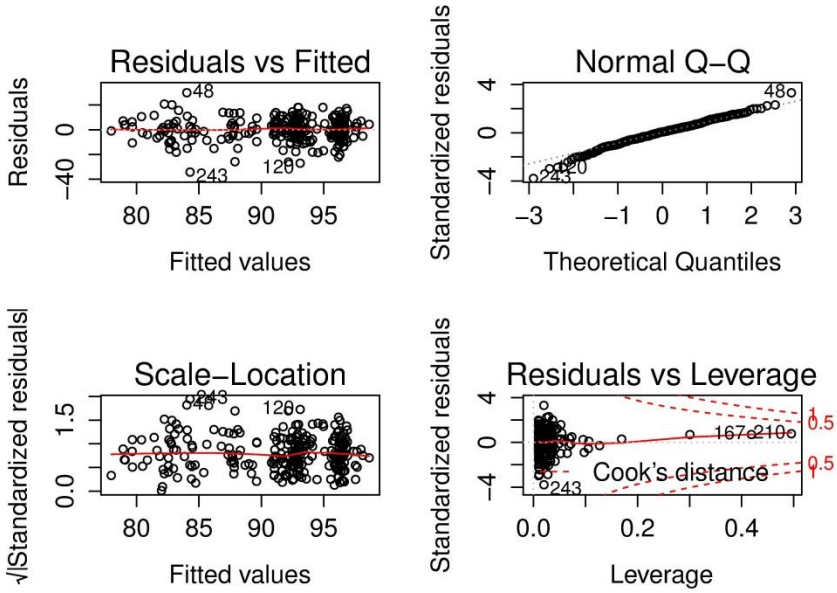
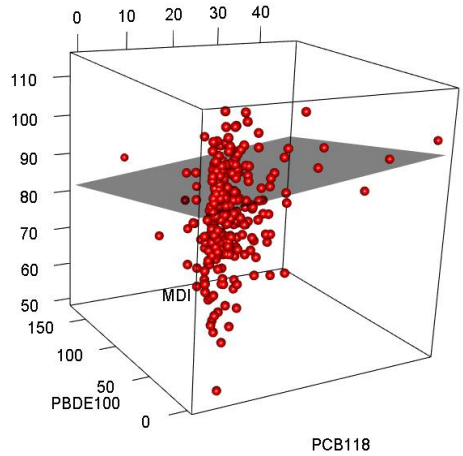


Figure 6: Fitted surface plot of MDI by exposures (PCB118 and PBDE100) stratified by mother's education while adjusting for the remaining covariates in the final model.

Mother's education>12 years



Mother's education≤12 years

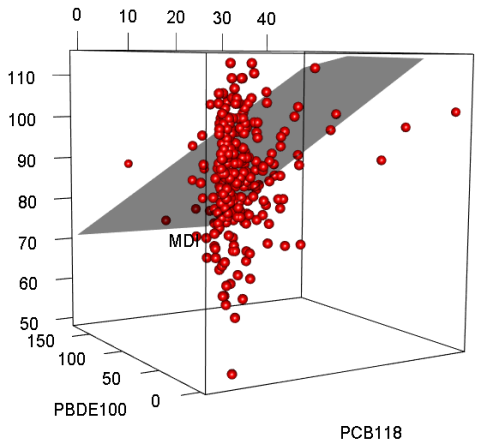


Figure 7: Selection of the top PC based on the scree plot.

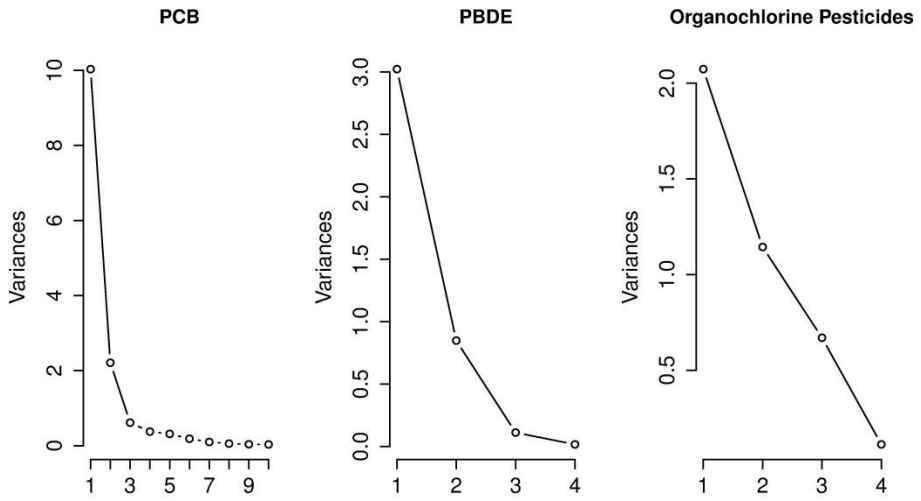


Figure 8: Factor loading plot of the top 3 PCs in PCB group.

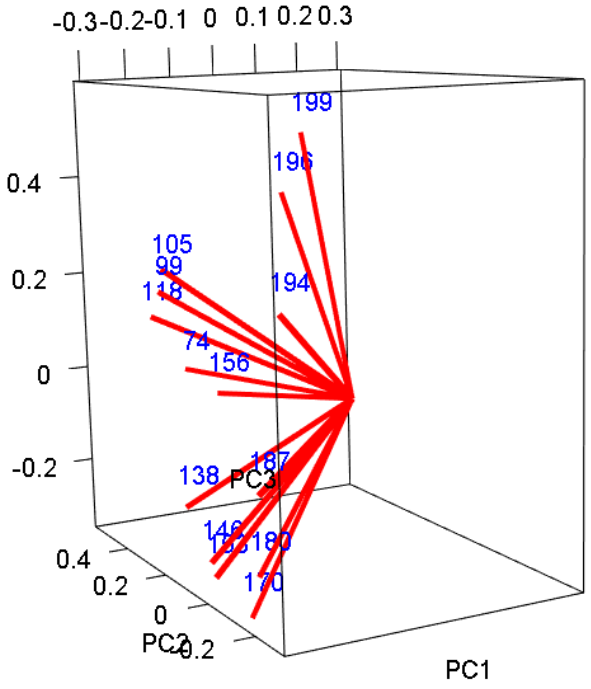


Figure 9: Factor loading plot of the top 3 PCs in PBDE group.

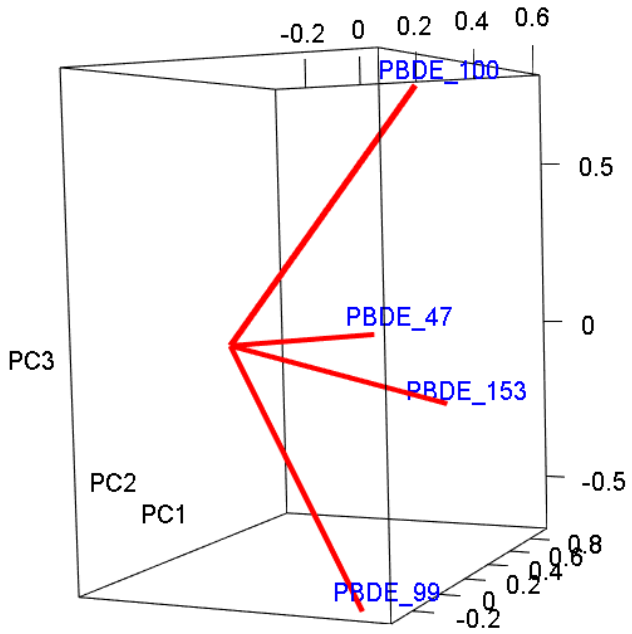


Figure 10: Factor loading plot of the top 3 PCs in organochlorine pesticides group.

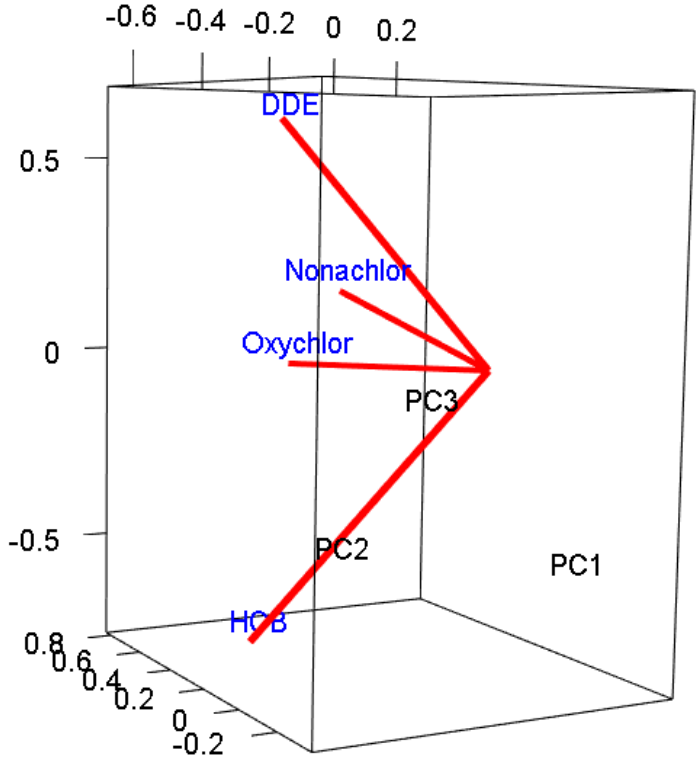


Table 1: Regression analysis results from Approach 1.

Variable	β (95% CI)	Std. β (95%CI)
PCB118	0.15 (-0.09, 0.38)	0.72 (-0.45, 1.90)
PBDE100	-0.06 (-0.14, 0.01)	-0.90 (-2.00, 0.20)
Mom Educ	-10.32*** (-15.37, -5.27)	-6.12*** (-9.18, -3.06)
Mom Race	-4.93*** (-7.67, -2.18)	-4.93***(-7.67, -2.18)
Child Sex	-3.55**(-5.81, -1.28)	-3.55**(-5.81, -1.28)
Interaction		
PCB118*Mom Educ	0.69 (-0.04, 1.41)	3.43 (-0.19, 7.04)
VIF	1.30	
R^2	0.23	
RMSE	9.14	
AIC	1970.02	

*<0.05, ** <0.01, ***<0.001

Table 2: Regression analysis results from Approach 2.

Variable	β (95% CI)	Std. β (95%CI)
PCB PC2	0.70 (-0.49, 1.88)	0.70 (-0.49, 1.88)
Mom Educ	-6.06*** (-9.13, -3.00)	-6.06*** (-9.13, -3.00)
Mom Race	-5.20*** (-7.94, -2.47)	-5.20*** (-7.94, -2.47)
Child Sex	-3.56**(-5.83, -1.29)	-3.56**(-5.83, -1.29)
Interaction		
PCB PC2*Mom Educ	3.33 (-0.18, 6.85)	3.33 (-0.18, 6.85)
VIF	1.29	
R^2	0.22	
RMSE	9.17	
AIC	1970.67	

*<0.05, ** <0.01, ***<0.001. PCB PC2 loads on PCB99, PCB105 and PCB118 with loadings of 0.40, 0.52 and 0.46.

30. Analysis of the First Simulated Dataset using Nonlinear and Weighted Quantile Sum (WQS) Regression

Presenting Author: Chris Gennings

Organization: Icahn School of Medicine at Mount Sinai

Contributing Authors: Chris Gennings

Abstract:

Preliminary analyses of the first simulated dataset indicated nonlinear marginal associations between exposure and Y with different maximum effects defined by the binary variable Z (See Figure 1). Thus, analyses either controlled for Z or were stratified by Z.

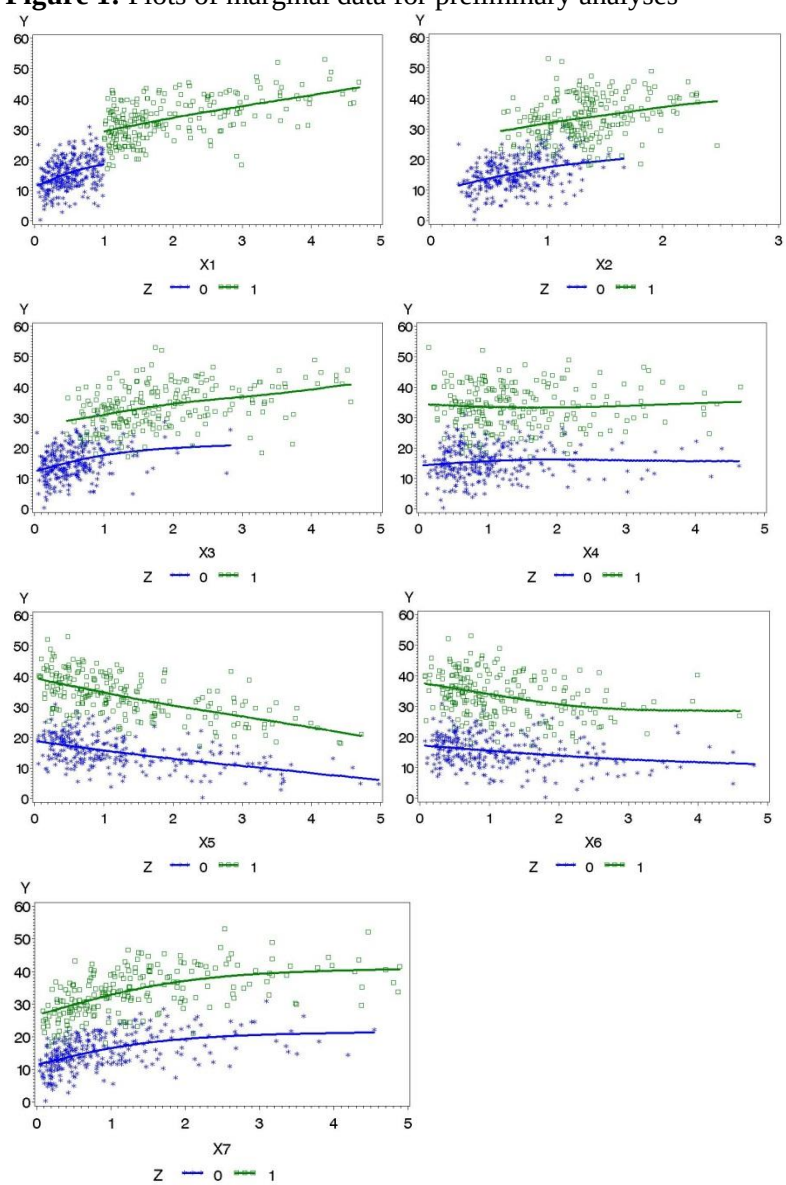
The bivariate correlation pattern among the X variables included a cluster among X1, X2 and X3 (correlation estimates ranging between 0.7 and 0.82); and between X5 and X6 (correlation estimate of 0.6); other correlations were less than 0.5.

Preliminary nonlinear regression analyses were based on a nonlinear exponential model: i.e.,

$$\mu = \alpha + \gamma\left[1 - \exp(-\mathbf{X\beta})\right],$$

parameterized to allow a different maximum effect for different levels of Z (i.e., $\gamma = \gamma_0(Z=0) + \gamma_1(Z=1)$). The model included linear and all pairwise cross-product terms. The resulting analysis indicated evidence of bivariate interaction between: X1 and X3 (negative); X1 and X5 (negative coefficient); X1 and X7 (positive coefficient); and X5 and X7. These interactions are associated with a generalized definition of dose

Figure 1: Plots of marginal data for preliminary analyses



addition where an interaction is associated with a change in slope.

Due to the complex correlation pattern among the X's, we evaluated the joint mixture effect using WQS regression embedded within nonlinear exponential functions.

Nonlinear WQS Regression: Extending the work of Carrico et al (2014), we embedded the WQS regression model in a nonlinear exponential model and allowed for specific interactions as suggested from preliminary analyses. For an increasing association (where β_1 >0):

$$\mu = \gamma\left[1 - \exp(-\left(\beta_0 + \beta_1(\sum_{j=1}^{c} w_j x_j + \sum_{(r,s\in S)} w_{rs} x_r x_s) + \theta z\right))\right]$$

where *S* is the set of interaction terms considered. For a decreasing association (where β_1 <0):

$$\mu = \alpha + \gamma\left[\exp\left(\beta_1(\sum_{j=1}^{c} w_j x_j + \sum_{(r,s\in S)} w_{rs} x_r x_s) + \theta z\right)\right].$$

In either case, the WQS index was defined as $WQS = \sum_{j=1}^{c} \bar{w}_j x_j + \sum_{(r,s\in S)} \bar{w}_{rs} x_r x_s$ where the weights are constrained to sum to 1 and are weighted averages across 100 bootstrap analyses. When the corresponding β_1 is positive and significant, the corresponding weights may be interpreted as associated with a positive mixture effect; when the corresponding β_1 is negative and significant, the corresponding weights may be interpreted as associated with a negative mixture effect. Analyses were conducted in both directions by constraining the beta coefficient to be either positive of negative. Simplified analyses were conducted with only linear terms for comparison and ease of interpretation.

Results: Analyses were conducted in the positive and negative directions (Table 1) by constraining the beta coefficient associated with the WQS indices. In the positive direction, X1 (weight of 59%) and X7 (35%) dominate the index with a positive and significant regression coefficient. Allowing for interaction as indicated from preliminary analyses, there appears to a synergy between X1 and X7 (weight of 21%). There is also an indication that X5

Table 1: Results from nonlinear WQS regression using all the data and adjusting for Z

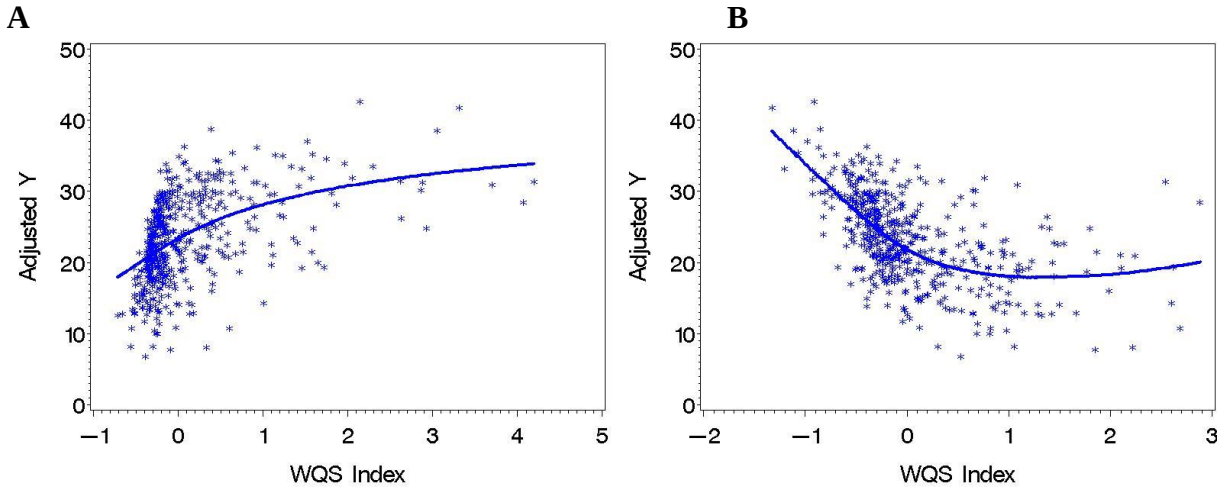
	Overall (N=500); adjusted by Z			
Constraint:	**Beta1>0**		**Beta1<0**	
Beta1	0.14	1.1	-0.10	-0.10
95% CI	0.05, 0.23	0.94, 1.3	-0.15, -0.04	-0.18, -0.03
Linear Weights				
W1	0.59	0.33	0	0
W2	<0.01	0.02	0	0
W3	0.06	0.03	0	0
W4	<0.01	0	0.07	0.06
W5	0	0	0.90	0.69
W6	0	0	0.04	0.03
W7	0.35	0.26	0	<0.01
Interactions				
W13		0.06		<0.01
W15		0.09		0.08
W17		0.21		<0.01
W57		0.01		0.14

increases the effect of X1 (and vice versa).

In the negative direction, X5 dominates the association (weight of 90%). Allowing for interaction as indicated from preliminary analyses, there is some evidence of a synergy between X1 and X5(weight of 8%) and between X5 and X7 (weight of 14%).

An advantage of WQS regression is the ease of presentation of the association between the estimated weighted index and the adjusted response variable (defined by the residuals from the covariate only model plus the mean of y). The data in Figure 2 are LOESS fits from the indices with interaction terms defined in Table 1. Clearly, there is a nonlinear association between the indices and the adjusted response variable.

Figure 2: Association between the estimated weighted indices (A) with positive association; and (B) negative association using the weights with interaction components from Table 1.



Sensitivity analyses: A sensitivity analysis was conducted in stratified analyses of nonlinear WQS regression. Due to the reduced sample sizes, we consider only linear weights. When Z=0, X1 and X7 accounted for 81% of the weight, with some indication of an effect due to X2 and X3 (15% of the weight). Although not significant, in the negative direction, X5 dominated the association. When Z=1, X1

Table 2: Results from sensitivity analyses.

	Z=0 (N=286)		**Z=1 (N=214)**	
Constraint:	**Beta1>0**	**Beta1<0**	**Beta1>0**	**Beta1<0**
Beta1	0.95	-0.27	0.64	-0.22
95% CI	0.49,1.4	-0.67, 0.14	0.27, 1.0	-0.60, 0.15
Linear Weights				
W1	0.12	0	0.41	0
W2	0.07	0	<0.01	0
W3	0.08	0	0.02	0
W4	0.04	0.06	<0.01	0.09
W5	0	0.94	0	0.89
W6	0	0	0	0.02
W7	0.69	0	0.56	0

and X7 accounted for 97% if the weight in the positive direction. Again, although not significant, in the negative direction, X5 accounted for 89% of the weight. These results are somewhat similar to those in the overall analyses (Table 1).

Reference:

Carrico C, Gennings C, Wheeler DC, Factor-Litvak P (2014) Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. Journal of Agricultural, Biological and Environmental Statistics, [epub: Dec 24, 2014].

31. Bayesian Methods for Assessing Health Effects of Chemical Mixtures

Presenting Author: David Dunson

Organization: Duke University

Contributing Authors: David Dunson and Amy Herring

Abstract:

In assessing the joint effects of multiple components of a chemical mixture on human health, several statistical difficulties arise: the possibility of synergy or antagonism among chemicals, collinearity in different exposures and inefficiency in estimating non-linear dose response surfaces. We propose Bayesian semiparametric methods for addressing these issues, adaptively reducing dimensionality in estimating the dose response "surface" in multiple chemicals, while enabling detection of additive, synergistic, or antagonistic effects among groups of chemicals. By using a flexible Bayesian approach, we avoid making restrictive assumptions while allowing for the incorporation of prior information on dose response shape, chemical class, and information from previous studies. The methods are applied to several epidemiology applications.

32. Assessing Health Effects of Environmental Chemical Mixtures Using Stepwise Multiple Linear Regression

Presenting Author: James Nguyen

Organization: U.S. Environmental Protection Agency, Office of Pesticide Programs

Contributing Authors: James Nguyen

Abstract:

While there are numerous advanced statistical approaches that could be used to investigate the effect of an exposure to environmental chemical mixtures on an outcome, these can be difficult to implement and interpret, and application of those methods has been a high barrier for individuals who may not have sufficient familiarity with those methods to use them. Also, there is a limitation in using the results from those analyses due to the fact that the chemicals or chemical groups are sometimes regulated individually. For this analysis, stepwise multiple linear regression has been selected to evaluate and study the effects of mixture of exposures on the outcome Y in Dataset 1 and Dataset 2. First, univariate analyses are performed to select a list of candidate confounders whose p-value < 0.15. Then, a stepwise procedure is done using the list of candidate confounders to select a semi-final model (model 1) that includes only confounders. Next, analyses with models including the confounders in model 1 and one of the exposures are performed to select exposures whose p-values < 0.15 as candidate exposures for the next stepwise procedure. Finally, the stepwise procedure is conducted to select final model, using model 1 as the initial model. At any step during the stepwise procedures, a confounder or exposure remains in the model if the resulting model has a lower Akaike Information Criterion (AIC) value. Interactions of two variables are included to the list of candidates for selection only if they are already selected in the model. A variable is removed from the model if its p-value > 0.15 and the resulting model has a lower AIC value.

In the analysis of dataset 1, we found the combination of exposures X1, X2, X5, and X7 describes the variation of Y well (r-squared = 0.919). In addition to our finding that the effect of X5 on Y is systematically different depending on the status Z of the subject (i.e., there is an interaction between X5 and Z), there were interactions between X1 and X2 and between X5 and X7.

In the analysis of dataset 2, we found that exposures X1, X5, X6, X10, X12, and X14 contribute to the outcome. Exposures X3, X4, X8, and X11 are strongly correlated to X5, X6, and X14, so their contributions to Y are potentially expressed by X5, X6 and X14. X13 is strongly correlated to X12, so its contribution to Y is potentially expressed by X12. Exposures X2, X7, and X9 do not contribute to the Y. There are no interactions between the exposures; however, the effects of X5, X6 and X12 on Y depend on the Z3 status of the subject; the effect of X10 depends on the status Z2; and the effect of X14 depends on the statuses of both Z2 and Z3. It appears that the selected model could explain more than 50% of the variation in Y (r-squared = 0.564).

Results:

Dataset 1: Exposures X1, X2, X5, and X7 contribute to the outcome. Since X3 is highly correlated to X1 and X2 and X6 are moderately correlated to X5, their contribution to Y is adequately expressed and accounted for by X1, X2, and X5 in the final model. X4 does not contribute to the variation of Y. Interaction occurs when the effect of an exposure on the outcome Y depends on the value(s) of other exposure(s). There is evidence of interactions between exposures X1 and X2 and between exposures X5 and X7. The selected model performs well in characterizing the variation of Y, with r-squared value = 0.919 and RMSE = 3.118.

Dataset 2: Exposures X1, X5, X6, X10, X12, and X14 contribute to the outcome. Exposures X3, X4, X8, and X11 are strongly correlated to X5, X6, and X14, so their contribution to Y is sufficiently expressed by X5, X6 and X14 in the final model. Exposure X13 is strongly correlated to X12, so its contribution to Y is potentially expressed by X12 in the final model. Exposures X2, X7, and X9 do not contribution to the variation of Y. There is no evidence of interaction between the exposures. However, there are interactions between confounder Z2 with X10 and X14 and also between Z3 with X5, X6, X12, and X14. The selected model explains more than 50% of the variation of Y, with r-squared value = 0.564 and RMSE = 0.437.

Table 1 and Figure 1 (for Dataset 1) and Table 2 and Figure 2 (for Dataset 2) present the final model parameter estimates and the Residual vs. Predicted Plots (rvp plots) for each of the two datasets. As can be seen, the rvp plots suggest that Z in Dataset 1 and Z3 in Dataset 2 are important predictors and that the distributions of the modeled residuals are reasonable. Overall, the model r-squared values and the diagnostic plots seem to imply a reasonably-specified baseline model to which it would be appropriate to compare to models developed using more advanced statistical techniques.

33. Traditional Epidemiological Approaches to Analyze Chemical Mixtures and Human Health

Presenting Author: Joseph M. Braun

Organization: Brown University

Contributing Authors: Joseph M. Braun

Abstract:

Introduction: Polychlorinated biphenyls (PCBs), polybrominated diphenyl ethers (PBDEs), and organochlorine pesticides are three classes of persistent environmental chemicals detected almost universally in human serum. Prenatal exposure to these chemicals may be associated with decreased cognitive abilities and behavior problems in childhood.

Objective: I used a traditional epidemiological approach to determine if prenatal exposure to individual PCBs, PBDEs, or organochlorine pesticides was associated with child cognitive development in a prospective birth cohort of 270 women and their children.

Methods: The dataset for this analysis includes concentrations of 14 PCBs, 4 PBDEs, and 4 organochlorine pesticides measured in maternal serum collected during the 2nd trimester of pregnancy. Children's cognitive development was assessed at 1-3 years of age using the Mental Development Index (MDI) of the Bayley Scales for Infant Development-II. We adjusted for child sex, maternal age, race, education, and smoking during pregnancy.

I started the analysis by describing the correlation between individual $\log_{10}$-transformed chemical concentrations. Then I employed two traditional epidemiological approaches to analyze these data. First, I examined the association between each chemical and MDI scores using 22 individual covariate-adjusted linear regression models. Second, I examined the relationship between MDI scores and two simple cumulative measures of PCB and PBDE exposure that I created by summing the concentrations of individual congeners (ΣPCB and ΣPBDE).

Results: The correlation between individual chemicals was higher for chemicals within a class than between classes (Figure 1). For example, the Pearson correlation coefficient between individual PCB congeners ranged from 0.03 to 0.82 (median=0.53), whereas the correlation between PBCs and PBDEs or organochlorine pesticides ranged from -0.24 to 0.66 (median=0.18). Unadjusted associations between serum PCBs, PBDEs, or organochlorine concentrations were biased away from the null and became smaller in magnitude after adjustment for covariates (Figure 2). After covariate adjustment, each 10-fold increase in three of the PBDEs (PBDEs 47, 99, and 100) was associated with a 1.9 to 2.4 point decrease in MDI scores. In contrast, each 10-fold increase in seven of the PCBs (PCBs 105, 146, 153, 156, 170, 138/158, and 196/203) was associated with a 2.1 to 3.9 point increase in MDI scores. Concentrations of three organochlorine pesticides was associated with <2 point increase in MDI scores, while each 10-fold increase in p'p'-dichlorodiphenyldichloroethylene (DDE) concentrations was associated with a 2.0 point increase in MDI scores. The two simple cumulative measures of PCB and PBDE exposure provided similar

results to the analysis of the individual congeners, where the ΣPCB concentrations were positively associated with MDI scores and ΣPBDE concentrations were inversely associated with MDI scores (Figure 4, 1st panel). Joint adjustment for ΣPCB, ΣPBDE, and DDE in the same model revealed that the positive association between DDE and MDI scores was biased in unadjusted models. This was because of positive confounding from the positive correlation between ΣPCBs and DDE (Pearson r=0.6) and the positive PCB-MDI association (Figure 2, 2nd panel). The associations between ΣPCB and ΣPBDE concentrations and MDI scores did not change appreciably in magnitude or precision when all three chemicals were included in the same model.

Conclusions: In these data, we show that prenatal serum PCB and PBDE concentrations are associated with subtle increases and decreases in early childhood cognitive development, respectively. These results provide a basis to compare the findings from more complex statistical approaches that consider these chemicals as a mixture. Future studies of chemical exposures and human health should consider biases that arise from co-pollutant confounding given the correlated nature of many environmental exposures.

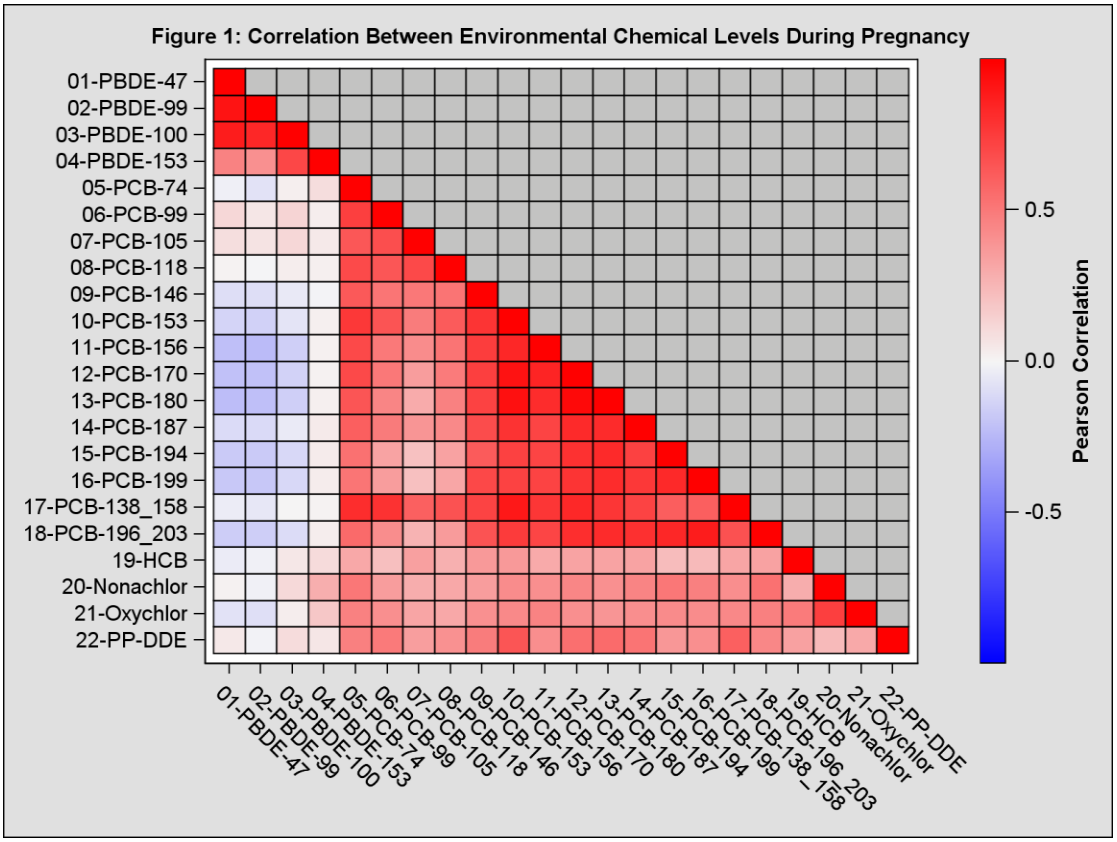


Figure 1: Correlation Between Environmental Chemical Levels During Pregnancy
01-PBDE-47
02-PBDE-99
03-PBDE-100
04-PBDE-153
05-PCB-74
06-PCB-99
07-PCB-105
08-PCB-118
09-PCB-146
10-PCB-153
11-PCB-156
12-PCB-170
13-PCB-180
14-PCB-187
15-PCB-194
16-PCB-199
17-PCB-138_158
18-PCB-196_203
19-HCB
20-Nonachlor
21-Oxychlor
22-PP-DDE
0.5
0.0
-0.5
Pearson Correlation

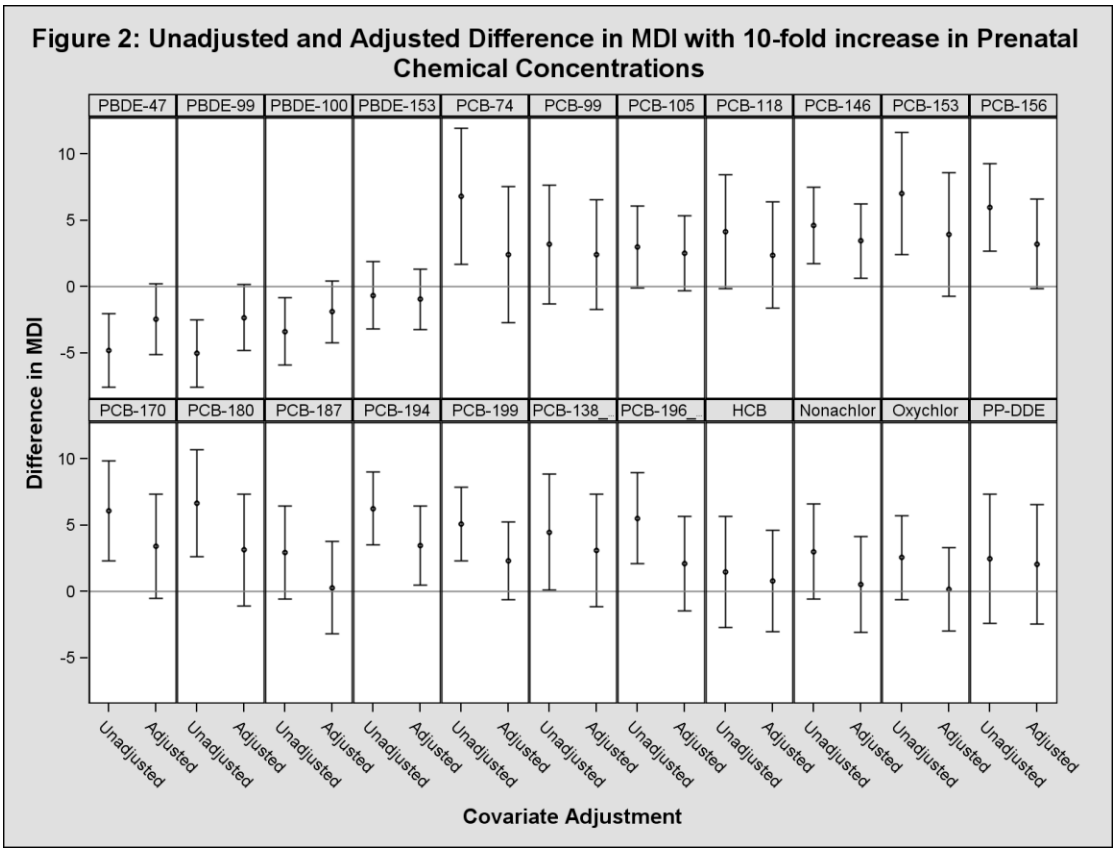


Figure 2: Unadjusted and Adjusted Difference in MDI with 10-fold increase in Prenatal Chemical Concentrations
PBDE-47
PBDE-99
PBDE-100
PBDE-153
PCB-74
PCB-99
PCB-105
PCB-118
PCB-146
PCB-153
PCB-156
PCB-170
PCB-180
PCB-187
PCB-194
PCB-199
PCB-138_
PCB-196_
HCB
Nonachlor
Oxychlor
PP-DDE
Difference in MDI
10
5
0
-5
Unadjusted
Adjusted
Covariate Adjustment

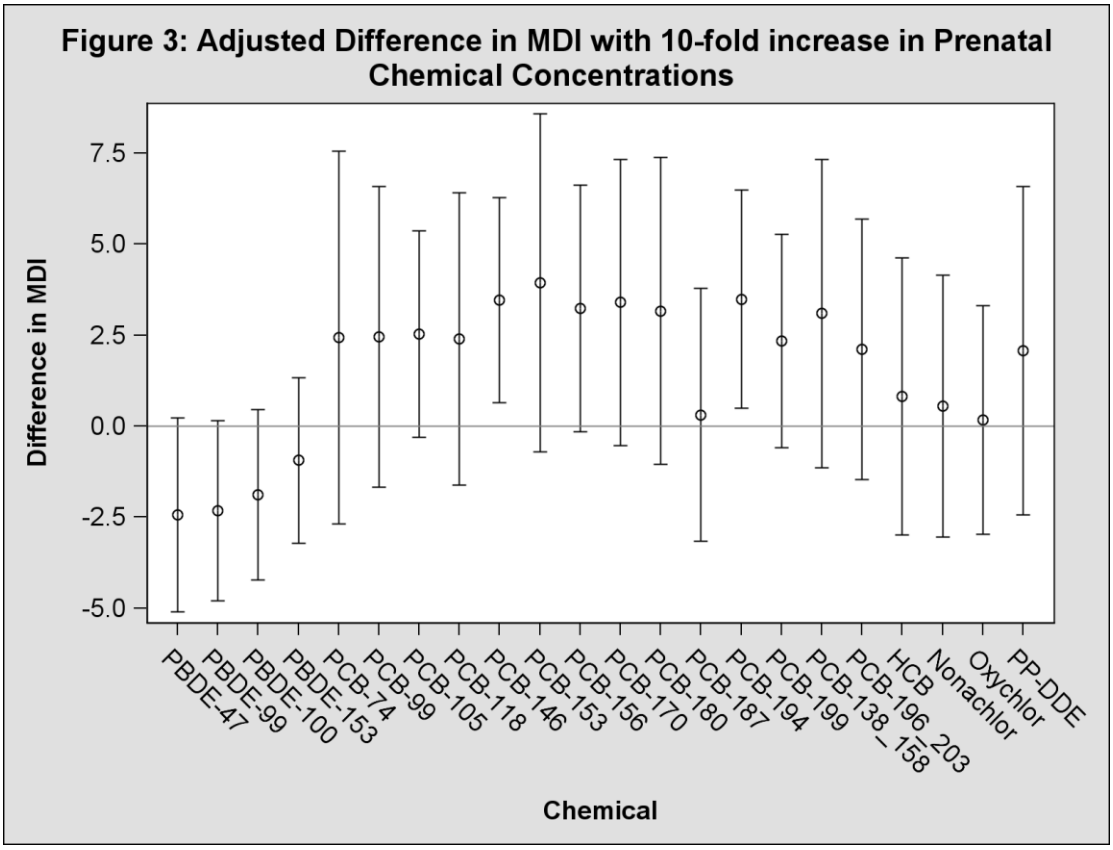


Figure 3: Adjusted Difference in MDI with 10-fold increase in Prenatal Chemical Concentrations
Difference in MDI
7.5
5.0
2.5
0.0
-2.5
-5.0
PBDE-47
PBDE-99
PBDE-100
PBDE-153
PCB-74
PCB-99
PCB-105
PCB-118
PCB-146
PCB-153
PCB-156
PCB-170
PCB-180
PCB-187
PCB-194
PCB-199
PCB-138 _ 158
PCB-196 _ 203
HCB
Nonachlor
Oxychlor
PP-DDE
Chemical

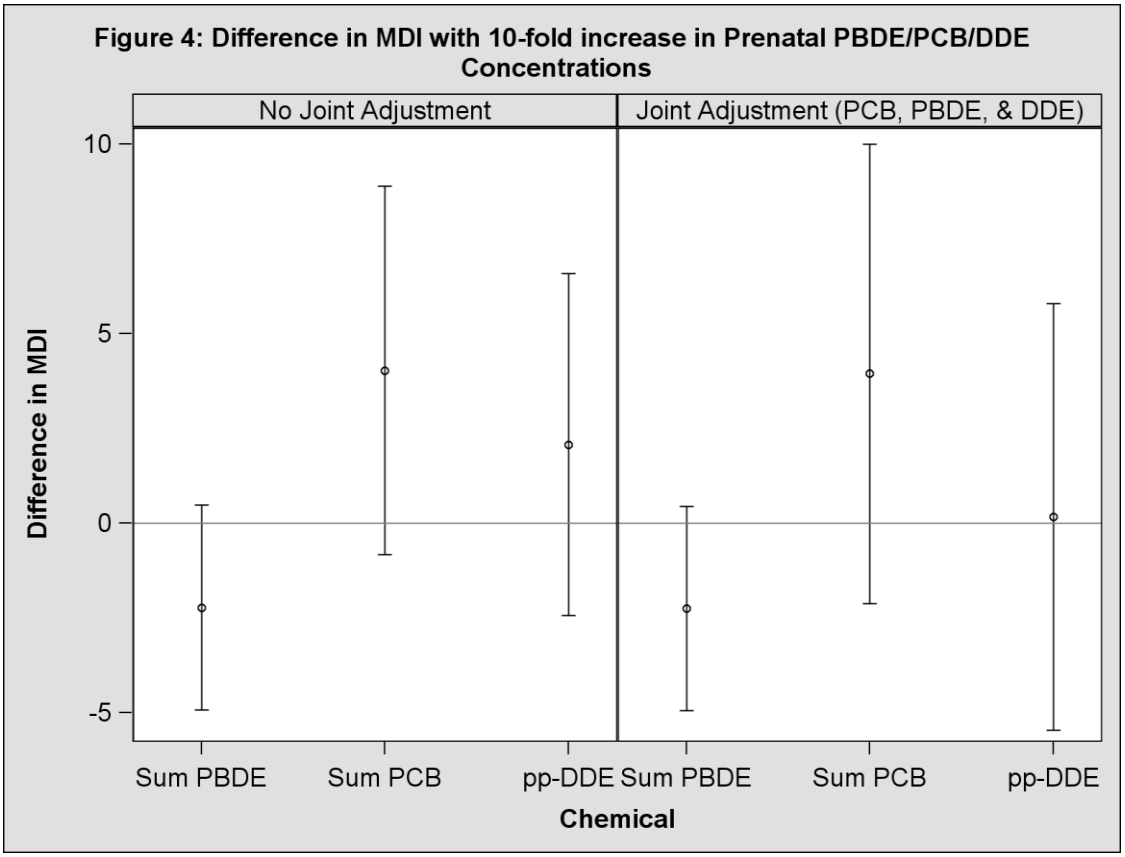


Figure 4: Difference in MDI with 10-fold increase in Prenatal PBDE/PCB/DDE Concentrations
No Joint Adjustment
Joint Adjustment (PCB, PBDE, & DDE)
Difference in MDI
10
5
0
-5
Sum PBDE
Sum PCB
pp-DDE
Sum PBDE
Sum PCB
pp-DDE
Chemical