HHS Public Access

Author manuscript

J Agric Biol Environ Stat. Author manuscript; available in PMC 2018 November 28.

Published in final edited form as:

J Agric Biol Environ Stat. 2015 March ; 20(1): 100–120. doi:10.1007/s13253-014-0180-3.

Characterization of Weighted Quantile Sum Regression for Highly Correlated Data in a Risk Analysis Setting

Caroline Carrico,
Department of Biostatistics, School of Medicine, Virginia Commonwealth University, Richmond, VA, USA

Chris Gennings,
Department of Biostatistics, School of Medicine, Virginia Commonwealth University, Richmond, VA, USA

David C. Wheeler, and
Department of Biostatistics, School of Medicine, Virginia Commonwealth University, Richmond, VA, USA

Pam Factor-Litvak
Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA

Abstract

In risk evaluation, the effect of mixtures of environmental chemicals on a common adverse outcome is of interest. However, due to the high dimensionality and inherent correlations among chemicals that occur together, the traditional methods (e.g. ordinary or logistic regression) suffer from collinearity and variance inflation, and shrinkage methods have limitations in selecting among correlated components. We propose a weighted quantile sum (WQS) approach to estimating a body burden index, which identifies “bad actors” in a set of highly correlated environmental chemicals. We evaluate and characterize the accuracy of WQS regression in variable selection through extensive simulation studies through sensitivity and specificity (i.e., ability of the WQS method to select the bad actors correctly and not incorrect ones). We demonstrate the improvement in accuracy this method provides over traditional ordinary regression and shrinkage methods (lasso, adaptive lasso, and elastic net). Results from simulations demonstrate that WQS regression is accurate under some environmentally relevant conditions, but its accuracy decreases for a fixed correlation pattern as the association with a response variable diminishes. Nonzero weights (i.e., weights exceeding a selection threshold parameter) may be used to identify bad actors; however, components within a cluster of highly correlated active components tend to have lower weights, with the sum of their weights representative of the set.

Keywords

Correlation; Nonlinear model; WQS; Subset selection; Variable selection

Supplementary materials accompanying this paper appear on-line.

1. INTRODUCTION

Several recent reviews have described the conceptual and analytical challenges involved in assessing the health effects of exposures to mixtures of chemical contaminants (Billionnet et al. 2012; Dominici et al. 2010; Vedal and Kaufman 2011). Complex mixtures are considered to be part of the "exposome" (Wild 2005, 2012; Rappaport and Smith 2010), and are of current interest, for example, in perinatal (Buck Louis et al. 2013) and air pollution (Brunekreef 2013) epidemiology. Conceptually, cumulative exposure to multiple compounds, each with concentrations below a set regulatory dose, may be associated with health endpoints, even though associations are not found at permissible levels of individual mixture components. A study of the food supply in New York State (Schecter et al. 2013) found that individual phthalate concentrations were all below the EPA regulatory dose, yet cumulative exposure was substantially higher.

Difficulties arise in the analysis of health effects of chemical mixtures when high correlations occur either between different chemicals within a class of contaminants (e.g. phthalates), between classes of contaminants (e.g. phthalates and phenols), and/or between contaminant metabolite concentrations measured in urine. For example, Fig. 1 presents a heat map of the complex observed correlation pattern among urinary phthalate monoesters as measured in the National Health and Nutrition Examination Survey (2005–2008), with estimates ranging between near 0 to near perfect (0.98). These correlations likely derive from shared exposure routes or sources (e.g. diet), as well as shared metabolic processes.

The task of identifying the key etiologically relevant compounds and/or mixtures associated with adverse health outcomes challenges standard regression-based analytical techniques, due to the strong correlation structure of the exposures as well as the hypothesized correlation between individual exposures and outcomes. Regularization methods, including ridge regression, lasso, adaptive lasso, and elastic net, have been proposed for conducting regression with a correlated set of predictors. The methods generally reduce regression coefficient variance at a cost of increased bias. In ridge regression, an L2 penalty is imposed on the regression coefficients that can shrink the coefficients, depending on the value of the ridge shrinkage parameter, to decrease the variance of the p coefficient estimates (e.g., Hoerl and Kennard 1970):

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\beta}\left[\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p-1}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right].$$

The lasso imposes an L1 penalty on the regression coefficients that can shrink some regression coefficients to exactly zero (Tibshirani 1996):

$$\hat{\beta}_{\text{lasso}} = \arg\min_{\beta}\left[\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p-1}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right].$$

Hence, the lasso performs model selection in addition to the coefficient shrinkage that is performed in ridge regression. As a result, lasso can lead to a more parsimonious model that is simpler to interpret than a ridge regression model. The adaptive lasso is a weighted lasso with data-dependent weights, often chosen to be the inverse absolute maximum likelihood estimate of the regression parameter (Zou 2006):

$$\hat{\beta}_{\text{adaptive lasso}} = \arg\min_{\beta}\left[\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p-1}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\hat{w}_j|\beta_j|\right].$$

The elastic net applies a combination of the lasso and ridge regression L1 and L2 penalties to the coefficients, where a parameter determines the amount of weight each of the lasso and ridge penalties receives (Zou and Hastie 2005):

$$\hat{\beta}_{\text{elastic net}} = \arg\min_{\beta}\left[\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p-1}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}(\alpha|\beta_j|+(1-\alpha)\beta_j^2)\right].$$

Both the adaptive lasso and elastic net can shrink regression coefficients to zero to perform model selection.

Shrinkage methods are commonly used for prediction and may be particularly useful in the $p >> n$ problem, but they have limitations for use in risk evaluation of environmental chemical mixtures. Traditional ridge regression does not reduce the dimensionality of the problem. In the presence of high correlations among predictor variables, the lasso method has been shown to select an arbitrary member from the group of correlated predictors (Zou and Hastie 2005). This is particularly problematic for risk evaluation of environmental chemicals, where the implication is that those not selected are not associated with the adverse health outcome. The elastic net method encourages a “grouping effect” that causes correlated predictors to either all be eliminated from the model or all be used in the model (Zou and Hastie 2005). The grouping effect is exhibited if the regression coefficients of highly correlated variables tend to be equal (Zou and Hastie 2005) and may be applicable in the case of genes along a common biological pathway. However, a grouping effect is problematic in the risk evaluation of environmental chemicals where correlations among the chemicals is due to exposure and/or behavior patterns and is not necessarily associated with the health outcome. For example, it may be the case that only one of two highly correlated chemicals has a biological association with the health effect and, thus, both chemicals should not be either jointly selected in the model or removed from the model. Further, simulation studies show that elastic net often outperforms lasso in terms of prediction accuracy (Zou and Hastie 2005). When the variables are orthogonal, lasso is consistent in variable selection (Zou 2006); however, lasso, adaptive lasso and elastic net have been shown in simulation studies to have poor specificity in variable selection with correlated data (i.e., select incorrect variables; Table 3 in Zou and Hastie 2005; Table 3 in Zou 2006).

In addition to these shrinkage methods, principal components analysis has been used to create synthetic variables to represent mixtures of chemicals (e.g., Mustapha et al. 2011).

However, the principal components are constructed based solely on the correlation pattern among the predictor variables, without regard to the outcome variable. Thus, the principal components do not identify a set of components associated with a selected health effect, as the loadings are the same regardless of the health effect. Further, when a principal component is significantly associated with a health effect, the components with high loadings may not all be associated with the outcome.

Recent articles on complex chemical mixtures and on the potential approaches and general issue of analyzing collinear variables have reviewed the various statistical approaches to model fitting and variable selection (Billionnet et al. 2012; Hastie et al. 2009). One such strategy is the empirical construction of a weighted index, or score of exposure, for use in a regression model (Gennings et al. 2010; Christensen et al. 2013). Such an index focuses a test for association of the mixture with a health effect to a single degree of freedom test with increased power. The index is readily interpretable as an estimation of the mixture effect (i.e., the slope associated with the index) where the weights identify the bad actors and "zero out" components with no (or negligible) association (Billionnet et al. 2012; Christensen et al. 2013). In addition, a weighted index evaluates components in the direction of increased risk, thereby averting the focus from the environmental toxins that may be found to have a protective effect (Roberts and Martin 2006). It may be the case that some components have positive associations and others have negative associations with the selected health outcome; however, we propose that focusing in the direction of increased risk improves the interpretability of the weighted index. Further, by focusing the inference in a single direction, the problem of the reversal paradox, where regression coefficients have opposite signs due to the correlation among the predictor variables, is avoided (Tu et al. 2008). The weights are constrained to sum to 1 and be between 0 and 1, reducing the dimensionality and the issues with collinearity while the weights identify important components, thereby making the index interpretable. Interestingly, regularization using a nonnegativity constraint (suggested in Breiman 1996) has been shown to "perform as well as, and sometimes better than, shrinkage estimators in terms of average prediction error" (Leblanc and Tibshirani 1993). Our focus herein is on accurate variable selection (instead of prediction error).

In this paper, we extend earlier work (Christensen et al. 2013; Gennings et al. 2010) using a weighted index by adding a bootstrap step to estimate the weights and naming the approach weighted quantile sum (WQS) regression. In the next section, we describe WQS regression, followed by a motivating example. Section 4 includes simulation studies to demonstrate the benefit of adding bootstrapping in WQS regression, and for comparing the performance of WQS regression to that of shrinkage methods in terms of accurate variable selection. The simulation studies are based on environmentally relevant correlation patterns among the components. We summarize the ability of the methods to correctly classify chemicals as bad actors in the simulation studies.

2. METHODS: WQS REGRESSION

Consider data with c correlated components that are reasonable to combine into an index (e.g., environmental contaminants with a potential common adverse health effect). Let the values for the c components be scored into quantiles, denoted q_i (e.g., for quartiles, $q_i = 0, 1,$

2, or 3 for values in the 1st, 2nd, 3rd, or 4th quartile, respectively) for $i = 1$ to c. The basic weighted index model (Christensen et al. 2013) is

$$g(\mu) = \beta_0 + \beta_1\left(\sum_{i=1}^{c} w_i q_i\right) + \mathbf{z}'\boldsymbol{\phi}, \quad (1)$$

where w_i is the unknown weight for the ith component, β_0 is the intercept, β_1 the regression coefficient for the weighted quantile sum (constraining its association with the mean to be either nonpositive or nonnegative), $\mathbf{z}$ is a vector of covariates (risk factors and confounders) determined prior to estimating the weights, $\boldsymbol{\phi}$ is a vector of regression coefficients for the covariates, and g represents any monotonic, differentiable link function as in a generalized linear model, which links the mean, μ, to the predictor variables. The term $\sum_{i=1}^{c} w_i q_i$ represents the weighted index for the set of c chemicals of interest. The weights are constrained to sum to 1, $\sum_{i=1}^{c} w_i = 1$, and are constrained by the limit $0 \leq w_i \leq 1$.

Following Christensen et al. (2013), the data (size N) may be split into a training (N_T) and a validation dataset (N_V) to estimate the WQS index weights in the training set, and then test the effect of the index for statistical significance in the validation set. In order to empirically and simultaneously estimate the weights and the parameters using the training data, we employ optimization algorithms that maximize a continuous nonlinear function subject to the constraints on the weights. We use the trust region method optimization algorithm for estimation because it allows for a linear constraint on a nonlinear objective function. A description of this optimization strategy is given in Nocedal and Wright (2006). The NLP procedure in SAS 9.2 treats the constrained optimization in the Lagrange format under the Kuhn-Tucker Conditions (SAS Institute Inc 2008).

In the spirit of recent work on stability selection (Meinshausen and Buhlmann 2010), we propose adding a bootstrap step to the WQS regression estimation to increase sensitivity in detecting important predictors through perturbing the data. A fixed number (B) of bootstrap samples of size N_T are generated from the training dataset (typically $B = 100$ or 1,000) and are used to estimate the unknown weights that maximize the likelihood for the model. That is, the unknown parameters $\boldsymbol{\theta} = (\beta_0, \beta_1, w_1, \ldots, w_c, \boldsymbol{\phi})$ are estimated in the nonlinear model given in equation (1) using maximum likelihood estimation for each bootstrap sample, $b = 1, \ldots, B$, where the weights are constrained to the unit interval and to sum to one within each bootstrap sample. The set of estimated weights are tested for significance in each bootstrap sample through the significance of β_1. The weights are used to estimate the weighted quantile sum index, given by:

$$\text{WQS} = \sum_{j=1}^{c} \overline{w}_j q_j \quad (2)$$

where $\bar{w}_j = \frac{1}{B}\sum_{b=1}^{B} w_{j(b)} f\left(\hat{\beta}_{1(b)}\right)$ and $f\left(\hat{\beta}_{1(b)}\right)$ is a pre-specified "signal function" (constrained to sum to 1) of the estimated slope parameter associated with WQS from the b^{th} bootstrap sample, i.e., a measure of the signal strength. The signal function is defined such that samples with higher signal have higher relative weight in creating *WQS*. For example, the signal function may be defined by the relative test statistic associated with β_1: i.e., define S_b as the test statistic from the bth bootstrap sample; then, the signal function is $S_b / \sum(S_b)$. Alternatively, the signal function may be based on an indicator of whether β_1 is significant in the b$^{\text{th}}$ bootstrap sample.

Ideally, the significance of the WQS index is determined using the validation data and the model

$$g(\mu) = \beta_0 + \beta_1 \text{WQS} + \mathbf{z}'\boldsymbol{\phi}. \quad (3)$$

The index in equation (2) is defined using average weights across the bootstrap samples. When the dataset is large enough to split into training and validation sets, this test for significance of β_1 is based on independent data. Otherwise, the model in equation (3) can be based on the full dataset. Considerations of whether the sample is large enough to split may be based on whether the results are robust to the random splitting of the data.

The proposed weighted index model in equation (1) can be written in the Lagrangian format. Without loss of generality, for $g(\mu) = \mu$ with the form of the model in equation (1) in least squares optimization,

$$\hat{\theta}_{WQS} = \arg\min_{\theta}\left[\sum_{j=1}^{n}\left(y_i - \left(\beta_0 + \beta_1 \sum_{i=1}^{c} w_i q_i + \mathbf{z}'\boldsymbol{\phi}\right)\right)^2 + \lambda\left(\sum_{i=1}^{c} w_i - 1\right)\right] \quad (4)$$

or, equivalently for large samples, the maximum likelihood form is

$$\hat{\theta}_{WQS} = \arg\max_{\theta}\left[\ln(L(\theta; y)) + \lambda\left(\sum_{i=1}^{c} w_i - 1\right)\right]. \quad (5)$$

Under this form of the equation, for each bootstrap sample, the log-likelihood for the model in equation (5) is optimized subject to the constraint on the weights. A further constraint due to the structure of the model is that the association between the weighted quantile sum and the mean is either nonnegative or nonpositive, which constrains the correlation between the response and all the components to be of the same sign.

Interpretation of the estimated WQS regression model follows in two steps. First, a test for significance of β_1 determines whether there is an association between the index and the outcome variable. The significance of the regression coefficient associated with the WQS index permits the further interpretation of the weights. Second, the important components in

the index are identified by comparing the "average" (across the bootstrap samples) weight for each component in equation (2) to a selection threshold parameter, τ, chosen *a priori*. For the analyses and simulation study conducted herein, with 11 components, we used $\tau = 0.05$.

3. MOTIVATING EXAMPLE

Phthalates are chemicals used extensively in consumer products including soft toys, flooring, medical equipment, paints, plastic bags, cosmetics, and air fresheners (Wormuth et al. 2006). Because of their ubiquitous exposure, phthalates are commonly detected in human samples (Kim et al. 2014). Epidemiologic research has found associations between exposure to phthalates and various adverse health outcomes, including altered male reproductive development and function, altered thyroid function, increased waist circumference and insulin resistance, decreased gestational age or increased risk of premature birth, and respiratory symptoms and asthma (Ferguson et al. 2011). In a cross-sectional study, Ferguson et al (2011; NHANES, 1999–2006) reported an association between several phthalate monoester metabolites with increased serum markers of inflammation and oxidative stress; and also inverse associations between these markers and several oxidized phthalate metabolites. As these monoesters have a complex correlation pattern (Fig. 1), we were interested in using both WQS regression and shrinkage methods to determine which phthalates were jointly associated with oxidative stress as measured by gamma glutamyltransferase (GGT).

Using data from NHANES (2005–2008), we evaluated the association between concentrations of 11 urinary phthalate monoesters (i.e., metabolites defined in Table 1; natural log scale transformed), associated with 8 parent diesters (i.e., parent compounds), and oxidative stress as measured by log(GGT) in young and middle aged adults (18–50 years; $N = 1{,}439$), adjusting for age, gender, smoking status, urinary creatinine (logscale), and BMI. In univariable analyses, only MBP and MIB were significant ($p = 0.001$ and $p = 0.005$, respectively; Table 1) with positive slopes. We used three shrinkage methods (lasso, adaptive lasso, and elastic net) for model building. Analyses were conducted in R (R Development Core Team 2008) using *glmnet* (Friedman et al. 2010). The five adjustment covariates were forced to remain in the models. The weights in adaptive lasso were set to the inverse of the absolute least squares estimates for the regression coefficients. The lambda parameter was chosen through cross-validation as the largest value of lambda such that the prediction error was within one standard error of the minimum (i.e., 1-SE rule). The alpha parameter in elastic net was selected over a grid of values where both parameters minimized prediction error within one standard error of the minimum. The entire dataset (with nonmissing values) was used for the shrinkage methods.

Nine of the 11 phthalates were selected in the model using either lasso or elastic net (Table 1): five had positive coefficients (ECP, MBP, MHH, MHP, and MIB) and four had negative coefficients (CNP, COP, MNM, and MOH); only ECP, MBP, MHH, and MOH were significant. Using adaptive lasso, only MHH (positive and significant) and MOH (negative and significant) were selected.

Discrepancies between the univariable analyses and the shrinkage regression models include: (1) MIB was significant in the univariable analysis but was not included in the full model; (2) ECP, MHH, and MOH were not significant in the univariable analyses but were in the full model and were significant; and (3) lasso and elastic net included nine phthalate variables, while adaptive lasso selected only two of the 11.

We next conducted a WQS regression analysis, after splitting the sample size into a training ($N = 593$) and validation ($N = 846$) dataset using a 40:60% split leaving more in the validation dataset for increased power for the significance of the WQS index. Using $B = 100$ bootstrap samples, the weighted (based on relative signal of the test statistic for the WQS slope in each bootstrap sample) average weights for each monoester (Table 1) were used to construct the WQS index for the phthalates. The corresponding slope was positive and significant ($p = 0.036$) in the validation dataset. Ninety-three percent (93 %) of the weight was on MBP, MEP MIB, ECP, and MZP; and weights between 0.01 and 0.02 corresponded to 5 other phthalates. For comparison, when the bootstrap step was omitted, the weights in the training dataset included 0 weights on 7 of the 11 monoesters with a weight of 0.66 on MBP, 0.22 on MEP, 0.07 on MIB, and 0.05 on ECP. The slope associated with this index was also positive and borderline significant ($p = 0.060$) in the validation dataset. Thus, the bootstrap step provided a stronger signal due to the inclusion of additional monoesters in the index—a point further evaluated through simulation studies in the next section.

4. SIMULATION STUDIES

4.1. Simulating Four Cases Based on Observed GGT and Phthalate Data

We evaluated WQS regression and other regularization methods through simulation studies based on (i) enhanced values (3 or 5 times) of the observed correlations between the outcome variable and the phthalates in NHANES (2005–2008) (Table 1); and (ii) the observed or diminished correlation pattern among the components (Fig. 1). Specifically, the simulation is based on the observed correlation pattern among the 11 phthalates concentrations and the residuals (Y) from the covariates only model (adjusting for BMI, smoking status, gender, age, and urinary creatinine (log scale)). Details of simulating multivariate normal data from a correlation matrix are provided in the Appendix. In short, the response variable (Y) and predictor variables (X) are simultaneously generated, as multivariate normal data, from an assumed form of their correlation matrix. We assumed CNP, ECP, MEP, MHP, MNM, and MZP were not associated with Y and the other five monoesters have the absolute observed correlation as given in Table 1, ranging between 0.02 and 0.05, in the following cases:

1. Three times the assumed correlation between Y and the specified five components (ranging between 0.06 and 0.15) and the observed correlation pattern among the phthalates (Fig. 1);
2. Three times the assumed correlation between Y and the specified five components and half the observed correlation pattern among the phthalates (i.e., off-diagonal values multiplied by 0.5);

3. Five times the assumed correlation between Y and the specified five components (ranging between 0.10 and 0.25) and observed correlation pattern among the phthalates; and

4. Five times the assumed correlation between Y and the specified five components and half the observed correlation pattern among the phthalates.

These scenarios provide a range of anticipated difficulty in correctly classifying predictors: a difficult scenario with weak association with Y but highly correlated components (Case 1); a less difficult scenario with weak association with Y and less correlated components (Case 2); a scenario with stronger association with Y and highly correlated components (Case 3); and an easier scenario with stronger signal and weaker correlation among the components (Case 4). Further, shrinkage methods (lasso, adaptive lasso, and elastic net) were evaluated using the predictor variables as continuous random variables and as scored variables using quartiles ($q = 0, 1, 2,$ or 3). WQS regression was evaluated (i) without a bootstrap step; (ii) with a signal function defined by the relative test statistic for β_1; and (iii) with a signal function defined as a binary function of whether β_1 was significant (value of 1) or not (value of 0).

One hundred (100) simulated studies were generated with a total sample size of 500, a study size considered more practical for cohort studies than that observed in the motivating example. The full sample size was used in the univariable analyses and the shrinkage methods. For WQS regression, the sample was randomly divided into a training dataset ($N = 250$) and a validation dataset ($N = 250$); a selection threshold parameter of 0.05 was used to determine whether or not a component was selected. Finally, a range of selection threshold parameters were evaluated graphically based on the (median) number of correctly selected and incorrectly selected variables across the 100 simulated studies.

4.2. Evaluation of WQS Regression and Accurate Variable Selection

As stated previously, construction of the WQS index is based on a weighted average of the empirical weights across the bootstrap samples. In the bootstrap samples where the signal of an association between the index and outcome variable (i.e., as measured by the signal-to-noise ratio associated with the regression parameter for the index) from the data is strong, the estimated weights are more informative compared to when the regression parameter is near zero. Two signal functions were used to construct the index—i.e., a weighted average across the bootstrap samples defines the weight for each component: (1) relative test statistic (S) for β_1, i.e., $S_b/\sum(S_b)$; and (2) whether the regression parameter in the bth bootstrap sample was significant or not. For comparison, the case where the signal function was set to 1 was also evaluated. The number of correctly and incorrectly selected variables using these signal functions was similar (Table 2); thus, a simple (unweighted) average across the bootstrap distributions had similar accuracy in variable selection as more complicated signal functions.

The nonlinear model in equation (1) was estimated for each of the simulated datasets with and without a bootstrap step. Results (Table 2) show that without the bootstrap step, the median number of variables correctly selected (from 5) was 2 (with IQR = (1,2)) in the most difficult case (Case 1), and was 3 in Cases 2 and 4. This is compared to a median of 4

variables correctly selected with the bootstrap step in Cases 1, 2, and 4 and a median of 3 in Case 3, with negligible difference observed between the two signal functions for Cases 2, 3 and 4. The relative signal function was somewhat less sensitive than using the indicator function in Case 1—a median of 3 compared to 4. Specificity improved for WQS as the signal increased (comparing Cases 1 and 2 to Cases 3 and 4).

These results are based on the *a priori* chosen selection threshold parameter, here 0.05. The sensitivity and specificity of WQS regression depends on this parameter. Figure 2 presents the median number of correctly and incorrectly selected variables for the 4 cases. A selection threshold that is too low results in an increase in the incorrect selection of variables; whereas a selection threshold that is too high results in a decrease in the correct selection of variables—something available for review in a simulation study but not for data analysis.

For each simulation study in WQS regression, the training dataset was the base for the bootstrap samples, which were used to determine the WQS index. The validation data were used to test for the significance of the constructed index. The power was low in cases 1 and 2, i.e., 37 % and 48 %, respectively; the power was high in cases 3 and 4, 93 % and 98 %, respectively.

4.3. Comparison with Ordinary and Shrinkage Regression

For comparison to WQS regression, we conducted a univariable analysis in each simulation study across the four cases. Using median counts, of the six variables that should not have been selected, none were significant in these models (Table 2). The median number of correctly selected variables was 2 for cases 1 and 2. But when the signal increased (5 times the observed correlation) the median number of significant variables was 4. This indicates the signal in the data is quite low in two of the four cases.

We also looked at three shrinkage methods for comparison: lasso, adaptive lasso, and elastic net (Table 2). In each analysis, we counted the number of nonzero regression parameters. In Case 1, the median number of correctly selected variables was 2 for lasso and adaptive lasso, with between 1 and 2 incorrectly selected. Elastic net was greedier with 3 (median) correctly selected but 3 were also incorrectly selected. All three shrinkage methods failed to detect a signal in case 2, generally, 75 % of the time. Further, the reversal paradox seems to have played a role in Case 3 when the signal was increased but the correlation pattern among the components was complex. The median number ranged between 3 and 6 for those variables that should not have been selected and were generally negative (data not shown). When the correlation pattern diminished and the signal was strong (case 4), specificity improved for the shrinkage methods, but these methods were not as sensitive as WQS.

In summary, these studies indicate a loss of information when the shrinkage methods are conducted on quartiles of the concentration variables (Table 2). We make this comparison because WQS is based on quartiles and results indicate that differences in accurate variable selection observed for WQS compared with the shrinkage methods are not due to the use of quartiles. For the shrinkage methods, variables that are incorrectly selected often (incorrectly) have a negative regression coefficient, perhaps, as a result of the reversal

paradox (data not shown). In cases where the shrinkage methods have high sensitivity, the specificity is often low. For WQS regression, the bootstrap step improves sensitivity in all four cases with impact on specificity only in the low signal cases. The signal functions were similar in terms of accurate variable selection. WQS performed better when there was a stronger signal, in contrast to all three shrinkage methods, which had low specificity in Case 3. Finally, WQS has higher specificity compared to shrinkage methods in Case 3 and comparable specificity in Case 4.

4.4 Evaluation of WQS Regression Weights and Underlying Correlation Pattern

As stated previously, the estimated weights in the WQS index are used for variable selection by comparing the weighted average to the selection threshold value. The weights may also be used to interpret the importance of the association between each component and the outcome variable. However, the magnitude of the weights also depends on the correlation pattern. We investigated this by adjusting the simulation study described in Sect. 4.1 for homogeneous association between the active components and Y—i.e., where the correlation between the five ‘bad actors’ and Y was either 0.1 (Case A) or 0.2 (Case B). In this scenario, the shifts in the distributions among the active chemicals are due to the correlation pattern among the components.

Histograms of the estimated WQS weights are provided in Fig. 3, where the (active) chemicals set to be truly associated with the outcome are in green and the (nonactive) chemicals set to not be associated with the outcome are plotted in red. The heat map in Fig. 1 provides an indication of Spearman correlation estimates for each chemical (log transformed) with the other 10 monoesters. The distributions of the weights associated with the active chemicals in Case B are distinctly different from zero and different from the “tower-like” distributions of the nonactive chemicals. The distributions for Case A are not as distinct as in Case B with the exception of ECP and MHP, where more than 90% of the weight is negligible (i.e., bar at 0). The two assumed active chemicals with high correlations with another (above 0.9) are MHH and MOH; both have lower average weights (0.11 for both in Case A and 0.13 in Case B) than other active chemicals. The nonactive chemicals, generally, have more weight at zero when the correlation with other chemicals is higher (e.g., ECP and MEP in Case A). Thus, chemicals with higher pairwise correlations compared to others tend to have lower weights. The distinction between inactive and active chemicals is improved as the signal increases (Fig. 3 and Table S1).

5. DISCUSSION

In summary, we have proposed the addition of a bootstrap step in the use of WQS regression for improved accuracy. Through simulation studies, we have observed an improvement in sensitivity with only minor changes in specificity (in the low signal cases) as a result of perturbing the data through bootstrapping. We have argued that regularization through different constraints than those used in shrinkage methods result in improved specificity when using components with a complex correlation pattern, as observed in environmental chemical exposures. The strategy is conducted in steps, where the identification of important

components, as determined by their estimated weights being above a selection threshold value, follows only when the regression parameter for the index is significant.

The strategy is based on the empirical estimation of a weighted quantile sum index as associated with an outcome variable of interest. We have proposed to use quantiles (e.g., quartiles due to their common usage in epidemiology studies) instead of the continuous variables for several reasons. Estimation of weights without bounds on the components results in extreme values having influence that grows with the weights. Second, translating the variables (here, concentrations) to the same scale allows for correction of different potencies for the components. However, a limitation to using quantiles is the loss of the full exposure range of the components. Through the simulation study, the use of quartiles in shrinkage methods was indeed associated with a loss of information. Additional limitations of the current formulation of the WQS regression model are the assumptions that there is no interaction among the exposures and that there is a constant change in risk between the quantiles. These are current lines of inquiry for our group in extending the WQS regression model.

The results from the simulation studies focus on the identification of bad actors when the estimated weights exceed a specified cutoff, here 0.05. Figure 2 elucidates the tension between a small threshold value, where too many variables are incorrectly selected, and a large threshold value, where too many variables are missed. Our experience indicates that the cutoff should be smaller as the number of components increases. For example, in an analysis of 34 highly correlated PCB congeners, a cutoff of 0.005 is more useful as many of the weights fall between 0.01 and 0.05, with the equally weighted average weight of 0.03. Unlike in a simulation study, the optimum choice of a selection threshold parameter cannot be determined in a real data analysis. However, we propose to use values between 0.01 and 0.05 inversely related to the number of components. Realistically, with a large number of components (say, dozens), estimated weights below a selection threshold of 0.01 (perhaps rounded from 0.005) seem negligible. The selection of a threshold parameter is somewhat analogous to the choice of the criterion for selecting the tuning parameter from cross validation in lasso. The lambda associated with the minimum prediction error may allow too few variables in the model, while, somewhat arbitrarily, the one associated with the 1-standard error rule is considered to impose more regularization.

Recent work by Meinshausen and Buhlmann (2010) on ‘stability selection’ combines subsampling with high dimensional selection algorithms resulting in “markedly improved” selection methods. In short, for a given tuning parameter, stability selection is defined as a strategy that randomly selects a subsample (of size $n/2$) drawn without replacement from a sample of size n many times and estimates the probability of each variable being in the selected set through repeated subsampling. Instead of looking at a single model, the data are perturbed (e.g., by subsampling) many times. Variables are selected that occur in a large fraction of the resulting selection sets. In fact, although they describe subsampling, the authors note “bootstrapping would behave similarly.” Interestingly, Meinshausen and Buhlmann (2010) prove that randomized lasso (similar to adaptive lasso where the weights are random values) is consistent in variable selection using stability selection with weaker conditions on the design than required for regular lasso. Thus, characteristics of variable

selection are improved by including a data perturbation step. In WQS regression, variables are associated with weights, i.e., a linear combination of the quantile scores, instead of the binary case of selection or not selection. In the spirit of stability selection, we propose that the bootstrap step improves variable selection. Instead of basing inference on the weights from a single model, the data are perturbed through bootstrapping and variables are selected with average weights that exceed a selection threshold value.

A standard analysis strategy in epidemiology studies of mixtures of chemical exposures is to evaluate each chemical alone while adjusting for covariates/confounders (e.g., Colt et al. 2005). However, humans are exposed to multiple chemicals that may impact human health. Interestingly, the simulation study conducted here demonstrated strong sensitivity and specificity in univariable models when the signal was high (Cases 3 and 4 in Table 2). In Cases 1 and 2, the univariable regression approach appears to have not only lower sensitivity (i.e., less power to detect truly important predictors) but also lower type I error (fewer number of false positives) than WQS. In addition, other metrics besides accurate variable selection are important in evaluating the impact of exposures to mixtures on human health. The univariable regression approach does not account for the full health impact of the mixture.

The WQS nonlinear model in equation (1) links the regression coefficient β_1 with the weighted quantile sum. Components in the sum not associated with the outcome variable have low weights that approach zero. In practice, these are often 10^{-12} or lower. However, the sum of the weights is constrained to one; thus, the model as stated does not support the case where none of the components have nonzero weights. In practice, we initialize the iterative nonlinear estimation algorithm with equal values for the weights (i.e., $1/c$ for c components). When the regression coefficient is close to zero, the signal from these weights is masked, leaving the weights near their starting values. The interpretation of the weights should follow only when the corresponding regression coefficient demonstrates an association with the outcome variable.

We have simulated data by concatenating the correlation matrix among the chemicals with the assumed bivariate correlations to the outcome variable and using a Cholesky decomposition to generate multivariate normal data (see Appendix). This is in contrast to the usual approach of simulating the mean through a vector of assumed regression coefficients given the predictor variables. With a specified correlation matrix among the predictors, **R**, the choice is whether to fix the bivariate correlations between the predictors and the response variable (i.e., **r**) or the regression coefficients, β; i.e.,

$$\mathbf{r} = \mathbf{R}\beta$$
$$\beta = \mathbf{R}^{-1}\mathbf{r}.$$

The two approaches for simulating data are similar when the predictor variables are independent (i.e., R is an identity matrix). However, with environmentally relevant correlation patterns (e.g., as observed in the phthalate data), the generated bivariate correlations with the response differ from the association as generated through the regression

parameters. We chose to control the bivariate correlation in the data by specifying those that are related to the response and those that are not.

In summary, WQS regression is useful for risk analysis because simulation study results suggest it has good specificity and adequate sensitivity for identifying predictors, with improvements in performance as the correlation with the outcome increases. When interpreting the weights, one should keep in mind both the pairwise correlations among the components and the correlation with the outcome variable. If the pairwise correlations are high relative to the correlation with the outcome, then this could lead to a breakdown case. In that situation, the weights should be considered in conjunction with the pairwise correlations that a given component has with other components. If a component has a minimal weight (i.e. less than 0.05 or 0.01 if a large number of components or a complex correlation structure is present) and is highly correlated with another component assigned minimal weight, the two are likely important, but have smaller weights as a result of their high pairwise correlation. From this type of analysis, we are able to detect components that are associated with a given health outcome and assess the total body burden they impose on an individual.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge support from #T32 ES0007334 and #UL1TR000058.

APPENDIX

SIMULATING CORRELATED DATA

Our objective is to simulate normally distributed data $N(M, \sum)$ with a given correlation structure for an outcome y and predictors $x_1, x_2, \ldots x_c$. Let ρ be the correlation matrix between and among y and the components in X and ∑ be the corresponding covariance matrix with diagonal values in vector S and sample means in vector **m**. To impose the correlation structure, we first use the relationship between the correlation and the variance that yields:

$$\square = \text{diag}(S) * \rho * \text{diag}(S)$$

Then follow the simulation steps below where p = c + 1:

1) Calculate the Cholesky decomposition of ∑ (p × p dimension), such that $\sum = \mathbf{U}'_{\text{pxp}}\mathbf{U}_{\text{pxp}}$. (see Harville (1997))

2) Simulate $\mathbf{Z_i} \sim \text{N}(\mathbf{0_{px1}}, \mathbf{I_p})$. $\mathbf{Z}' = [\mathbf{Z_1 Z_2} \ldots. \mathbf{Z_n}]$, i.e., **Z** is nxp where each row is a *p*-variate standard normal distribution.

3) Let $\mathbf{M} = (\mathbf{m}*\mathbf{1}_{1\text{xn}})'$ and $\mathbf{Y}_{\text{nxp}} = \mathbf{M}_{\text{nxp}} + \mathbf{Z}^*_{\text{nxp}}\mathbf{U}_{\text{pxp}}$

a) $E(\mathbf{Y}) = E(\mathbf{M} + \mathbf{Z}^*\mathbf{U}) = \mathbf{M} + E(\mathbf{Z}) = \mathbf{M}$

b) $Var(\mathbf{Y}) = Var(\mathbf{M} + \mathbf{Z}^*\mathbf{U}) = Var(\mathbf{M}) + Var(\mathbf{Z}^*\mathbf{U}) = \mathbf{0} + \mathbf{U}'Var(\mathbf{Z})\mathbf{U} = \mathbf{U}'\mathbf{U} = \sum$

So, **Y** is nxp and has the distribution $N_p(\mathbf{M}, \sum)$

In the first step, in order to calculate **U**, $\sum$ must be positive definite. To evaluate relevant cases with highly correlated data, $\sum$ may be nearly singular. In this case, we use matrix ridging to stabilize the matrix.

References

Billionnet C, Sherrill D, Annesi-Maesano I; GERIE Study (2012). Estimating the health effects of exposure to multi-pollutant mixture. Annals of Epidemiology 22(2): 126–141. [PubMed: 22226033]

Breiman L (1996). Stacked regressions. Machine Learning 24:49–64.

Brunekreef B Exposure science, the exposome, and public health. Environmental and molecular mutagenesis. 2 26 2013.

Buck Louis GM, Yeung E, Sundaram R, Laughon SK, Zhang C. The exposome-exciting opportunities for discoveries in reproductive and perinatal epidemiology. Paediatr Perinat Epidemiol. 5 2013;27(3):229–236. [PubMed: 23574410]

Center for Disease Control. National Health and Nutrition Examination Study. http://www.cdc.gov/nchs/nhanes.htm.

Christensen KLY, Carrico CK, Sanyal AJ, Gennings C (2013). Multiple classes of environmental chemicals are associated with liver disease: NHANES 2003-04. International Journal of Hygiene and Environmental Health. [epub March 8, 2013].

Colt J, Severson R, Lubin J, Rothman N, Camann D, Davis S, Cerhan JR, Cozen W, Hartge P. (2005). Organochlorines in carpet dust and non-Hodgkin lymphoma. Epidemiology 16(4): 516–525. [PubMed: 15951670]

Dominici F, Peng RD, Barr CD, Bell ML. (2010) Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. Epidemiology. 21(2):187–194. [PubMed: 20160561]

Ferguson KK, Loch-Caruso R, Meeker JD (2011) Exploration of oxidative stress and inflammatory markers in relation to urinary phthalate metabolites: NHANES 1999-2006. Environ Sci Technol. 2012 1 3;46(1):477–85. Epub 2011 Dec 1. [PubMed: 22085025]

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J of Statistical Software, 33(1), 1–22. [http://www.jstatsoft.org/v33/i01/]

Gennings C, Sabo RT, Carney E. (2010). Identifying subsets of complex mixtures most associated with complex diseases polychlorinated biphenyls and endometriosis as a case study. Epidemiology, 21, S77–S84. [PubMed: 21422968]

Hastie T, Tibshirani R, Friedman J (2009). The Elements of Statistical Learning: Data mining, Inference, and Prediction, 2nd edn, Springer Series in Statistics.

Harville DA (1997). Matrix algebra from a statistician's perspective. Dordrecht: Dordrecht Springer-Verlag New York Inc.

Hoerl AE and Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1): 55–67.

Kim S, Kang S, Lee G, Lee S, Jo A, Kwak K, Kim D, Koh D, Kho YL, Kim S, Choi K (2014). Urinary phthalate metabolites among elementary school children of Korea: sources, risks, and their association with oxidative stress marker. Sci Total Environment, 472:49–55.

Leblanc M and Tibshirani R (1993). Combining estimates in regression and classification. J American Statistical Association, 91:1641–1650.

Mustapha BA, Blangiardo M, Briggs DJ, Hansell AL (2011). Traffic ari pollution and other risk factors for respiratory illness in schoolchildren in the Niger-Delta region of Nigeria. Environ Health Perspect. 119:1478–1482. [PubMed: 21719372]

Meinshausen N and Buhlmann P (2010) Stability selection. Journal of the Royal Statistical Society, 72, 417–473.

Nocedal J and Wright S (2006). Numerical optimization. New York: New York: Springer

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0 [http://www.R-project.org]

Rappaport SM, Smith MT (2010) Epidemiology. Environment and disease risks. Science. 330(6003): 460–461. [PubMed: 20966241]

Roberts S and Martin MA (2006) Investigating the mixture of airpollutants associated with adverse health outcomes. Atmospheric Environment 40(5):984–991.

SAS Institute Inc (2008). SAS 9.2 Help and Documentation. Cary, NC: SAS Institute Inc.

Schecter A, Lorber M, Guo Y, Wu Q, Yun SH, Kannan K, Hommel M, Imran N, Hynan LS, Cheng D, Colacino JA, Birnbaum LS (2013) Phthalateconcentrations and dietary exposure from food purchased in New YorkState. Environ Health Perspect. 121(4):473–494. [PubMed: 23461894]

Tibshirani R (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Vol. 58, No. 1, pages 267–288.

Tu YK, Gunnell D, Gilthorpe MS (2008). Simpson's Paradox, Lord's Paradox, and Suppression Effects are the same phenomenon – the Reversal Paradox. Emerging Themes in Epidemiology, 5:2 [http://www.ete-online.com/content/5/1/2]. [PubMed: 18211676]

Vedal S, Kaufman JD (2011). What does multi-pollutant air pollution research mean? Am J Respir Crit Care Med. 183(1):4–6. [PubMed: 21193783]

Wild CP (2005). Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev. 14(8):1847–1850. [PubMed: 16103423]

Wild CP (2012). The exposome: from concept to utility. International journal of epidemiology. 41(1): 24–32. [PubMed: 22296988]

Wormuth M, Scheringer M, Vollenweider M, Hungerbuhler K (2006) What are the sources of exposure to eight frequently used phthalic acid esters in Europeans? Risk Analysis, 26(3):803–824. [PubMed: 16834635]

Zou H (2006). The adaptive lasso and its oracle properties. J American Statistical Association. 101:1418–1429.

Zou H, Hastie T (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society.Series B (Statistical Methodology), 67(2), 301–320.

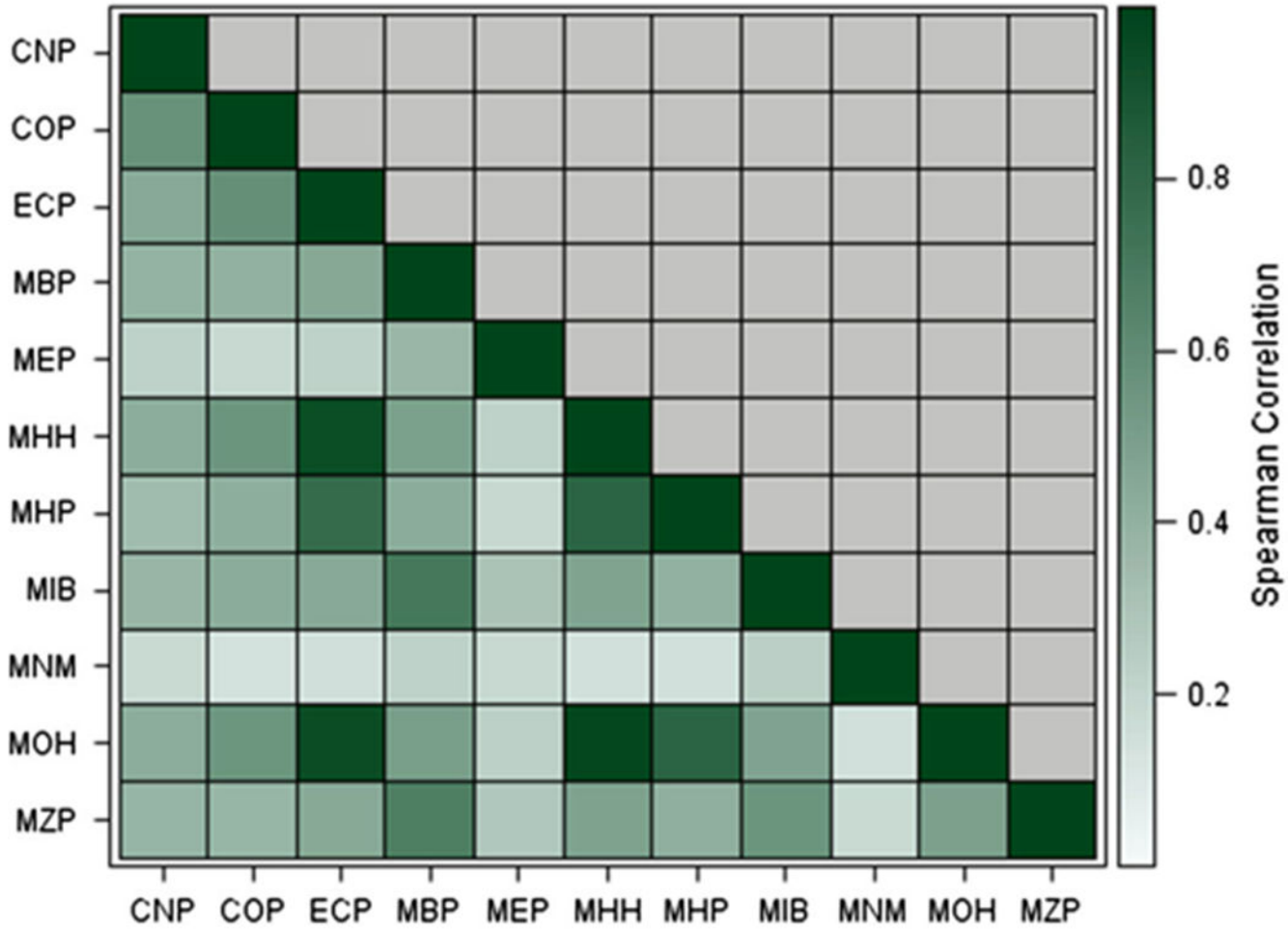


Figure 1.

Heat map of Spearman correlation estimates between the urinary phthalate monoesters (logscale transformed; $N = 1439$) as measured in the National Health and Nutrition Examination Survey (NHANES, 2005–2008).

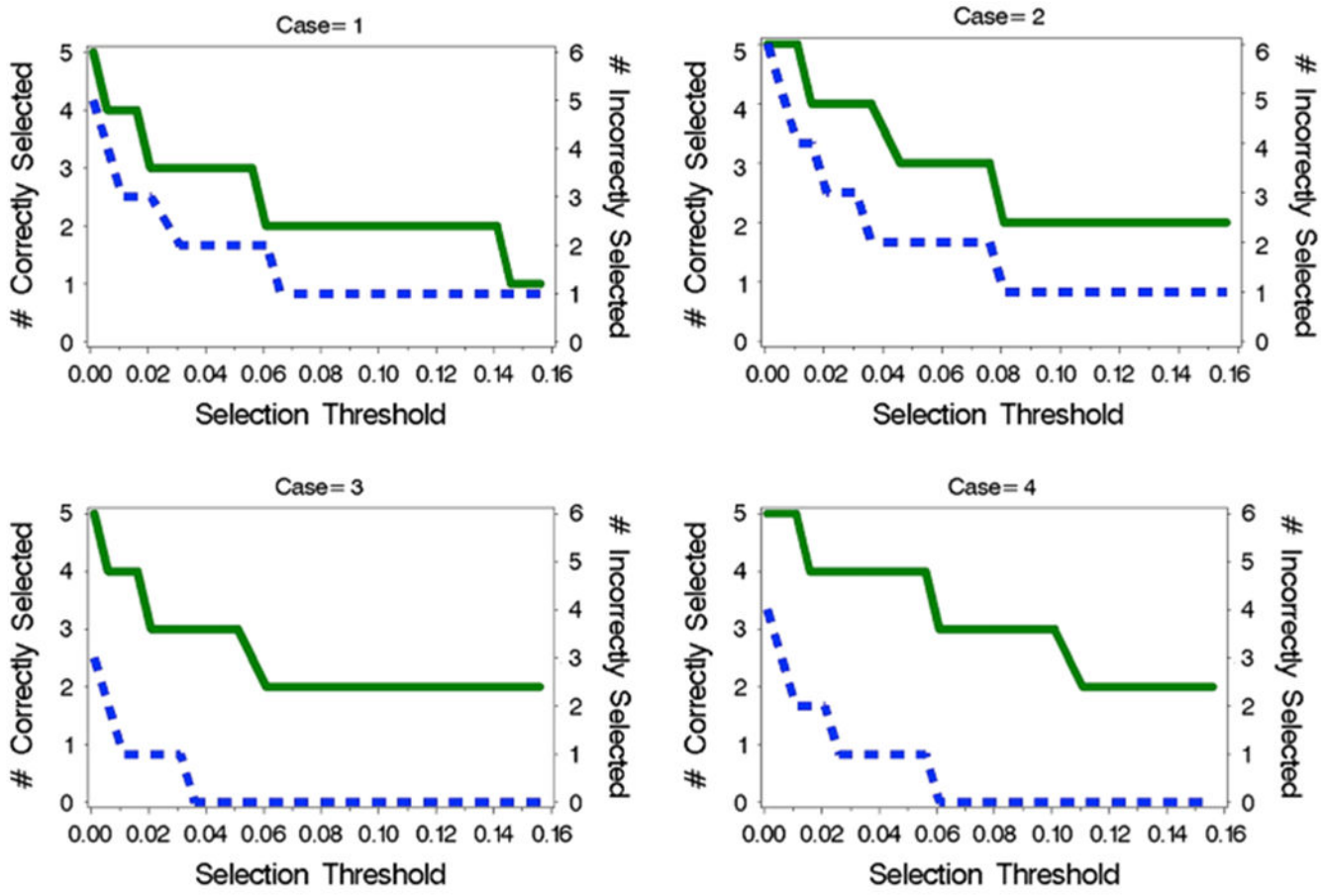
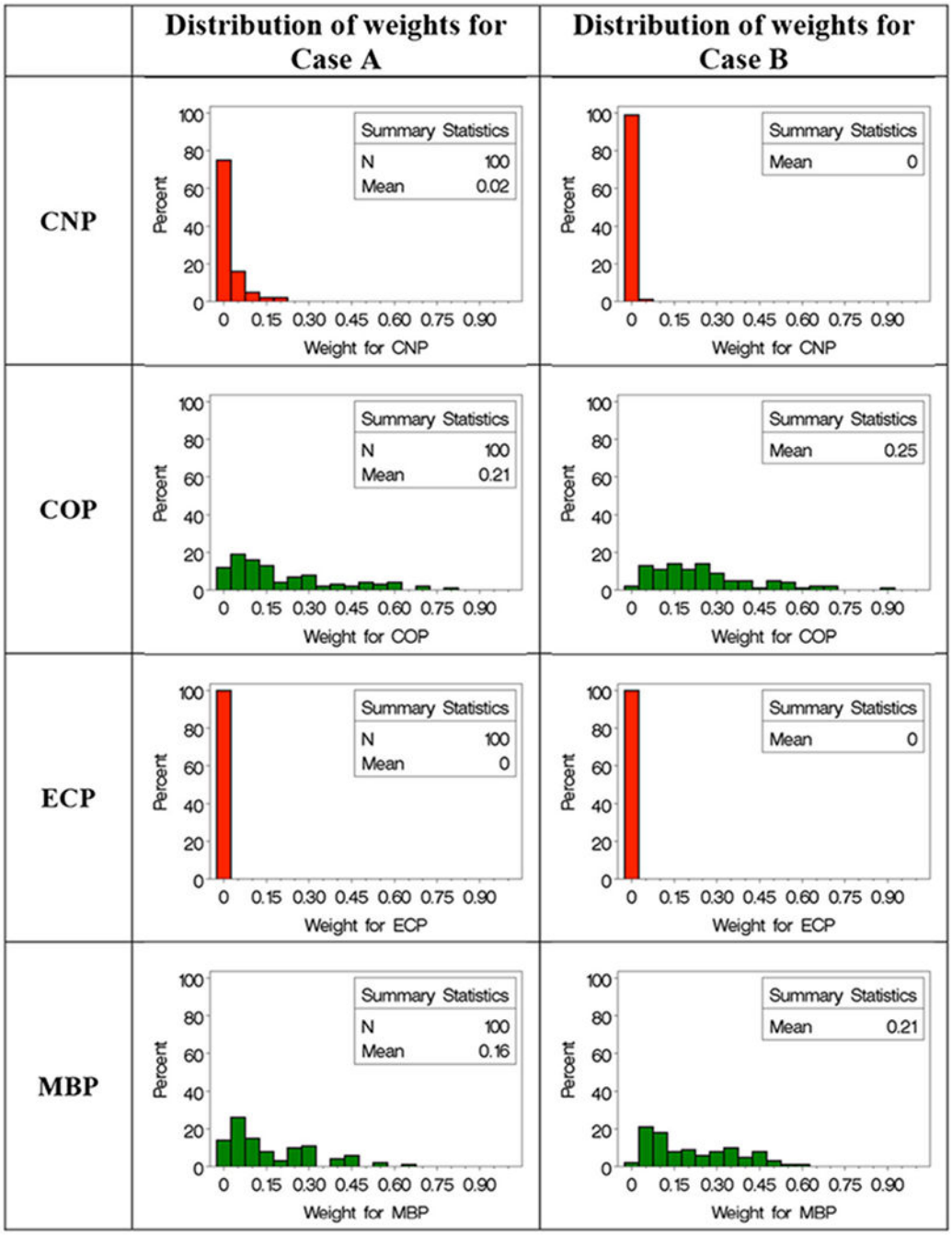
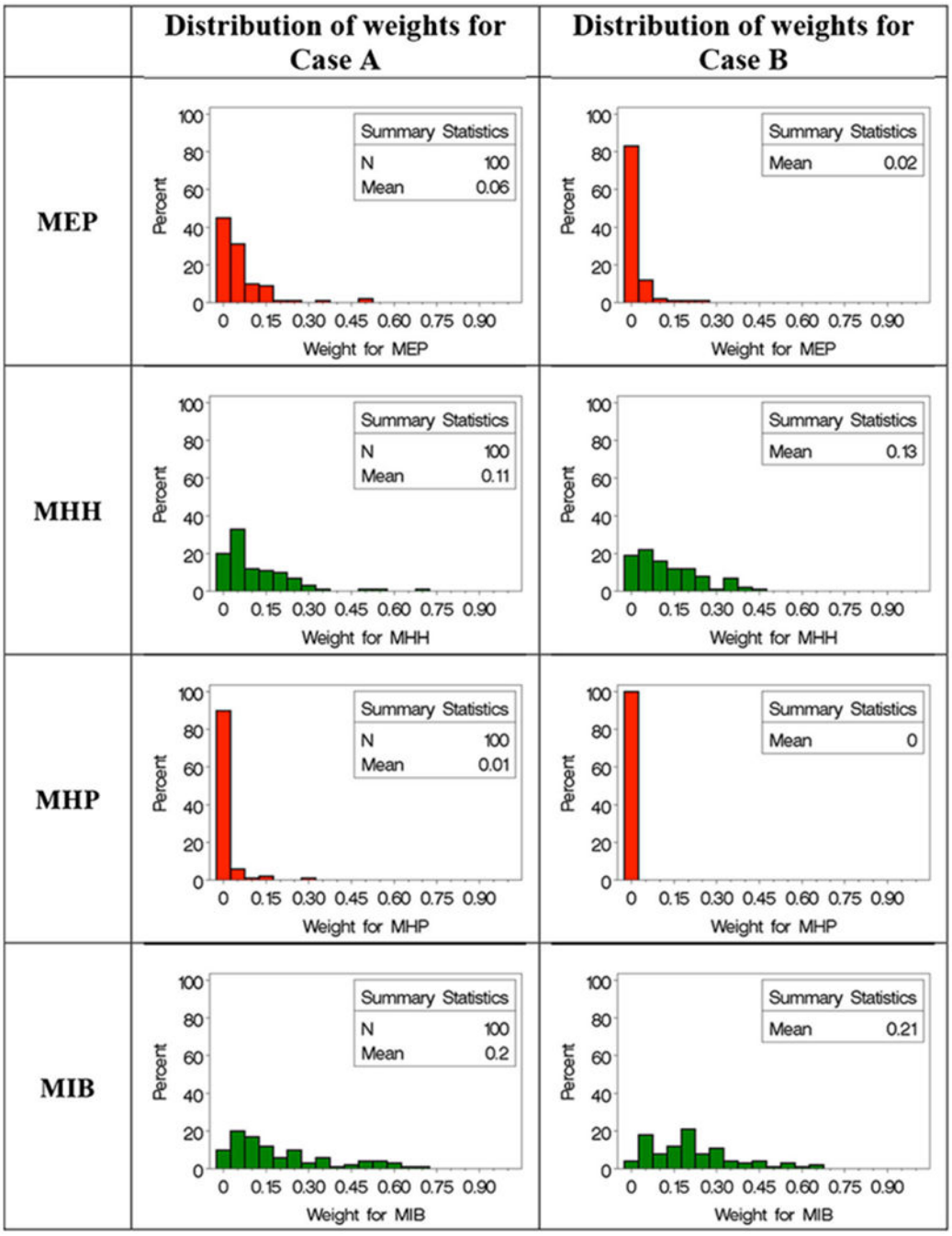


Figure 2.

Median number of correctly (*solid*) and incorrectly (*dashed*) selected variables over values of the selection threshold parameter for each of the four cases (Table 1) in the simulation study.





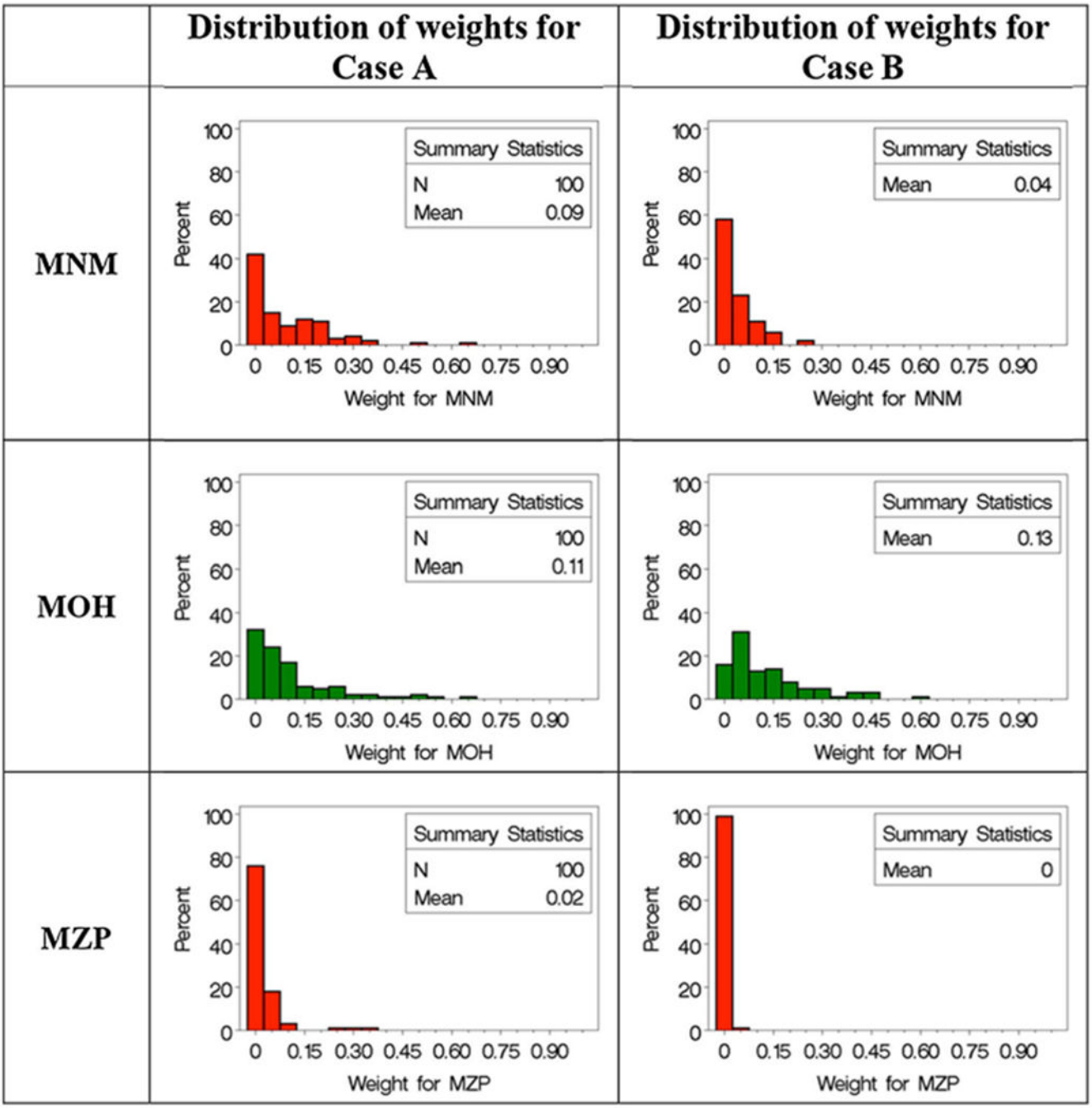


Figure 3.

Weights across 100 simulation studies (Case A) similar to Case 1, with observed correlation pattern among the chemicals where the correlation between the outcome and each active component is 0.1; and (Case B) similar to Case 4 with the observed correlation pattern diminished by half and where the correlation between the outcome and the active component is 0.2. Histograms for active chemicals are *green* and for inactive chemicals are *red*.

Table 1.

Summary statistics from the analysis of GGT and phthalate monoesters (natural log scale) in NHANES (2005–2008) using WQS regression, lasso, adaptive lasso, and elastic net adjusted for covariates (BMI, smoking status, age, urinary creatinine, and gender).

Monoester#	Correlation with residuals	Univariable analysis		WQS regression			Shrinkage methods (coefficient sign)		
		coefficient sign	*p* value	Weighted mean	Minimum	Maximum	Lasso	Adaptive lasso	Elastic net
CNP	−0.01	Negative	0.235	0.02	0	0.28	Negative	0	Negative
COP*	−0.02	Negative	0.368	0.01	0	0.24	Negative	0	Negative
ECP	0.01	Positive	0.917	0.08	0	0.41	Positive	0	Positive
MBP*	0.04	Positive	0.001	0.46	0	1.0	Positive	0	Positive
MEP	0.01	Positive	0.708	0.18	0	0.63	0	0	0
MHH*	0.02	Positive	0.450	0.02	0	0.42	Positive	Positive	Positive
MHP	<0.01	Positive	0.518	0.01	0	0.18	Positive	0	Positive
MIB*	0.05	Positive	0.005	0.14	0	0.76	Positive	0	Positive
MNM	−0.01	Negative	0.583	0.01	0	0.43	Negative	0	Negative
MOH*	−0.02	Negative	0.117	0	0	0	Negative	Negative	Negative
MZP	0.01	Positive	0.267	0.07	0	0.58	0	0	0

The table includes the correlation with residuals from the covariate only model; the coefficient sign and p-value from univariable regression with each monoester (log scale) adjusted for the five covariates; and the average (across 100 simulated studies) adjusted (due to the signal function) estimated weights from 100 bootstrap samples sparabreak # *CNP* mono(carboxynonyl) phthalate, *COP* mono (carboxyoctyl) phthalate, *ECP* mono-2-ethyl-5-carboxypentyl phthalate, *MBP* mono-n-butyl phthalate, *MEP* mono-ethyl phthalate, *MHH* mono-(2-ethyl-5-hydroxyhexyl) phthalate, *MHP* mono-(2-ethyl)-hexyl phthalate, *MIB* mono-isobutyl phthalate, *MNM* mono-n-methyl phthalate, *MOH* mono-(2-ethyl-5 oxohexyl) phthalate, *MZP* mono-benzyl phthalate.

*Assumed nonzero value in the simulation study.

Table 2.

Summary statistics (median and IQR of the number of variables correctly and incorrectly selected) for univariable analyses, regularization methods (lasso, adaptive lasso, and elastic net), and WQS regression.

Method	Case1		Case2		Case3		Case4	
	# Correct	# Incorrect	# Correct	# Incorrect	# Correct	# Incorrect	# Correct	# Incorrect
TRUTH	5	0	5	0	5	0	5	0
11 univariable analyses	2 (1,3)	0 (0,0)	2 (2,3)	0 (0,1)	4 (2,4)	0 (0,0)	4 (3,4)	0 (0,1)
Lasso: Nonzero	2 (0,3)	2 (0,4)	0 (0,0)	0 (0,0)	5 (4,5)	6 (5,6)	2 (1,2)	0 (0,0)
Lasso (quartiles): Nonzero	0 (0,0)	0 (0,0)	0 (0,0)	0 (0,0)	3 (3,4)	4 (3,5)	1 (0,2)	0 (0,0)
Adaptive lasso: Nonzero	2 (1,2)	1 (1,2)	0 (0,0)	0 (0,0)	4 (3,4)	3 (3,3)	1 (1,2)	0 (0,1)
Adaptive lasso (quartiles): Nonzero	2 (1,2)	1 (1,2)	0 (0,0)	0 (0,0)	4 (3,4)	3 (3,3)	1 (1,2)	0 (0,1)
Elastic net: Nonzero	3 (1,4)	3 (0,5)	1 (0,1)	0 (0,0)	5 (5,5)	6 (5,6)	2 (2,4)	0 (0,3)
Elastic net (quartiles): Nonzero	1 (0,1)	0 (0,0)	0 (0,1)	0 (0,0)	4 (3,5)	5 (4,5)	2 (1,3)	0 (0,1)
WQS regression (no bootstrap step)	2 (1,2)	0 (0,1)	3 (2,3)	1 (0,1)	2 (2,2)	0 (0,0)	3 (2,4)	0 (0,1)
WQS regression (with bootstrap step; signal function = 1)	3 (2,3)	1 (1,2)	3 (3,4)	2 (1,2)	3 (2,3)	0 (0,1)	4 (3,4)	1 (0,1)
WQS regression (with bootstrap step; signal function: relative test statistic)	3 (2,3)	1 (1,2)	4 (3,4)	2 (1,2)	3 (2,3)	0 (0,1)	4 (3,4)	1 (0,1)
WQS regression (with bootstrap step; signal function: significant β_1)	4 (3,4)	2 (1,2)	4 (3,4)	2 (1,2)	3 (2,3)	0 (0,1)	4 (3,4)	1 (0,1)

One hundred datasets were simulated based on correlation patterns observed in NHANES (2005-08) between phthalate concentrations and residuals from a regression of oxidative stress as measured by log(GGT) on age, gender, smoking status, urinary creatinine (log scale), and BMI. Five of the 11 phthalates were set to be truly associated with the residuals. Case 1: observed correlation pattern among the chemicals and 3 times the observed absolute correlation with the residuals (Table 1); Case 2: observed correlations diminished by half with 3 times the observed absolute correlation with residuals; Case 3: observed correlation pattern among the chemicals with 5 times the observed absolute correlation with residuals; and Case 4: observed correlation diminished by half with 5 times the observed absolute correlation with residuals. A selection threshold of 0.05 was used for WQS regression.