NIH Public Access
Author Manuscript
Stat Med. Author manuscript; available in PMC 2013 December 20.

Published in final edited form as:
Stat Med. 2012 December 20; 31(29): 4102–4113. doi:10.1002/sim.5520.

Adverse Subpopulation Regression for Multivariate Outcomes with High-Dimensional Predictors

Bin Zhu[a,b,*], **David B. Dunson**[a], and **Allison E. Ashley-Koch**[b]

[a]Department of Statistical Science, Duke University, Durham, North Carolina 27708, U.S.A.

[b]Center for Human Genetics, Duke University Medical Center, Durham, North Carolina 27710, U.S.A.

Abstract

Biomedical studies have a common interest in assessing relationships between multiple related health outcomes and high-dimensional predictors. For example, in reproductive epidemiology, one may collect pregnancy outcomes such as length of gestation and birth weight and predictors such as single nucleotide polymorphisms in multiple candidate genes and environmental exposures. In such settings, there is a need for simple yet flexible methods for selecting true predictors of adverse health responses from a high-dimensional set of candidate predictors. To address this problem, one may either consider linear regression models for the continuous outcomes or convert these outcomes into binary indicators of adverse responses using pre-defined cutoffs. The former strategy has the disadvantage of often leading to a poorly fitting model that does not predict risk well, while the latter approach can be very sensitive to the cutoff choice. As a simple yet flexible alternative, we propose a method for adverse subpopulation regression (ASPR), which relies on a two component latent class model, with the dominant component corresponding to (presumed) healthy individuals and the risk of falling in the minority component characterized via a logistic regression. The logistic regression model is designed to accommodate high-dimensional predictors, as occur in studies with a large number of gene by environment interactions, through use of a flexible nonparametric multiple shrinkage approach. The Gibbs sampler is developed for posterior computation. The methods are evaluated using simulation studies and applied to a genetic epidemiology study of pregnancy outcomes.

Keywords

Bayesian; Genetic epidemiology; Latent class model; Logistic regression; Mixture model; Model averaging; Nonparametric; Variable selection

1. Introduction

Biomedical studies routinely collect multiple quantitative health outcomes and investigate how the risk of having adverse values for these outcomes is associated with predictors. The typical approach in such setting is to 1) use multivariate normal linear regression in which the mean of the response distribution varies linearly with predictors; 2) first categorize the responses based on pre-specified cutoffs and then fit a logistic regression. The former

Copyright © 2011 John Wiley & Sons, Ltd.

*Correspondence to: Center for Human Genetics, Duke University Medical Center, Durham, North Carolina 27710, U.S.A. bin.zhu@duke.edu.

7. Supplementary Materials Web Appendix and Web Figure referenced in Sections 5 and the MCMC codes for ASPR model are available at the *Statistics in Medicine* website http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1097-0258.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

approach is insufficiently flexible to accommodate settings in which the predictors do not simply shift the response by a fixed amount for all individuals, while the latter approach is extremely sensitive to cut-point choices. In this article, we propose a simple alternative approach for adverse subpopulation regression (ASPR) relying on a two component mixture model that incorporates a logistic regression for the risk of falling into the minority component in the mixture, with the logistic regression model accommodating high-dimensional predictors. We focus on the case when researchers are interested in dichotomizing the subjects into two classes: healthy and unhealthy group (corresponding to the majority and minority of the population); and each component is modeled by the multivariate normal distribution whose mean vector and covariance matrix change with latent class membership. This model is purposefully chosen to be simple to facilitate analyses and interpretations in settings involving high-dimensional predictors, though generalizations to multiple latent classes is straightforward as discussed in Section 6.

Our approach is motivated by the Healthy Pregnancy, Healthy Baby (HPHB) Study, a prospective cohort study of pregnant women residing within Durham County, NC with the goal of identifying environmental, social and genetic factors that contribute to racial disparities in birth outcomes [1]. Here we focus on assessing how predictors - a large number of maternal candidate gene single nucleotide polymorphisms (SNPs), environmental exposures, and their interactions - impact the risk of low values of infant birth weight and gestational age at delivery. Such research questions cannot be addressed by the standard linear regression with continuous responses, where one models the predictor effects on the response means. The standard approach is instead to dichotomize the quantitative outcomes into binary indicators, such as low birth weight (LBW, birth weight<2500g), preterm birth (PTB, gestational age at delivery<37weeks) and small for gestational age (SGA, birth weight less than 10th percentile for that gestational age), and then apply logistic regression. While such analyses are easily implemented, they rely on pre-defining thresholds with the analysis results varying significantly according to the threshold choice [2].

We propose an alternative method for adverse subpopulation regression, which relies on a two component latent class model [3, 4, 5], with the component weights dependent on predictors via logistic regression. Related approaches are considered by Gage [6], Gage et al. [7] and Schwartz et al. [8], but they focused on models with fixed component weights and with the means varying with predictors. In addition, our emphasis is on applications involving high-dimensional predictors in which maximum likelihood can be expected to have poor performance. Stegle et al. [9] presented a Bayesian model for mapping expression quantitative trait loci (eQTLs) jointly contributed from genotype as well as known and hidden confounding factors. This approach is suitable for the subjects sampled from one population group, while our approach focuses on the population with subpopulation admixture (such as healthy versus unhealthy groups).

In such settings, it has become quite common to rely on either Bayesian methods or penalized likelihoods with penalties incorporated to favor having many coefficients estimated at or near zero, leading to variable selection and an effectively lower dimensional model. In linear regression and generalized linear models, such methods have become standard, with the Lasso [10], elastic net [11] and relevance vector machine [12] providing popular examples. These methods have been applied in genome-wide association studies to cope with large number of SNPs and to select multiple SNPs simultaneously [13, 14, 15]. The penalized likelihood estimators have a Bayesian interpretation in corresponding to the mode of the posterior distribution obtained under carefully chosen priors on the coefficients, with the Laplace leading to the Lasso [16] and a *t*-distribution with low degrees of freedom leading to the relevance vector machine. MacLehose and Dunson [17] recently proposed a new class of multiple shrinkage priors that allow shrinkage towards not only zero but also

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

other values, leading to improved performance in estimating non-zero coefficients. We will consider these and other shrinkage priors in the context of our ASPR model.

The remainder of the paper is organized as follows. In Section 2, we introduce the proposed Bayesian adverse subpopulation regression model, describing both fully Bayes and fast two-stage approaches for inference. Section 3 provides details of an Markov chain Monte Carlo (MCMC) algorithm. Section 4 presents the simulation results, evaluating and comparing the proposed methods with existing methods. In Section 5, we apply the model to pregnancy outcome data. The article concludes with a discussion in Section 6.

2. Bayesian Adverse Subpopulation Regression

2.1. Model Formulation

Suppose we collect the data $(\boldsymbol{y}_i', \boldsymbol{x}_i)$, for subject i, $i = 1, 2, \ldots, n$, where $\boldsymbol{y}_i$ is an $s \times 1$ vector of outcomes and $\boldsymbol{x}_i$ is a $p \times 1$ vector of predictors. We make the simplifying assumption that there are two types of individuals, with $z_i = 0$ denoting healthy individuals and $z_i = 1$ for potentially unhealthy individuals. In addition, we assume for identifiability that the unhealthy individuals are in the minority, with the specific constraints and prior information included for identifiability discussed in detail in Section 2.2. This is a simplification which is made for ease in interpretation, assessment of risk, and scaling to higher dimensions while accommodating the curse of dimensionality that arises. In many cases, such a simplification is made in advance of the analysis by taking one or more response variables and defining cutoffs to dichotomize the data prior to analysis. However, it is well known that results are quite sensitive to the choice of cutoff [18], and hence we prefer allowing z_i to be an adverse health status latent variable. By using a Bayesian approach, we can fully accommodate uncertainty in imputing z_i and avoid forcing any hard threshold on the observed quantitative traits.

Denote $\omega_1(\boldsymbol{x}_i) = \Pr(z_i = 1 \mid \boldsymbol{x}_i)$ as the probability of allocating subject i to the unhealthy population and let $\omega_2(\boldsymbol{x}_i) = 1 - \omega_1(\boldsymbol{x}_i)$. We then express the conditional density of the response $\boldsymbol{y}_i$ given predictors $\boldsymbol{x}_i$ as

$$f(\boldsymbol{y}_i|\boldsymbol{x}_i) = \sum_{h=1}^{2} \omega_h(\boldsymbol{x}_i)\, \mathrm{N}_s(\boldsymbol{y}_i|\theta_h, \Sigma_h), \quad (1)$$

where $\mathbf{N}_s(\boldsymbol{y}_i \mid \boldsymbol{\theta}_h, \boldsymbol{\Sigma}_h)$ is the s-dimensional Normal distribution with mean vector $\boldsymbol{\theta}_h$ and covariance matrix $\boldsymbol{\Sigma}_h$. We note that the traits will not have a multivariate normal distribution *marginally*, but will instead have a mixture of normal distributions. The primary assumption we have made, to facilitate interpretation and implementation, is that the conditional density of the response is well characterized as a mixture of two multivariate normals with the weights dependent on predictors. If we instead allowed many components, we could fit any conditional density but would lose interpretability and encounter challenging identifiability issues. Even if the mixture of two normals assumption is violated, we expect that the proposed approach will nonetheless be highly robust in terms of inferences on the impact of the predictors. We fully expect this to be a rough approximation but a better one than existing practice that dichotomizes or assumes a single normal regression. In practice, the assumption can be checked by first estimating posterior means of the standardized residuals, $e_i = \Sigma_h^{1/2}(\boldsymbol{y}_i - \theta_{z_i})$, for each subject and then applying typical tests for normality. Standard transformations can be applied (e.g., Box-Cox) to improve fit. $\omega_1(\boldsymbol{x}_i)$ in expression (1) depends on predictors $\boldsymbol{x}_i$ through a logistic regression model:

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

$$\omega_1(\boldsymbol{x}_i) = \frac{\exp\left(\gamma + \boldsymbol{x}_i'\beta\right)}{1 + \exp\left(\gamma + \boldsymbol{x}_i'\beta\right)}, \quad (2)$$

where the coefficient vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)'$ characterizes the effect of predictors on the risk of falling in the minority subpopulation. Due to the logistic regression form, the exponentiated β_j coefficients can be interpreted as odds ratios.

2.2. Prior Specification

For the ASPR model in (1)-(2), identifiability of the unhealthy subpopulation necessarily relies on prior information. In the absence of some prior knowledge, the two subgroups would be exchangeable, and we would encounter a label-ambiguity problem. Removal of this problem through appropriate priors is one of the advantages of the simple two component framework over more complex latent class regression models having unknown numbers of components. The most common approach would place restrictions on the means of the components; for example, ordering the components in advance by letting $\theta_{11} < \theta_{21}$. This approach assumes that low values of the first response variable are adverse, which may be reasonable for a given study but is not so in general. Moreover, placing restriction on the means will fail to solve the label ambiguity problem if the components are not separated sufficiently.

Thus, we consider alternative strategies depending on the application. The first is to elicit informative values for the means and covariance matrix in the two components from prior empirical knowledge of the typical distribution of the responses in healthy and unhealthy groups. In the absence of such extra knowledge, we may fit a mixture of two multivariate normals to the data using EM for maximum likelihood estimation, defining the minority component to be adverse. This can be done either from historical data or the current data. Then, fix the $\boldsymbol{\theta}_h$, $\boldsymbol{\Sigma}_h$ at these estimates in the subsequent analysis. This runs the risk of under-estimating uncertainty but has the advantage of simplifying interpretation and completely eliminating identifiability concerns.

Another alternative is to specify conditionally conjugate prior distributions for $\boldsymbol{\theta}_h$, $\boldsymbol{\Sigma}_h$ and γ as follows,

$$(\theta_h, \Sigma_h) \sim \mathrm{NIW}_s(\theta_h, \Sigma_h | \theta_0, \psi_0, \rho_0, \Sigma_0), \quad h = 1, 2,$$

where $\mathrm{NIW}_s(\boldsymbol{\theta}_h, \boldsymbol{\Sigma}_h \mid \boldsymbol{\theta}_0, \psi_0, \rho_0, \boldsymbol{\Sigma}_0)$ is the Normal-inverse-Wishart distribution proportional to $|\Sigma_h|^{-(s+\rho_0+2)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\Sigma_0\Sigma_h^{-1}\right) - \frac{\psi_0}{2}(\theta_h - \theta_0)'\Sigma^{-1}(\theta_h - \theta_0)\right\}$. As weakly informative empirical Bayes priors, the hyperparameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\Sigma}_0$ are chosen to be the sample means and covariance matrix for all subjects, and we set $\psi_0 = 1$ and $\rho_0 = s + 2$ to reduce the prior information. Additionally, we place an informative prior on the intercept in the logistic regression model (2) after centering the predictors, $\gamma \sim \mathbf{N}_1(\gamma \mid \gamma_0, \lambda_0)$. For example, by choosing $\gamma_0 = -2.20$ and $\lambda_0 = 2.42$ for the intercept, the expected baseline probability (*a priori*) of an adverse response is 10% and falls in the range between 3% and 30% with 0.95 probability.

As for the priors for β, if $\boldsymbol{x}_i$ is low-dimensional, we can rely on standard choices, such as independent Gaussian distributions with modest variance. However, as the number of predictors increases, we need some approach for addressing the high dimensionality. A common strategy in the frequentist literature is to use sparse penalized regression (e.g.,

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Lasso, elastic net, etc) to favor many elements of β that are equal to zero while shrinking the non-zero elements toward zero. In the Bayesian literature, a rich variety of shrinkage priors have been proposed for high-dimensional regression coefficients, with most approaches relying on priors that are centered at zero, potentially with a point mass incorporated to allow variable selection. Hierarchical shrinkage priors that are centered at zero can potentially lead to over-shrinkage of coefficients that are not close to zero. Such over-shrinkage can be reduced by choosing a prior which is concentrated near zero with very heavy tails, but in that case there is no borrowing of information or incorporation of prior knowledge in estimating the coefficients that are not close to zero.

As an alternative approach that had excellent performance in high-dimensional logistic regression, MacLehose and Dunson [17] proposed a multiple shrinkage prior (MSP) for the jth coefficient $\beta_j \sim \int \mathrm{DE}(\beta_j \mid \mu_j, \tau_j) d\mathrm{P}(\mu_j, \tau_j)$, where $\mathrm{DE}(\beta_j \mid \mu_j, \tau_j)$ is the double exponential (Laplace) distribution with location parameter μ_j and scale parameter τ_j; the mixture distribution P is assigned a modified Dirichlet process prior that incorporates a mass at $\mu_j = 0$ for the first component. Specifically, the MSP is expressed as

$$\begin{aligned} \mathrm{P}(\cdot) &= \pi_1 \delta_{(0,\tau_1^*)}(\cdot) + \sum_{t=2}^{\infty} \pi_t \delta_{(\mu_t^*,\tau_t^*)}(\cdot), \\ \mu_t^* &\sim \mathrm{N}_1(\mu_t^* | c, d), \\ \tau_1^* &\sim \mathrm{Gamma}\left(\tau_1^* | a_0, b_0\right), \quad \tau_t^* \sim \mathrm{Gamma}(\tau_t^* | a_1, b_1), \\ \pi_t &= V_t \prod_{l<t} (1 - V_l), \quad V_t \sim \mathrm{Beta}(1, \alpha), \end{aligned} \tag{3}$$

where $\mathrm{Gamma}(\tau \mid a, b) = 1/[b^a \Gamma(a)] \tau^{a-1} \exp(-\tau/b)$ with mean $a \times b$; $\delta_\theta(\cdot)$ is the probability measure with all its mass at θ; following MacLehose and Dunson [17], we choose $c = 0$, $a_0 = b_0 = 30$, $a_1 = b_1 = 6.5$ and $\alpha = 1$. Note that the MSP (3) is represented in the stick-breaking form [19], which starts with a unit probability "stick" and sequentially breaks off random proportions of the stick, with each of these pieces corresponding to the probability π_t placed on one mixture component (μ_t^*, τ_t^*). This formulation allows infinitely many components, with only a relatively small number occupied by the p predictors, effectively bypassing the difficult issue of estimating the number of mixture components. In addition, the stick-breaking form facilitates MCMC sampling. The discrete form of P leads to ties between (μ_j, τ_j), $j = 1, 2, \ldots, p$ and hence clusters corresponding to multiple (μ_j, τ_j) equal to (μ_t^*, τ_t^*). Consequently, through MSP, the coefficients β will be shrunk to multiple locations μ_t^*s, including zero in the first cluster ($t = 1$), corresponding to the usual Bayesian Lasso prior, while the other components are centered at unknown locations away from zero. For β_j and $\beta_{j'}$ belonging to the same cluster t, $\beta_j \neq \beta_{j'}$ with $E(\beta_j) = E(\beta_{j'}) = \mu_t^*$ and $Var(\beta_j) = Var(\beta_{j'}) = \tau_t^{*2}$. In our application of the ASPR model, it is unlikely that any of the predictors being considered have a log-odds ratio of falling in the adverse sub-group outside of $\beta_j \in [-1, 1]$, corresponding to an interval of [0.37, 2.72] for the odds ratio. In most genetic epidemiology studies involving complex health conditions, one expects at most a modest deviation from a log-odds near zero for single SNPs or SNP × environment interactions. This small signal-to-noise ratio is one aspect that makes detection of important variants so challenging. To express this prior information, while inflating the prior variance somewhat to corresponding to a "weakly informative" prior [20], we let $d = 0.1507$, which leads to $\Pr(\mu_t^* \in [-1, 1]) = 0.99$ *a priori*.

3. Posterior Computation

In describing an approach for posterior computation, we focus on the approach described in Section 2 that places a normalinverse-Wishart prior on the component-specific parameters,

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

an informative prior on the intercept γ for identifiability, and a mixture of double exponential shrinkage prior on the high-dimensional vector of β coefficients. This approach is straightforward to modify to accommodate the other approaches described in Section 2. For example, to instead use the two-stage plug-in approach, we would run the EM algorithm first to estimate μ_h, Σ_h for $h = 1, 2$ and then would hold these component-specific parameters fixed in the proposed data augmentation Gibbs sampling algorithm to be described below. In addition, if an alternative shrinkage prior were used for the coefficients β_j, then one could simply modify the sampling steps for updating the β_j appropriately. For scale mixture of normal priors, such as double exponentials, t priors or other standard choices, this is straightforward.

If we observe the latent subpopulation index z_i directly for each individual and are interested in the coefficients β, then we could apply the MCMC algorithm of MacLehose and Dunson [17] directly. However, because we do not observe z_i for any of the subjects, we instead modify their algorithm to include steps for imputing z_i from the corresponding full Bernoulli conditional posterior distribution and sampling the mean and covariance specific to each component. We start by relating the latent subpopulation index $z_i = I(g_i > 0)$ to an auxiliary random variable g_i, where $I(\cdot)$ is the indicator function, which equals 1 when $g_i > 0$ and 0 otherwise. To induce expression (2) through marginalizing g_i, we assume g_i follows a logistic distribution centered on $\gamma+\boldsymbol{x}_i'\beta$. Holmes and Held [21] proposed a data augmentation MCMC algorithm for posterior computation in logistic regression models relying on characterizing the latent g_i as a scale mixture of normals, with the square root of the scale parameters following a Kolmogorov-Smirnov (KS) distribution. Due to lack of conjugacy of the conditional posteriors of scale parameters specific to each subject, they recommend using rejection sampling. However, use of a large number of rejection sampling steps can lead to inefficiencies, so we instead apply an alternative data augmentation scheme. Following O'Brien and Dunson [22], the logistic distribution can be almost exactly approximated by a noncentral t-distribution $t_\nu\left(g_i|\gamma+\boldsymbol{x}_i'\beta,\sigma^2\right)=\int_0^\infty \mathrm{N}_1\left(g_i|\gamma+\boldsymbol{x}_i'\beta,\sigma^2/\phi_i\right)$ Gamma$(\phi_i \mid \nu/2, 2/\nu)d\phi_i$, when we set $\sigma^2 = \pi^2(\nu-2)/2\nu$ with degree of freedom $\nu = 7.3$. Kinney and Dunson [23] showed that posterior distributions of g_i estimated with the Holmes and Held [21] and O'Brien and Dunson [22] algorithms are essentially completely indistinguishable given sufficient numbers of MCMC samples. We outline the Gibbs sampler for Bayesian adverse subpopulation regression in the following steps:

a. Draw $\boldsymbol{\theta}_h$ and $\boldsymbol{\Sigma}_h$ from $\mathrm{NIW}_s\left(\theta_h,\Sigma_h|\widehat{\theta}_h,\widehat{\psi}_h,\widehat{\rho}_h,\widehat{\Sigma}_h\right)$, $h=1,2$, where

$$\begin{aligned}
\widehat{\theta}_h &= \tfrac{n_h}{n_h+\psi_0}\bar{\boldsymbol{y}}_h+\tfrac{\psi_0}{n_h+\psi_0}\theta_0,\\
\widehat{\psi}_h &= n_h+\psi_0,\\
\widehat{\rho}_h &= n_h+\rho_0,\\
\widehat{\Sigma}_h &= \Sigma_0+\boldsymbol{S}_h+\tfrac{n_h}{1+n_h\psi_0^{-1}}\left(\bar{\boldsymbol{y}}_h-\theta_0\right)\left(\bar{\boldsymbol{y}}_h-\theta_0\right)',
\end{aligned}$$

with $n_1=\sum_{i=1}^n I(z_i=1)$, $\bar{\boldsymbol{y}}_1=\frac{1}{n_1}\sum_{i:z_i=1}\boldsymbol{y}_i$ and $\boldsymbol{S}_1=\sum_{i:z_i=1}\left(\boldsymbol{y}_i-\bar{\boldsymbol{y}}_1\right)\left(y_i-\bar{\boldsymbol{y}}_1\right)'$; $n_2=\sum_{i=1}^n I(z_i=0)$, $\bar{\boldsymbol{y}}_2=\frac{1}{2}\sum_{i:z_i=0}\boldsymbol{y}_i$ and $\boldsymbol{S}_2=\sum_{i:z_i=0}\left(\boldsymbol{y}_i-\bar{\boldsymbol{y}}_2\right)\left(\boldsymbol{y}_i-\bar{\boldsymbol{y}}_2\right)'$.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

b. Impute component indicator z_i from the conditional Bernoulli distribution by setting $z_i = 1$ with probability $\frac{\omega_1(\boldsymbol{x}_i)\,\mathrm{N}_s(\boldsymbol{y}_i|\theta_1,\Sigma_1)}{\sum_{h=1}^{2}\omega_h(\boldsymbol{x}_i)\,\mathrm{N}_s(\boldsymbol{y}_i|\theta_h,\Sigma_h)}$ for $i=1,2,\ldots,n$.

c. Augment auxiliary variable g_i, $i = 1, 2,\ldots, n$, sampled from the normal distribution $\mathrm{N}_1\left(g_i|\gamma+\boldsymbol{x}_i'\beta,\sigma^2/\phi_i\right)$, which is truncated above (below) by zero when $z_i = 0$ ($z_i = 1$).

d. Update $\boldsymbol{\phi}$ from Gamma $\left(\phi_i|\frac{\nu+1}{2},\frac{2}{\nu+\left(g_i-\gamma-\boldsymbol{x}_i'\beta\right)^2/\sigma^2}\right), i=1,2,\ldots,n$.

e. Update the regression coefficients $\beta^* = (\gamma, \beta')'$ given g_i, ϕ_i and other parameters, following the MCMC algorithm of MacLehose and Dunson [17].

Although we illustrate the algorithm focusing on the multiple shrinkage prior, the above algorithm could be easily modified for different shrinkage priors of β by using the corresponding sampling algorithm in the step (e). Moreover, we could combine the shrinkage and selection method (e.g. Lasso and elastic net) for logistic regression with step (a) and (b) to get a Monte Carlo EM algorithm [24] for the adverse subpopulation regression.

4. Simulations

In this section, we examine the performance of our approach along with alternative simple two-stage methods through simulation studies. The two-stage methods generate the binary indicators for the adverse subpopulation in the first stage. For example, indicators can be chosen as the true binary indicators (known for simulation data), estimated by the maximum a posteriori (MAP) allocation from a simple two component mixture model with the EM algorithm, specified by using preselected cutoffs, or identified by K-means clustering for two clusters. In the second stage, we fit both the standard logistic regression model without penalization (Logit-Standard) and the penalized logistic regression models with the shrinkage methods Lasso and elastic net (Logit-Lasso and Logit-ElasticNet) [25].

One hundred datasets were simulated to represent the data observed in the HPHB data set. In particular, we simulated 813 women with two response variables corresponding to infant birth weight and gestational age at delivery. We used maternal genotype for 100 SNPs as predictors that were fixed across the simulations, with only the response variable generation varying. By using the real SNP data, we obtained simulated datasets with a realistic dependence structure among the predictors, which is important given that the dependence structure can have a fundamental impact on variable selection and estimation performance. We simulated data under the model proposed in Section 2.1, with β chosen so that the first ten elements were set equal to 0.5 (corresponding to odds ratios for the minor allele of exp(0.5) = 1.65) and the remaining elements were set equal to zero (corresponding to no association with risk of falling in the adverse subpopulation for SNPs 11,…, 100). In addition, we considered another scenario with the first ten coefficients equal to 0.8 (exp(0.8) = 2.23) and others to zero. We simulated $\boldsymbol{y}_i$ based on expression (1), where the $\boldsymbol{\theta}_h$ and $\boldsymbol{\Sigma}_h$ were set equal to the maximum likelihood estimates from the HPHB dataset by using a two component latent class model without predictors.

We applied the proposed ASPR model with default priors specified in Section 2.2 and the two-stage methods to the simulated datasets. For ASPR model, we implemented the data augmentation Gibbs sampling algorithm outlined in Section 3. The sampling ran for 11,000 iterations, 1,000 iterations were discarded as a burn-in and every 10th sample was saved to thin the chain. The trace and autocorrelation plots of the posterior samples were examined to

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

determine the convergence. We used the cyclical coordinate descent algorithm by Friedman et al. [25] to find the Lasso or elastic net regularization paths for penalized logistic regression models. Ten-fold cross-validations were used to select the optimal shrinkage parameter which gave the minimal deviance.

We first compared the estimation performance measured by mean squared errors (MSEs) which were calculated for each coefficient across 100 datasets. Table 1 presents the averaged MSEs obtained across the first ten non-null coefficients and the remaining ninety null coefficients. The averaged MSEs of non-null coefficients by ASPR model are smaller than those given by the two-stage methods even with true indicators. More importantly, when the true indicators are unknown, a common scenario in practice, the two-stage methods rely on indicators generated either by classification algorithms (here the MAP allocation and K-means clustering) or by medical cutoffs will inflate the MSEs of non-null coefficients significantly. This observation indicates that if part of subjects are mis-classified in the two-stage methods, the coefficient estimates would be affected in the standard and penalized logistic regressions. A better solution is thus to avoid the classification or using cutoffs in the first place.

We also compared the variable selection ability for different methods. Based on substantive knowledge, we choose an interval null hypothesis as $H_{0j}: |\beta_j| \leq \epsilon$; we chose $\epsilon = 0.1$ in practice as log odds ratios within ±0.1 of zero are clearly not significant from a public health viewpoint in our motivating applications. We can then calculate the posterior probability of the alternative $H_{1j}: |\beta_j| * * * * \epsilon$ as $\tilde{\pi}_j = \sum_{g=1}^{G} I\left(|\beta_j^g| * * * * \epsilon\right)/G$ under the ASPR model, with β_j^g the gth MCMC draw from the posterior after discarding a burn-in and $I(\cdot)$ the indicator function. Larger values of $\tilde{\pi}_j$ suggest a greater weight of evidence that the jth predictor is significant from a public health viewpoint based on our elicited ϵ value. We further chose a threshold c for the $\tilde{\pi}_j$ with the goal of controlling the false discovery rate (FDR)[26]; selecting predictors having $\tilde{\pi}_j * * * * c$ as statistically significant adjusting for multiple comparisons. For a given c, we could estimate the posterior expected FDA [27] as $\sum_{j=1}^{p}\left(1-\tilde{\pi}_j\right) I\left(\tilde{\pi}_j * * * * c\right) / \sum_{j=1}^{p} I\left(\tilde{\pi}_j * * * * c\right)$, and chose a c as the smallest value such that the FDR was less than or equal to a desired level. For the two-stage methods, the predictor β_j was selected if its 90% confidence interval did not contain zero for the standard logistic regression model, or if the estimate $\widehat{\beta}_j \neq 0$ for the penalized logistic regression models.

Based on above selection criteria, we are able to assess the variable selection ability by using the averaged true positive rate (TPR) and false positive rate (FPR) for different methods across multiple datasets. For a given method and simulated dataset, TPR is defined as the number of predictors correctly selected as significant predictors divided by the number of true non-null predictors, and similarly FPR as number of predictors falsely selected as significant predictors divided by the number of true null predictors. As listed in Table 1, the ASPR model achieves higher TPR and lower FPR (controlling FDR at 50%) than the alternative methods for most scenarios. The only exception is the Logit-ElasticNet method in the (unrealistic) case in which the true sub-population indicators are assumed known. One may argue that it is arbitrary to control FDR at 50% in the ASPR approach and use 90% confidence intervals in the two-stage method, with different values leading to different comparisons of TPRs and FDRs. To obtain a more fair comparison, we varied the values of FDR, confidence interval value, and threshold ϵ' (previously assumed to be zero) for penalized logistic regression and calculated a series of TPRs and FDRs, which are plotted in Figure 1 as receiver operating curves (ROCs). It it clear that the ROC curves by

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

ASPR model stay above the other curves and are closer to the top left corner, indicating a better tradeoff between the TPRs and FPRs. In addition, we may calculate AUCs (area under the curves), which is approximated by applying the trapezoidal rule for a series of TPRs and FPRs. The AUCs by various methods under different scenarios are given in Table 1 with ASPR model showing the largest AUC. It suggests that in general the method by ASPR model achieves a better performance in variable selection than the alternative methods.

5. Application to Pregnancy Outcomes

There is increasing appreciation that interactions between the genetic and environmental factors contribute to adverse birth outcomes. In this analysis, we investigated the effects of maternal genotype and their interaction with lead and tobacco exposure on adverse birth outcomes in the infant, adjusting for several confounding factors. The dataset included 813 non-Hispanic black pregnant women who had singleton pregnancy and were less than 28 weeks gestation at the time of enrollment in HPHB study. Based on published studies, we focused on 31 candidate genes which are involved with maladaptive inflammatory regulation, maternal-fetal circulation, stress response, and environmental contaminant metabolism. For those candidate genes, we selected 275 haplotype tagging SNPs which effectively capture the genetic diversity of these genes. Please see Swamy et al. [28] and Ashley-Koch et al. [29] for further details on genotyping approaches. A detailed description of the SNPs and genes used in this analysis can be found in the Web Appendix. For the purpose of this analysis, we assumed that the risk for adverse birth outcomes would be associated with minor alleles. The value of each SNP was recorded as one if the mother carried the less frequent allele and as zero otherwise. In addition to the genetic data, we measured maternal blood levels of lead and cadmium. The interaction of lead and cadmium with the SNPs in relation to gestational age and birth weight is an important research question. We also controlled confounding factors by including them in the analysis. These confounders are mother's age, recorded as age group 18-20, 21-35 vs 35+; education, as no college vs some college; insurance, as private vs others; parity, as zero vs others; infant sex, as male vs female.

We fit the ASPR model with default priors after checking the composite normal assumption. The MCMC algorithm was run for 11,000 iterations with the first 1,000 iterations discarded as burn-in and every 10th remaining draw retained for analysis. The trace plots and the autocorrelation plots suggested the algorithm converged fast and mixes well. Table 2 presents the posterior summary for the component parameter $\boldsymbol{\theta}_h = (\theta_{h1}, \theta_{h2})'$ and $\Sigma_h = \begin{pmatrix} \Sigma_{h11} & \Sigma_{h12} \\ \Sigma_{h21} & \Sigma_{h22} \end{pmatrix}$. The table indicates that the healthy group in general has longer gestational age with higher birth weight, compared to the unhealthy group. In addition, the subjects in the healthy group are more homogeneous with the smaller values in the components of $\boldsymbol{\Sigma}_h$. Figure 2(a) shows shaded circles at the raw data points, with the darkness of the shading being proportional to the estimated posterior probability of allocation to the healthy subgroup. Standard cutoffs for defining preterm birth and low birth weight are also shown. Although most of the children that are in the preterm and low birth weight bin have small posterior probabilities of allocation to the healthy subgroup, there is substantial uncertainty around the boundary region in particular. This uncertainty is taken into the account by the ASPR model but not by the other two-stage approaches. Figure 2(b) plots the raw data and also demonstrates the contours of posterior predictive density based on the MCMC samples of ASPR model. There seems no systematic discrepancy between the observations and the contours of posterior predictive density, suggesting the ASPR model fits the data well.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Figure 3 shows the posterior means and 90% credible intervals for the coefficients of SNPs and their interactions with lead and cadmium. Although all the credible intervals cover zero, the posterior means of coefficients of SNPs rs2420620, rs10107390 and rs7017402, given in Table 3, stay clearly away from zero in the main effect (the top panel). This result is also supported by Web Figure 1, which plots $\tilde{\pi}_j = \sum_{g=1}^{G} I\left(|\beta_j^g| * * * * \epsilon\right)/G$, the posterior probability that jth predictor have an effect based on G samples of β_j. The larger the posterior probability $\tilde{\pi}_j$ suggests a stronger effect. When we take $\epsilon = 0.1$, coefficient of SNP rs2420620 stands out with the value 0.123 and coefficients of SNPs rs10107390 and rs7017402 with the value 0.098 and 0.092, while the majority of other posterior probabilities are around 0.05. SNP rs2420620 is located in gene GRK5 and SNPs rs10107390 and rs7017402 in gene NAT1. GRK5 is a member of the G protein-coupled receptor kinase family which is involved in regulating the activity levels of G protein-coupled receptors. Polymorphisms in GRK5 have been previously linked to risk for heart failure in African Americans [30]. The N-acetyltransferase genes (NAT1 and NAT2) are involved in the metabolism of xenobiotics. NAT1 has been shown to be expressed in early placenta [31]. We also analyzed the data using two-stage methods, in which the penalized logistic regressions were applied with the indicators generated by maximum a posteriori method. The results are presented in Table 3. Both penalized logistic regression methods identifies the SNP rs2420620 in GRK5 and SNP rs7017402 in NAT1, since their coefficients are not equal to zero. These two SNPs are also selected by the ASPR model. For ASPR model, additional SNPs in genes CR1 and IGF1 are interesting but not consistent across the three approaches. This may suggest that these are false positive results.

6. Discussion

In this article, we propose an adverse subpopulation regression model for investigating the relationship among multiple quantitative outcomes and high-dimensional predictors. Unlike the traditional two-stage methods, the proposed method does not require dichotomizing the continuous outcomes into binary indicators and thus avoids information loss. Two stage methods are outperformed with smaller MSE and higher area under the ROC curve for variable selection as demonstrated by the simulation studies. The new model has been applied to examine the effect of gene and environment interaction on adverse pregnancy outcomes. The results suggest the gene GRK5 and NAT1 may influence the occurrence of low birth weight and preterm delivery. Our focus is on defining a simple approach for assessing the impact of high-dimensional predictors on the risk of an adverse outcome when data consist of multiple quantitative traits. By using a two component mixture model, we can use a binary response logistic regression model, a framework that is very familiar to epidemiologists, to characterize non-linear genetic and environment associations with potentially complex multivariate quantitative traits. The proposed framework provides a parsimonious alternative to normal linear regression and logistic regressions based on preliminary categorization of quantitative traits, and should be able to detect associations that would not be detected with these methods. The proposed ASPR framework has purposefully been chosen to be a simple and parsimonious model that is easy to interpret and is scalable to high-dimensional predictors. We are aware that dealing with hundreds of thousands or millions of SNPs will cause serious computational burden for ASPR model. This limitation is shared by most of existing shrinkage methods. The popular solution is to reduce the dimensionality of the predictors first by using sure independence screening [32] or strong rules for discarding predictors [33]. We choose Normal-inverse-Wishart distribution as a conjugate prior for the parameters of multivariate normal distribution. The closed forms in the MCMC step a) lead to computational efficiency. With a large number of traits, it however may be computational expensive to evaluate the posterior Normal-inverse-Wishart density. For sake of simplicity and parsimony, we have avoided fully nonparametric

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Bayesian density regression models [34] that allow unknown numbers of latent classes. Although generalizations in such directions are conceptually straightforward, for each additional latent class, one introduces an additional p regression coefficients and corresponding hyperparameters, and difficult issues in identifiability, label switching and computational complexity arise. For our pregnancy outcome application, the ASPR model provides a good fit to the data, as illustrated by the posterior predictive density.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by Award Number R01ES017436 from the National Institute of Environmental Health Sciences, and by funding from the National Institutes of Health (5P2O-RR020782-O3) and the U.S. Environmental Protection Agency (RD-83329301-0). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences, the National Institutes of Health or the U.S. Environmental Protection Agency. We also thank the editor and two anonymous reviewers for very helpful comments and suggestions.

References

1. Miranda ML, Maxson P, Edwards S. Environmental contributions to disparities in pregnancy outcomes. Epidemiologic Reviews. 2009; 31(1):67. [PubMed: 19846592]
2. Boucher KM, Slattery ML, Berry TD, Quesenberry C, Anderson K. Statistical Methods in Epidemiology: A Comparison of Statistical Methods to Analyze Dose-Response and Trend Analysis in Epidemiologic Studies. Journal of Clinical Epidemiology. 1998; 51(12):1223–1233. [PubMed: 10086814]
3. Clogg, CC. Handbook of statistical modeling for the social and behavioral sciences. Springer; New York: 1995.
4. Lindsay, BG. Mixture models: theory, geometry, and applications. IMS; Beachwood: 1995.
5. McLachlan, GJ.; Peel, D. Finite mixture models. Wiley Blackwell; New York: 2000.
6. Gage TB. Classification of births by birth weight and gestational age: an application of multivariate mixture models. Annals of Human Biology. 2003; 30(5):589–604. [PubMed: 12959900]
7. Gage TB, Fang F, Stratton H. Modeling the pediatric paradox: Birth weight by gestational age. Biodemography and Social Biology. 2008; 54(1):95–112. [PubMed: 19350763]
8. Schwartz SL, Gelfand AE, Miranda ML. Joint Bayesian analysis of birthweight and censored gestational age using finite mixture models. Statistics in Medicine. 2010; 29(16):1710–1723. [PubMed: 20575047]
9. Stegle O, Parts L, Durbin R, Winn J. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. PLoS computational biology. 2010; 6(5):e1000 770.
10. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B. 1996; 58(1):267–288.
11. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B. 2005; 67(2):301–320.
12. Tipping ME. Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research. 2001; 1:211–244.
13. Wu TT, Chen YF, Hastie T, Sobel E, Lange K. Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics. 2009; 25(6):714–721. [PubMed: 19176549]
14. Cho, S.; Kim, H.; Oh, S.; Kim, K.; Park, T. BMC proceedings. Vol. vol. 3. BioMed Central Ltd; 2009. Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis; p. S25
15. Ayers KL, Cordell HJ. Snp selection in genome-wide and candidate gene studies via penalized logistic regression. Genetic Epidemiology. 2010; 34(8):879–891. [PubMed: 21104890]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

16. Park T, Casella G. The bayesian lasso. Journal of the American Statistical Association. 2008; 103(482):681–686.
17. MacLehose RF, Dunson DB. Bayesian Semiparametric Multiple Shrinkage. Biometrics. 2010; 66(2):455–462. [PubMed: 19508244]
18. Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology. 1995; 6(4):450. [PubMed: 7548361]
19. Sethuraman J. A constructive definition of Dirichlet priors. Statistica Sinica. 1994; 4(2):639–650.
20. Gelman A, Jakulin A, Grazia M. A weakly informative default prior distribution for logistic and other regression models. Annals of Applied Statistics. 2008; 2(4):1360–1383.
21. Holmes CC, Held L. Bayesian auxiliary variable models for binary and multinomial regression. Bayesian Analysis. 2006; 1(1):145–168.
22. O'Brien SM, Dunson DB. Bayesian multivariate logistic regression. Biometrics. 2004; 60(3):739–746. [PubMed: 15339297]
23. Kinney SK, Dunson DB. Fixed and random effects selection in linear and logistic models. Biometrics. 2007; 63(3):690–698. [PubMed: 17403104]
24. Wei GCG, Tanner MA. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. Journal of the American Statistical Association. 1990; 85(411): 699–704.
25. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010; 33(1):1. [PubMed: 20808728]
26. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B. 1995; 57(1):289–300.
27. Müller P, Parmigiani G, Robert C, Rousseau J. Optimal sample size for multiple testing. Journal of the American Statistical Association. 2004; 99(468):990–1001.
28. Swamy GK, Garrett ME, Miranda ML, Ashley-Koch AE. Maternal vitamin D receptor genetic variation contributes to infant birthweight among black mothers. American Journal of Medical Genetics Part A. 2011; 155(5):1264–1271. [PubMed: 21548019]
29. Ashley-Koch, AE.; Garrett, ME.; Edwards, S.; Quinn, KS.; Swamy, GK.; Miranda, ML. Maternal genetic variation in genes involved in the inflammatory response interact with measures of air pollution exposure to affect infant birthweight among non-Hispanic black women. 2011. Submitted
30. Liggett SB, Cresci S, Kelly RJ, Syed FM, Matkovich SJ, Hahn HS, Diwan A, Martini JS, Sparks L, Parekh RR, et al. A GRK5 polymorphism that inhibits β-adrenergic receptor signaling is protective in heart failure. Nature medicine. 2008; 14(5):510–517.
31. Smelt VA, Upton A, Adjaye J, Payton MA, Boukouvala S, Johnson N, Mardon HJ, Sim E. Expression of arylamine N-acetyltransferases in pre-term placentas and in human pre-implantation embryos. Human molecular genetics. 2000; 9(7):1101. [PubMed: 10767335]
32. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B. 2008; 70(5):849–911.
33. Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, Tibshirani RJ. Strong rules for discarding predictors in lasso-type problems. Journal of the Royal Statistical Society: Series B. 2012; 74(2):245–266.
34. Dunson DB, Pillai N, Park JH. Bayesian density regression. Journal of the Royal Statistical Society: Series B. 2007; 69(2):163–183.

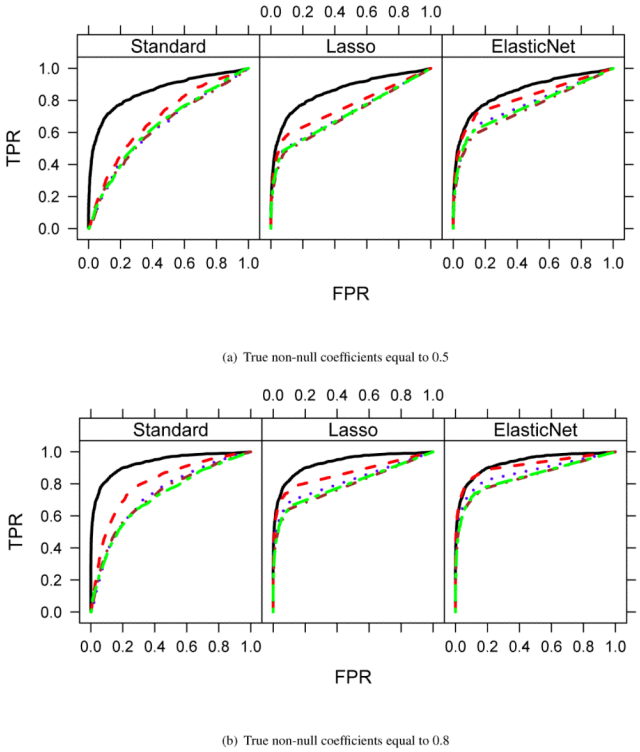

Figure 1.

Receiver Operating Characteristic (ROC) curves for different methods: —— by ASPR-MSP method, – – – by the two-stage methods with true indicators, · · · by the two-stage methods with subjects allocated by maximum a posteriori, – · – by the two-stage methods with indicators defined by cutoffs and — – — by the two-stage methods with indicators generated by K-means.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

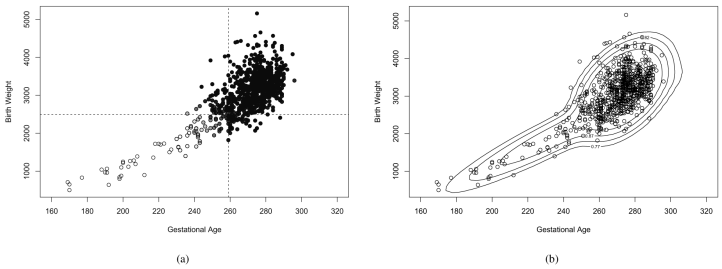

Figure 2.

Scatter plots of birth weight (grams) and gestational age (days) overlaid with (a) the posterior mean of allocation weights $\omega_2(\mathbf{x}_i)$ with the cutoffs (dash lines) at 257 days for the gestational age and 2500 grams for the birth weight; (b) posterior predictive density contours of observations (○) at the levels of 0.77, 0.82, 0.87, 0.92 and 0.97.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

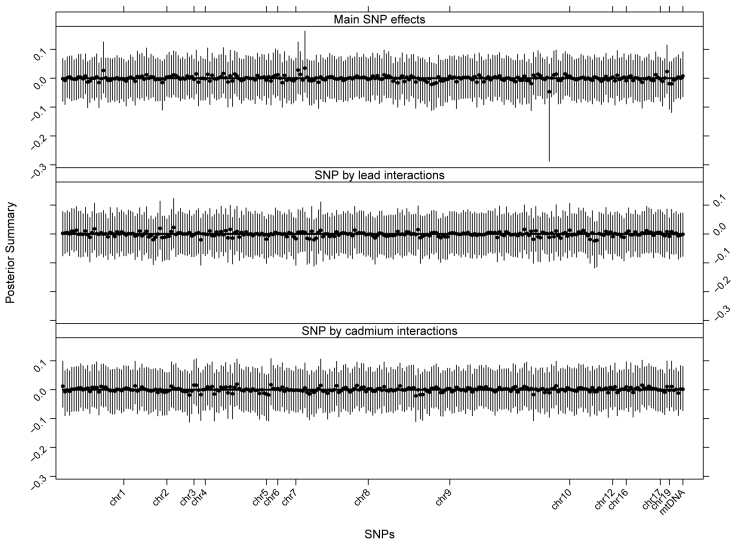

Figure 3.

Posterior mean and 90% credible interval for coefficients β in ASPR model. The coefficients are illustrated for the SNPs main effects and their interactions with lead and cadmium respectively.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Simulation results to compare coefficient estimation and variable selection by the ASPR model, the standard logistic regression models without penalization and the penalized standard logistic regression models with shrinkage methods for true non-null coefficients being equal to 0.5 and 0.8 (in parentheses), respectively.

	ASPR-MSP	Truth	Classification	Cutoff	K-means
		Logit-standard			
MSE					
Non-null/Null[a]	0.117/0.005 (0.223/0.007)	0.283/0.620 (0.630/0.692)	0.572/0.818 (1.503/1.065)	0.212/0.340 (0.280/0.476)	0.881/0.541 (1.593/0.555)
TPR[b]	0.642 (0.867)	0.428 (0.685)	0.378 (0.571)	0.346 (0.527)	0.360 (0.519)
FPR[c]	0.095 (0.170)	0.149 (0.165)	0.151 (0.172)	0.139 (0.144)	0.146 (0.154)
AUC[d]	0.858 (0.932)	0.686 (0.806)	0.635 (0.736)	0.632 (0.730)	0.643 (0.718)
		Logit-Lasso			
MSE					
Non-null/Null		0.146/0.004 (0.274/0.0050)	0.224/0.003 (0.519/0.004)	0.174/0.003 (0.385/0.003)	0.310/0.002 (0.751/0.004)
TPR		0.595 (0.790)	0.498 (0.697)	0.477 (0.642)	0.489 (0.662)
FPR		0.111 (0.140)	0.085 (0.113)	0.084 (0.097)	0.079 (0.107)
AUC		0.760 (0.867)	0.717 (0.820)	0.705 (0.792)	0.714 (0.800)
		Logit-ElasticNet			
MSE					
Non-null/Null		0.132/0.003 (0.236/0.005)	0.208/0.004 (0.475/0.005)	0.167/0.002 (0.359/0.003)	0.287/0.002 (0.704/0.004)
TPR		0.729 (0.896)	0.652 (0.829)	0.587 (0.772)	0.619 (0.778)
FPR		0.158 (0.235)	0.162 (0.213)	0.130 (0.175)	0.137 (0.181)
AUC		0.817 (0.915)	0.772 (0.872)	0.747 (0.844)	0.763 (0.847)

[a]The MSEs are presented for the non-null predictors whose coefficients are not equal to zero (known in the simulation) and null predictors separately.

[b]True positive rate (TPR) is defined as the number of predictors correctly selected (under a criterion) as significant ones divided by the number of non-null predictors.

[c]False positive rate (FPR) is defined as the number of predictors falsely selected (under a criterion) as significant ones divided by the number of null predictors.

[d]AUC stands for area under the receiver operating characteristic (ROC) curve, which consists of a series of TPRs and FPRs under various selection criteria.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Posterior summary for component parameters in the ASPR model. $\theta_{\cdot 1}$ and $\theta_{\cdot 2}$ are the average gestational age and birthweight in the subgroup with corresponding variance and covariance denoted as $\Sigma_{\cdot 11}$, $\Sigma_{\cdot 22}$ and $\Sigma_{\cdot 12}$.

Group	Parameters	Mean	SD	2.5%	50%	97.5%
Unhealthy	θ_{11}	237.52	4.54	228.06	237.67	245.93
	θ_{12}	2001.55	118.19	1751.22	2002.20	2219.74
	Σ_{111}	829.19	134.00	590.10	822.11	1122.02
	Σ_{112}	19322.84	3255.32	13671.92	19100.50	26389.50
	Σ_{122}	508531.02	88244.72	367278.25	499645.00	695598.00
Healthy	θ_{21}	273.25	0.40	272.49	273.25	274.05
	θ_{22}	3182.41	20.15	3141.49	3181.70	3223.60
	Σ_{211}	96.78	6.52	85.01	96.32	110.15
	Σ_{212}	2174.75	239.50	1719.19	2168.40	2628.22
	Σ_{222}	235640.32	13202.36	211617.25	235185.00	262059.25

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 3

List of SNPs which have largest estimated SNP effects for different methods.

ASPR-MSP				Logit-Lasso				Logit-ElasticNet			
SNP ID	Chr	Gene	Estimate	SNP ID	Chr	Gene	Estimate	SNP ID	Chr	Gene	Estimate
rs2420620	10	GRK5	−0.048	rs2420620	10	GRK5	−0.114	rs2420620	10	GRK5	−0.065
rs10107390	8	NAT1	0.035	rs7017402	8	NAT1	0.104	rs7017402	8	NAT1	0.059
rs7017402	8	NAT1	0.028								
rs4844599	1	CR1	0.026								
rs5742629	12	IGF1*Lead	−0.024								