Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

HHS Public Access
Author manuscript
Stat Methods Med Res. Author manuscript; available in PMC 2017 August 01.

Published in final edited form as:
Stat Methods Med Res. 2016 August ; 25(4): 1736–1752. doi:10.1177/0962280213499679.

Variable selection in semi-parametric models

Hongmei Zhang[1], **Arnab Maity**[2], **Hasan Arshad**[3,4], **John Holloway**[5], and **Wilfried Karmaus**[1]

[1]Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, TN, USA

[2]Department of Statistics, North Carolina State University, NC, USA

[3]The David Hide Asthma and Allergy Research Center, St. Marys Hospital, Isle of Wight, UK

[4]Allergy and Clinical Immunology, University of Southampton, Southampton, UK

[5]Faculty of Medicine, University of Southampton, Southampton, UK

Abstract

We propose Bayesian variable selection methods in semi-parametric models in the framework of partially linear Gaussian and problit regressions. Reproducing kernels are utilized to evaluate possibly non-linear joint effect of a set of variables. Indicator variables are introduced into the reproducing kernels for the inclusion or exclusion of a variable. Different scenarios based on posterior probabilities of including a variable are proposed to select important variables. Simulations are used to demonstrate and evaluate the methods. It was found that the proposed methods can efficiently select the correct variables regardless of the feature of the effects, linear or non-linear in an unknown form. The proposed methods are applied to two real data sets to identify cytosine phosphate guanine methylation sites associated with maternal smoking and cytosine phosphate guanine sites associated with cotinine levels with creatinine levels adjusted. The selected methylation sites have the potential to advance our understanding of the underlying mechanism for the impact of smoking exposure on health outcomes, and consequently benefit medical research in disease intervention.

Keywords

Bayesian methods; Gaussian kernel; non-linear effects; partially linear regression; probit regression; reproducing kernel; variable selection

1 Introduction

The work proposed in this article was motivated by an epidemiological project aimed to choose important epigenetic variants (predictor variables) potentially associated with smoking exposures (outcome) including exposure to maternal smoking and posnatal

Corresponding author: Hongmei Zhang, Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC 29208, USA. hzhang@sc.edu.

Conflicts of Interest

The authors declare that there is no conflict of interest.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

smoking exposure. The epigenetic variants are represented by deoxyribonucleic acid (DNA) methylation sets of cytosine phosphate guanine (CpG) sites. DNA cytosine methylation plays a critical role in modulating the transcriptional potential of the genome and may influence the development of complex human diseases.[1] Maternal smoking during pregnancy and postnatal smoking exposure are important risk factors for adverse health outcomes in children including cancer and respiratory illnesses such as asthma.[2–5] The underlying mechanisms for the diverse impacts of smoking exposure may involve epigenetic modifications such as DNA methylation.[6] Identifying important methylation sites from a pool of candidates possibly associated with smoking exposure will advance our understanding of the underlying mechanisms and consequently benefit medical research in disease intervention.

Existing variable selection methods in general are not applicable to select variables such as genetic or epigenetic variants (GEVs). Most methods in the Frequentist framework were proposed for parametric linear models by use of penalty functions.[7–9] Recently, some methods for feature selections in non-linear models were developed.[10,11] These methods generally built on splines or Taylor series expansions and may have difficulty in accommodating a large number of predictors and describing complex interaction effects. However, in genetic or epigenetic studies, very often, the number of possible predictors in the candidate pool is not small due to biological uncertainty. It is also considered that the effects of variants on the outcome are not linear and can be in any unknown form. Genes or epigenes (genes associated with epigenetic variants) do not necessarily function individually; rather, they work in concert with others to manifest a disease condition. Furthermore, these approaches are not appropriate for discrete variables such as single nucleotide polymorphism (SNP) genotypes. In the area of machine learning, methods of variable selection in semi-parametric models constructed by reproducing kernels have been discussed,[12] although it is also limited to continuous variables and requires intensive computing. Similarly, in the Bayesian framework, most variable selection methods are built upon model selections in parametric linear models. These methods utilize indicator variables for variable inclusion,[13–16] Zellner's *g*-prior controlling the importance of variables,[17–19] or Bayes factors comparing posterior probabilities between models.[20] Bayesian variable selection methods in semi-parametric models are rather limited. In the area of genetic or epigenetic studies, this type of methods is particularly appealing simply because researchers can embed prior knowledge of genetic or epigenetic factors into the selection process in order to obtain more meaningful selection results.

In this article, under a Bayesian framework, we propose a variable selection method built into reproducing kernels to evaluate the effect of a set of variables (e.g. GEVs). These variables can be continuous such as measurements of DNA methylation or gene expression. They can also be discrete such as SNP genotypes. The evaluation of a set effect is built upon the method of set analysis[21] to capture the significance of a group of variants allowing (for possible) unknown non-linear effects. The set analysis has the ability to capture the overall contribution from a whole group of variants, which may involve convoluted unknown interactions between the variants. The result from this type of analysis, however, is influenced by the choice of candidate variants, and it can be misleading if variants included in the kernel lack proper justification.[22]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In our method, two types of statistical models with reproducing kernels included are considered, partially linear regressions and probit regressions. To select important variables, we introduce an indicator variable into the reproducing kernel for the inclusion of a variable in the kernel. Compared to the existing methods, the proposed method has the ability to choose important variables regardless of the effect forms, e.g. linear or non-linear with unknown complex interactions. It thus has strong potential for application in gene and epigenetic studies to detect potentially important genetic and/or epigenetic variants associated with health outcomes, but the general association trend is unknown. There are many other possible applications of this method including, for example, studies examining the effects of nutrition or physical activity on health outcomes.

The structure of the article is as follows. The modeling and the Bayesian framework are presented in Section 2 including the selection of prior distributions and a discussion on posterior distributions and sampling. Simulations are presented in Section 3, where different structures of variable effects are considered. We apply the methods to select important CpG sites of which DNA methylation is potentially associated with maternal smoking and cotinine levels. This is discussed in Section 4. Finally, we summarize our findings and propose future work in Section 5.

2 The statistical models

Suppose we observe a vector of responses $\boldsymbol{Y}_{n\times 1}$, a matrix of variables $\boldsymbol{g}_{n\times p}$ whose joint effect is of interest (e.g. DNA methylation in a pathway), and a covariate matrix $\boldsymbol{X}_{n\times p_0}$. Here n is the sample size, p is the number of variables of interest such as genetic or epigenetic variants, and p_0 is the number of covariates. Note that it is possible $p + p_0 > n$. We assume that the mean of the response is modeled as $E(Y_i|\boldsymbol{X}_i, \boldsymbol{g}_i) = f^{-1}\{\boldsymbol{X}_i\boldsymbol{\beta} + h(\boldsymbol{g}_i)\}$, where $h(\cdot)$ is an unknown function, and $\boldsymbol{\beta}_{p_0\times 1}$ describes the additive linear effects of p_0 covariates $\boldsymbol{X}$. Define $\boldsymbol{h}(\boldsymbol{g})_{n\times 1}$ to be a vector of unknown functions evaluating the joint effect of p variables $\boldsymbol{g}$ that is possibly non-linear and may involve complex interactions between $\boldsymbol{g}$; $\boldsymbol{h}(\boldsymbol{g})$ can be modeled parametrically or non-parametrically. Function $f(\cdot)$ is a known link function. For instance, $f(\cdot)$ being the identity function results in a partially linear model and the inverse of a probit function gives a probit regression model. Our goal is to select a set of important variables from $\boldsymbol{g}$ that have legitimate contributions to the joint effect and exclude variables with no contributions.

As noted above, we allow the p variables $\boldsymbol{g}$ to have a complex (interaction) effect on the response variable. In practice, this is particularly true among genes or epigenes functioning in the same pathway. To this end, we incorporate reproducing kernels into the modeling process in appreciation of their ability to describe any underlying unknown patterns and the ability of handling high-dimensional data with $p + p_0 > n$.[21,23] Specifically, we represent $h(\cdot)$ using a kernel function $K(\cdot, \cdot)$. By the Mercer's theorem,[24,25] under some regularity conditions, the kernel function $K(\cdot, \cdot)$ specifies a unique function space $\mathcal{H}$ spanned by a particular set of orthogonal basis functions. The orthogonality is defined with respect to L_2 norm. Following the Mercer's theorem, any function $h(\cdot)$ in the function space $\mathcal{H}$ can be represented as a linear combination of reproducing kernels,[25,26]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$h(\boldsymbol{g}_i)=\sum_{k=1}^{n}K(\boldsymbol{g}_i,\boldsymbol{g}_k)\alpha_k=\boldsymbol{K}'{}_i\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\alpha_k, k = 1, \ldots, n)'$ is a vector of unknown parameters and $\boldsymbol{K}_i'$ is the ith row of kernel matrix $\boldsymbol{K}$. Defining $h(\cdot)$ non-parametrically as above has two major advantages in that it can handle large number of covariates and can capture potentially complex interaction between variables $\boldsymbol{g}$ via the specified kernel function.

The kernel function $K(\cdot, \cdot)$ determines the space of functions used to approximate the function $h(\cdot)$. For instance, $K(\boldsymbol{g}_{\boldsymbol{i}},\boldsymbol{g}_{\boldsymbol{j}})=(1+\boldsymbol{g}_{\boldsymbol{i}}'\boldsymbol{g}_{\boldsymbol{j}})^{d}$ generates a space of functions $\mathcal{H}$ spanned by all possible dth order monomials of $\boldsymbol{g}$. It corresponds to models with dth-order polynomials including the cross product terms. Another example is the Gaussian kernel, which is a function of smoothness parameter ρ and we denote by $\boldsymbol{K}(\rho)$. The (i, j) th entry of the kernel matrix $\boldsymbol{K}(\rho)$ is defined as $k_{ij}(\rho) = \exp\{-\Sigma_m \| g_{im} - g_{jm}\|^2/\rho\}$, with $i, j = 1, \ldots, n$, $m = 1, \ldots, p$, where g_{im} is the measure of variable m of subject i. This kernel acts as a correlation matrix. The functionality of Gaussian kernels is similar to that of exponential and Laplacian kernels.[27,28] All these kernels are constructed for continuous variables. For discrete variables, a commonly used kernel is the IBS (identity by state) kernel, which is constructed based on the agreement between variables and usually used for genetic variants such as SNPs.[23] In this article, to ensure a clear presentation of the proposed methods, we take $\boldsymbol{g}$ to be continuous and adopt a Gaussian kernel because of its flexibility and its ability in modeling complex functions.[21] However, other kernels can be used as well. In addition, we emphasize that the methods can be easily extended to fit discrete variables.

Turning back to the Gaussian kernel, we note that different values of ρ with different sets of selected variables can result in the same $\boldsymbol{K}(\rho)$ and consequently the same likelihood. In the context of variable selection, we fix ρ at $\rho = \rho_0$ with ρ_0 being the value at the full model with all variable included, considering that unimportant variables do not significantly contribute to the joint effect of $\boldsymbol{g}$. We also examine the sensitivity of variable selection results with respect to the choice of ρ_0. In the following sections, we particularly discuss variable selections in the setting of two models, partially linear regressions and probit regressions, in that these are used most often in practice.

2.1 Partially linear regression model

Consider the following setting to evaluate the effect of a set of variables:

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{h}(\boldsymbol{g})+\boldsymbol{\varepsilon} \quad (1)$$

where $\boldsymbol{Y}$, $\boldsymbol{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{h}(\cdot)$ are defined as before. Random error $\boldsymbol{\varepsilon}$ is with dimension $n \times 1$ and we assume $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$, where I is an identity matrix. Hereinafter, to simplify the notation, we use $\boldsymbol{h}$ instead of $\boldsymbol{h}(\boldsymbol{g})$ without ambiguity that $\boldsymbol{h}$ evaluates the joint effect of $\boldsymbol{g}$. To select variables from $\boldsymbol{g}$, we introduce an indicator variable $\boldsymbol{\delta} = \{\delta_m | m = 1 \cdots, p\}$ into the kernel matrix with $\delta_m = 1$ denoting the inclusion of variable m and 0 otherwise. For instance, with $p = 3$, $\boldsymbol{\delta} = \{1, 1, 0\}$ means that the first two variables are selected. Accordingly, we update the notation of kernel matrix as $\boldsymbol{K}(\rho_0, \boldsymbol{\delta})$ with its th entry defined as (i, j)-th entry defined as

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$k_{ij}(\rho_0,\boldsymbol{\delta})=\exp\left\{-\sum_m \|(g_{im}-g_{jm})\delta_m\|^2/\rho_0\right\}$$

If variable m is excluded, then it will not appear in any entry of the kernel matrix. The idea of using indicator variables for the inclusion or exclusion of a variable has been applied in previous studies.[13] Selecting covariates $\boldsymbol{X}$ in linear models is not the focus of our work and interested readers are referred to the Bayesian or Frequentist methods discussed in the literature.[13,16,19,20,29,30] Assuming $\boldsymbol{K}(\rho_0, \boldsymbol{\delta})$ is positive definite, as shown in an earlier study,[21] parameter estimation in the semi-parametric model (1) based on penalized least squares is equivalent to the estimates from a linear mixed model with random effects $\boldsymbol{h} \sim N(\boldsymbol{0}, \tau\boldsymbol{K}(\rho_0, \boldsymbol{\delta}))$, where τ is an unknown variance component.

Denote by $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \sigma^2, \tau, \boldsymbol{\delta}\}$ a collection of parameters to be inferred. In the following sections, we discuss a fully Bayesian method to infer these parameters.

2.1.1 The prior distributions—

The prior distribution of $\boldsymbol{\beta}$ will be selected as normal distributions with vague hyper-parameters $\boldsymbol{\beta}\sim N(\boldsymbol{0},\sigma_\beta^2\boldsymbol{I})$, with σ_β^2 selected to be large to ensure a vague prior for $\boldsymbol{\beta}$.

The prior distributions for σ^2 and τ are chosen as inverse gamma distributions with vague hyper-parameters, $\sigma^2 \sim \mathrm{Inv}-\mathrm{Gam}(a_{\sigma^2}, b_{\sigma^2})$, and $\tau \sim \mathrm{Inv}-\mathrm{Gam}(a_\tau, b_\tau)$, where the shape and scale parameters a_{σ^2}, b_{σ^2}, a_τ, b_τ are chosen to be small. As noted in an earlier study,[31] cautions should be taken on the choice of these hyper-parameters. Note that our goal is to remove non-important variables, which are expected not to be associated with the response variable. Thus, τ measures the effect of a collection of important variables, if there are any.

The prior distribution of δ_m is assumed to be Bernoulli with parameter q_m. We can take $q_m = 0.5$ or assign a hyper-prior distribution to q_m such as Beta $(\eta, 1)$ with η known. Our simulations indicate that adopting a hyper-prior distribution for q_m does not necessarily improve our results. So far, all the prior distributions are chosen to be vague, which is aimed to draw inferences fully based on data. In some situations, prior knowledge may guide us to select more rigorous and informative prior distributions. For instance, if it is known that some variables are potentially important, then we can set $q_m > 0.5$ for those variables to indicate their importance a priori.

2.1.2 Posterior distributions and their calculations—

The joint posterior density is

$$p(\boldsymbol{\Theta}|\boldsymbol{Y}) \propto p(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\sigma}^2,\boldsymbol{\delta})p(\boldsymbol{\beta})p(\boldsymbol{\tau})p(\boldsymbol{\sigma}^2)\Pr(\boldsymbol{\delta}|\boldsymbol{q})p(\boldsymbol{q}) \quad (2)$$

where $\boldsymbol{\Theta} = \{\beta, \sigma^2, \tau, \boldsymbol{\delta}, \boldsymbol{q}\}$ denotes an expanded set of parameters and $\boldsymbol{q} = \{q_1, \ldots, q_p\}$. To differentiate between continuous and discrete random variables, "$Pr(\cdot)$" is used to denote probability mass functions for discrete random variables. We use the Gibbs sampler to sample from this joint posterior distribution.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

In the following presentation, $(\cdot)$ denotes the parameters and data to be conditioned on. The full conditional posterior of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}|(\cdot) \sim N(\Sigma_{\beta}/\sigma^2(\boldsymbol{Y} - \boldsymbol{h}), \Sigma_{\beta})$, where $\sum_{\beta}=\{\boldsymbol{X}'\boldsymbol{X}/\sigma^2+\boldsymbol{I}(\sigma_{\beta}^2)^{-1}\}^{-1}$, and that of $\boldsymbol{h}$ is $\boldsymbol{h}|(\cdot) \sim N(\Sigma_h(\boldsymbol{Y} - \boldsymbol{X\beta})/\sigma^2, \Sigma_h)$ with $\Sigma;_h = \{\boldsymbol{\tau}^{-1}\boldsymbol{K}^{-1} + \boldsymbol{\sigma}^2\boldsymbol{I}\}^{-1}$. The full conditional posterior distributions of $\boldsymbol{\sigma}^2$ and $\boldsymbol{\tau}$ are inverse gamma,

$$\sigma^2|(\cdot)\sim\mathrm{Inv-Gam}(n/2+a_{\sigma^2},\left\{(\boldsymbol{Y}-\boldsymbol{X\beta}-\boldsymbol{h})'(\boldsymbol{Y}-\boldsymbol{X\beta}-\boldsymbol{h})+2b_{\sigma^2}\right\}/2)$$
$$\tau|(\cdot)\sim\mathrm{Inv-Gam}(n/2+a_{\tau},(\boldsymbol{h}'\{\boldsymbol{K}(\rho_0,\boldsymbol{\delta})\}^{-1}\boldsymbol{h}+2b_{\tau})/2)$$

The full conditional posterior distribution of q_m is Beta $(\delta_m + \eta, 2 - \delta_m)$, where η is the prior distribution of q_m such that the prior mean of q_m is $\eta/(1 + \eta)$. The full conditional posterior distribution of δ_m is

$$\Pr(\delta_m|(\cdot)) \propto \exp\left\{-(\boldsymbol{Y}-\boldsymbol{X\beta}-\boldsymbol{h})'(\boldsymbol{Y}-\boldsymbol{X\beta}-\boldsymbol{h})/(2\sigma^2)-[\boldsymbol{h}'\{\boldsymbol{K}(\rho_0,\boldsymbol{\delta})\}^{-1}h]/(2\tau)\right\}\times|\boldsymbol{K}(\rho_0,\boldsymbol{\delta})|^{-1/2}q_m^{\delta_m}(1-q_m)^{1-\delta_m} \tag{3}$$

The possible values of δ_m are 1 and 0. Denote the right side of equation (3) by $cPr(\delta_m)$, with c being the proportion. Let a be $cPr(\delta_m = 0|(\cdot))$, proportional to the conditional posterior probability of $\delta_m = 0$ and b denote $cPr(\delta_m = 1|(\cdot))$, proportional to the conditional posterior probability of $\delta_m = 1$. Then the full conditional posterior distribution of δ_m is Bernoulli with parameter $b/(b + a)$ measuring the conditional posterior probability of including variable m in the kernel. Clearly, all the conditional posterior distributions are standard. It is worthy of note that this feature holds regardless of the choice of kernels or the continuity of $\boldsymbol{g}$. Extension to discrete $\boldsymbol{g}$ is thus expected to be straightforward, so is in the probit regression model to be discussed in the next section. To obtain posterior samples of $\boldsymbol{\Theta}$, we sequentially draw from the standard full conditional posterior distributions, which are expected to converge quickly. For the situation that $\boldsymbol{K}$ is singular, the conditional posterior of $\boldsymbol{h}$ and $\boldsymbol{\tau}$ do not exist. This can be solved by bypassing direct posterior sampling of $\boldsymbol{h}$; instead, we infer the parameters in the prior distribution of $\boldsymbol{h}$. By doing so, valid but non-standard posterior distributions of $\boldsymbol{\tau}$ and $\boldsymbol{\sigma}^2$ will be obtained and the Metropolis-Hastings algorithm has to be used to draw the posterior samples.

Determination of important variables: To conclude the importance of each variable, we summarize the posterior probabilities (posterior mean of $\boldsymbol{q}$) of including each variable in the model. To determine which variable should be kept, we apply the concept in scree plot to the posterior probabilities, calculated as the percentage of times that a variable is selected among a certain number of uncorrelated MCMC samples. Scree plots are often used in principal component analysis to determine the number of components, where a sharp decrease in eigenvalues indicates less importance for the rest of the components.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Analogously, in our application of scree plots, a sharp decrease in probabilities indicates less importance of the remaining variables. Variables identified by this rule are treated as the most important variables. In addition, a reference probability 0.50 will be used to identify a group of possibly important variables. This reference probability represents the expected frequency if a variable is chosen randomly.

2.2 Probit regression model

In this section, we consider probit regression models for binary outcomes. Like logistic regressions, probit regressions are commonly used in case–control studies to infer factor effects on the risk of disease. Denote by Z_i a binary 0/1 response on the ith observation and let $\boldsymbol{Z} = \{Z_1, \ldots, Z_n\}$ be the collection of response on n subjects. A generalized linear model with probit link function that links $\boldsymbol{Z}$ to the $\boldsymbol{g}$ variables and $\boldsymbol{X}$ covariates can be conveniently formulated through data augmentation via Gaussian latent variables. Define a latent variable Y_i with $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ such that

$$Z_i=1_{\{Y_i>0\}}, \quad \boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{h}+\boldsymbol{\varepsilon} \qquad (4)$$

where 1_A denotes the indicator function of the event A, and $\boldsymbol{h}$ and $\boldsymbol{\varepsilon}$ are defined as in partially linear regression models. Note that the second part of equation (4) is a partially linear model. To be consistent with the partially linear model discussed earlier, we keep using $\boldsymbol{Y}$ to denote response although $\boldsymbol{Y}$ is latent in this framework. The probability of $\boldsymbol{Z} = \boldsymbol{z}$ satisfies

$$\Pr(\boldsymbol{Z}=z|\beta,\tau,\sigma^2)=\int_{A(Z)}\left(\frac{1}{\sigma^2+\tau}\right)^{n/2}\left|\sum\nolimits_0(\rho_0,\boldsymbol{\delta})\right|^{-1/2}\exp\left\{-\frac{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'\sum\nolimits_0(\rho_0,\boldsymbol{\delta})^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})}{2(\sigma^2+\tau)}\right\}\mathrm{d}Y \qquad (5)$$

where $\boldsymbol{A}(\boldsymbol{Z}) = \{A(Z_1), \ldots, A(Z_n)\}$ with

$$\boldsymbol{A}(Z_i)=\begin{cases}(-\infty,0], & \text{if } Z_{\mathrm{i}}=0\\ (0,\infty), & \text{if } Z_{\mathrm{i}}=1\end{cases}$$

and $\Sigma_0(\rho_0, \boldsymbol{\delta})$ is an $n \times n$ matrix with (i, j)-th entry being $\tau/(\tau + \sigma^2)k_{ij}(\rho_0, \boldsymbol{\delta})$ when $i \neq j$, and 1 when $i = j$. Note that under the model given in equation (5), the parameters $(\boldsymbol{\beta}, \sigma^2, \tau)$ are not identifiable. It can be shown that for any positive constant a_0, the parameter vectors $(a_0\boldsymbol{\beta}, a_0^2\sigma^2, a_0^2\tau)$ will give the same likelihood. To avoid this problem, we fix σ^2 and τ at known values σ_0^2 and τ_0.[32] In this article, we take $\sigma_0 = 1$ and $\tau_0 = 0.8$. The selection of σ_0^2 and τ_0 in principle is arbitrary, but their ratios will influence the estimates of $\boldsymbol{\beta}$. Furthermore, since τ

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

measures the contribution of a set of variables, e.g. expression of genes or DNA methylation of different CpG sites, setting $\tau_0 = 0.8$ a priori assumes that the variable set has some effect. Our simulations indicate that choosing τ_0 small compared to σ_0 can possibly lead to under selection, which is likely due to the pre-assumed large random error (large σ_0) in $\boldsymbol{Y}$.

As in partially linear regressions, we define $\boldsymbol{\delta}$ as a vector of 1 or 0 indicating the inclusion/exclusion of a variable. The likelihood of $\boldsymbol{\beta}$, $\boldsymbol{\delta}$ is given as

$$\begin{gathered} L(\boldsymbol{\beta},\boldsymbol{\delta}|\boldsymbol{Z}) \propto p(\boldsymbol{Z}|\boldsymbol{\beta},\boldsymbol{\delta}) = \int p(\boldsymbol{Z},\, \boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\delta})\mathrm{d}\,Y \\ \propto \int_{\boldsymbol{A}(\boldsymbol{Z})} |\textstyle\sum_0|^{-1/2} \exp\{-(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'\{\textstyle\sum_0(\rho_0,\boldsymbol{\delta})\}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})/4\}\mathrm{d}\,Y \end{gathered} \qquad (6)$$

We assign the same prior distribution as in partially linear regressions to $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$, and to the hyper parameter $\boldsymbol{q}$ in the probability mass function of $\boldsymbol{\delta}$. The collection of parameters is denoted as $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{q}\}$.

2.2.1 The joint posterior distribution—To derive the joint posterior distribution of $\boldsymbol{\Theta}$, we start from equation (6). However, the set of integrations in equation (6) can slow down the whole estimating process and may bring in computing difficulty. To avoid this, instead of integrating out $\boldsymbol{Y}$, we formulate the posterior distribution as the joint distribution of $\boldsymbol{\Theta}$ and the latent variable $\boldsymbol{Y}$,[33,34]

$$\begin{gathered} p(\boldsymbol{\beta},\boldsymbol{\delta},\boldsymbol{q},\,\boldsymbol{Y}|\boldsymbol{Z}) \propto \Pr(\boldsymbol{Z}|\,\boldsymbol{Y},\boldsymbol{\beta},\boldsymbol{\delta})p(\,\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\delta})p(\boldsymbol{\beta})\Pr(\boldsymbol{\delta}|\boldsymbol{q})p(\boldsymbol{q}) \\ = I\{\,\boldsymbol{Y} \in \boldsymbol{A}(\boldsymbol{Z})\}p(\,\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\delta})p(\boldsymbol{\beta})\Pr(\boldsymbol{\delta}|\boldsymbol{q})p(\boldsymbol{q}) \\ p(\,\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\delta}) \propto \,|\textstyle\sum_0(\rho_0,\boldsymbol{\delta})|^{-1/2} \exp\left\{-(\,\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'\{\textstyle\sum_0(\rho_0,\boldsymbol{\delta})\}^{-1}(\,\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})/4\right\} \end{gathered} \qquad (7)$$

The second line in equation (7) is a result of equation (4).

2.2.2 Full conditional posterior distributions and their calculations—We apply the Gibbs sampler to the full conditional posterior distributions to draw posterior samples. From equation (7), the full conditional posterior distribution of $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}|\boldsymbol{\delta},\,\boldsymbol{Y},\boldsymbol{Z}) = p(\boldsymbol{\beta}|\boldsymbol{\delta},\,\boldsymbol{Y}) \propto \exp\left\{-(\,\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'\{\textstyle\sum_0(\rho_0,\boldsymbol{\delta})\}^{-1}(\,\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})/4-\beta'/\sigma_\beta^2\right\}$$

which is $N(\frac{1}{2}\boldsymbol{V}_0^{-1}\boldsymbol{X}'[\sum_0(\rho_0,\boldsymbol{\delta})]^{-1}\boldsymbol{Y}, \sigma_\beta^2 I+\frac{1}{2}\boldsymbol{X}'[\sum_0(\rho_0,\boldsymbol{\delta})]^{-1}\boldsymbol{X})$. The full conditional posterior distribution of Y_i is a truncated normal, given as

$$Y_i|Y_{(i)},\boldsymbol{\beta},\boldsymbol{\delta} \sim 1_{\{Y_i \in A(Z_i)\}} N(\beta'\boldsymbol{X}_i+V_i'\textstyle\sum_0(i)^{-1}(Y_{(i)}-\boldsymbol{X}_{(i)}\boldsymbol{\beta}), 2(1-V_i'\sum_0(i)^{-1}V_i))$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

where $Y_{(i)}$ is with ith observation removed, $\Sigma_0(i)^{-1}$ is $[\Sigma_0(\rho_0, \boldsymbol{\delta})]^{-1}$ with elements in the ith row and ith column removed, $\boldsymbol{X}_{(\boldsymbol{i})}$ is the matrix of $\boldsymbol{X}$ with the ith row removed, and V_i' is the ith row of $\Sigma_0(\rho_0, \boldsymbol{\delta})$ with its ith element removed.

The full conditional posterior distribution of $\boldsymbol{q}$ is only related to $\boldsymbol{\delta}$ and is the same as that in partially linear regressions. The derivation of the full conditional posterior of $\boldsymbol{\delta}$ also follows the same path as in Section 2.1.2. From equation (7), we have

$$\begin{aligned} & \Pr(\delta_m|(\cdot)) \propto p(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\delta})\Pr(\delta_m|q_m) \\ = & |\textstyle\sum_0(\rho_0,\boldsymbol{\delta})|^{-1/2}\exp\left\{-(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'\{\textstyle\sum_0(\rho_0,\boldsymbol{\delta})\}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})/4\right\} q_m^{\delta_m}(1-q_m)^{1-\delta_m} \end{aligned}$$

Then following the same procedure as in Section 2.1.2, we draw posterior samples of δ_m from a Bernoulli distribution. All the full conditional posterior distributions are in the family of standard distributions, which ensures an efficient sampling process. To obtain posterior samples of $\boldsymbol{\Theta}$, we apply the Gibbs sampler to sequentially draw from the full conditional posterior distributions.

3 Simulation studies

3.1 Simulation scenarios

We consider 500 Monte Carlo (MC) replicates each with sample size n. Each MC replicate includes 12 predictors (e.g. DNA methylation of different CpG sites), generated from uniform distributions with lower bound 0.0001 and upper bound $12/(2m)$, $m = 1, \ldots, 12$, plus one covariate X, generated from $N(0, 2^2)$. The random errors are assumed to be independently and identically distributed with $N(0, \sigma^2)$. We consider the following three types of models:

1. Model 1 (linear): $E(\mathbf{Y}_i|X_i, \boldsymbol{g}_i) = f^{-1}\{X_i + 3g_{i1} - 2.5g_{i2} + 3.5g_{i4}\}$.
2. Model 2 (quadratic): $E(\mathbf{Y}_i|X_i, \boldsymbol{g}_i) = f^{-1}\{X_i + 3(g_{i1} - g_{i2})^2 + 2g_{i3}\}$.
3. Model 3 (non-linear): $E(\mathbf{Y}_i|X_i, \boldsymbol{g}_i) = f^{-1}\{X_i + 3\cos(g_{i1} \times g_{i2}) + 2g_{i3}\}$.

In the above, $\boldsymbol{g}_i = \{g_{i1}, \ldots, g_{ip}\}$ are the measures of the predictors and $p = 12$. For partially linear regressions, we consider $n = 100$ and take $\sigma^2 = 0.5^2$. Recall that by using reproducing kernels, we are able to evaluate the joint effect of a set of predictors, which may result from main effects plus any unknown interactions between the predictors. Even if it was known that only additive main effects and two-way interactions were possibly present, with 12 predictors plus one covariate X, in a standard linear regression model, we will have 80 effect terms (one for X, 12 main effects, 66 interactions, and one intercept) plus one variance parameter in the model. For probit regressions, because of the use of binary data, we consider a relatively large sample size $n = 300$ and assume $\sigma^2 = 1.9^2$.

The literature of variable selection in semi-parametric models is rather limited. To demonstrate the performance of the method, we compare our method with a recently developed variable selection approach, the adaptive least absolute shrinkage and selection

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

operator (ALASSO).[9] ALASSO is applied to linear additive models and with the feature of enjoying the oracle property,[9] that is, the method will correctly select the model as if the correct submodel were known. ALASSO thus serves as a benchmark for models 1 and 2. The early developed variable selection method LASSO,[7] on which ALASSO is built, is also included in our comparison. To compare between different methods, the variables selected using the proposed methods are the most important variables identified using the rule given in Section 2.1.2. Our focus is on variable selection, and thus percentages of correct selection (all important variables and no inclusion of unimportant ones), under selection (a subset of important variables and no unimportant ones), and over selection (all important variables plus at least one unimportant variables), along with model size (the number of variables selected) are recorded. The method is coded in statistical computing package R and the programs along with instructions and sample data are available on the first author's website (http://www.sph.sc.edu/epid_bios/facultystaffdetails.php?ID=574).

3.2 Results

We run two chains with overdispersed starting values for each data set and the convergence is evaluated by comparing the between and within sequence variations as proposed by Gelman et al.[35] This diagnostic step is included in our R programs. Converged MCMC simulations are usually represented by well-mixed MCMC sequences from different chains. As an illustration, Figure 1 shows the time series for the two key parameters (β, τ) in the partially linear regression setting for model 3 for the first 2000 iterations. The sequences converge quickly within about 1200 iterations. The inference discussed below is based on MCMC samples from one chain after burn in. To ensure true convergence, we in total run 10,000 iterations and use the last 5000 to draw the inferences. Table 1 lists the selection results in partially linear regression models. For each data set, the parameter ρ_0 is estimated based on the full model and the posterior mean is used in the subsequent variable selection steps. Also listed in the table are the results from LASSO and ALASSO.

For linear additive effect models (Model 1), the three methods all perform reasonably well (Table 1). Since LASSO and adaptive LASSO are both for linear additive effect models, this result is expected. This observation also indicates that the selection method built into reproducing kernels has the ability to capture additive linear effects and select the truly important variables. Although the LASSO and adaptive LASSO also do well for Model 2 (Table 1), essentially still a linear model with additive effect, these two methods lose their power in Model 3. The proposed method performs much better. In Model 3, the adaptive LASSO severely under-selects variables, while the LASSO tends to over-select, which is consistent with previous findings.[36] These findings are not surprising. LASSO and adaptive LASSO were designed specifically for linear regression models, which fits nicely the linearity property in the first two models. In Model 3, the main effects of g_1 and g_2 are absent and the effect of their interaction is non-linear. Both LASSO and adaptive LASSO do poorly in Model 3, while the proposed method does roughly equally well in all the three models.

Table 2 lists the results of variable selection under probit regression models. Following the same direction for variable selection in partially linear regressions, we compare the results

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

with those from LASSO and adaptive LASSO in the framework of logistic regressions. It is clear that the proposed method has a better model selection results compared to those two methods. LASSO tends to select more variables, while adaptive LASSO severely under select important variables, consistent with the findings in the setting of partially linear models.

We also compared our approach with the traditional AIC- and BIC-based forward, backward, and step-wise selection methods, and our approach always performed better than these standard approaches (results not shown). In summary, the new method overall out-performs both LASSO-related methods and other standard approaches, and chooses variables effectively regardless of linear or non-linear complex variable effects. In terms of computing time, however, the proposed method is relatively slower compared to the competing methods, which is due to the nature of almost all simulation-based Bayesian methods.

The results in Tables 1 and 2 are drawn using the rule utilized in scree plots, which aims to identify the most important variables. It is possible that some variables are still of interest but not selected by using this approach. An alternative way is to utilize the reference probability of 0.5 to identify a pool of potentially important variables such that their posterior probabilities of being included in the reproducing kernel are greater than the reference probability. Figure 2 illustrates this approach using model 3 in the probit regression setting, which clearly shows that the first three variables are the most important ones by use of scree plot. However, the posterior probability of including variable 11 in the model is higher than 0.5, which indicates that we might want to treat variable 11 as a potentially important variable. We did not identify any strong correlation between variable 11 and the other three most important variables. Its selection could be due to complex interactions with the most important ones. Thus, this potentiality in practice is likely to be data specific, but in our simulations, it is random because this variable is not chosen often among all the 500 data replicates. If we want to use the reference probability of 0.5 to identify whether variable 11 is a truly important variable, permutations on the variables can be considered to empirically test its significance of being an important variable, although this approach may impose strong computing burden.

3.3 The influence of ρ_0

The value of ρ_0 in the above section is taken as the posterior mean of ρ based on the full model, that is, all variables are included in the reproducing kernel. In the following, we numerically evaluate the sensitivity of selection results to different choices of ρ_0. We consider additional three values of ρ_0. One choice is to set $\rho_0 = 1$, and the other two are either 10% smaller (ρ_s) or larger (ρ_l) than the estimated ρ_0 from the full model. We choose the first 100 data sets used in the above section from each scenario under each model setting (partially linear regressions and probit regressions) to examine the sensitivity to ρ_0. Empirical intervals at the level of 95% are derived based on estimates of ρ_0 from the 100 data sets.

Overall, the variable selection results are not substantially influenced by the choice of ρ_0 (Table 3). Except for model 1 in the setting of probit regressions, all the correctness

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

percentages based on $\rho_0 = 1$ are higher than 63% and close to the percentages from the other two choices of ρ_0 (ρ_l and ρ_s), indicating the robustness of the approach with respect to the value of ρ_0. Furthermore, for the three choices of ρ_0, the model sizes are all comparable and close to the truth.

We shall point out that for a given model, the value of ρ_0 affects the fit of the reproducing kernel. For instance, model 1 in the setting of linear regressions is a regular linear regression model. In this case, a reproducing kernel built upon first order polynomial kernels shall provide the best fit. Since a Gaussian kernel approaches to a first order polynomial kernel when ρ_0 goes to infinity, taking large values of ρ_0 in Gaussian kernels will provide a reasonable approximation. Our overall large estimates of ρ_0, shown by the 95% empirical intervals (third column of Table 3), reflect the effort of approximation. Furthermore, the choice of ρ_0 will influence the estimate of a variable set effect. Given that unimportant variables provide negligible effects on the outcome, we expect ρ_0 under the true model to be comparable to that under the full model. It is thus recommended that, if estimating the effect of the whole group of variables is also desired besides selecting the variables, we should use the estimate of ρ_0 under the full model to provide simultaneous results on variable selection and group effect estimate.

3.4 Handling high-dimensional data

Earlier simulations indicated the feasibility of the method in dealing with complicated interaction effects. By design, the method has the ability to handle high-dimensional data, in particular, data with large p and small n. To illustrate this, we use model 3 in the setting of partially linear regression. Model 3 is chosen because it represents a more realistic association common in many high-dimensional data, e.g. microarray gene expression data or DNA methylation data. Instead of 12 predictors as noted in Section 3.1, we simulate 100 data sets with each including 120 predictors and of sample size n=100. These 120 predictors are generated from uniform distributions with lower bound 0.0001 and upper bound 120/(2m), a range the same as that in Section 3.1.

We use the same sampling methods noted in earlier sections to infer the parameters and select the variables. Based on the 100 data sets, the percentages of correct selections and over selections are 96% and 4%, respectively, and there is no under selection. The average model size is 3.04, which is consistent with the high percentage of correctness. These findings are similar to those listed in Table 1. We also considered the probit regression model and observed a similar pattern as in Table 2 (results not shown). On the other hand, the competing methods, LASSO and adaptive LASSO, missed almost all the correct models. These findings demonstrate that the new methods not only have the potential to deal with complicated interaction effects but also have the ability to handle high-dimensional data with the number of variables larger than sample size.

4 Real data application

We apply the proposed methods to identify epigenetic factors potentially associated with environmental tobacco smoke exposure. The epigenetic factors considered in this application are DNA methylation of 12 CpG sites (columns 1 to 3 in Table 4). These 12 CpG sites are

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

chosen based on their potential association with asthma and maternal smoking drawn from our preliminary study and other studies.[37,38] There is evidence that DNA methylation is associated with maternal smoking,[38] a strong risk factor of asthma.[39] Furthermore, maternal smoking might be linked to early onset of offspring smoking.[40] Assuming a CpG site is stable once it is methylated as indicated in some recent studies,[38,41] it will be of interest to identify CpG sites whose methylation level is associated with maternal smoking during pregnancy and those whose methylation level is associated with tobacco smoke exposure in postnatal life. For the first case, we use the method designed for continuous outcomes, i.e. the setting of partially linear regressions, and use cotinine level as the outcome variable. Cotinine is an alkaloid detected in tobacco and has been used as a biomarker of smoke exposure.[42] In this model, we further include creatinine as an adjusting factor to adjust the effect of fluid intake on the urinary concentration of cotinine.[43] The sample size for the first application is 114. For the second application, we select the CpG sites in the setting of probit regressions and use maternal smoking status as the response. In total, 245 observations are available for this application.

For the selection of CpG sites based on cotinine levels, the estimate of ρ_0 is 745.245, implying a possible linear association of methylation with cotinine level. We run two chains with 10,000 iterations each to estimate the number of iterations needed for convergence. The sequences converged within about 800 iterations for β and τ, similar as those observed in the simulations. The inference given below is based on one chain of 10,000 iterations with 5000 iterations used as burn in to ensure true convergence. By using the rule in scree plots, we identified the most important CpG site, cg11924019, which is in the CYP1A1 gene. Additional seven CpG sites are identified with posterior probabilities larger than 0.5 and should be treated as important sites as well (Figure 3(a), column 4 in Table 4). We further apply the LASSO and ALASSO methods to select the CpG sites. Both methods give the same selection results with two CpG sites (cg05549655 and cg05575921) selected (columns 5 and 6 in Table 4). These two sites are selected by the proposed method as well. A further calculation on the correlations in methylation indicates that methylation of these two sites are highly correlated with the methylation of CpG sites selected using the reproducing kernel based method. Except for site cg11679455 showing a weak correlation with cg05549655 and cg05575921 ($|r|<0.2$), most other correlations are higher than 0.91 indicating strong collinearity. This possibly explains the under-selection from the LASSO and the ALASSO methods.

When selecting CpG sites based on maternal smoking status, ρ_0 in the probit model is estimated as 0.675. As expected, the convergence of the two Markov chains is observed around 1700 iterations, slower than that under the partial linear regression models. The results are from 5000 iterations after 5000 iterations for burn in. One most important CpG site is identified by the scree plot rule (cg05575921) and seven additional CpG sites are identified after comparing with the reference probability of 0.5 (Figure 3(b), column 7 in Table 4). The methods of LASSO and ALASSO are also applied and identified 10 and three CpG sites (last two columns in Table 4), respectively. Based on previous studies[36] and from the findings in our simulation studies in terms of overselection and underselection patterns of these two methods, we conclude that the selection results from the proposed method are closer to the truth.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Overall, the proposed methods seem to conclude a more accurate selection of CpG sites in both situations, continine level-based and maternal smoking-based. In addition, among the selected CpG sites using the proposed methods, three sites are unique to the selection process based on maternal smoking, three are unique based on cotinine levels, and five are in common. This indicates that exposure to maternal smoking and postnatal smoking (reflected by cotinine levels) may affect DNA methylation differently to some extent.

5 Summary

We present Bayesian methods for variable selection in semi-parametric models assuming possibly non-linear in an unknown form of associations between candidate variables and a response variable. The association is described using reproducing kernels, which allow linear or non-linear effects in any form. An indicator variable is introduced to the reproducing kernels to indicate the inclusion or exclusion of a variable. Two model settings are considered: partially linear regressions and probit regressions.

The method is demonstrated and evaluated through simulations. The simulation results show that the proposed method can efficiently identify the correct variables regardless of association patterns. We compare the methods with the LASSO and adaptive LASSO methods. In the simulations, we assumed the variables are not correlated. Thus, for regular linear regression models (models 1 and 2), the LASSO and adaptive LASSO give similar results to those from the proposed methods, but they become inferior when the variable effects are non-linear (model 3). When applying the methods to real data sets, the selection results from the proposed methods seem to be more reliable.

The proposed methods are easy to implement and expected to have quick convergence because all the full conditionals are standard distributions. On the other hand, as for all Bayesian methods, the inferences generally are drawn from MCMC simulations. Thus, in terms of computing speed, when the number of predictors is large, the Bayesian methods are relatively slower than the Frequentist approaches considered in this article (LASSO and ALASSO), especially under the probit regression models due to the computing time used to infer the latent variables $\boldsymbol{Z}$. However, the results from the proposed Bayesian methods are much improved compared to those from the Frequentist approaches. In addition, extending the two methods to discrete $\boldsymbol{g}$ variables is straightforward. We believe these advantages will benefit medical researchers to efficiently identify important risk factors that may have convoluted effects on a certain type of health outcome. Finally, our methods can be easily modified to fit other types of statistical models including log-linear models and models applied to survival data analysis. On the other hand, the methods have some limitations that warrant a discussion. Recall that the variable selection approaches are built upon the evaluation of an overall set effect measured through reproducing kernels. The selection of each important predictor is based on the evaluation of its contribution to the overall set effect on an outcome variable instead of each individual variable's direct effect on the outcome. In some situations, it may be desired to evaluate the direct effect of each selected variable, besides their overall contribution as a group. Furthermore, it is possible that the candidate variables are a mixture of categorical and continuous variables. The kernels normally applied are either suitable for categorical variables or continuous variables. Developing an

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

approach that has the ability to handle the mixture of two types of variables is our on-going work.

Acknowledgments

Funding

This project was supported by NIH grants R01AI091905 (to HZ, HA, JH, and WK) and R00ES017744 (AM).

References

1. Feinberg AP. Genome-scale approaches to the epigenetics of common human disease. Virchows Archiv. 2010; 456:13–21. [PubMed: 19844740]
2. Frischer T, Kuehr J, Meinert R, et al. Maternal smoking in early childhood: a risk factor for bronchial responsiveness to exercise in primary-school children. J Pediatr. 1992; 121:17–22. [PubMed: 1625083]
3. Jarvis D, Burney P. The epidemiology of allergic disease. BMJ. 1998; 316:607–610. [PubMed: 9518918]
4. Office of the Surgeon General. The health consequences of involuntary exposure to tobacco smoke: a report of the surgeon general. Rockville, MD: US Department of Health and Human Services, Public Health Service, Office of the Surgeon General; 2006.
5. Karmaus W, Dobai AL, Ogbuanu I, et al. Long-term effects of breastfeeding, maternal smoking during pregnancy, and recurrent lower respiratory tract infections on asthma in children. J Asthma. 2008; 45:688–695. [PubMed: 18951262]
6. Suter M, Ma J, Harris AS, et al. Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression. Epigenetics. 2011; 6:1284–1294. [PubMed: 21937876]
7. Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat Soc Ser B Methodol. 1996; 58:267–288.
8. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001; 96:1348–1360.
9. Zou H. The adaptive LASSO and its oracle properties. J Am Stat Assoc. 2006; 101:1418–1429.
10. Castle, JL.; Hendry, DF. Automatic selection for nonlinear models. In: Garnier, H.; Wang, L.; Jackman, T., editors. System identification, environmental modelling and control. London: Springer Verlag; 2010. p. 229-250.
11. Radchenko P, James GM. Variable selection using adaptive nonlinear interaction structures in high dimensions. J Am Stat Assoc. 2010; 104:1541–1553.
12. Rosasco, L.; Mosci, S.; Santoro, MS., et al. A regularization approach to nonlinear variable selection. In: Yee, W.; Teh; Titterington, DM., editors. Proceedings of the 13 international conference on artificial intelligence and statistics, Journal of Machine Learning Research – Workshop and Conference Proceedings Proceeding. Vol. 9. Chia Laguna Resort; Sardinia, Italy: 2010. p. 653-660.
13. George EI, McCulloch RE. Variable selection via Gibbs sampling. J Am Stat Assoc. 1993; 88:881–889.
14. George EI, McCulloch RE. Approaches for Bayesian variable selection. Stat Sini. 1997; 7:339–374.
15. George EI. The variable selection problem. J Am Stat Assoc. 2000; 95:1304–1308.
16. Ishwaran H, Rao JS. Spike and slab variable selection: Frequentist and Bayesian strategies. Ann Stat. 2005; 33:730–773.
17. Zellner, A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Goel, PK.; Zellner, A., editors. Bayesian inference and decision techniques: essays in honor of Bruno de Finetti. New York: Elsevier/North-Holland Elsevier Science Publishing Co; Amsterdam: North-Holland Publishing Co; 1986. p. 233-243.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

18. Smith M, Kohn R. Nonparametric regression using Bayesian variable selection. J Econometr. 1996; 75:317–343.

19. Liang F, Paulo R, Molina G, et al. Mixtures of *g* priors for Bayesian variable selection. J Am Stat Assoc. 2008; 103:410–423.

20. Maruyama Y, George EI. Fully Bayes factor with a generalized *g*-prior. Ann Stat. 2011; 39:2740–2765.

21. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007; 63:1079–1088. [PubMed: 18078480]

22. He H, Zhang H, Maity A, et al. Power of a reproducing kernel-based method for testing the joint effect of a set of single-nucleotide polymorphisms. Genetica. 2012; 40:421–427. [PubMed: 23180006]

23. Wu MC, Kraft P, Epstein MP, et al. Powerful SNP-set analysis for case-control genome-wide association studies. Am J Human Genet. 2010; 86:929–942. [PubMed: 20560208]

24. Mercer J. Functions of positive and negative type, and their connection with the theory of integral equations. Philos Trans R Soc Lond Ser A. 1909; 209:415–446.

25. Cristianini, N.; Shawe-Taylor, J. An introduction to support vector machines and other kernel-based learning methods. New York USA: Cambridge University Press; 2000.

26. Gianola D, Van Kaam JB. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics. 2008; 178:2289–2303. [PubMed: 18430950]

27. Shawe-Taylor, J.; Cristianini, N. Kernel methods for pattern analysis. Cambridge, UK; New York: Cambridge University Press; 2004.

28. Zhang H, Gan J. A reproducing kernel-based spatial model in Poisson regressions. Int J Biostat. 2012; 8:28.

29. Liang H, Li R. Variable selection for partially linear models with measurement errors. J Am Stat Assoc. 2009; 104:234–248. [PubMed: 20046976]

30. Ma Y, Li R. Variable selection in measurement error models. Bernoulli. 2010; 16:274–300. [PubMed: 20209020]

31. Gelman A. Prior distributions for variance parameters in hierarchical models. Bayesian Anal. 2006; 1:515–533.

32. Berrett C, Calder CA. Data augmentation strategies for the Bayesian spatial probit regression model. Comput Stat Data Anal. 2012; 56:478–490.

33. Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc. 1993; 88:669–679.

34. Chib S, Greenberg E. Analysis of multivariate probit models. Biometrika. 1998; 85:347–361.

35. Gelman, A.; Carlin, JB.; Stern, HS., et al. Bayesian data analysis. Boca Raton, Florida, USA: Chapman & Hall; 2003.

36. Horowitz, JL.; Huang, J. The adaptive LASSO under a generalized sparsity condition. Northwestern University; 2010. Manuscript

37. Chatterjee R, Batra J, Das S, et al. Genetic association of acidic mammalian chitinase with atopic asthma and serum total IgE levels. J Allergy Clin Immunol. 2008; 122:202–208. e7. [PubMed: 18602573]

38. Joubert BR, Håberg SE, Nilsen RM, et al. 450k epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environ Health Perspect. 2012; 120:1425–1431. [PubMed: 22851337]

39. Weitzman M, Gortmaker S, Walker DK, et al. Maternal smoking and childhood asthma. Pediatrics. 1990; 85:505–511. [PubMed: 2314963]

40. Huizink AC, Mulder EJ. Maternal smoking, drinking or cannabis use during pregnancy and neurobehavioral and cognitive functioning in human offspring. Neurosci Biobehav Rev. 2006; 30:24–41. [PubMed: 16095697]

41. Karmaus, W.; Zhang, H.; Holloway, JW., et al. Maternal smoking during pregnancy is associated with DNA methylation in female offspring at age 18 – results of an epigenome-wide scan. University of South Carolina; 2012. Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

42. Benowitz NL. Cotinine as a biomarker of environmental tobacco smoke exposure. Epidemiol Rev. 1996; 18:188–204. [PubMed: 9021312]

43. Fried PA, Perkins SL, Watkinson B, et al. Association between creatinine-adjusted and unadjusted urine cotinine values in children and the mother's report of exposure to environmental tobacco smoke. Clin Biochem. 1995; 28:415–420. [PubMed: 8521596]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

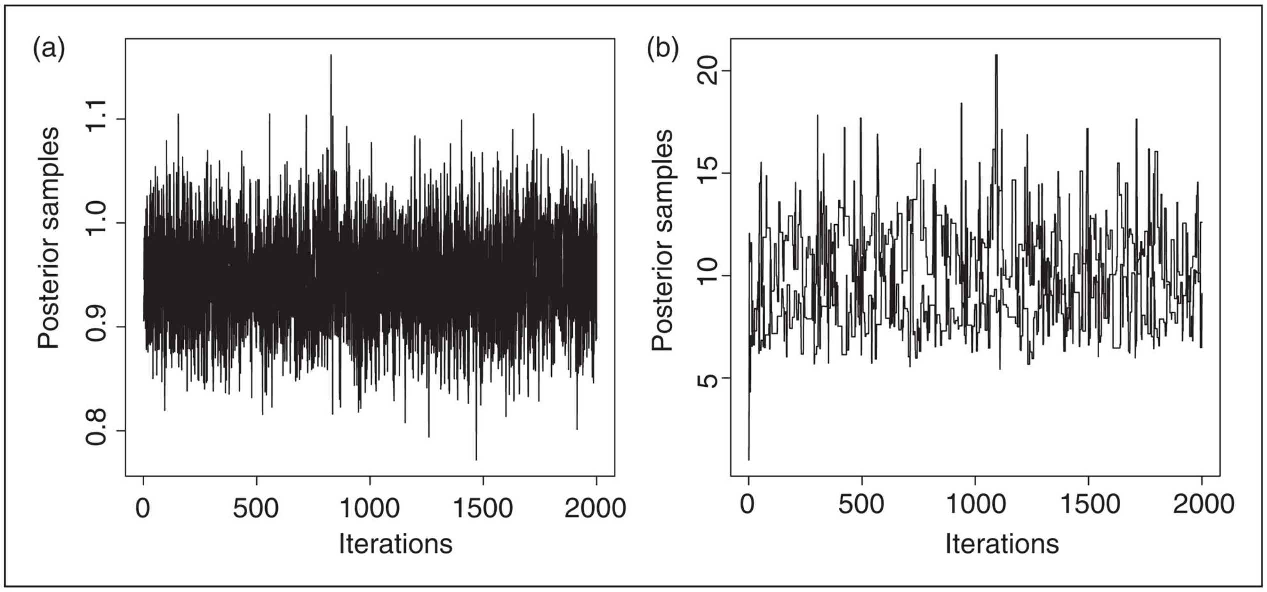

Figure 1.

Plot of posterior samples for the coefficient β (left) and set effect τ (right).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

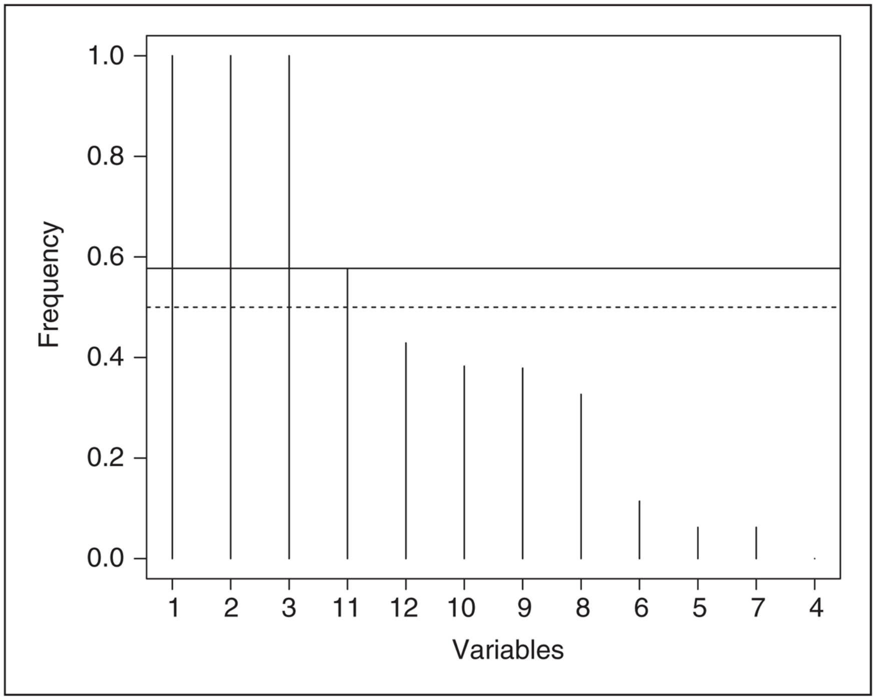

Figure 2.

Plot of posterior probability of each variable to be included in the model in the probit regression setting. The horizontal axis is for variable indices. The solid horizontal line indicates where the sharp decrease occurs in posterior probabilities. The dotted line is the 0.5 line representing the probability of selecting a variable randomly.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

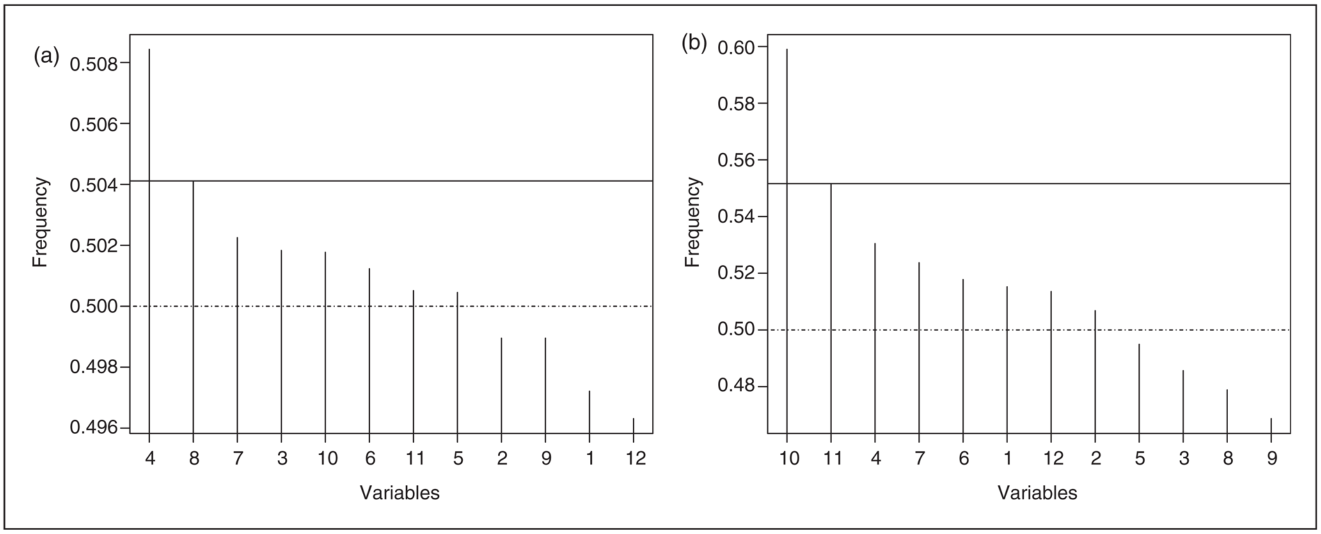

Figure 3.

Plot of posterior probability of each variable to be included in the model. (a) Cotinine level-based (partially linear regression setting). (b) Maternal smoking-based (probit regression setting). The horizontal axis is for variable indices. The solid horizontal line indicates where the sharp decrease occurs in posterior probabilities. The dotted line is the 0.5 line representing the probability of selecting a variable randomly.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Summary of variable selection for partially linear regressions.

Method	% Correctness	% Under selection	% Over selection	Model size
Model 1				
RKB	94.0	0.0	6.0	3.06
LASSO	100.0	0.0	0.0	3.00
ALASSO	82.2	0.0	17.8	3.25
Model 2				
RKB	88.0	0.0	12.0	3.12
LASSO	81.0	0.0	19.0	3.20
ALASSO	100.0	0.0	0.0	3.00
Model 3				
RKB	98.0	0.0	2.0	3.02
LASSO	0.0	33.0	67.0	4.90
ALASSO	0.0	100.0	0.0	1.00

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

Summary of variable selection for probit regressions.

Method	% Correctness	% Under selection	% Over selection	Model size
Model 1				
RKB	74.0	6.0	20.0	2.86
LASSO	15.0	0.0	85.0	5.65
ALASSO	10.0	90.0	0.0	0.45
Model 2				
RKB	56.0	44.0	0.0	2.59
LASSO	1.6	60.8	37.6	4.50
ALASSO	0.4	99.6	0.0	1.63
Model 3				
RKB	77.0	14.0	9.0	3.02
LASSO	2.4	37.0	60.6	4.3
ALASSO	0.0	100.0	0.0	1.01

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

Variable selection results for different choices of ρ_0.

Model	Choice of ρ_0	95% EI of ρ_0	% CS	% US	% OS	Model size
Partially linear regression setting						
1		(6.64, 2095.09)				
	1		67.0	33.0	0.0	2.60
	ρ_l		93.8	0	6.2	3.08
	ρ_s		100.0	0.0	0.0	3.00
2		(2.49, 451.10)				
	1		100.0	0.0	0.0	3.00
	ρ_l		94.0	0.0	6.0	3.06
	ρ_s		92.4	0.0	7.6	3.08
3		(2.34, 21.86)				
	1		85.0	0.6	14.4	3.10
	ρ_l		98.2	0	1.8	3.04
	ρ_s		100.0	0.0	0.0	3
Probit regression setting						
1		(0.99, 14.92)				
	1		34.8	60.6	4.61	2.41
	ρ_l		76.0	23.0	1.0	2.79
	ρ_s		74.6	25.4	0.0	2.74
2		(1.06, 2.70)				
	1		63.8	36.0	0.2	2.68
	ρ_l		55.0	44.4	0.6	2.61
	ρ_s		62.0	38.0	0.0	2.64
3		(0.91, 5.18)				
	1		85.6	9.0	5.4	2.98
	ρ_l		77.0	12.2	10.8	3.02
	ρ_s		80.4	13.0	6.6	3.07

ρ_S is 10% lower and ρ_I is 10% higher than the posterior mean based on the full model.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

EI: empirical interval of $\rho0$ inferred from the full model; CS: correct selection; US: under selection; OS: over selection.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4

Information of the 12 CpG sites, ordered by gene names and CpG ID, and the selection results (selected CpG sites are marked with “✓”).

			Partially linear models			Probit models		
CpG ID	**Reference gene name**	**Chromosome**	**RKB**	**LASSO**	**ALASSO**	**RKB**	**LASSO**	**ALASSO**
cg05575921	AHRR	5	✓	✓	✓	✓	✓	–
cg17924476	AHRR	5	✓	–	–	–	✓	–
cg21161138	AHRR	5	✓	–	–	✓	✓	✓
cg23067299	AHRR	5	–	–	–	–	✓	–
cg05549655	CYP1A1	15	✓	✓	✓	–	✓	–
cg11924019	CYP1A1	15	✓	–	–	✓	✓	–
cg17852385	CYP1A1	15	✓	–	–	✓	✓	✓
cg18092474	CYP1A1	15	✓	–	–	✓	–	–
cg25949550	CNTNAP2	7	–	–	–	✓	–	–
cg07989490	GATA3	10	–	–	–	✓	✓	–
cg11679455	GATA3	10	✓	–	–	–	✓	✓
cg09791102	IL4R	16	–	–	–	✓	✓	–

RKB: the reproducing kernel based method.