RESEARCH ARTICLE

Bayesian variable selection in additive partial linear models

S. Banerjee∗ and S. Ghosal

Department of Statistics, North Carolina State University, Raleigh, NC, USA

(Received 00 Month 20XX; final version received 00 Month 20XX)

Many studies in recent time include a large number of predictor variables, but typically only a few of the predictors have significant roles. Variable selection techniques have been developed using both non-Bayesian and Bayesian approaches. Additive partial linear models (APLM) provide a flexible yet manageable extension of linear models, where some variables can have non-linear effects. We develop a Bayesian method for variable selection for APLM by expanding the non-linear functions in a polynomial basis and introducing sparsity by allowing point masses in the prior distribution of regression coefficients. We address variable selection for both linear and non-linear parts. The nonsingular part of the prior is given by a Laplace or multivariate Laplace density depending on whether the predictor has only a linear effect or a general effect. However, instead of using Markov Chain Monte Carlo methods, which are extremely slow in high dimensional models, we use Laplace approximation technique around posterior mode, which can be identified with the group lasso solution. We conduct a simulation study and present real data analysis for a nutritional epidemiology study and prostate cancer data.

Keywords: Group lasso; Laplace approximation; variable selection.

Classification codes: 62G08; 62J02; 62J07

1. Introduction

Additive regression models are widely used in data analysis when the response variable and the predictors need not be related by a linear function. Fully non-parametric regression models have been studied well in this regard, but do not have the same degree of interpretability compared with linear models. Additive models act as a balance between the two, assuming that the contribution of each covariate to the mean response can be modeled by a smooth function of that covariate. Linear relationship may be still very important and many predictors may have only linear effects on the response variable. To incorporate more flexibility, additive partial linear models (APLM) are introduced, where a subset of the covariates are modeled to have linear effects, while the remaining covariates are modeled using additive non-parametric functions similar to that in fully non-parametric models. APLMs are preferred in applications where there is a prior belief that some covariates have strictly linear effect only, and also in situations where some covariates are not continuous. For instance, a binary predictor can have no more than a linear effect.

APLMs are generalization of partial linear models, where there is only one covariate having a non-parametric (i.e., unknown non-linear) effect. Statistical inference for APLMs have been well studied in literature; for example, see Stone [29], Opsomer and Ruppert [22]. Kernel based procedures (Liang et al. [14], Opsomer and Ruppert [23]), and spline-based procedures (Li [13]) have been developed in this regard. Liang et al. [14] studied APLMs when the linear covariate is measured with error.

∗Corresponding author: Email: sbanerj5@ncsu.edu
Variable selection in regression models are plentiful in the literature; see Bondell and Reich [3], Breiman [4], Fan and Li [7], George [9], Hwang et al. [12], Miller [20], Tibshirani [30], Zou [34], Zou and Hastie [35], for linear regression models, and Avalos et al. [1], Belitz and Lang [2], Chen [5], Curtis et al. [6], Gustafson [10], Huang et al. [11], Lin and Zhang [15], Marra and Wood [18], Meier et al. [19], Ravikumar et al. [24], Reich et al. [25], Shi and Tsai [26], Shively et al. [27], Wood et al. [31], for additive models. Variable selection in APLMs has been studied much recently. Liu et al. [16] developed a SCAD-based variable selection procedure using a spline based approximation for the non-parametric components. In contrast to a number of Bayesian methods for variable selection in additive models, to the best of our knowledge, there is no Bayesian variable selection method for APLMs available in the literature. Bayesian methods available in variable selection problems provide measures of model uncertainty. However, in high dimensional setting, Bayesian variable selection methods are computationally expensive, since commonly used procedures like Markov Chain Monte Carlo (MCMC) do not scale well in high dimension. Moreover, even when such a procedure is implemented, estimates of posterior probabilities from MCMC visits to different models can be extremely unreliable when the number of parameters is high. Curtis et al. [6] developed a Bayesian variable selection method in non-parametric additive regression models by approximating different model posterior probabilities using Laplace approximation, thus avoiding any time-consuming MCMC based procedures. The goal of the present paper is to extend the idea to APLMs, thus providing a fast Bayesian variable selection technique for the same.

We use a polynomial approximation for each non-parametric component and allow the model selection procedure to include only the linear effect, only the non-linear effect, or both, of those covariates whose effects are modeled non-parametrically. For covariates with strictly linear effects, our model selection procedure selects a subset of corresponding relevant covariates. For the linear effects, we use a Laplace prior on the corresponding regression coefficients, and for the non-linear effects, we use a multivariate Laplace distribution on the coefficients corresponding to each of the covariates. The priors on the coefficients (or a group of coefficients for non-linear effects) are assumed to be apriori independent. The non-linear effects are subject to a stronger penalty compared with the linear ones. The posterior mode can be identified as the group lasso (Yuan and Lin [33]), for which the ‘grplasso’ package in R is available. We use the group lasso solution to approximately compute various model probabilities and provide measures of model uncertainty. In addition to this, we also compute the median probability model, the model with those covariates whose marginal inclusion probability exceeds half.

We organize our paper as follows. In Section 2, we introduce the model and specify the prior distributions for the coefficients. In the next section, we develop the Laplace approximation for the posterior probabilities of different models, and also provide estimating procedures of the error variance and the penalty parameter for the group lasso. In Section 4, we illustrate our method through a simulation study. Finally, in Section 5, we provide two real data analysis based on a nutritional epidemiology study and prostate cancer data.

2. Model and prior specification

Consider an additive partial linear regression model

\[ Y = X^T \beta + \sum_{j=1}^{s} g_j(Z_j) + \varepsilon, \]

where \( X = (X_1, \ldots, X_p)^T \) and \( Z = (Z_1, \ldots, Z_s)^T \) are \( p + s \) predictors with response \( Y \), and random error term \( \varepsilon \sim N(0, \sigma^2) \). The set of predictors with only linear effects possible (such as binary predictors) are collected in \( X \), while predictors with both linear or non-linear
effects possible are collected in $Z$. The function $g_j(\cdot)$, $j = 1, \ldots, s$, can be approximated by a polynomial function of degree $k$ corresponding to each predictor in $Z_j$, that is,

$$g_j(Z_j) = \sum_{l=1}^k \eta_{jl} Z_{lj}, \quad j = 1, \ldots, s.$$  \hspace{1cm} (2)

Consider $n$ independent observations $Y_1, \ldots, Y_n$ with corresponding predictor variables given by $X_i, Z_i$, $1 \leq i \leq n$. Thus, the responses can be modeled as

$$Y_i = \sum_{j=1}^p \beta_j X_{ij} + \sum_{j=1}^s \sum_{l=1}^k \eta_{jl} Z_{lj}^i + \varepsilon_i, \quad \varepsilon_i \sim iid \sim N(0, \sigma^2), \quad 1 \leq i \leq n.$$  \hspace{1cm} (3)

The above model can be written in matrix-vector notation as

$$Y = X\beta + Z\eta + \varepsilon.$$  \hspace{1cm} (4)

We denote the $n \times (p + ks)$ matrix of covariates as

$$\Psi_{n \times (p + ks)} = (X : Z)$$  \hspace{1cm} (5)

where

$$X_{n \times p} = \begin{pmatrix} X_{11} \cdots X_{1p} \\ X_{21} \cdots X_{2p} \\ \vdots \\ X_{n1} \cdots X_{np} \end{pmatrix},$$  \hspace{1cm} (6)

and,

$$Z_{n \times ks} = \begin{pmatrix} Z_{11} \cdots Z_{1s} \cdots Z_{1s}^k \\ Z_{21} \cdots Z_{2s} \cdots Z_{2s}^k \\ \vdots \\ Z_{ns} \cdots Z_{ns}^k \end{pmatrix}.$$  \hspace{1cm} (7)

We denote the sub-matrix corresponding to the first $s$ columns of $Z$ as $Z^L$ and the rest as $Z^N$. The vector of co-efficients is $(\beta^T, \eta^T)^T$, where

$$\beta_{p \times 1} = (\beta_1, \ldots, \beta_p)^T,$$  \hspace{1cm} (8)

and,

$$\eta_{k \times 1} = (\eta_1, \ldots, \eta_{s1}, \eta_{1}^T, \ldots, \eta_{k-1}^T)^T,$$  \hspace{1cm} (9)

where

$$\eta_j = (\eta_{j2}, \ldots, \eta_{jk})^T, \quad 1 \leq j \leq s.$$  \hspace{1cm} (10)
We are interested in variable selection in the above model, and in particular allow for the possibility that only the linear or non-linear effect of a particular covariate is present in the model. For this, we define the indicator vector

$$\gamma = \left( \gamma_{i1}, \ldots, \gamma_{ip}, \gamma_{i1}, \ldots, \gamma_{is}, \gamma_{i1}, \ldots, \gamma_{is} \right)^T,$$

where

$$\gamma_{ij} = \begin{cases} 1, & \text{if } X_j \text{ is in the model, } 1 \leq j \leq p, \\ 0, & \text{otherwise} \end{cases}$$

and,

$$\gamma_{ij} = \begin{cases} 1, & \text{if linear part of } Z_j \text{ is in the model, } 1 \leq j \leq s, \\ 0, & \text{otherwise} \end{cases}.$$ 

and,

$$\gamma_{ij} = \begin{cases} 1, & \text{if non-linear part of } Z_j \text{ is in the model, } 1 \leq j \leq s, \\ 0, & \text{otherwise} \end{cases}.$$

We denote the vector of co-efficients selected by $$\gamma$$ as $$(\beta^T \gamma, \eta^T \gamma)$$ and the corresponding matrix of covariates as $$\Psi = (X \gamma, Z \gamma)$$. Thus the responses can be represented as

$$Y \sim N(X \gamma \beta + Z \gamma \eta, \sigma^2 I_n).$$

We put priors on the co-efficients $$\beta$$ and $$\eta$$ separately for the linear part and non-linear part. Let $$\Pi_A$$ stand for the indicator function of a set $$A$$. The prior on $$\beta_j$$, $$1 \leq j \leq p$$ is degenerate at 0 or has a Laplace density, depending on $$\gamma^L_j = 0$$ or $$\gamma^L_j = 1$$ respectively, that is,

$$p(\beta_j \mid \gamma) = (1 - \gamma^X_j) \Pi_{\{0\}}(\beta_j) + \gamma^X_j \frac{\lambda}{4\sigma^2} \exp \left\{ -\frac{\lambda}{2\sigma^2} |\beta_j| \right\}. \tag{12}$$

Similarly, the prior on $$\eta_{j1}$$, $$1 \leq j \leq s$$ is

$$p(\eta_{j1} \mid \gamma) = (1 - \gamma^L_j) \Pi_{\{0\}}(\beta_{j1}) + \gamma^L_j \frac{\lambda}{4\sigma^2} \exp \left\{ -\frac{\lambda}{2\sigma^2} |\eta_{j1}| \right\}. \tag{13}$$

For the co-efficients corresponding to the non-linear part, given by $$\eta_j$$, $$1 \leq j \leq s$$, we have point-mass at 0 corresponding to $$\gamma^N_j = 0$$ and a multivariate Laplace density corresponding to $$\gamma^N_j = 1$$, so that the prior density of $$\eta_j$$ given $$\gamma$$, denoted by $$p(\eta_j \mid \gamma)$$ is given by,

$$(1 - \gamma^N_j) \Pi_{\{0\}}(\eta_j) + \gamma^N_j \frac{\Gamma((k - 1)/2)}{2\pi^{(k-1)/2}\Gamma(k-1)} \left( \frac{\lambda \sqrt{k - 1}}{2\sigma^2} \right)^{k-1} \exp \left\{ -\frac{\lambda \sqrt{k - 1}}{2\sigma^2} \|\eta_j\| \right\}. \tag{14}$$

We have used the same tuning parameter $$\lambda$$ in the above priors, though it can be made more general by choosing different tuning parameters corresponding to different sets of coefficients. The idea behind the choice of such priors is to get the group lasso solution as the posterior mode corresponding to a normal likelihood. However, the computation of the group lasso solution available in $$\mathbb{R}$$ can be performed only when the tuning parameters are the same. Hence we resort
to the specification of a single tuning parameter $\lambda$ only.

Let $|\gamma|$ stand for the $\ell_1$-norm of $\gamma$, or equivalently the number of 1’s in $\gamma$, for a binary vector $\gamma$. The prior on the variable selection indicator $\gamma$ is given by

$$p(\gamma) \propto d_{\gamma^X} d_{\gamma^L} d_{\gamma^\mathbb{N}} q^{|\gamma|} (1-q)^{(p+2s)-|\gamma|},$$  

(15)

where $q \in (0,1)$, $d_{\gamma^X} = \text{det}(\mathbf{X}_X^T \mathbf{X}_X)$, $d_{\gamma^L} = \text{det}((\mathbf{Z}_L)^T \mathbf{Z}_L)$, and $d_{\gamma^\mathbb{N}} = \text{det}(\mathbf{K}_\gamma)$ is the determinant of the sub-matrix corresponding to the Kendall’s tau matrix $\mathbf{K}$ computed corresponding to the variables in $\mathbb{Z}^N$, that is, $K_\gamma = \{ (\tau(X_i, X_j); \gamma_i, \gamma_j = 1) \}$, where

$$\tau(X, Y) = \frac{1}{(n^2-n)/2} \sum_{i \neq j} \text{sign}(X_i - X_j)\text{sign}(Y_i - Y_j)$$

is the Kendall’s tau-coefficient between two variables $X$ and $Y$.

### 3. Posterior computation

The joint posterior density $p(\beta_\gamma, \eta_\gamma, \gamma | Y)$ for $\beta_\gamma$ and $\gamma$, given $Y$ is proportional to

$$p(Y|\beta_\gamma, \eta_\gamma, \gamma)p(\beta_\gamma, \eta_\gamma | \gamma)p(\gamma) = (1-q)^{p+2s} (2\pi \sigma^2)^{-n/2} d_{\gamma^X} d_{\gamma^L} d_{\gamma^\mathbb{N}} \left( \frac{q}{2(1-q)} \right)^{|\gamma|} \left( \frac{\lambda}{2\sigma^2} \right)^{|\gamma^X|+|\gamma^L|+(k-1)|\gamma^\mathbb{N}|}$$

$$\times \left( \frac{\Gamma((k-1)/2)}{\pi^{(k-1)/2} \Gamma(k-1)} \right)^{|\gamma^\mathbb{N}|} \exp \left\{ -\frac{1}{2\sigma^2} h(\beta_\gamma, \eta_\gamma) \right\},$$  

(16)

where

$$h(\beta_\gamma, \eta_\gamma) = \| Y - \mathbf{X}_\gamma \beta_\gamma - \mathbf{Z}_\gamma \eta_\gamma \|^2 + \lambda \left( \sum_{\{\gamma_i^X = 1:1 \leq j \leq p\}} |\beta_j| + \sum_{\{\gamma_i^L = 1:1 \leq j \leq s\}} |\eta_{j1}| \right)$$

$$+ \lambda \sqrt{k-1} \sum_{j: \gamma_j^\mathbb{N} = 1} ||\eta_j||$$  

(17)

The marginal posterior probability for model $\gamma$ can be obtained by integrating out $\beta_\gamma$ and $\eta_\gamma$, that is,

$$p(\gamma|Y) \propto C(Y) B(\gamma) \int_{\mathbb{R}^{m_\gamma}} \exp \left\{ -\frac{1}{2\sigma^2} h(\beta_\gamma, \eta_\gamma) \right\} d\beta_\gamma d\eta_\gamma,$$  

(18)

with

$$m_\gamma = |\gamma^X| + |\gamma^L| + (k-1)|\gamma^\mathbb{N}|,$$

$$C(Y) = (1-q)^{p+2s} (2\pi \sigma^2)^{-n/2},$$

$$B(\gamma) = d_{\gamma^X} d_{\gamma^L} d_{\gamma^\mathbb{N}} \left( \frac{q}{2(1-q)} \right)^{|\gamma|} \left( \frac{\lambda}{2\sigma^2} \right)^{m_\gamma} \left( \frac{\Gamma((k-1)/2)}{\pi^{(k-1)/2} \Gamma(k-1)} \right)^{|\gamma^\mathbb{N}|}.$$  

(19)
The integral in (18) can be approximated using the Laplace's approximation. Let \((\beta^{T*}_\gamma, \eta^{T*}_\gamma)^T\) denote the group lasso solution, that is,

\[
(\beta^{T*}_\gamma, \eta^{T*}_\gamma)^T = \arg\min_{\beta, \eta} h(\beta, \eta).
\]

(20)

Put \(u = (\beta^{T*}_\gamma, \eta^{T*}_\gamma)^T - (\beta^{T*}_\gamma, \eta^{T*}_\gamma)^T\). Substituting this quantity into (18) gives the expression

\[
C(Y)B(\gamma) \exp \left\{-\frac{1}{2}\sigma^2 h(\beta^{*}_\gamma, \eta^{*}_\gamma)\right\} \int_{\mathbb{R}^{m_\gamma}} \exp \left\{-\frac{1}{2} f(u)\right\} du,
\]

(21)

where,

\[
f(u) = \frac{1}{\sigma^2} \left[ \|\Psi \gamma u\|^2 - 2u^T \Psi \gamma Y^* + \lambda \sum_{\{\gamma_j = 1: 1 \leq j \leq p\}} (|\beta_j^* + u_j| - |\beta_j^*|) 
+ \lambda \sum_{\{\gamma_j = 1: 1 \leq j \leq s\}} (|\eta_{j1}^* + u_{j1}| - |\eta_{j1}^*|) 
+ \lambda \sqrt{k-1} \sum_{j: \gamma_j^N = 1} \left( \|\eta_j^* + u_j\| - \|\eta_j^*\| \right) \right].
\]

(22)

Clearly \(f(u)\) is minimized at \(u = 0\) by definition, and

\[
\frac{\partial f(u)}{\partial u^T} \bigg|_{u=0} = \frac{1}{\sigma^2} (2\Psi \gamma^T \Psi \gamma + \lambda A_\gamma),
\]

(23)

where the \(m_\gamma \times m_\gamma\) matrix \(A_\gamma\) is given by

\[
A_\gamma = \begin{bmatrix}
O & O \\
O & \sqrt{k-1}A^N_\gamma
\end{bmatrix},
\]

(24)

\(O\) stands for a zero matrix of appropriate order,

\[
A^N_\gamma = \begin{bmatrix}
-\eta_1^* \eta_1^T \frac{I_{k-1}}{\|\eta_1^*\|^2} + I_{k-1} & O_{12} & \cdots & O_{1t} \\
O_{21} & -\eta_2^* \eta_2^T \frac{I_{k-1}}{\|\eta_2^*\|^2} + I_{k-1} & \cdots & O_{2t} \\
\vdots & \vdots & \ddots & \vdots \\
O_{t1} & O_{t2} & \cdots & -\eta_t^* \eta_t^T \frac{I_{k-1}}{\|\eta_t^*\|^2} + I_{k-1}
\end{bmatrix},
\]

(25)

\(O_{ij}\) is the zero matrix corresponding to the variables in the \(i\)th row and \(j\)th column selected by \(\gamma^N\), and \(t = |\gamma^N|\).

Note that for a fixed model \(\gamma\), the function \(f(u)\) is differentiable at \(u = 0\) only if the group lasso solution for all the coefficients corresponding to the predictors included in the model as indicated by \(\gamma\) are non-zero. These models are referred to as regular models. We approximate the model posterior probabilities for the regular models only.

The above equations can be used to apply Laplace approximation to the quantity in (21), which
gives

\[ p(\gamma | Y) \propto C(Y)B(\gamma) \exp \left\{ -\frac{1}{2\sigma^2} h(\beta^*_\gamma, \eta^*_\gamma) \right\} \]

\[ \times \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2} f(u) \right\} du \]

\[ \approx C(Y)B(\gamma) \exp \left\{ -\frac{1}{2\sigma^2} h(\beta^*_\gamma, \eta^*_\gamma) \right\} \]

\[ \times \exp \left\{ -\frac{1}{2} f(0) \right\} \left(2\pi\right)^{m_\gamma/2} \frac{1}{\sigma^2} \frac{1}{2} \left| \frac{\partial f(0)}{\partial u \partial u^T} \right|^{-1/2} . \]

Substituting (23) in the above quantity, the marginal posterior probability \( p(\gamma | Y) \) for \( \gamma \) is approximately proportional to

\[ Q(\gamma | Y) = C(Y)B(\gamma) \exp \left\{ -\frac{1}{2\sigma^2} h(\beta^*_\gamma, \eta^*_\gamma) \right\} \]

\[ \times \left(2\pi\right)^{m_\gamma/2} \frac{1}{\sigma^2} \left( \Psi_\gamma^T \Psi_\gamma + \frac{\lambda}{2} A_\gamma \right) \left| \Psi_\gamma \right| \left| \gamma \right| N \left( \left| \gamma \right| \right) . \]

The posterior approximation is not valid if the model is non-regular, that is, if the group lasso solution is zero for a subset of variables already in the model. This issue of non-regularity also arises in variable selection for linear models and for non-parametric additive regression models; see Yuan and Lin [32], Curtis et al. [6]. Similar to Curtis et al. [6], the model posterior probabilities of the non-regular models are negligible compared with the regular ones when \( q \) is chosen to be small, and hence it suffices to compute the model posterior probabilities of the regular models only. We re-normalize the approximate posterior probabilities taking the regular models only in our computation.

### 3.1 Estimation of \( \lambda \) and \( \sigma^2 \)

The joint density of the response and the regression coefficient vector given the other model parameters is given by

\[ p(Y, \beta_\gamma, \eta_\gamma | \gamma, \lambda, \sigma^2) = (2\pi)\sigma^{-n/2} \left( \frac{1}{2} \right)^{m_\gamma} \frac{1}{\sigma^2} \left( \frac{\lambda}{\pi^{(k-1)/2}\Gamma(k-1)} \right)^{m_\gamma} \]

\[ \times \exp \left\{ -\frac{1}{2\sigma^2} h(\beta^*_\gamma, \eta^*_\gamma) \right\} \left( \frac{\Gamma((k-1)/2)}{\Gamma(k-1)} \right)^{m_\gamma} . \]

Integrating out \( \beta_\gamma \) and \( \eta_\gamma \), and using Laplace approximation, we get,

\[ p(Y | \gamma, \lambda, \sigma^2) \approx 2^{-(n+2)|\gamma|+m_\gamma/2} \pi^{-(n-m_\gamma+(k-1)|\gamma|)/2} \sigma^{-(n+m_\gamma)} \lambda^{m_\gamma} \]

\[ \times \left( \frac{\Gamma((k-1)/2)}{\Gamma(k-1)} \right)^{m_\gamma} \left| \Psi_\gamma \right| \left| \gamma \right| N \left( \left| \gamma \right| \right) . \]

\[ \times \exp \left\{ -\frac{1}{2\sigma^2} h(\beta^*_\gamma, \eta^*_\gamma) \right\} \left( \Psi_\gamma^T \Psi_\gamma + \frac{\lambda}{2} A_\gamma \right)^{-1/2} . \]

7
Denoting the model chosen by the group lasso solution for a fixed \( \lambda \) to be \( \hat{\gamma}_{\lambda} \), then maximizing (27) with respect to \( \sigma^2 \), an estimate of \( \sigma^2 \) is given by
\[
\hat{\sigma}^2_{\lambda} = h(\beta_{\hat{\gamma}_{\lambda}}^*, \eta_{\hat{\gamma}_{\lambda}}^*)/(n + m_{\gamma}). \tag{28}
\]
Similar to the penalized maximum likelihood criterion as in Curtis et al. [6], we can choose \( \lambda \) by penalizing negative 2 times the log-likelihood. Plugging in the estimate of \( \sigma^2 \) from equation (28) in equation (27), and taking negative 2 times the logarithm of (27) gives,
\[
r(\lambda) = (n + 2|\gamma| + m_{\gamma}) \log 2 + (n - m_{\gamma}) \log \pi - 2m_{\gamma} \log \lambda + |\gamma|^N \left( \log \Gamma((k - 1)/2) - \log \Gamma(k - 1) \right) + \log |\Psi_T \Psi_{\gamma} + \lambda/2 A_{\gamma}| \nonumber
\]
\[
+ (n + m_{\gamma}) \left[ \log \left( h(\beta_{\hat{\gamma}_{\lambda}}^*, \eta_{\hat{\gamma}_{\lambda}}^*)/(n + m_{\gamma}) \right) + 1 \right]. \tag{29}
\]
We estimate \( \lambda \) by minimizing
\[
r(\lambda) + m_{\gamma} \log n. \tag{30}
\]

4. Simulation study

To evaluate the performance of our method, we conduct two simulation studies using \( \mathbb{R} \). We simulate 100 data sets from a model having 2 strictly linear predictors and 2 predictors with both linear and non-linear effects, and refer to this model as the ‘true model’. In the first case, we analyze the data using a total of 10 strictly linear predictors and 10 predictors with both linear and non-linear effect. In the second case, we increase the above numbers to 100 in both set of predictors. We consider polynomial effects of degree 5 in the non-linear predictors in our simulations. More specifically, the model is given by,
\[
Y_i = X_i^T \beta + \sum_{j=1}^{s} g(Z_{ij}) + \varepsilon_i, i = 1, \ldots, n, \tag{31}
\]
where \( X_i \) is a \( p \)-dimensional vector of strictly linear predictors, and \( g(Z_{ij}) = \sum_{l=1}^{5} \eta_j Z_{ij}^l, j = 1, \ldots, s \). We take \( p = s = 10 \) in the first case and \( p = s = 100 \) in the second one. The true model is given by,
\[
Y_i = \sum_{l=1}^{5} 2X_{i1}^l + 4X_{i2}^l + Z_{i1} + 2Z_{i2} + 4Z_{i1}^2 + 3Z_{i1}^3 + 2Z_{i1}^4 + Z_{i1}^5 + \exp(5Z_{i2})/15 + \varepsilon_i, \tag{32}
\]
\( i = 1, \ldots, n \). We generate data of sample sizes \( n = 100, 200, 500 \) and the random error terms \( \varepsilon_i \) from a normal distribution with mean zero and variance \( \sigma^2 = 1 \). Each of the variables \( X_{ij} \) and \( Z_{ij} \) are generated from a standard uniform distribution, independent of each other. The value of \( q \) is chosen to be 0.5 throughout. Corresponding to each of the generated data sets, we calculate the approximate posterior probabilities of various models using our method. We record the first three models which have the highest posterior probabilities in order, and also the median probability model. We find the average number of true predictors which are not included in these models (denoted by ‘False neg’) and also the average number of predictors which are included in the model, but absent in the true model (denoted by ‘False pos’). Apart from these, we also find out the proportion of times the various models are identical with the true model (denoted by ‘True model’). We record the same for the model selected by the group lasso itself. To incorporate model uncertainty, we also perform Bayesian model averaging and compare the
Table 1. Table corresponding to independent predictors, $p = 10$, $s = 10$

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Pred error</th>
<th>False neg</th>
<th>False pos</th>
<th>True model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPPM</td>
<td>100</td>
<td>2.209(0.036)</td>
<td>0.190(0.039)</td>
<td>0.740(0.097)</td>
<td>0.400(0.049)</td>
</tr>
<tr>
<td>Median model</td>
<td>100</td>
<td>2.211(0.036)</td>
<td>0.200(0.040)</td>
<td>0.740(0.097)</td>
<td>0.400(0.049)</td>
</tr>
<tr>
<td>Group lasso</td>
<td>100</td>
<td>2.183(0.034)</td>
<td>0.010(0.010)</td>
<td>0.770(0.098)</td>
<td>0.530(0.050)</td>
</tr>
<tr>
<td>BMA</td>
<td>100</td>
<td>1.716(0.047)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPPM</td>
<td>200</td>
<td>1.632(0.019)</td>
<td>0.000(0.000)</td>
<td>0.580(0.117)</td>
<td>0.680(0.047)</td>
</tr>
<tr>
<td>Median model</td>
<td>200</td>
<td>1.632(0.019)</td>
<td>0.000(0.000)</td>
<td>0.580(0.117)</td>
<td>0.680(0.047)</td>
</tr>
<tr>
<td>Group lasso</td>
<td>200</td>
<td>1.632(0.019)</td>
<td>0.000(0.000)</td>
<td>0.580(0.117)</td>
<td>0.680(0.047)</td>
</tr>
<tr>
<td>BMA</td>
<td>200</td>
<td>1.556(0.034)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPPM</td>
<td>500</td>
<td>1.446(0.009)</td>
<td>0.000(0.000)</td>
<td>0.260(0.052)</td>
<td>0.780(0.041)</td>
</tr>
<tr>
<td>Median model</td>
<td>500</td>
<td>1.446(0.009)</td>
<td>0.000(0.000)</td>
<td>0.260(0.052)</td>
<td>0.780(0.041)</td>
</tr>
<tr>
<td>Group lasso</td>
<td>500</td>
<td>1.446(0.009)</td>
<td>0.000(0.000)</td>
<td>0.260(0.052)</td>
<td>0.780(0.041)</td>
</tr>
<tr>
<td>BMA</td>
<td>500</td>
<td>1.442(0.010)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Table corresponding to independent predictors, $p = 100$, $s = 100$

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Pred error</th>
<th>False neg</th>
<th>False pos</th>
<th>True model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPPM</td>
<td>100</td>
<td>2.594(0.040)</td>
<td>0.320(0.060)</td>
<td>3.310(0.207)</td>
<td>0.040(0.020)</td>
</tr>
<tr>
<td>Median model</td>
<td>100</td>
<td>2.596(0.038)</td>
<td>0.340(0.059)</td>
<td>3.360(0.208)</td>
<td>0.040(0.020)</td>
</tr>
<tr>
<td>Group lasso</td>
<td>100</td>
<td>2.565(0.036)</td>
<td>0.080(0.031)</td>
<td>3.370(0.207)</td>
<td>0.040(0.020)</td>
</tr>
<tr>
<td>BMA</td>
<td>100</td>
<td>1.170(0.053)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPPM</td>
<td>200</td>
<td>1.855(0.035)</td>
<td>0.000(0.000)</td>
<td>2.620(0.237)</td>
<td>0.190(0.039)</td>
</tr>
<tr>
<td>Median model</td>
<td>200</td>
<td>1.855(0.035)</td>
<td>0.000(0.000)</td>
<td>2.640(0.238)</td>
<td>0.190(0.039)</td>
</tr>
<tr>
<td>Group lasso</td>
<td>200</td>
<td>1.855(0.034)</td>
<td>0.000(0.000)</td>
<td>2.650(0.238)</td>
<td>0.190(0.039)</td>
</tr>
<tr>
<td>BMA</td>
<td>200</td>
<td>1.374(0.069)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HPPM</td>
<td>500</td>
<td>1.299(0.011)</td>
<td>0.000(0.000)</td>
<td>2.710(0.221)</td>
<td>0.210(0.041)</td>
</tr>
<tr>
<td>Median model</td>
<td>500</td>
<td>1.299(0.011)</td>
<td>0.000(0.000)</td>
<td>2.710(0.225)</td>
<td>0.220(0.041)</td>
</tr>
<tr>
<td>Group lasso</td>
<td>500</td>
<td>1.299(0.011)</td>
<td>0.000(0.000)</td>
<td>2.730(0.224)</td>
<td>0.210(0.041)</td>
</tr>
<tr>
<td>BMA</td>
<td>500</td>
<td>0.989(0.039)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Prediction accuracy with other models. We compute the estimated squared prediction errors for each model averaged over the replications (denoted by ‘Pred error’). The results are presented in Table 1 and 2 for the low dimensional and high dimensional examples respectively. The model chosen by our Bayesian method with the highest posterior probability is denoted by ‘HPPM’. The median probability model is denoted by ‘Median model’ and the model selected by the group lasso solution is denoted by ‘Group lasso’. Prediction results obtained by Bayesian model averaging are listed corresponding to the row ‘BMA’. In the lower dimensional example, we find that the performance of the Bayesian method in terms of selecting the true model is comparable with the group lasso for different sample sizes, but the median probability model selects lesser number of predictors which are not present in the true model. The performance of the approximate Bayesian method improves with increasing sample size. In the high dimensional situation, the model selection performance of the Bayesian method is superior compared with that of the group lasso, both in terms of selecting the true model in greater proportion and selecting lesser number of predictors not present in the true model. Prediction performances of Bayesian model averaging is better in all the cases, both in lower and higher dimensions. The computation times for the Bayesian method corresponding to 100 replications of the data are presented in Table 3. All the computations have been done using R in DELL Dual Processor Xeon Six Core 3.6 GHz machines with 60GB RAM running 64Bit CentOS Linux 5.0. Interestingly, in the high dimensional situation, the computation time reduces with the increase in sample size. On a closer look, we find that when $n \ll p + s$, the group lasso tends to select models having a larger number of predictors not included in the true model, and hence the Bayesian method has to evaluate a larger number of models in each case.
5. Real data analysis

5.1 Nutritional epidemiology study

There is an increased risk of developing certain types of cancer including lung, colon, breast and prostrate cancer when the blood plasma concentration of beta-carotene is low, as indicated by some epidemiological studies. It is known that beta-carotene has remarkable anti-oxidant properties and regular dietary intake of fruits and vegetables rich in beta-carotene helps the body’s auto-immune system to fight cancer. It is of a lot of interest for clinical practitioners to know how the plasma concentrations of beta-carotene depend on certain regulatory factors like age, gender, regular use of vitamins, dietary intake, smoking status, alcohol consumption, etc. A number of diverse results have been found regarding the relation to these factors; for example, see Nierenberg et al. [21], Faure et al. [8].

We use the data-set based on a cross-sectional study provided by Therese Stukel of Dartmouth Hitchcock Medical Center, available at http://lib.stat.cmu.edu/datasets/PlasmaRetinol. Details of the data can be found in the above link. The response variables are plasma concentrations of beta-carotene and retinol obtained from 315 patients. Observations are made relating to 12 other factors, namely,

(1) AGE: Age (years).
(2) SEX: Sex (1=Male, 2=Female).
(3) SMOKSTAT: Smoking status (1=Never, 2=Former, 3=Current Smoker).
(4) BMI: Body Mass Index (weight/height^2).
(5) VITUSE: Vitamin Use (1=Yes, fairly often, 2=Yes, not often, 3=No).
(6) CALORIES: Number of calories consumed per day.
(7) FAT: Grams of fat consumed per day.
(8) FIBER: Grams of fiber consumed per day.
(9) ALCOHOL: Number of alcoholic drinks consumed per week.
(10) CHOL: Cholesterol consumed (mg per day).
(11) BETADIET: Dietary beta-carotene consumed (mcg per day).
(12) RETDIET: Dietary retinol consumed (mcg per day).

Liu et al. [16] used an additive partial linear model for this kind of problem and apply their variable selection method for APLM after doing some primary elicitation of effects which may be presumed to be linear and some effects which do not seem to have a linear effect. Similar to their model, we consider an additive partial linear model with ‘AGE’ and ‘CHOL’ having non-linear effects and all other predictors having linear effects only. We perform variable selection using the proposed Bayesian method by computing approximate posterior probabilities of various models.

The median probability model selects the linear effects of ‘BMI’, ‘VITUSE’ and ‘FIBER’ and the non-linear effects of ‘CHOL’ as the effective set of variables related to the beta-carotene levels. The median probability model is also the maximum a posteriori model. In comparison, the group lasso selects the linear effects of all the predictors excluding ‘SEX’ and ‘SMOKESTAT’, and the non-linear effects of both ‘AGE’ and ‘CHOL’. Table 4 lists the marginal inclusion probabilities of the predictors as obtained from our Bayesian procedure.

<table>
<thead>
<tr>
<th>n</th>
<th>p = 10, s = 10</th>
<th>p = 100, s = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>01:12:33</td>
<td>14:06:18</td>
</tr>
<tr>
<td>200</td>
<td>00:13:38</td>
<td>08:15:45</td>
</tr>
<tr>
<td>500</td>
<td>00:25:24</td>
<td>04:45:08</td>
</tr>
</tbody>
</table>
Table 4. Marginal inclusion probabilities of predictors for Nutritional Epidemiology study

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Inclusion Probability</th>
<th>Predictor</th>
<th>Inclusion Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEX</td>
<td>0.000</td>
<td>ALCOHOL</td>
<td>0.064</td>
</tr>
<tr>
<td>SMOKESTAT</td>
<td>0.000</td>
<td>BETADIET</td>
<td>0.011</td>
</tr>
<tr>
<td>BMI</td>
<td>0.996</td>
<td>RETDIET</td>
<td>0.002</td>
</tr>
<tr>
<td>VITUSE</td>
<td>0.507</td>
<td>AGE (linear)</td>
<td>0.375</td>
</tr>
<tr>
<td>CALORIES</td>
<td>0.002</td>
<td>CHOL (linear)</td>
<td>0.015</td>
</tr>
<tr>
<td>FAT</td>
<td>0.038</td>
<td>AGE (non-linear)</td>
<td>0.000</td>
</tr>
<tr>
<td>FIBER</td>
<td>0.994</td>
<td>CHOL (non-linear)</td>
<td>0.999</td>
</tr>
</tbody>
</table>

5.2 Prostate cancer data

We consider the prostate cancer data (Stamey et al. [28]) for analysis using the additive partial linear model framework. The data consists of clinical measures of 97 men who were about to receive a radical prostatectomy (see ‘lasso2’ package in R, Lokhorst et al. [17]). The response variable is log level of prostate specific antigen (lpsa), corresponding to 8 other predictors, namely,

(1) lcavol: log cancer volume.
(2) lweight: log prostate weight.
(3) age: age.
(4) lbph: log benign prostatic hyperplasia amount.
(5) svi: seminal vesicle invasion.
(6) lcp: log capsular penetration.
(7) gleason: Gleason score.
(8) pgg45: percentage Gleason scores 4 or 5.

The linear regression model using the lasso selects the predictors ‘lcavol’, ‘lweight’ and ‘svi’ as the important linear predictors (Tibshirani [30]). In our analysis, we consider an additive partial linear model using ‘lcavol’ and ‘lweight’ as the predictors with both linear and non-linear effects, and all remaining predictors having strictly linear effect. The median probability model selects ‘lbph’, ‘svi’, ‘gleason’ and ‘lcavol’ among the linear effects and non-linear effects of both ‘lcavol’ and ‘lweight’. The marginal inclusion probabilities of the various predictors are tabulated below.

Table 5. Marginal inclusion probabilities of predictors for Prostate Cancer data

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Inclusion Probability</th>
<th>Predictor</th>
<th>Inclusion Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.124</td>
<td>pgg45</td>
<td>0.084</td>
</tr>
<tr>
<td>lbph</td>
<td>0.510</td>
<td>lcavol (linear)</td>
<td>0.997</td>
</tr>
<tr>
<td>svi</td>
<td>0.826</td>
<td>lweight (linear)</td>
<td>0.000</td>
</tr>
<tr>
<td>lcp</td>
<td>0.000</td>
<td>lcavol (non-linear)</td>
<td>0.571</td>
</tr>
<tr>
<td>gleason</td>
<td>0.948</td>
<td>lweight (non-linear)</td>
<td>0.966</td>
</tr>
</tbody>
</table>

In order to further validate our method, we also conducted an analysis with the above data but now including additional 50 junk variables. The Bayesian method correctly selects a subset of the original predictors discarding the additional predictors. The marginal inclusion probabilities of the original predictors are almost same as in Table 5.

References


