Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Fast Bayesian model assessment for nonparametric additive regression[☆]

S. McKay Curtis, Sayantan Banerjee *, Subhashis Ghosal

North Carolina State University, United States

ARTICLE INFO

Article history:
Received 30 May 2012
Received in revised form 15 April 2013
Accepted 15 May 2013
Available online xxxx

Keywords:
Group LASSO
Laplace approximation
Model uncertainty
Penalized regression
Variable selection

ABSTRACT

Variable selection techniques for the classical linear regression model have been widely investigated. Variable selection in fully nonparametric and additive regression models has been studied more recently. A Bayesian approach for nonparametric additive regression models is considered, where the functions in the additive model are expanded in a B-spline basis and a multivariate Laplace prior is put on the coefficients. Posterior probabilities of models defined by selection of predictors in the working model are computed, using a Laplace approximation method. The prior times the likelihood is expanded around the posterior mode, which can be identified with the group LASSO, for which a fast computing algorithm exists. Thus Markov chain Monte-Carlo or any other time consuming sampling based methods are completely avoided, leading to quick assessment of various posterior model probabilities. This technique is applied to the high-dimensional situation where the number of parameters exceeds the number of observations.

1. Introduction

The literature abounds in variable selection methods for the linear model; see, for example, Miller (2002) and George (2000). One particular method that has generated a substantial amount of research is the Least Absolute Shrinkage and Selection Operator or LASSO (Tibshirani, 1996). This method involves minimizing penalized sums of squares where the penalty is the sum of the absolute values of the coefficients. For certain values of a tuning parameter, the minimizer of this penalized sum of squares can set one or more coefficients exactly to zero, and thus remove those variables from the model. A fast computing algorithm for the LASSO is given by a modification of the Least Angle Regression (LARS) algorithm (Efron et al., 2004). Many other variable selection approaches are variations on this penalized regression theme and typically differ from the LASSO by varying the form of the penalty; see, for example, Breiman (1995), Fan and Li (2001), Zou and Hastie (2005), Zou (2006), Bondell and Reich (2008), Hwang et al. (2009) and so on.

In many practical applications, the linear model setting is too restrictive and nonparametric regression models are preferred. In the recent years, several authors have proposed variable selection techniques for fully nonparametric regression. Friedman (1991) uses a forward stepwise regression procedure to construct a regression function from "reflected pairs" of basis functions. Linkletter et al. (2006) define the covariance function of a Gaussian process to be a function of individual predictors. Variables are selected by inclusion or exclusion from the covariance function. Lafferty and Wasserman (2008) use derivatives of the nonparametric function estimates with respect to smoothing parameters to find sparse solutions to the nonparametric variable selection problem.

[☆] Research is partially supported by NSF grant number DMS - 1106570.

[*] Correspondence to: Department of Statistics, NCSU, 2311 Stinson Drive, SAS Hall, Raleigh, NC 27695-8203, United States. Tel.: +1 919 699 8773.
E-mail address: sbanerj5@ncsu.edu (S. Banerjee).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.csda.2013.05.012

Although, fully nonparametric regression models are attractive in that they make relatively few assumptions about the regression function, they also lack the interpretability of the classical linear model. Additive models (Buja et al., 1989; Hastie and Tibshirani, 1990; Stone, 1985) provide a nice compromise between the restrictive linear model and the fully flexible nonparametric model. The additive model assumes that each predictor's contribution to the mean of the response can be modeled by an unspecified smooth function, thereby retaining some of the benefits of fully nonparametric regression. Additive models retain some of the benefits of interpretability found in classical linear models because each predictor has its own functional effect on the response. In addition, the simplifying assumptions of additive functional effects allow additive models to avoid the curse of dimensionality. Additive models can also be extended to smoothing-spline ANOVA (SS-ANOVA) models that allow for higher order interactions among the predictors (Barry, 1986; Gu, 2002; Wahba, 1990).

A handful of variable selection techniques exist for additive models. Chen (1993) develops a bootstrap procedure for model selection in SS-ANOVA models. Shively et al. (1999) develop a Bayesian model where the functional effect of each predictor is given a prior with a linear component and a nonlinear Wiener process component. Shi and Tsai (1999) give a modified version of Akaike's Information Criterion (AIC) (Akaike, 1974) suitable for selection of regression models with linear and additive components. Gustafson (2000) presents a Bayesian variable selection technique for regression models that allow predictors to have linear or functional effects and two-way interactions. Wood et al. (2002) develop a Bayesian method, based on the Bayesian Information Criterion (BIC) (Schwarz, 1978), for selecting between a linear regression model, a model with additive functional effects, or a fully nonparametric regression model. Lin and Zhang (2006) present the Component Selection and Smoothing Operator (COSSO) which is a generalization of the LASSO based on fitting a penalized SS-ANOVA model where the penalty is the sum of norms of the projection of each functional component into a partition of the model space. Belitz and Lang (2008) propose algorithms for variable selection and choosing the degree of smoothness in the regression models with structured additive predictors. Marra and Wood (2011) use shrinkage methods along with an extension of the non-negative garotte estimator for generalized additive models. Reich et al. (2009) develop a Bayesian variable selection technique for SS-ANOVA models with Gaussian process priors.

Yuan and Lin (2006) present a variable selection technique, called the group LASSO, for predictors that form natural groupings (e.g., sets of dummy variables for factors). Avalos et al. (2003) also develop a similar procedure for the special case of additive models using a B-spline basis. The group LASSO is a penalized least-squares method that uses a special form of penalty to eliminate redundant variables from the model simultaneously in pre-specified groups of variables. More specifically, let $\boldsymbol{Y}$ be an $n \times 1$ vector of responses, $\boldsymbol{X}_j$ is an $n \times m_j$ matrix of variables associated with the jth predictor (which may be stochastic or nonstochastic) and $\boldsymbol{\beta}_j$ is an $m_j \times 1$ vector of coefficients. Then group LASSO minimizes

$$\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\| \boldsymbol{Y} - \sum_{j=1}^{g} \boldsymbol{X}_j \boldsymbol{\beta}_j \right\|^2 + \lambda \sum_{j=1}^{g} \|\boldsymbol{\beta}_j\|, \tag{1}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1{}^T, \ldots, \boldsymbol{\beta}_g{}^T)^T$ and g is the number of groups. Yuan and Lin (2006) show that for some values of the tuning parameter λ, the solution to (1) includes $\boldsymbol{\beta}_j = \boldsymbol{0}$ for some subset of $j = 1, \ldots, g$. Other penalized approaches for variable selection in nonparametric additive regression models are also available (Meier et al., 2009; Ravikumar et al., 2009). Huang et al. (2010) consider variable selection in nonparametric additive regression models using a generalization of the adaptive LASSO (Zou, 2006) to the group LASSO (Yuan and Lin, 2006), called the adaptive group LASSO, and give conditions for consistent selection of the components in the underlying model.

One drawback of most variable selection methods is that they do not provide a measure of model uncertainty. Variable selection methods typically give one model as the best, without giving some measurement of uncertainty for this estimated model. The exceptions to this are methods that follow the Bayesian paradigm. They typically provide a measure of model uncertainty by calculating the number of times a particular model is visited in the posterior draws from a Markov chain Monte Carlo (MCMC) simulation (George and McCulloch, 1993). However, MCMC methods are computationally expensive when a large number of variables are involved and it can be hard to assess convergence when MCMC methods must traverse a space of differing dimensions. In fact, when the model dimension is very high, most MCMC based methods break down.

In this paper, we present a method for calculating approximate posterior model probabilities without having to draw MCMC samples. We use a multivariate Laplace prior on the coefficients of the functions in the model. In a linear model with normal errors, it is well known that when using independent univariate Laplace priors, the posterior mode coincides with the LASSO. Similarly the group LASSO can be viewed as the posterior mode with respect to some appropriate multivariate Laplace prior. The prior dependence in the components induces the grouping structure in the group LASSO. In additive models, we expand functions in a suitable basis such as the spline basis, and put a multivariate Laplace prior on the coefficients of the model. The coefficients of functions of the same predictors are taken to be a priori dependent, but coefficients of functions referring to different predictors are taken to be a priori independent. This introduces a natural grouping of variables formed by basis expansion of function of original predictor variables, for which the group LASSO is the posterior mode. We use the Laplace method of approximation of integrals by expanding the integrand around its maxima, thus avoiding costly MCMC simulations. Our method may be viewed as a generalization of the method of Yuan and Lin (2005), who develop a similar method for the classical linear regression model, by using the Laplace approximation around the standard LASSO. However, the main focus of Yuan and Lin (2005) was to obtain an empirical Bayes estimate of the tuning parameter of LASSO using the Bayesian approach. In contrast, our interest is truly in obtaining posterior probabilities of various models.

By obtaining posterior probabilities of various models, we can also perform Bayesian model averaging, which is typically preferred in prediction due to its ability to incorporate uncertainty in model selection. Some other models, such as the median probability model are often of interest as well. The median probability model is defined as the collection of all variables whose individual selection probabilities are at least one half, and is known to possess better prediction properties than the maximum a posteriori model (Barbieri and Berger, 2004).

We organize the paper as follows. In Section 2, we formally discuss the model and prior distribution, and describe the Laplace approximation method, along with the method of estimation of error variance and the tuning parameter in the multivariate Laplace prior in Section 3. In Section 4, through a simulation study we investigate which models carry appreciable posterior probabilities. A real data analysis is presented in Section 5.

2. Model formulation and prior specification

We consider a regression model $Y = f(\boldsymbol{X}) + \varepsilon$, where $\boldsymbol{X} = (X_1, \dots, X_p)$ is a set of p-predictors and random errors $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, where $\mathcal{N}$ stands for the normal distribution. We assume that the regression function has an additive form $f(\boldsymbol{x}) = \sum_{j=1}^{p} f_j(x_j)$. We suspect that all predictors $X_1, \dots, X_p$ may not be relevant, so we consider various submodels corresponding to each subcollection of predictors. Let $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)$ stand for the vector containing the p variable selection parameters γ_j, where $\gamma_j = 1$ if predictor j is in the model and $\gamma_j = 0$ otherwise. Let $k = \sum_{j=1}^{p} \gamma_j$ stand for the number of predictors included in the model described by $\boldsymbol{\gamma}$. Then we may represent the joint density of $(Y, \boldsymbol{\gamma})$ given $\boldsymbol{X} = \boldsymbol{x}$ in a hierarchical fashion as

$$p(y, \boldsymbol{\gamma}|\boldsymbol{x}) = p(y|\boldsymbol{x}_{\boldsymbol{\gamma}})p(\boldsymbol{\gamma}), \tag{2}$$

where $\boldsymbol{x}_{\boldsymbol{\gamma}}$ denote the vector of the values of the selected predictors, that is, $\boldsymbol{x}_{\boldsymbol{\gamma}} = \{x_j : \gamma_j = 1\}$.

If the individual regression functions $f_j(x_j)$'s are reasonably smooth, they can be expanded in a convenient basis $\{\psi_{j,1}, \psi_{j,2}, \dots\}$ up to sufficiently many terms, leading to representations of the form

$$f_j(x_j) = \sum_{l=1}^{m_j} \beta_{j,l}\psi_{j,l}(x_j), \quad j = 1, \dots, p. \tag{3}$$

We shall specifically work with the flexible and convenient B-spline basis functions. The number of terms m_j corresponding to x_j here acts as a tuning parameter—larger values of m_j reduce bias, but the increased variability of the estimates of corresponding regression coefficients may reduce the accuracy of the estimated function. Let $m_0 = \sum_{j=1}^{p} m_j$.

We obtain n independent observations whose values are denoted by $\boldsymbol{Y} = (Y_1, \dots, Y_n)^T$ and the corresponding values of the p predictor variables as $X_{i1}, \dots, X_{ip}$, $i = 1, \dots, n$. We can write the basis functions in a matrix as

$$\underset{n\times m_0}{\boldsymbol{\Psi}} = \begin{pmatrix} \psi_{11}(X_{11}) & \cdots & \psi_{1m_1}(X_{11}) & \cdots & \psi_{p1}(X_{1p}) & \cdots & \psi_{pm_p}(X_{1p}) \\ \psi_{11}(X_{21}) & \cdots & \psi_{1m_1}(X_{21}) & \cdots & \psi_{p1}(X_{2p}) & \cdots & \psi_{pm_p}(X_{2p}) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \psi_{11}(X_{n1}) & \cdots & \psi_{1m_1}(X_{n1}) & \cdots & \psi_{p1}(X_{np}) & \cdots & \psi_{pm_p}(X_{np}) \end{pmatrix} \tag{4}$$

and the coefficients as a vector

$$\underset{m_0\times 1}{\boldsymbol{\beta}} = (\underset{m_1\times 1}{\boldsymbol{\beta}_1^T}, \dots, \underset{m_p\times 1}{\boldsymbol{\beta}_p^T})^T = (\beta_{11}, \dots, \beta_{1m_1}, \dots, \beta_{p1}, \dots, \beta_{pm_p})^T. \tag{5}$$

The coefficients of the basis expansion of functions f_j not selected by $\boldsymbol{\gamma}$ are all zero. Let $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ denote non-zero coefficient values and let $\boldsymbol{\Psi}_{\boldsymbol{\gamma}}$ denote the matrix obtained from $\boldsymbol{\Psi}$ by discarding the columns corresponding to the irrelevant predictors. Then the model is representable as

$$\underset{n\times 1}{\boldsymbol{Y}} \sim \mathcal{N}(\underset{n\times m_{\boldsymbol{\gamma}}}{\boldsymbol{\Psi}_{\boldsymbol{\gamma}}}\ \underset{m_{\boldsymbol{\gamma}}\times 1}{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}, \sigma^2 \underset{n\times n}{\boldsymbol{I}_n}), \tag{6}$$

where $J_{\boldsymbol{\gamma}} = \{j : \gamma_j = 1\}$, $\boldsymbol{I}_n$ is the identity matrix of order n and $m_{\boldsymbol{\gamma}} = \sum_{j\in J_{\boldsymbol{\gamma}}} m_j$.

Without additional information, we view the functions $f_1, \dots, f_p$ as twice continuously differentiable, the usual level of smoothness people are visually able to confirm. In such a case, the bias with m terms decays like m^{-2}, while the variance decays like m/n, leading to the optimal rate for the tuning parameter m to be $n^{1/5}$. In practice, the value of m is chosen using cross-validation, which we shall follow as well. Let $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^T\boldsymbol{x}}$ and $\mathbb{1}_A(\cdot)$ is the indicator function of a set A. We consider the prior for $\boldsymbol{\beta}_j$ to be degenerate at $\boldsymbol{0}$, or to have Lebesgue density given by a multivariate Laplace distribution (Ernst, 1998), depending on whether $\gamma_j = 0$ or $\gamma_j = 1$, that is,

$$p(\boldsymbol{\beta}_j|\boldsymbol{\gamma}) = (1 - \gamma_j)\mathbb{1}_{\{\boldsymbol{0}\}}(\boldsymbol{\beta}_j) + \gamma_j \frac{\Gamma(m_j/2)}{2\pi^{m_j/2}\Gamma(m_j)} \left(\frac{\lambda}{2\sigma^2}\right)^{m_j} \exp\left\{-\frac{\lambda}{2\sigma^2}\|\boldsymbol{\beta}_j\|\right\}. \tag{7}$$

Thus, for the full coefficient vector $\boldsymbol{\beta}$, the prior density $p(\boldsymbol{\beta}|\boldsymbol{\gamma})$ (with respect to the product of sums of the counting measure at $\mathbf{0}$ and the Lebesgue measure) is

$$\left(\prod_{j\notin J_{\boldsymbol{\gamma}}} \mathbb{1}_{\{\mathbf{0}\}}(\boldsymbol{\beta}_j)\right)\left(\prod_{j\in J_{\boldsymbol{\gamma}}} \frac{\Gamma(m_j/2)\lambda^{m_j}}{2(2\sigma^2)^{m_j}\pi^{m_j/2}\Gamma(m_j)}\right)\exp\left\{-\frac{\lambda}{2\sigma^2}\sum_{j\in J_{\boldsymbol{\gamma}}}\|\boldsymbol{\beta}_j\|\right\}. \tag{8}$$

The final piece of our hierarchical specification is a prior distribution on all models $\boldsymbol{\gamma}$. We let the prior probabilities be

$$p(\boldsymbol{\gamma}) \propto d_{\boldsymbol{\gamma}} q^{|\boldsymbol{\gamma}|}(1-q)^{p-|\boldsymbol{\gamma}|}, \tag{9}$$

where $q \in (0, 1)$ and $d_{\boldsymbol{\gamma}}$ is a measure of dependence among the $|\boldsymbol{\gamma}|$ variables in the model. The value of q indicates the propensity of a variable being selected in the model. While in lower dimensional models, $q = \frac{1}{2}$ seems to be a reasonable default choice, much smaller values should be used in high dimensional models due to sparsity considerations. In other words, q stands for the prior guess for the proportion of relevant variables in the model. In our analysis, we shall take q always as given rather than an unknown hyperparameter. The quantity $d_{\boldsymbol{\gamma}}$ in our specification is similar in spirit to the term $\det(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}})$ in the model formulation of Yuan and Lin (2005), where $\boldsymbol{X}_{\boldsymbol{\gamma}} = ((x_{ij}))_{1\le i\le n,\, j\in J_{\boldsymbol{\gamma}}}$. In their formulation, the term $\det(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}})$ is small for models which contains at least a pair of highly correlated, and therefore redundant, predictors.

Because we are looking beyond linear models, correlation is no longer the most appropriate measure to look at. A useful analog of the correlation coefficient in the nonlinear setting is given by Kendall's tau coefficient, which is particularly good at picking up monotone relationship. We shall therefore work with the choice $d_{\boldsymbol{\gamma}}$ the determinant of the matrix of Kendall's tau for all pairings of predictors in model $\boldsymbol{\gamma}$. More formally, let κ_{jl} be Kendall's tau for the pair of vectors $\boldsymbol{x}_j$ and $\boldsymbol{x}_l$ and let $\boldsymbol{K} = ((\kappa_{jl}))$. Then we choose $d_{\boldsymbol{\gamma}} = \det(\boldsymbol{K}_{\boldsymbol{\gamma}})$, where $\boldsymbol{K}_{\boldsymbol{\gamma}}$ is a submatrix of $\boldsymbol{K}$ corresponding to non-zero elements of $\boldsymbol{\gamma}$. The following result shows that the matrix $\boldsymbol{K}$, and hence all submatrices $\boldsymbol{K}_{\boldsymbol{\gamma}}$, are non-negative definite. Therefore the factor $d_{\boldsymbol{\gamma}} = \det(\boldsymbol{K}_{\boldsymbol{\gamma}}) \ge 0$ justifies the specification of model prior probabilities by relation (9).

Lemma 2.1. *The matrix* $\boldsymbol{K} = ((\kappa_{jl}))$ *is always non-negative definite.*

Proof. By definition of Kendall's tau coefficient, the (j, l)th element of $\boldsymbol{K}$ is given by

$$\begin{aligned}\kappa_{jl} &= \frac{\sum\limits_{i=1}^{n}\sum\limits_{i'=i+1}^{n}\operatorname{sign}(X_{ij}-X_{i'j})\operatorname{sign}(X_{il}-X_{i'l})}{\binom{n}{2}}\\ &= \frac{\sum\limits_{i=1}^{n}\sum\limits_{i'=1}^{n}\operatorname{sign}(X_{ij}-X_{i'j})\operatorname{sign}(X_{il}-X_{i'l})}{n(n-1)},\end{aligned}$$

$j, l = 1, 2, \ldots, p$, since $\operatorname{sign}(0) = 0$.

It suffices to show that for any $a_1, \ldots, a_p$,

$$\sum_{j=1}^{p}\sum_{l=1}^{p}a_ja_l\sum_{i=1}^{n}\sum_{i=1}^{n}\operatorname{sign}(X_{ij}-X_{i'j})\operatorname{sign}(X_{il}-X_{i'l}) \ge 0.$$

The expression is equal to

$$\sum_{i=1}^{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}a_j\operatorname{sign}(X_{ij}-X_{i'j})\right)\left(\sum_{l=1}^{p}a_l\operatorname{sign}(X_{il}-X_{i'l})\right) = \sum_{i=1}^{n}\sum_{i'=1}^{n}\left(\sum_{j=1}^{p}a_j\operatorname{sign}(X_{ij}-X_{i'j})\right)^2 \ge 0.$$

This shows that $\boldsymbol{K}$ is always non-negative definite. □

The term $d_{\boldsymbol{\gamma}}$ penalizes redundant models that have a high degree of dependence among the predictors. Measures of nonlinear association obtained from the empirical copula between pairs of predictors, may also be used instead of Kendall's tau.

We note that there have arisen two philosophies with regard to redundant predictors. The first is that if two predictors are highly related, then one or the other should be included in the model but not both. This philosophy is exemplified by the approach of Yuan and Lin (2005). The other philosophy is that if two predictors are highly related, then they should both be included in the model as a group (or excluded from the model as a group). This approach is exemplified by Zou and Hastie (2005) and Bondell and Reich (2008).

Thus, we explored a few other variations to the prior on $\boldsymbol{\gamma}$. For example, one method assumed an ordering to the predictors – for example, least costly to measure to most costly to measure – and penalized models that included "higher-cost" predictors that were highly correlated with excluded predictors of "lower cost". In our simulation studies, however, we did not find significant differences arising out of these different priors and hence those results are not presented.

Please cite this article in press as: McKay Curtis, S., et al., Fast Bayesian model assessment for nonparametric additive regression. Computational Statistics and Data Analysis (2013), http://dx.doi.org/10.1016/j.csda.2013.05.012

3. Posterior computation

With the model formulation and prior specification as in the last section, using the normal likelihood for $\boldsymbol{Y}$ as in Eq. (6) and Eqs. (8) and (9), we can now write the joint posterior density $p(\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}|\boldsymbol{Y})$ for $\boldsymbol{\beta}_\gamma$ and $\boldsymbol{\gamma}$ given $\boldsymbol{Y}$ as proportional to

$$p(\boldsymbol{Y}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma})p(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})p(\boldsymbol{\gamma}) = (1-q)^p(2\pi\sigma^2)^{-n/2}d_\gamma\left(\frac{q}{2(1-q)}\right)^{|\gamma|}\left(\prod_{j\in J_\gamma}\frac{\Gamma(m_j/2)\lambda^{m_j}}{(2\sigma^2)^{m_j}\pi^{m_j/2}\Gamma(m_j)}\right) \times \exp\left\{-\frac{\|\boldsymbol{Y}-\boldsymbol{\Psi}_\gamma\boldsymbol{\beta}_\gamma\|^2 + \lambda\sum_{j\in J_\gamma}\|\boldsymbol{\beta}_j\|}{2\sigma^2}\right\}. \quad (10)$$

The marginal posterior probability for model $\boldsymbol{\gamma}$ can be obtained by integrating out $\boldsymbol{\beta}_\gamma$, that is,

$$p(\boldsymbol{\gamma}|\boldsymbol{Y}) \propto C(\boldsymbol{Y})B(\boldsymbol{\gamma})\int_{\mathbb{R}^{m_\gamma}}\exp\left\{-\frac{\|\boldsymbol{Y}-\boldsymbol{\Psi}_\gamma\boldsymbol{\beta}_\gamma\|^2 + \lambda\sum_{j\in J_\gamma}\|\boldsymbol{\beta}_j\|}{2\sigma^2}\right\}d\boldsymbol{\beta}_\gamma, \quad (11)$$

with

$$C(\boldsymbol{Y}) = (1-q)^p(2\pi\sigma^2)^{-n/2},$$
$$B(\boldsymbol{\gamma}) = \left(\frac{q}{2(1-q)}\right)^{|\gamma|}\left(\prod_{j\in J_\gamma}\frac{\Gamma(m_j/2)\lambda^{m_j}}{(2\sigma^2)^{m_j}\pi^{m_j/2}\Gamma(m_j)}\right). \quad (12)$$

The integral in (11) can be approximated using the Laplace approximation. Let $\boldsymbol{\beta}^*_\gamma$ denote the group LASSO solution, that is,

$$\boldsymbol{\beta}^*_\gamma = \underset{\boldsymbol{\beta}_\gamma}{\operatorname{argmin}}\|\boldsymbol{Y}-\boldsymbol{\Psi}_\gamma\boldsymbol{\beta}_\gamma\|^2 + \lambda\sum_{j\in J_\gamma}\|\boldsymbol{\beta}_j\|. \quad (13)$$

Put $\boldsymbol{u} = \boldsymbol{\beta}_\gamma - \boldsymbol{\beta}^*_\gamma$. Substituting this quantity into (11) gives the expression

$$C(\boldsymbol{Y})B(\boldsymbol{\gamma})\exp\left\{-\frac{\min_{\boldsymbol{\beta}_\gamma}\left(\|\boldsymbol{Y}-\boldsymbol{\Psi}_\gamma\boldsymbol{\beta}_\gamma\|^2 + \lambda\sum_{j\in J_\gamma}\|\boldsymbol{\beta}_j\|\right)}{2\sigma^2}\right\} \times \int_{\mathbb{R}^{m_\gamma}}\exp\left\{-\frac{1}{2\sigma^2}\left[\|\boldsymbol{\Psi}_\gamma\boldsymbol{u}\|^2 - 2\boldsymbol{u}^T\boldsymbol{\Psi}_\gamma^T\boldsymbol{Y}^* + \lambda\sum_{j\in J_\gamma}\left(\|\boldsymbol{\beta}^*_j+\boldsymbol{u}_j\| - \|\boldsymbol{\beta}^*_j\|\right)\right]\right\}d\boldsymbol{u}, \quad (14)$$

where $\boldsymbol{Y}^* = \boldsymbol{Y} - \boldsymbol{\Psi}_\gamma\boldsymbol{\beta}^*_\gamma$ and $\boldsymbol{\beta}^*_j$ and $\boldsymbol{u}_j$ are the elements of $\boldsymbol{\beta}^*_\gamma$ and $\boldsymbol{u}$ that correspond to the basis functions of the jth predictor, $j \in J_\gamma$.

Let

$$f(\boldsymbol{u}) = \frac{1}{\sigma^2}\left[\|\boldsymbol{\Psi}_\gamma\boldsymbol{u}\|^2 - 2\boldsymbol{u}^T\boldsymbol{\Psi}_\gamma^T\boldsymbol{Y}^* + \lambda\sum_{j\in J_\gamma}\left(\|\boldsymbol{\beta}^*_j+\boldsymbol{u}_j\| - \|\boldsymbol{\beta}^*_j\|\right)\right]. \quad (15)$$

Clearly $f(\boldsymbol{u})$ is minimized at $\boldsymbol{u} = \boldsymbol{0}$ by definition, and

$$\left.\frac{\partial f(\boldsymbol{u})}{\partial\boldsymbol{u}\partial\boldsymbol{u}^T}\right|_{\boldsymbol{u}=\boldsymbol{0}} = \frac{1}{\sigma^2}(2\boldsymbol{\Psi}_\gamma^T\boldsymbol{\Psi}_\gamma + \lambda\boldsymbol{A}_\gamma), \quad (16)$$

where the $m_\gamma \times m_\gamma$ matrix $\boldsymbol{A}_\gamma$ is given by

$$\boldsymbol{A}_\gamma = \begin{bmatrix} -\frac{\boldsymbol{\beta}^*_1\boldsymbol{\beta}^{*T}_1}{\|\boldsymbol{\beta}^*_1\|^3} + \frac{\boldsymbol{I}_{11}}{\|\boldsymbol{\beta}^*_1\|} & \boldsymbol{O}_{12} & \cdots & \boldsymbol{O}_{1k} \\ \boldsymbol{O}_{21} & -\frac{\boldsymbol{\beta}^*_2\boldsymbol{\beta}^{*T}_2}{\|\boldsymbol{\beta}^*_2\|^3} + \frac{\boldsymbol{I}_{22}}{\|\boldsymbol{\beta}^*_2\|} & \cdots & \boldsymbol{O}_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{O}_{k1} & \boldsymbol{O}_{k2} & \cdots & -\frac{\boldsymbol{\beta}^*_k\boldsymbol{\beta}^{*T}_k}{\|\boldsymbol{\beta}^*_k\|^3} + \frac{\boldsymbol{I}_{kk}}{\|\boldsymbol{\beta}^*_k\|} \end{bmatrix}, \quad (17)$$

$\boldsymbol{I}_{jj}$ is the identity matrix of order equal to the length of the jth predictor and $\boldsymbol{O}_{jl}$ is a matrix of zeros with the number of rows equal to the length of the jth predictor and the number of columns equal to that of the lth predictor, $j, l \in J_{\boldsymbol{\gamma}}$.

The above equations can be used to apply the Laplace approximation to the quantity in (14), which gives

$$\begin{aligned} p(\boldsymbol{\gamma}|\boldsymbol{Y}) &\propto C(\boldsymbol{Y})B(\boldsymbol{\gamma}) \exp\left\{-\frac{\min\limits_{\boldsymbol{\beta}_{\gamma}}\left(\|\boldsymbol{Y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum\limits_{j\in J_{\gamma}}\|\boldsymbol{\beta}_j\|\right)}{2\sigma^2}\right\} \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}f(\boldsymbol{u})\right\} d\boldsymbol{u} \\ &\approx C(\boldsymbol{Y})B(\boldsymbol{\gamma}) \exp\left\{-\frac{\min\limits_{\boldsymbol{\beta}_{\gamma}}\left(\|\boldsymbol{Y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum\limits_{j\in J_{\gamma}}\|\boldsymbol{\beta}_j\|\right)}{2\sigma^2}\right\} \exp\left\{-\frac{1}{2}f(\boldsymbol{0})\right\} (2\pi)^{m_{\gamma}/2}\left|\frac{1}{2}\frac{\partial f(\boldsymbol{0})}{\partial\boldsymbol{u}\partial\boldsymbol{u}^T}\right|^{-1/2}. \end{aligned}$$

Substituting (16) in the above quantity, the marginal posterior probability $p(\boldsymbol{\gamma}|\boldsymbol{Y})$ for $\boldsymbol{\gamma}$ is approximately proportional to

$$Q(\boldsymbol{\gamma}|\boldsymbol{Y}) = C(\boldsymbol{Y})B(\boldsymbol{\gamma}) \exp\left\{-\frac{\min\limits_{\boldsymbol{\beta}_{\gamma}}\left(\|\boldsymbol{Y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum\limits_{j\in J_{\gamma}}\|\boldsymbol{\beta}_j\|\right)}{2\sigma^2}\right\} (2\pi)^{m_{\gamma}/2}\left|\frac{1}{\sigma^2}\left(\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T\boldsymbol{\Psi}_{\boldsymbol{\gamma}}+\frac{\lambda}{2}\boldsymbol{A}_{\boldsymbol{\gamma}}\right)\right|^{-1/2}.$$

Plugging in the expressions for $C(\boldsymbol{Y})$ and $B(\boldsymbol{\gamma})$ from (12) in the equation above, we get

$$\begin{aligned} Q(\boldsymbol{\gamma}|\boldsymbol{Y}) &= (1-q)^p(2\pi\sigma^2)^{-n/2}d_{\boldsymbol{\gamma}}\left(\frac{q}{2(1-q)}\right)^{|\boldsymbol{\gamma}|}\left(\prod_{j\in J_{\gamma}}\frac{\Gamma(m_j/2)\lambda^{m_j}}{(2\sigma^2)^{m_j}\pi^{m_j/2}\Gamma(m_j)}\right) \\ &\quad\times(2\pi)^{m_{\gamma}/2}\left|\frac{1}{\sigma^2}\left(\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T\boldsymbol{\Psi}_{\boldsymbol{\gamma}}+\frac{\lambda}{2}\boldsymbol{A}_{\boldsymbol{\gamma}}\right)\right|^{-1/2}\exp\left\{-\frac{\min\limits_{\boldsymbol{\beta}_{\gamma}}\left(\|\boldsymbol{Y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum\limits_{j\in J_{\gamma}}\|\boldsymbol{\beta}_j\|\right)}{2\sigma^2}\right\}. \end{aligned} \tag{18}$$

The approximation in (18) holds only if all components of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}^*$ are non-zero—else the derivative in (16) does not exist. This happens when the group LASSO sets one or more of the elements of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ to $\boldsymbol{0}$. Yuan and Lin (2005), in the context of linear models, define a "nonregular" model as any model where at least one coefficient is set to zero by the LASSO, and they show that in the special case of an orthogonal design matrix, for every nonregular model $\boldsymbol{\gamma}$, there exists a submodel $\boldsymbol{\gamma}^*$ of $\boldsymbol{\gamma}$ with only those predictors in $\boldsymbol{\gamma}$ whose coefficients were not set to zero by the LASSO, with higher asymptotic posterior probability. Thus such nonregular models may be ignored for the purpose of posterior model probability maximization. Note that any nonregular model is also counted as a regular model corresponding to the index $\boldsymbol{\gamma}^*$.

Similarly, we define a nonregular additive model as any model $\boldsymbol{\gamma}$ for which $\boldsymbol{\beta}_j^* = \boldsymbol{0}$ for at least one $j \in J_{\boldsymbol{\gamma}}$. For a given λ, any nonregular model is essentially equivalent to the submodel that has removed predictors whose coefficients were set to zero by the group LASSO. Therefore, we need not calculate posterior probabilities of the nonregular models. Since we discount these nonregular models, while normalizing to obtain the expression for $p(\boldsymbol{\gamma}|\boldsymbol{Y})$ (see Eq. (18)) over different $\boldsymbol{\gamma}$'s, we consider only the regular models. Thus for any (regular) model $\boldsymbol{\gamma}$,

$$p(\boldsymbol{\gamma}|\boldsymbol{Y}) \approx \frac{Q(\boldsymbol{\gamma}|\boldsymbol{Y})}{\sum\limits_{\boldsymbol{\gamma}'\text{regular}} Q(\boldsymbol{\gamma}'|\boldsymbol{Y})}. \tag{19}$$

3.1. Estimation of λ and σ^2

The joint density of the observation and the coefficient vectors conditional on all other model parameters is given by

$$\begin{aligned} p(\boldsymbol{Y}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \lambda, \sigma^2) &= (2\pi\sigma^2)^{-n/2}\left(\frac{1}{2}\right)^{|\boldsymbol{\gamma}|}\left(\frac{\lambda}{2\sigma^2\pi^{1/2}}\right)^{m_{\gamma}}\left(\prod_{j\in J_{\gamma}}\frac{\Gamma(m_j/2)}{\Gamma(m_j)}\right) \\ &\quad\times\exp\left\{-\frac{\|\boldsymbol{Y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum\limits_{j\in J_{\gamma}}\|\boldsymbol{\beta}_j\|}{2\sigma^2}\right\}. \end{aligned} \tag{20}$$

Integrating out $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ we get

$$p(\boldsymbol{Y}|\boldsymbol{\gamma},\lambda,\sigma^2) = (2\pi\sigma^2)^{-n/2}\left(\frac{1}{2}\right)^{|\boldsymbol{\gamma}|}\left(\frac{\lambda}{2\sigma^2\pi^{1/2}}\right)^{m_{\boldsymbol{\gamma}}}\left(\prod_{j\in J_{\boldsymbol{\gamma}}}\frac{\Gamma(m_j/2)}{\Gamma(m_j)}\right) \times \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\exp\left\{-\frac{\|\boldsymbol{Y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum_{j\in J_{\boldsymbol{\gamma}}}\|\boldsymbol{\beta}_j\|}{2\sigma^2}\right\}d\boldsymbol{\beta}_{\boldsymbol{\gamma}}. \tag{21}$$

Using the Laplace approximation as in (18), we have,

$$p(\boldsymbol{Y}|\boldsymbol{\gamma},\lambda,\sigma^2) \approx (2\pi)^{-(n-m_{\boldsymbol{\gamma}})/2}\sigma^{-(n+m_{\boldsymbol{\gamma}})}2^{-(m_{\boldsymbol{\gamma}}+|\boldsymbol{\gamma}|)}\lambda^{m_{\boldsymbol{\gamma}}}\left(\prod_{j\in J_{\boldsymbol{\gamma}}}\frac{\Gamma(m_j/2)}{\Gamma(m_j)}\right) \times \exp\left\{-\frac{\min_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}\left(\|\boldsymbol{Y}-\boldsymbol{\Psi}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|^2+\lambda\sum_{j\in J_{\boldsymbol{\gamma}}}\|\boldsymbol{\beta}_j\|\right)}{2\sigma^2}\right\}\left|\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T\boldsymbol{\Psi}_{\boldsymbol{\gamma}}+\frac{\lambda}{2}\boldsymbol{A}_{\boldsymbol{\gamma}}\right|^{-1/2}. \tag{22}$$

If we set $\boldsymbol{\gamma}$ in (22) equal to $\hat{\boldsymbol{\gamma}}_\lambda$, the model chosen by the group LASSO for a given λ, then maximizing (22) with respect to σ^2, we obtain an estimate of σ^2 as

$$\hat{\sigma}^2_\lambda = \frac{\|\boldsymbol{Y}-\boldsymbol{\Psi}_{\hat{\boldsymbol{\gamma}}_\lambda}\boldsymbol{\beta}^*_{\hat{\boldsymbol{\gamma}}_\lambda}\|^2+\lambda\sum_{j\in J_{\hat{\boldsymbol{\gamma}}_\lambda}}\|\boldsymbol{\beta}^*_j\|}{n+m_{\hat{\boldsymbol{\gamma}}_\lambda}}. \tag{23}$$

Substituting (23) back into (22) and taking −2 times the natural logarithm of (22) give

$$h(\lambda) = (n-m_{\boldsymbol{\gamma}})\log(2\pi) - 2\left(\sum_{j\in J_{\hat{\boldsymbol{\gamma}}_\lambda}}\log\Gamma(m_j/2)-\sum_{j\in J_{\hat{\boldsymbol{\gamma}}_\lambda}}\log\Gamma(m_j)\right)+2(m_{\hat{\boldsymbol{\gamma}}_\lambda}+|\hat{\boldsymbol{\gamma}}_\lambda|)\log 2 - 2m_{\hat{\boldsymbol{\gamma}}_\lambda}\log\lambda + (n+m_{\hat{\boldsymbol{\gamma}}_\lambda})\left[\log\left(\frac{\|\boldsymbol{Y}-\boldsymbol{\Psi}_{\hat{\boldsymbol{\gamma}}_\lambda}\boldsymbol{\beta}^*_{\hat{\boldsymbol{\gamma}}_\lambda}\|^2+\lambda\sum_{j\in J_{\hat{\boldsymbol{\gamma}}_\lambda}}\|\boldsymbol{\beta}^*_j\|}{n+m_{\hat{\boldsymbol{\gamma}}_\lambda}}\right)+1\right]+\log\left|\boldsymbol{\Psi}^T_{\hat{\boldsymbol{\gamma}}_\lambda}\boldsymbol{\Psi}_{\hat{\boldsymbol{\gamma}}_\lambda}+\frac{\lambda}{2}\boldsymbol{A}_{\hat{\boldsymbol{\gamma}}_\lambda}\right|. \tag{24}$$

For $m_j = m$ for all $j = 1, 2, \ldots, p$, the above expression for $h(\lambda)$ becomes

$$(n-m|\hat{\boldsymbol{\gamma}}_\lambda|)\log(2\pi) - 2|\hat{\boldsymbol{\gamma}}_\lambda|[\log\Gamma(m/2)-\log\Gamma(m)] + 2|\hat{\boldsymbol{\gamma}}_\lambda|(m+1)\log 2 + (n+m|\hat{\boldsymbol{\gamma}}_\lambda|)\left[\log\left(\frac{\|\boldsymbol{Y}-\boldsymbol{\Psi}_{\hat{\boldsymbol{\gamma}}_\lambda}\boldsymbol{\beta}^*_{\hat{\boldsymbol{\gamma}}_\lambda}\|^2+\lambda\sum_{j\in J_{\hat{\boldsymbol{\gamma}}_\lambda}}\|\boldsymbol{\beta}^*_{\hat{\boldsymbol{\gamma}}_\lambda}\|}{n+m|\hat{\boldsymbol{\gamma}}_\lambda|}\right)+1\right] - 2m|\hat{\boldsymbol{\gamma}}_\lambda|\log\lambda + \log\left|\boldsymbol{\Psi}^T_{\hat{\boldsymbol{\gamma}}_\lambda}\boldsymbol{\Psi}_{\hat{\boldsymbol{\gamma}}_\lambda}+\frac{\lambda}{2}\boldsymbol{A}_{\hat{\boldsymbol{\gamma}}_\lambda}\right|. \tag{25}$$

In applications, we shall, for simplicity, restrict to the situation $m_j = m$ for all $j = 1, 2, \ldots, p$, and choose m by cross-validation. An estimate of λ can then be found by minimizing (25) by a grid search, for instance.

Simulations have shown that choosing λ based on (25) results in overparametrized models (see Table 1 in Section 4). Therefore, we suggest using the BIC criterion for selecting λ (Schwarz, 1978). The BIC criterion in our case for normally distributed errors is given by

$$\mathrm{BIC}(\lambda) = \log\|\boldsymbol{Y}-\boldsymbol{\Psi}_{\hat{\boldsymbol{\gamma}}_\lambda}\hat{\boldsymbol{\beta}}^{\mathrm{LS}}_{\hat{\boldsymbol{\gamma}}_\lambda}\|^2 + \frac{m|\hat{\boldsymbol{\gamma}}_\lambda|}{n}\log n, \tag{26}$$

where $\hat{\boldsymbol{\beta}}^{\mathrm{LS}}_{\hat{\boldsymbol{\gamma}}_\lambda}$ is the least squares estimate of the regression coefficients based on the model selected by the group LASSO. Alternatively, we can slightly modify (25) by adding a logarithmic penalty term as in BIC, so that λ is chosen by minimizing

$$h(\lambda) + m|\hat{\boldsymbol{\gamma}}_\lambda|\log n. \tag{27}$$

Table 1
Table corresponding to independent predictors, $p = 10$ and $q = 0.5$.

	n	Error I	Error II	True.model
Approx.Bayes1	100	2.442 (0.066)	0.064 (0.012)	0.144 (0.016)
Reich.method	100	0.150 (0.004)	0.720 (0.008)	0.409 (0.005)
G.Lasso1	100	2.442 (0.066)	0.064 (0.012)	0.144 (0.016)
Approx.Bayes2	100	1.052 (0.053)	0.052 (0.010)	0.412 (0.022)
RJMCMC	100	0.002 (0.000)	4.971 (0.000)	0.000 (0.000)
G.Lasso2	100	1.052 (0.053)	0.052 (0.010)	0.412 (0.022)
Approx.Bayes3	100	0.002 (0.002)	3.600 (0.058)	0.014 (0.005)
G.Lasso3	100	0.002 (0.002)	3.616 (0.060)	0.012 (0.005)
Approx.Bayes1	200	0.110 (0.024)	0.586 (0.041)	0.574 (0.022)
Reich.method	200	0.100 (0.003)	0.820 (0.009)	0.372 (0.005)
G.Lasso1	200	0.110 (0.024)	0.586 (0.041)	0.574 (0.022)
Approx.Bayes2	200	0.058 (0.012)	0.020 (0.006)	0.932 (0.011)
RJMCMC	200	0.000 (0.000)	5.000 (0.000)	0.000 (0.000)
G.Lasso2	200	0.058 (0.012)	0.020 (0.006)	0.932 (0.011)
Approx.Bayes3	200	0.000 (0.000)	3.618 (0.055)	0.014 (0.005)
G.Lasso3	200	0.000 (0.000)	4.196 (0.043)	0.004 (0.003)
Approx.Bayes1	500	0.000 (0.000)	1.944 (0.059)	0.170 (0.017)
Reich.method	500	0.130 (0.003)	0.830 (0.008)	0.352 (0.005)
G.Lasso1	500	0.000 (0.000)	2.462 (0.077)	0.168 (0.043)
Approx.Bayes2	500	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
RJMCMC	500	–	–	–
G.Lasso2	500	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
Approx.Bayes3	500	0.000 (0.000)	1.648 (0.052)	0.168 (0.017)
G.Lasso3	500	0.000 (0.000)	4.690 (0.024)	0.000 (0.000)

We refer to the criterion in the above equation as the penalized maximum likelihood criterion. The difference in the two is that in the BIC, maximization is performed with respect to $\boldsymbol{\beta}_\gamma$, while in (27), $\boldsymbol{\beta}_\gamma$ is integrated out. If the posterior density of $\boldsymbol{\beta}_\gamma$ concentrates near the maximizer in large samples, as it happens in fixed dimensions, the integral in (21) can be approximated by the value of the integrand at the maximizer. Hence the two criteria become asymptotically equivalent.

Remark 3.1. Below we give a heuristic argument why the Laplace approximation may be trusted even when the dimension of the parameter space goes to infinity.

The original Laplace approximation was developed for a fixed dimensional setting, but in the high dimensional context, it is natural to think that the dimension $p_n \to \infty$. Shun and McCullagh (1995) show that in many common situations, the error in the Laplace approximations converges to zero even when $p_n \to \infty$, provided that $p_n = o(n^{1/3})$. In the present context, although p_n can be much larger than n, sparsity of the true model typically will lead to a sparse structure of the model $\hat{\boldsymbol{\gamma}}_n$ selected by the group LASSO. Since we apply the Laplace approximation to regular models only, which are necessarily submodels of $\hat{\boldsymbol{\gamma}}_n$, it follows that we only need to control the size $|\hat{\boldsymbol{\gamma}}_n|$ of $\hat{\boldsymbol{\gamma}}_n$ appropriately. More formally, assume that

1. sparsity of the true model $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0 : |\boldsymbol{\gamma}_0| = s_n \ll n \ll p_n$;
2. group LASSO is screening consistent for model selection, in the sense that $\mathrm{P}(\boldsymbol{\gamma}_0 \subset \hat{\boldsymbol{\gamma}}_n) \to 1$;
3. $|\hat{\boldsymbol{\gamma}}_n| = O_P(r_n)$, where $r_n = o(n^{1/3})$.

Then under the above assumptions, the error in the Laplace approximation converges to zero in probability, and all genuine predictors are included in the class of models being considered, with probability tending to one.

4. Simulation study

To examine the performance of our method of computing approximate posterior probabilities, we conduct two simulation studies, where all computation is executed in the R statistical programming language. We simulate data sets from a model with 5 "active" predictors and 5 "inactive" predictors in the first case, and with 5 "active" predictors and 495 "inactive" predictors in the second case. The purpose of the second study is to check the performance of the proposed method in the high-dimensional situation under sparsity. Thus the model may be written as

$$Y_i = \sum_{j=1}^{p} f_j(X_{ij}) + \varepsilon_i, \tag{28}$$

where $p = 10$ in the first scenario and $p = 500$ in the second, true value of σ is 1, and

$$f_1(x) = \exp(1.1x^3) - 2,$$
$$f_2(x) = 2x - 1,$$
$$f_3(x) = \sin(4\pi x),$$

Table 2
Table corresponding to AR(1) predictors, $p = 10$ and $q = 0.5$.

	n	Error I	Error II	True.model
Approx.Bayes1	100	2.018 (0.056)	0.068 (0.012)	0.128 (0.015)
G.Lasso1	100	2.018 (0.056)	0.068 (0.012)	0.128 (0.015)
Reich.method	100	1.270 (0.007)	2.000 (0.010)	0.012 (0.001)
Approx.Bayes2	100	2.124 (0.045)	0.020 (0.006)	0.058 (0.010)
G.Lasso2	100	2.124 (0.045)	0.020 (0.006)	0.058 (0.010)
Approx.Bayes1	200	0.720 (0.043)	0.306 (0.027)	0.354 (0.021)
G.Lasso1	200	0.720 (0.043)	0.306 (0.027)	0.354 (0.021)
Reich.method	200	1.140 (0.008)	0.710 (0.008)	0.084 (0.001)
Approx.Bayes2	200	1.328 (0.039)	0.010 (0.004)	0.154 (0.016)
G.Lasso2	200	1.328 (0.039)	0.010 (0.004)	0.154 (0.016)
Approx.Bayes1	500	0.076 (0.013)	0.930 (0.046)	0.378 (0.022)
G.Lasso1	500	0.076 (0.013)	0.936 (0.047)	0.378 (0.022)
Reich.method	500	–	–	–
Approx.Bayes2	500	0.572 (0.031)	0.006 (0.003)	0.532 (0.022)
G.Lasso2	500	0.572 (0.031)	0.006 (0.003)	0.532 (0.022)

$$f_4(x) = \log\{(e^2 - 1)x + 1\} - 1,$$
$$f_5(x) = -32(x - 0.5)^2/4 + 1, \quad \text{and}$$
$$f_j(x) = 0 \quad \text{for } j = 6, \ldots, p.$$

Note that each f_j is scaled to lie in $[-1, 1]$ when $x \in [0, 1]$. This simulation model is taken from Shively et al. (1999). We generate n samples where $n = 100, 200, 500$.

The X_{ij} variables are generated from two different sampling schemes. The first scheme – the independent X scheme – generates each X variable independently from the standard uniform distribution. The second scheme – the AR(1) scheme – generates the ith row of the X matrix from a multivariate normal distribution with an AR(1) covariance structure with variance–covariance matrix $\mathbf{\Sigma}_{ij} = 0.7^{|i-j|}$. The value of q is taken to be 0.5 throughout for low dimensional examples and 0.2 for the high dimensional example. Smaller value of q is chosen for the latter case in order to induce more sparsity through the model selection prior. We also perform a sensitivity analysis for the prior of the model by choosing q to be 0.2 and 0.8 for the low dimensional examples.

For each of the $\boldsymbol{X}$-matrix-generating schemes, we simulate 500 data sets and calculate approximate posterior model probabilities. We record the proportion of times that the model with the highest posterior probability (denoted by "Approx.Bayes" in the tables) is the true model. For the low dimensional examples, we present results with the value of the group LASSO penalty parameter λ selected using two different methods, one by minimizing the penalized marginal likelihood for λ as given by Eq. (27) (denoted by "Approx.Bayes1"), and the other using the BIC criterion (26) (denoted by "Approx.Bayes2"). For the high dimensional example, we only present results for λ chosen by minimizing (27). We also record the average number of "active predictors" the model with the highest posterior probability failed to include ("Error I") and the average number of "inactive predictors" the model incorrectly included ("Error II"). We note the proportion of times the method selected the correct model (denoted by "True.model"). We record this same information for the competing method proposed by Reich et al. (2009), and for the model selected by the group LASSO alone (denoted by "G.Lasso1" and "G.Lasso2" respectively corresponding to the approximate Bayes methods). The results are presented in Tables 1–3. Table 1 also presents results corresponding to the approximate Bayes method with the penalty parameter chosen by minimizing Eq. (25) (denoted by "Approx.Bayes3" for the Bayes method and "G.Lasso3" for the corresponding group LASSO method), and results corresponding to the Reversible Jump MCMC (RJMCMC) algorithm. As discussed in Section 3, we neglect the nonregular models, that is, the models for which some of the predictors with positive prior probabilities were not selected by the group LASSO. The posterior probabilities are re-normalized accounting only for the regular models. Tables 4 and 5 present the sensitivity analysis results for different choices of the parameter q using both the penalized likelihood criterion and the BIC criterion for selecting the penalty parameter λ.

Overall, the group LASSO and the model with the highest posterior probability are similar (almost exactly the same) in the number of "active" predictors that are selected. The model with the highest posterior probability tends to select less "inactive" predictors than the group LASSO. The difference between the two methods in this regard is not large, but the trend is persistent across all the simulations and across all sample sizes. For the low dimensional examples, we note that the approximate Bayes method seems to produce better results when the BIC criterion is used for selecting λ in comparison to the same using the penalized marginal likelihood criterion in the independent covariate structure, although the latter criterion produces better results for smaller sample sizes in the AR(1) covariate structure. Though the two criteria given as in (26) and (27) are supposed to be asymptotically equivalent for fixed dimensions, their fixed sample properties may differ significantly. The approximate Bayes method results in over-parametrized models when λ is chosen using Eq. (25) as seen from Table 1, and hence we do not adopt this technique for choosing λ in the other examples. For the high dimensional situation, we note that though the true model is selected only in approximately 4%–11% for our proposed method, the corresponding rates for the fully Bayes method of Reich et al. (2009) are zero with much larger Error II rates. Further, the

Table 3
Table corresponding to AR(1) predictors, $p = 500$ and $q = 0.2$, choosing penalty parameter λ using penalized marginal likelihood criterion.

	n	Error I	Error II	True.model
Approx.Bayes	100	2.335 (0.076)	0.120 (0.025)	0.040 (0.009)
Reich.method	100	0.980 (0.009)	392.020 (0.093)	0
G.Lasso	100	2.335 (0.076)	0.120 (0.025)	0.040 (0.009)
Approx.Bayes	200	1.460 (0.127)	0.060 (0.017)	0.110 (0.014)
Reich.method	200	1.330 (0.010)	356.610 (0.103)	0
G.Lasso	200	1.460 (0.127)	0.060 (0.017)	0.110 (0.014)
Approx.Bayes	500	0.405 (0.043)	0.175 (0.030)	0.540 (0.022)
Reich.method	500	–	–	–
G.Lasso	500	0.405 (0.043)	0.175 (0.030)	0.540 (0.022)

Table 4
Sensitivity analysis table for approximate Bayesian methods corresponding to independent predictors, $p = 10$.

	n	q	Error I	Error II	True.model
Approx.Bayes1	100	0.2	2.442 (0.066)	0.060 (0.012)	0.146 (0.016)
	100	0.5	2.442 (0.066)	0.064 (0.012)	0.144 (0.016)
	100	0.8	2.442 (0.066)	0.064 (0.012)	0.144 (0.016)
	200	0.2	0.110 (0.024)	0.534 (0.035)	0.574 (0.022)
	200	0.5	0.110 (0.024)	0.586 (0.041)	0.574 (0.022)
	200	0.8	0.110 (0.024)	0.586 (0.041)	0.574 (0.022)
	500	0.2	0.000 (0.000)	0.952 (0.038)	0.344 (0.021)
	500	0.5	0.000 (0.000)	1.944 (0.059)	0.170 (0.017)
	500	0.8	0.000 (0.000)	2.462 (0.077)	0.168 (0.017)
Approx.Bayes2	100	0.2	1.052 (0.053)	0.052 (0.010)	0.412 (0.022)
	100	0.5	1.052 (0.053)	0.052 (0.010)	0.412 (0.022)
	100	0.8	1.052 (0.053)	0.052 (0.010)	0.412 (0.022)
	200	0.2	0.058 (0.012)	0.020 (0.006)	0.932 (0.011)
	200	0.5	0.058 (0.012)	0.020 (0.006)	0.932 (0.011)
	200	0.8	0.058 (0.012)	0.020 (0.006)	0.932 (0.011)
	500	0.2	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
	500	0.5	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)
	500	0.8	0.000 (0.000)	0.000 (0.000)	1.000 (0.000)

Table 5
Sensitivity analysis table for approximate Bayesian methods corresponding to AR(1) predictors, $p = 10$.

	n	q	Error I	Error II	True.model
Approx.Bayes1	100	0.2	2.022 (0.056)	0.068 (0.012)	0.124 (0.015)
	100	0.5	2.018 (0.056)	0.068 (0.012)	0.128 (0.015)
	100	0.8	2.018 (0.056)	0.068 (0.012)	0.128 (0.015)
	200	0.2	0.728 (0.043)	0.302 (0.027)	0.354 (0.021)
	200	0.5	0.720 (0.043)	0.306 (0.027)	0.354 (0.021)
	200	0.8	0.720 (0.043)	0.306 (0.027)	0.354 (0.021)
	500	0.2	0.082 (0.013)	0.768 (0.036)	0.384 (0.022)
	500	0.5	0.076 (0.013)	0.930 (0.046)	0.378 (0.022)
	500	0.8	0.076 (0.013)	0.936 (0.047)	0.378 (0.022)
Approx.Bayes2	100	0.2	2.126 (0.045)	0.020 (0.006)	0.056 (0.010)
	100	0.5	2.124 (0.045)	0.020 (0.006)	0.058 (0.010)
	100	0.8	2.124 (0.045)	0.020 (0.006)	0.058 (0.010)
	200	0.2	1.328 (0.039)	0.010 (0.004)	0.154 (0.016)
	200	0.5	1.328 (0.039)	0.010 (0.004)	0.154 (0.016)
	200	0.8	1.328 (0.039)	0.010 (0.004)	0.154 (0.016)
	500	0.2	0.572 (0.031)	0.006 (0.003)	0.532 (0.022)
	500	0.5	0.572 (0.031)	0.006 (0.003)	0.532 (0.022)
	500	0.8	0.572 (0.031)	0.006 (0.003)	0.532 (0.022)

latter fails to give an output due to computer memory problems when both the sample sizes and the number of parameters are high. The BIC criterion in (26) suffered from memory problems in the high dimensional setting in at least one of the replications, when the group LASSO selects a model having dimension higher than the sample size. In such a situation, the penalized likelihood criterion works well as shown in the simulation results (Table 3). In lieu of the results obtained from the simulations, we propose to use the BIC criterion for the low dimensional examples and the penalized criterion for the high dimensional one. From the sensitivity analysis results we can see that the choice of q has hardly any effect on the results for the BIC criterion. For the criterion based on (27), the choice of q affects the two different errors, that is, failing to include an active predictor or incorrectly including an inactive predictor in the model. Lower value of q prefers lower dimensional models and hence incurs more error by excluding active predictors but at the same time has lower error rate for including inactive ones. The average number of times the true model is selected in such cases is almost comparable for different values of q. We have also tried to find out the model posterior probabilities using the Reversible Jump MCMC (RJMCMC) algorithm, but the corresponding results are far from being reliable, as this algorithm failed to visit the true model or any model close to that with appreciable probability. The corresponding results for the low dimension independent covariate structure are shown in Table 1. For $n = 500$, the algorithm failed to give an output. The RJMCMC algorithm produces similar results in the other situations, and hence we exclude the results for brevity.

5. Illustration with real data

We demonstrate our method on the NIR data set from Liebmann et al. (2009). The data set is also available in the R package 'chemometrics'. The NIR data consists of glucose and ethanol concentrations (in g/L) for 166 alcoholic fermentation mashes of different feedstock (rye, wheat and corn) to be modeled by 235 variables containing Near Infrared (NIR) spectroscopy absorbance values acquired in the wavelength range of 1115–2285 nanometer (nm) by a transflectance probe (Liebmann et al., 2009). We implement the proposed Bayesian method on the data set corresponding to the ethanol concentrations, and for comparison, we compute the group LASSO and implement the MCMC based method of Reich et al. (2009).

The group LASSO solution in this example selects only 3 predictors out of 235 variables, corresponding to the variables for wavelengths 1670, 1675 and 1705 nm. The approximate Bayesian method selects the same model with a posterior probability greater than 0.98. The median probability model selected by the MCMC based method has 148 variables including the three selected by our method. Also, the MCMC based method took 35 928 s for 10 000 iterations with 1000 burn-in samples in comparison to 377 s for our method (run on a DELL Dual Processor Xeon Six Core 3.6 GHz machine with 60 GB RAM running 64 Bit CentOS Linux 5.0), which justifies the use of the word "fast" Bayesian computation in the title of the paper.

Based on the simulation results in the previous section, it is not surprising that the group LASSO and the model with the highest approximate posterior probability were the same. The MCMC based method is very "liberal" in comparison, that is, it selects models with many more predictors than the proposed method. This appears to be owing to the inability of the MCMC algorithm to cover the entire model space within a reasonable number of runs in such high dimensional situations.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control AC-19, 716–723. System Identification and Time-Series Analysis.

Avalos, M., Grandvalet, Y., Ambroise, C., 2003. Regularization methods for additive models. In: Advances in Intelligent Data Analysis V. pp. 509–520.

Barbieri, M.M., Berger, J.O., 2004. Optimal predictive model selection. Ann. Statist. 32, 870–897.

Barry, D., 1986. Nonparametric Bayesian regression. Ann. Statist. 14, 934–953.

Belitz, C., Lang, S., 2008. Simultaneous selection of variables and smoothing parameters in structured additive regression models. Comput. Statist. Data Anal. 53, 61–81.

Bondell, H.D., Reich, B.J., 2008. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. Biometrics 64, 115–123.

Breiman, L., 1995. Better subset regression using the nonnegative garrote. Technometrics 37, 373–384.

Buja, A., Hastie, T., Tibshirani, R., 1989. Linear smoothers and additive models. Ann. Statist. 17, 453–555.

Chen, Z.H., 1993. Fitting multivariate regression functions by interaction spline models. J. R. Stat. Soc. Ser. B 55, 473–491.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32, 407–499.

Ernst, M.D., 1998. A multivariate generalized Laplace distribution. Comput. Statist. 13, 227–232.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.

Friedman, J.H., 1991. Multivariate adaptive regression splines. Ann. Statist. 19, 1–141.

George, E.I., 2000. The variable selection problem. J. Amer. Statist. Assoc. 95, 1304–1308.

George, E., McCulloch, R., 1993. Variable selection via Gibbs sampling. J. Amer. Statist. Assoc. 88, 881–889.

Gu, C., 2002. Smoothing Spline ANOVA Models. In: Springer Series in Statistics, Springer-Verlag, New York.

Gustafson, P., 2000. Bayesian regression modeling with interactions and smooth effects. J. Amer. Statist. Assoc. 95, 795–806.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. In: Monographs on Statistics and Applied Probability, vol. 43. Chapman and Hall Ltd., London.

Huang, J., Horowitz, J.L., Wei, F., 2010. Variable selection in nonparametric additive models. Ann. Statist. 38, 2282–2313.

Hwang, W.Y., Zhang, H.H., Ghosal, S., 2009. FIRST: combining forward iterative selection and shrinkage in high dimensional sparse linear regression. Stat. Interface 2, 341–348.

Lafferty, J., Wasserman, L., 2008. Rodeo: sparse, greedy nonparametric regression. Ann. Statist. 36, 28–63.

Liebmann, B., Friedl, A., Varmuza, K., 2009. Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics. Anal. Chim. Acta 642, 171–178.

Lin, Y., Zhang, H.H., 2006. Component selection and smoothing in multivariate nonparametric regression. Ann. Statist. 34, 2272–2297.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D., Ye, K.Q., 2006. Variable selection for Gaussian process models in computer experiments. Technometrics 48, 478–490.

Marra, G., Wood, S.N., 2011. Practical variable selection for generalized additive models. Comput. Statist. Data Anal. 55, 2372–2387.

Meier, L., van de Geer, S., Bühlmann, P., 2009. High-dimensional additive modeling. Ann. Statist. 37, 3779–3821.

Miller, A., 2002. Subset Selection in Regression, second ed. In: Monographs on Statistics and Applied Probability, vol. 95. Chapman & Hall/CRC, Boca Raton, FL.

Ravikumar, P., Lafferty, J., Liu, H., Wasserman, L., 2009. Sparse additive models. J. R. Stat. Soc. Ser. B Stat. Methodol. 71, 1009–1030.

Reich, B.J., Storlie, C.B., Bondell, H.D., 2009. Variable selection in Bayesian smoothing spline ANOVA models: application to deterministic computer codes. Technometrics 51, 110–120.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Shi, P., Tsai, C.L., 1999. Semiparametric regression model selections. J. Statist. Plann. Inference 77, 119–139.

Shively, T.S., Kohn, R., Wood, S., 1999. Variable selection and function estimation in additive nonparametric regression using a data-based prior. J. Amer. Statist. Assoc. 94, 777–806.

Shun, Z., McCullagh, P., 1995. Laplace approximation of high-dimensional integrals. J. R. Stat. Soc. Ser. B 57, 749–760.

Stone, C.J., 1985. Additive regression and other nonparametric models. Ann. Statist. 13, 689–705.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B 58, 267–288.

Wahba, G., 1990. Spline Models for Observational Data. In: CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Wood, S., Kohn, R., Shively, T., Jiang, W., 2002. Model selection in spline nonparametric regression. J. R. Stat. Soc. Ser. B 64, 119–139.

Yuan, M., Lin, Y., 2005. Efficient empirical Bayes variable selection and estimation in linear models. J. Amer. Statist. Assoc. 100, 1215–1225.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B 68, 49–67.

Zou, H., 2006. The adaptive Lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320.

Please cite this article in press as: McKay Curtis, S., et al., Fast Bayesian model assessment for nonparametric additive regression. Computational Statistics and Data Analysis (2013), http://dx.doi.org/10.1016/j.csda.2013.05.012