

Fast Bayesian Model Assessment for Nonparametric Additive Regression

S. McKay Curtis and Subhashis Ghosal

University of Washington and North Carolina State University

March 9, 2010

Abstract

The literature is replete with variable selection techniques for the classical linear regression model. It is only relatively recently that authors have begun to explore variable selection in fully nonparametric and additive regression models. In this paper, we consider a Bayesian approach for nonparametric additive regression models. We expand the functions in the additive model in a B-spline basis and put a multivariate Laplace prior on the coefficients. We approximately calculate posterior probability of models defined by selection of predictors in the working model, using a Laplace approximation method, where we expand the prior times the likelihood around the posterior mode. The posterior mode for this prior is the so called group LASSO, for which a fast computing algorithm exists. Thus we completely avoid Markov chain Monte-Carlo or any other time consuming sampling based method, leading to quick assessment of various posterior model probabilities. The resulting posterior probabilities may be utilized for prediction using Bayesian model averaging.

KEYWORDS: Group LASSO, Laplace approximation, model uncertainty, penalized regression, variable selection.

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), Zou and Zhang (2009), Bondell and Reich (2008), Hwang *et al.* (2009) and so on.

In many practical applications, the linear model assumptions are too restrictive and nonparametric regression models are preferred. In 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.

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 [Stone (1985), Buja *et al.* (1989), Hastie and Tibshirani (1990)] are a nice compromise between the restrictive linear model and the fully nonparametric formulation. 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), Wahba (1990), Gu (2002)].

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 norm of the projection of each functional component into a partition of the model space. Reich *et al.* (2008) 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 \mathbf{Y} be an $n \times 1$ vector of responses, \mathbf{X}_j is an $n \times m_j$ matrix of variables associated with the j th predictor (which may be stochastic or nonstochastic) and $\boldsymbol{\beta}_j$ is an $m_j \times 1$ vector of coefficients. Then group LASSO minimizes

$$\operatorname{argmin}_{\boldsymbol{\beta}} \|\mathbf{Y} - \sum_{j=1}^g \mathbf{X}_j \boldsymbol{\beta}_j\|^2 + \lambda \sum_{j=1}^g \|\boldsymbol{\beta}_j\|_{\mathbf{w}_j}, \quad (1)$$

where $\|\mathbf{z}\|_{\mathbf{A}} = (\mathbf{z}^T \mathbf{A} \mathbf{z})^{1/2}$ for $\mathbf{z} \in \mathbb{R}^d$ and \mathbf{A} is a $d \times d$ positive definite matrix, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_g^T)^T$ 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 $\boldsymbol{\beta}_j = \mathbf{0}$ for some subset of $j = 1, \dots, g$.

One drawback to 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 computa-

tionally expensive and it can be hard to assess convergence when MCMC methods must traverse a space of differing dimensions.

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 the 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 Laplace's 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 in the case of the classical linear regression model, by using Laplace's 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. Similar to Yuan and Lin (2005), the group LASSO solution turns out to be the model with highest approximate posterior probability. However, by obtaining posterior probabilities of other models, we can perform Bayesian model averaging, which is typically preferred in prediction due to its ability to incorporate uncertainty in model selection.

We organize the paper as follows. In Section 2, we formally discuss the model and prior distribution, and describe the Laplace approximation method in Section 3. The method of estimation of error variance and the tuning parameter in the multivariate Laplace prior is described in Section 4. In Section 5, through a simulation study we investigate which models carry appreciable posterior probabilities. A real data analysis is presented in Section 6.

2 Model Formulation and Prior Specification

We consider a regression model $Y = f(\mathbf{X}) + \varepsilon$, where $\mathbf{x} = (x_1, \dots, x_p)$ is a set of p -predictors and random errors $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. We assume that the regression function has an additive form $f(\mathbf{x}) = \sum_{j=1}^p f_j(x_j)$. 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 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 $k = \sum_{i=1}^p \gamma_i$ 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 $\mathbf{X} = \mathbf{x}$ in a hierarchical fashion as

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

where $\mathbf{x}_{\boldsymbol{\gamma}}$ denote the vector of the values of the selected predictors.

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

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

We shall specifically work with the flexible and convenient B -spline basis functions. The number of terms m here acts as a tuning parameter — larger values of m reduce bias, but the increased variability of the estimates of corresponding regression coefficients may reduce the accuracy of the estimated function.

We obtain n independent observations whose values are denoted by $\mathbf{y} = (y_1, \dots, y_n)$ 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

$$\boldsymbol{\Psi}_{n \times mp} = \begin{pmatrix} \psi_{11}(x_{11}) & \cdots & \psi_{1m}(x_{11}) & \cdots & \psi_{p1}(x_{1p}) & \cdots & \psi_{pm}(x_{1p}) \\ \psi_{11}(x_{21}) & \cdots & \psi_{1m}(x_{21}) & \cdots & \psi_{p1}(x_{2p}) & \cdots & \psi_{pm}(x_{2p}) \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \psi_{11}(x_{n1}) & \cdots & \psi_{1m}(x_{n1}) & \cdots & \psi_{p1}(x_{np}) & \cdots & \psi_{pm}(x_{np}) \end{pmatrix} \quad (4)$$

and the coefficients as a vector

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_p^T)^T = (\beta_{11}, \dots, \beta_{1m}, \dots, \beta_{p1}, \dots, \beta_{pm})^T. \quad (5)$$

The coefficients of the basis expansion of functions f_j not selected by $\boldsymbol{\gamma}$ are all zero. Let $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ denote and non-zero coefficient values and let $\boldsymbol{\Psi}_{\boldsymbol{\gamma}}$ denote

the matrix obtained from Ψ by discarding the columns corresponding to the irrelevant predictors. Then the model is representable as

$$\mathbf{Y}_{n \times 1} \sim \mathcal{N} \left(\begin{array}{cc} \Psi_{\gamma} & \beta_{\gamma} \\ n \times mk & mk \times 1 \end{array}, \sigma^2 \mathbf{I}_n \right). \quad (6)$$

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 $n^{1/5}$. We shall work with the choice, $m = \lfloor n^{1/5} \rfloor$, where $\lfloor z \rfloor$ is the largest integer not greater than z . Admittedly, a more justifiable approach would be to use $m = \lfloor Cn^{1/5} \rfloor$, and consider a prior on C . However, since our main goal is fast computation of model probabilities, we do not address the issue of choice of the constant. Let $\|\mathbf{x}\| = \sqrt{\mathbf{x}^T \mathbf{x}}$ and $I_A(\cdot)$ is the indicator function of a set A . We consider the prior for β_j to be degenerate at $\mathbf{0}$, 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,

$$p(\beta_j | \gamma) = (1 - \gamma_j) I_{\{\mathbf{0}\}}(\beta_j) + \gamma_j \frac{\Gamma(m/2)}{2\pi^{m/2} \Gamma(m)} \tau^m \exp \{-\tau \|\beta_j\|\}. \quad (7)$$

Thus, for the full coefficient vector β , the prior density (with respect to the product of sums of counting measure on $\mathbf{0}$ and Lebesgue measure) is

$$p(\beta | \gamma) = \left(\prod_{j \notin J_{\gamma}} I_{\{\mathbf{0}\}}(\beta_j) \right) \left(\frac{\Gamma(m/2) \tau^m}{2\pi^{m/2} \Gamma(m)} \right)^{|\gamma|} \exp \left\{ -\tau \sum_{j \in J_{\gamma}} \|\beta_j\| \right\}, \quad (8)$$

where $J_{\gamma} = \{k : \gamma_k = 1\}$.

The final piece of our hierarchical specification is a prior distribution on all models γ . We let the prior probabilities be

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

where $q \in (0, 1)$ and d_{γ} is a measure of dependence among the $|\gamma|$ variables in the model. The quantity d_{γ} in our specification is similar in spirit to the term $\det(\mathbf{X}_{\gamma}^T \mathbf{X}_{\gamma})$ in the model formulation of Yuan and Lin (2005), where $\mathbf{X}_{\gamma} = ((x_{ij}))_{1 \leq i \leq n, j \in \gamma}$. In their formulation, the term $\det(\mathbf{X}_{\gamma}^T \mathbf{X}_{\gamma})$ is small for models with 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 for the correlation

in the nonlinear setting is given by the Kendall’s tau coefficient, which is particularly good at picking up monotone relationship. We shall therefore work with the choice d_γ the determinant of the matrix of Kendall’s tau for all pairings of predictors in model γ . More formally, let κ_{ij} be Kendall’s tau for the pair of vectors \mathbf{x}_i and \mathbf{x}_j and let $\mathbf{K} = ((\kappa_{ij}))$, then $d_\gamma = \det(\mathbf{K}_\gamma)$, where \mathbf{K}_γ is a submatrix of \mathbf{K} corresponding to nonzero elements of γ . The term d_γ serves to penalize redundant models that have a high degree of dependence among the predictors. Measures of nonlinear association other than some summary measures 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 γ . 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.

3 Posterior Computation

With the model formulation and prior specification as in the last section, we can now write the joint posterior density $p(\boldsymbol{\beta}_\gamma, \gamma | \mathbf{y})$ for $\boldsymbol{\beta}_\gamma$ and γ as

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

The marginal posterior probability for model γ can be obtained by in-

tegrating out β

$$p(\gamma|\mathbf{y}) \propto C(\mathbf{y})B(\gamma) \int_{\mathbb{R}^{mp}} \exp \left\{ -\frac{\|\mathbf{y} - \Psi_\gamma \beta_\gamma\|^2 + \lambda \sum_{j \in J_\gamma} \|\beta_j\|}{2\sigma^2} \right\} d\beta \quad (11)$$

with

$$C(\mathbf{y}) = (1 - q)^p (2\pi\sigma^2)^{-n/2}, \quad B(\gamma) = \left(\frac{q}{1 - q} \frac{\Gamma(m/2)\tau^m}{2\pi^{m/2}\Gamma(m)} \right)^{|\gamma|}. \quad (12)$$

The integral in (11) can be approximated using the Laplace's approximation. Let β_γ^* denote the group LASSO solution, that is,

$$\beta_\gamma^* = \underset{\beta_\gamma}{\operatorname{argmin}} \|\mathbf{y} - \Psi_\gamma \beta_\gamma\|^2 + \lambda \sum_{j \in J_\gamma} \|\beta_j\|. \quad (13)$$

Put $\beta_\gamma = \beta_\gamma^* + \mathbf{u}$. Substituting this quantity into (11) gives the expression

$$\begin{aligned} C(\mathbf{y})B(\gamma) \exp \left\{ -\frac{\min_{\beta_\gamma} \left(\|\mathbf{y} - \Psi_\gamma \beta_\gamma\|^2 + \lambda \sum_{j \in J_\gamma} \|\beta_j\| \right)}{2\sigma^2} \right\} \\ \times \int_{\mathbb{R}^{mp}} \exp \left\{ -\frac{1}{2\sigma^2} \left[\|\Psi_\gamma \mathbf{u}\|^2 - 2\mathbf{u}^T \Psi_\gamma^T \mathbf{y}^* \right. \right. \\ \left. \left. + \lambda \sum_{j \in J_\gamma} (\|\beta_j^* + \mathbf{u}_j\| - \|\beta_j^*\|) \right] \right\} d\beta, \end{aligned} \quad (14)$$

where $\mathbf{y}^* = \mathbf{y} - \Psi_\gamma \beta_\gamma^*$ and β_j^* and \mathbf{u}_j are the elements of β_γ^* and \mathbf{u} that correspond to the basis functions of the j th predictor in model γ .

Let

$$f(\mathbf{u}) = \frac{1}{\sigma^2} \left[\|\Psi_\gamma \mathbf{u}\|^2 - 2\mathbf{u}^T \Psi_\gamma^T \mathbf{y}^* + \lambda \sum_{j \in J_\gamma} (\|\beta_j^* + \mathbf{u}_j\| - \|\beta_j^*\|) \right]. \quad (15)$$

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

$$\frac{\partial f(\mathbf{u})}{\partial \mathbf{u} \partial \mathbf{u}^T} \Big|_{\mathbf{u}=\mathbf{0}} = \frac{1}{\sigma^2} (2\Psi_\gamma^T \Psi_\gamma + \lambda \mathbf{A}), \quad (16)$$

where

$$\mathbf{A} = \begin{bmatrix} -\frac{\beta_1^* \beta_1^{*T}}{\|\beta_1^*\|^3} + \frac{\mathbf{I}_m}{\|\beta_1^*\|} & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & -\frac{\beta_2^* \beta_2^{*T}}{\|\beta_2^*\|^3} + \frac{\mathbf{I}_m}{\|\beta_2^*\|} & \cdots & \mathbf{O} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{O} & \mathbf{O} & \cdots & -\frac{\beta_k^* \beta_k^{*T}}{\|\beta_k^*\|^3} + \frac{\mathbf{I}_m}{\|\beta_k^*\|} \end{bmatrix}, \quad (17)$$

and \mathbf{O} is an $m \times m$ matrix of zeroes.

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

$$\begin{aligned}
p(\boldsymbol{\gamma}|\mathbf{y}) &\propto C(\mathbf{y})B(\boldsymbol{\gamma}) \exp \left\{ -\frac{\min_{\boldsymbol{\beta}_\gamma} \left(\|\mathbf{y} - \boldsymbol{\Psi}_\gamma \boldsymbol{\beta}_\gamma\|^2 + \lambda \sum_{j \in J_\gamma} \|\boldsymbol{\beta}_j\| \right)}{2\sigma^2} \right\} \\
&\times \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2\sigma^2} f(\mathbf{u}) \right\} d\mathbf{u} \\
&\approx C(\mathbf{y})B(\boldsymbol{\gamma}) \exp \left\{ -\frac{\min_{\boldsymbol{\beta}_\gamma} \left(\|\mathbf{y} - \boldsymbol{\Psi}_\gamma \boldsymbol{\beta}_\gamma\|^2 + \lambda \sum_{j \in J_\gamma} \|\boldsymbol{\beta}_j\| \right)}{2\sigma^2} \right\} \\
&\times \exp \left\{ -\frac{1}{2} f(\mathbf{0}) \right\} (2\pi)^{m|\boldsymbol{\gamma}|/2} \left| \frac{1}{2} \frac{\partial f(\mathbf{0})}{\partial \mathbf{u} \partial \mathbf{u}^T} \right|^{-1/2} \\
&= C(\mathbf{y})B(\boldsymbol{\gamma}) \exp \left\{ -\frac{\min_{\boldsymbol{\beta}_\gamma} \left(\|\mathbf{y} - \boldsymbol{\Psi}_\gamma \boldsymbol{\beta}_\gamma\|^2 + \lambda \sum_{j \in J_\gamma} \|\boldsymbol{\beta}_j\| \right)}{2\sigma^2} \right\} \\
&\times (2\pi)^{m|\boldsymbol{\gamma}|/2} \left| \frac{1}{\sigma^2} (\boldsymbol{\Psi}_\gamma^T \boldsymbol{\Psi}_\gamma + \frac{\lambda}{2} \mathbf{A}) \right|^{-1/2} \\
&= (1-q)^p (2\pi\sigma^2)^{-n/2} d_\gamma \left(\frac{q}{1-q} \frac{\Gamma(m/2)\tau^m}{2\pi^{m/2}\Gamma(m)} \right)^{|\boldsymbol{\gamma}|} \\
&\times (2\pi)^{m|\boldsymbol{\gamma}|/2} \left| \frac{1}{\sigma^2} (\boldsymbol{\Psi}_\gamma^T \boldsymbol{\Psi}_\gamma + \frac{\lambda}{2} \mathbf{A}) \right|^{-1/2} \\
&\times \exp \left\{ -\frac{\min_{\boldsymbol{\beta}_\gamma} \left(\|\mathbf{y} - \boldsymbol{\Psi}_\gamma \boldsymbol{\beta}_\gamma\|^2 + \lambda \sum_{j \in J_\gamma} \|\boldsymbol{\beta}_j\| \right)}{2\sigma^2} \right\}. \quad (18)
\end{aligned}$$

The approximation in (18) holds only if all components of $\boldsymbol{\beta}_\gamma^* = (\boldsymbol{\beta}_1^*, \dots, \boldsymbol{\beta}_p^*)$ 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 $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_p^T)^T$ to $\mathbf{0}$. 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. Yuan and Lin (2005) 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.

Similarly, we define a nonregular additive model as any model γ for which $\beta_j^* = \mathbf{0}$ for at least one $j \in J_\gamma$. 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 do not need calculate posterior probabilities of nonregular models.

4 Estimation of λ and σ^2

To select a value for λ we begin with an approach similar to the approach taken by Yuan and Lin (2005) for linear models. The joint density of the observation and the coefficient vectors conditional on all other model parameters is given by

$$p(\mathbf{y}, \beta_\gamma | \gamma, \lambda, \sigma^2) = (2\pi\sigma^2)^{-n/2} \left[\frac{\Gamma(m/2)}{2\Gamma(m)} \left(\frac{\lambda}{2\sigma^2\pi^{1/2}} \right)^m \right]^{|\gamma|} \times \exp \left\{ -\frac{\|\mathbf{y} - \Psi_\gamma \beta_\gamma\|^2 + \lambda \sum_{j \in J_\gamma} \|\beta_j\|}{2\sigma^2} \right\}.$$

Integrating out β_γ and then using Laplace's approximation as in (18),

$$\begin{aligned} p(\mathbf{y} | \gamma, \lambda, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \left[\frac{\Gamma(m/2)}{2\Gamma(m)} \left(\frac{\lambda}{2\sigma^2\pi^{1/2}} \right)^m \right]^{|\gamma|} \\ &\quad \times \int_{\mathbb{R}^{m|\gamma|}} \exp \left\{ -\frac{\|\mathbf{y} - \Psi_\gamma \beta_\gamma\|^2 + \lambda \sum_{j \in J_\gamma} \|\beta_j\|}{2\sigma^2} \right\} d\beta_\gamma \\ &\approx (2\pi)^{-n/2} \sigma^{-(n+m)|\gamma|} \left(\frac{\Gamma(m/2)}{\Gamma(m)} 2^{(m+2)/2} \lambda^m \right)^{|\gamma|} \\ &\quad \times \exp \left\{ -\frac{\min_{\beta_\gamma} \left(\|\mathbf{y} - \Psi_\gamma \beta_\gamma\|^2 + \lambda \sum_{j \in J_\gamma} \|\beta_j\| \right)}{2\sigma^2} \right\} \\ &\quad \times \left| \Psi_\gamma^T \Psi_\gamma + \frac{\lambda}{2} \mathbf{A} \right|^{-1/2} \end{aligned} \quad (19)$$

If we set γ in (19) equal to $\hat{\gamma}_\lambda$, the model chosen by the group LASSO for a given λ , then maximizing (19) with respect to σ^2 , we obtain

$$\hat{\sigma}^2 = \frac{\|\mathbf{y} - \Psi_{\hat{\gamma}_\lambda} \beta_{\hat{\gamma}_\lambda}^*\|^2 + \lambda \sum_{j \in J_{\hat{\gamma}_\lambda}} \|\beta_j^*\|}{n + |\hat{\gamma}_\lambda| m} \quad (20)$$

Substituting (20) back into (19) and taking -2 times the natural logarithm of (19) gives

$$\begin{aligned}
h(\lambda) &= -2|\hat{\gamma}_\lambda|[\log(m/2) - \log m] + |\hat{\gamma}_\lambda|(m+2)\log 2 - 2m|\hat{\gamma}_\lambda|\log \lambda \\
&\quad + (n + m|\hat{\gamma}_\lambda|) \left[\log \left(\frac{\|\mathbf{y} - \Psi_{\hat{\gamma}_\lambda} \boldsymbol{\beta}_{\hat{\gamma}_\lambda}^*\|^2 + \lambda \sum_{j \in J_{\hat{\gamma}_\lambda}} \|\boldsymbol{\beta}_{\hat{\gamma}_\lambda}^*\|}{n + m|\hat{\gamma}_\lambda|} \right) \right] \\
&\quad + \log \left| \Psi_{\hat{\gamma}_\lambda}^T \Psi_{\hat{\gamma}_\lambda} + \frac{\lambda}{2} \mathbf{A} \right|. \tag{21}
\end{aligned}$$

An estimate of λ can then be found by minimizing (21) by a grid search, for instance.

Simulations have shown that choosing λ based on (21) results in over-parametrized models. Therefore, we present an alternative BIC-type criterion for selecting λ .

Yuan and Lin (2006) use an estimate of the risk to select a value of the tuning parameter in their group LASSO algorithm. Their approach minimization of

$$C_p(\hat{\boldsymbol{\mu}}_\lambda) = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{\hat{\sigma}^2} - n + 2\hat{\text{df}}, \tag{22}$$

where $\hat{\boldsymbol{\mu}}_\lambda = \mathbf{X}\hat{\boldsymbol{\beta}}$ is an estimate of $E(\mathbf{Y}|\mathbf{X})$ that depends on λ through $\hat{\boldsymbol{\beta}}$, the group LASSO estimates and

$$\hat{\text{df}} = \sum_{j=1}^p I(\|\hat{\boldsymbol{\beta}}_j\| > 0) + \sum_{j=1}^p \frac{\|\hat{\boldsymbol{\beta}}_j\|}{\|\hat{\boldsymbol{\beta}}_j^{LS}\|} (m-1) \tag{23}$$

in our case.

Using the criterion suggested by Yuan and Lin (2006) also results in overparameterized models. To compensate for that, we use a slightly modified version of (22) that is similar in spirit to BIC to select a value for λ . The criterion is

$$C_p(\hat{\boldsymbol{\mu}}_\lambda) = \frac{\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2}{\hat{\sigma}^2} - n + \hat{\text{df}} \log n, \tag{24}$$

where

$$\hat{\text{df}} = m \sum_j^p I(\|\hat{\boldsymbol{\beta}}_j\| > 0). \tag{25}$$

This selection criterion uses the more severe term $\log n$ in the penalty of C_p . The idea of the logarithmic factor comes from the BIC of Schwarz (1978). Also, the second term on the right-hand side of (23) attenuates the penalty for inclusion of the j th predictor in the model by adjusting the

penalty through $\|\hat{\beta}_j\|/\|\hat{\beta}_j^{LS}\|$ which is always less than one. Our degrees-of-freedom term does not attenuate the penalty in this manner, because we are less concerned with the actual values of the coefficients and more concerned with the inclusion (or not) of each predictor in the model. Therefore, our chosen penalty penalizes each predictor the same amount for entering the model.

5 Simulation Study

To examine the performance of our method of computing approximate posterior probabilities, we conducted a simulation study, where all computation was executed in the R statistical programming language. We simulated data sets from a model with 5 “active” predictors and 5 “inactive” predictors

$$y_i = \sum_{j=1}^{10} f_j(x_{ij}) + \varepsilon_i, \quad (26)$$

where $f_1(x) = \exp(1.1x^3) - 2$, $f_2(x) = 2x - 1$, $f_3(x) = \sin(4\pi x)$, $f_4(x) = \log\{(e^2 - 1)x + 1\} - 1$, $f_5(x) = -32(x - 0.5)^2/4 + 1$, and $f_j(x) = 0$ for $j = 6, \dots, 10$, and $\varepsilon_i \sim N(0, 1)$ independently. Note that each f_j is scaled to lie in $[-1, 1]$ when $x \in [0, 1]$. This simulation model was taken from Shively *et al.* (1999).

The x_{ij} variables were generated from three 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 i th row of the x matrix from a multivariate normal distribution with an AR(1) covariance structure with variance-covariance matrix $\Sigma_{ij} = 0.7^{|i-j|}$. The third simulation scheme—the split-plot scheme—generates the i th row of the x matrix from a multivariate normal with a zero mean vector and a “split-plot” style covariance structure, i.e., Σ is a block diagonal matrix with the first block equal to

$$\Sigma_1 = \begin{bmatrix} 1.16 & 1.00 & 1.00 \\ 1.00 & 1.16 & 1.00 \\ 1.00 & 1.00 & 1.16 \end{bmatrix},$$

the second block is a 2×2 submatrix of the first block and the third block is the 5×5 identity matrix. After each x matrix is generated, the columns of the x matrix are scaled to lie in $[0, 1]$.

For each of the x -matrix-generating schemes, we simulated 1,000 data sets and calculated approximate posterior model probabilities. We recorded

Table 1: Table of simulation results for uncorrelated predictors. The column “true.mod” contains the proportion of times (out of 1000 simulated data sets) a method selected the true model. The column “false.neg” contains the average number of active variables that were not in the selected model of a given method. The column “false.pos” contains the average number of inactive variables that were included in the model of a given method.

	n	true.mod	false.neg	false.pos
bayes	100	0.269 (0.014)	0.187 (0.020)	1.261 (0.039)
bayes2	100	0.228 (0.013)	0.526 (0.024)	0.729 (0.034)
bayes3	100	0.029 (0.005)	0.761 (0.023)	0.946 (0.031)
G. LASSO	100	0.268 (0.014)	0.184 (0.020)	1.398 (0.047)
bayes	200	0.426 (0.016)	0.000 (0.000)	0.993 (0.035)
bayes2	200	0.297 (0.014)	0.421 (0.016)	0.458 (0.027)
bayes3	200	0.000 (0.000)	0.715 (0.014)	0.772 (0.026)
G. LASSO	200	0.421 (0.016)	0.000 (0.000)	1.096 (0.040)
bayes	500	0.509 (0.016)	0.000 (0.000)	0.763 (0.030)
bayes2	500	0.292 (0.014)	0.062 (0.008)	0.287 (0.021)
bayes3	500	0.001 (0.001)	0.024 (0.005)	0.300 (0.020)
G. LASSO	500	0.508 (0.016)	0.000 (0.000)	0.793 (0.032)

the proportion of times that the model with the highest posterior probability (denoted by Bayes in the tables) was the true model. We also recorded the number of times the model with the highest posterior probability failed to include an “active” predictor (a “false negative”) and each time the model incorrectly included and “inactive” predictor (a “false positive”). We recorded this same information for the models with the second and third highest posterior probability (denoted respectively by Bayes2 and Bayes3) and for the model selected by the group LASSO alone. The results are presented in tables 1, 2 and 3, where the column “true.mod” contains the proportion of times a method selected the true model, the column “false.neg” contains the average number active variables that were not in the selected model of a given method and the column “false.pos” contains the average number of inactive variables were included in the model of a given method.

Overall, the group LASSO and the model with the highest posterior probability were similar (almost exactly the same) in the number of “active” predictors that were selected. The model with the highest posterior probability tended to select less “inactive” predictors than the group LASSO. The difference between the two methods in this regard was not large, but the trend is persistent across all 3 simulations and across all sample sizes.

As we have pointed out before, the group LASSO (and other model selec-

Table 2: Table of simulation results for AR(1) predictors. The column “true.mod” contains the proportion of times (out of 1000 simulated data sets) a method selected the true model. The column “false.neg” contains the average number of active variables that were not in the selected model of a given method. The column “false.pos” contains the average number of inactive variables that were included in the model of a given method.

	n	true.mod	false.neg	false.pos
Bayes	100	0.191 (0.012)	1.094 (0.034)	0.442 (0.024)
Bayes2	100	0.026 (0.005)	1.729 (0.034)	0.296 (0.020)
Bayes3	100	0.030 (0.005)	1.750 (0.034)	0.313 (0.020)
G. LASSO	100	0.191 (0.012)	1.088 (0.034)	0.460 (0.026)
Bayes	200	0.363 (0.015)	0.491 (0.023)	0.460 (0.024)
Bayes2	200	0.067 (0.008)	1.298 (0.027)	0.288 (0.018)
Bayes3	200	0.067 (0.008)	1.287 (0.027)	0.300 (0.019)
G. LASSO	200	0.363 (0.015)	0.491 (0.023)	0.470 (0.025)
Bayes	500	0.573 (0.016)	0.160 (0.013)	0.376 (0.021)
Bayes2	500	0.169 (0.012)	0.916 (0.020)	0.134 (0.012)
Bayes3	500	0.041 (0.006)	1.048 (0.017)	0.274 (0.016)
G. LASSO	500	0.573 (0.016)	0.160 (0.013)	0.381 (0.021)

tion procedures) select only one model. In contrast, the Bayesian procedure that we have derived gives multiple models with a degree of confidence in each as measured by the posterior probability. The benefit of this feature is best demonstrated in the case of independent predictors. For a sample size of 100, the model with the highest posterior probability was the true model in 26.9% of the simulations. However, the model with the second highest posterior probability was the true model in 22.8% of the posterior simulations. Thus, the model with the highest or second highest posterior probability was the true model in close to 50% of the simulations. This trend also holds across all simulations, although it is more dramatic in the case of independent predictors.

6 Illustration with Real Data

We demonstrate our method on the NCAA data set from Mangold *et al.* (2003). We use a reduced version of the full data set where three observations with missing data have been removed from the 97 total observations Boos and Stefanski (2008). The data contain six-year graduation rates at Division I universities and 19 predictors which could affect graduation rates. Table 4 shows predictors selected by five variable selection rules given by the highest

Table 3: Table of simulation results for split-plot predictors. The column “true.mod” contains the proportion of times (out of 1000 simulated data sets) a method selected the true model. The column “false.neg” contains the average number of active variables that were not in the selected model of a given method. The column “false.pos” contains the average number of inactive variables that were included in the model of a given method.

	n	true.mod	false.neg	false.pos
Bayes	100	0.109 (0.010)	1.576 (0.035)	0.289 (0.020)
Bayes2	100	0.000 (0.000)	2.203 (0.034)	0.254 (0.019)
Bayes3	100	0.006 (0.002)	2.125 (0.035)	0.201 (0.017)
G. LASSO	100	0.109 (0.010)	1.573 (0.035)	0.291 (0.020)
Bayes	200	0.212 (0.013)	1.017 (0.029)	0.286 (0.019)
Bayes2	200	0.000 (0.000)	1.939 (0.029)	0.240 (0.018)
Bayes3	200	0.012 (0.003)	1.772 (0.029)	0.222 (0.017)
G. LASSO	200	0.212 (0.013)	1.013 (0.029)	0.296 (0.021)
Bayes	500	0.488 (0.016)	0.433 (0.020)	0.230 (0.017)
Bayes2	500	0.030 (0.005)	1.367 (0.022)	0.164 (0.014)
Bayes3	500	0.045 (0.007)	1.254 (0.021)	0.146 (0.013)
G. LASSO	500	0.488 (0.016)	0.433 (0.020)	0.230 (0.017)

and the second highest posterior probability models, the group LASSO, ordinary LASSO and COSSO.

The predictors Table 4 in are given by `top10` standing for % students in top 10% HS, `act25` for ACT composite 25th, `oncampus` for % students living on campus, `ft.grad` for % First-time undergraduates, `size` for Total enrollment/1000, `tateach` for % courses taught by TAs, `bbindex` for composite of basketball ranking, `board` for in-state tuition/1000, `board` for room and board/1000, `attend` for average basketball home attendance, `full.sal` for full professor salary, `sf.ratio` for ftudent to faculty ratio, `white` for % white, `ast.sal` for assistant professor salary, `pop` for population of city where located, `phd` for % faculty with PhD, `accept` for acceptance rate, `l.pct` for % receiving loans and `outstate` for % out of state.

We fit our Bayesian model to the data set and, for comparison, also fit the group LASSO, LASSO and the COSSO. The model selected by each method is listed in Table 4.

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 model probability were the same. The model with the second highest probability picked one less variable than the model with the highest

Table 4: Selected models for the NCAA data set using the method presented in this paper (in the Bayes column), the LASSO and the COSSO. Descriptions of the predictors in the NCAA data set are from Boos and Stefanski (2008).

Variables	Bayes	Bayes2	GLASSO	LASSO	COSSO
top10				✓	
act25	✓	✓	✓	✓	✓
oncampus	✓	✓	✓	✓	✓
ft.grad	✓		✓	✓	
size				✓	
tateach				✓	✓
bbindex					
tuition					
board				✓	
attend					✓
full.sal	✓	✓	✓	✓	✓
sf.ratio					✓
white				✓	✓
ast.sal	✓	✓	✓		✓
pop					
phd				✓	✓
accept	✓	✓	✓		✓
l.pct					✓
outstate					✓

posterior probability. The LASSO and the COSSO were both more “liberal” in this example, i.e., they both selected models with a larger number of predictors than the Bayesian method. The LASSO selected 10 of 19 and the COSSO selected 12 of 19 predictors compared to the Bayesian method that selected 6 of 19 predictors. Also, the Bayesian fit included one variable not included in the COSSO model (`ft.grad`) and two variables not included in the LASSO.

References

- Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723.
- Avalos, M., Grandvalet, Y. and Ambroise, C. (2003). Regularization meth-

- ods for additive models. In: *Lecture Notes in Computer Science* **2779**, 509–520.
- Barry, D. (1986). Nonparametric Bayesian regression. *Annals of Statistics* **14**, 934–953.
- Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics* **64**, 115–123.
- Boos, D. and Stefanski, L. (2008). Boos-Stefanski variable selection home page. <http://www4.stat.ncsu.edu/~boos/var.select/ncaa.html>.
- Breiman, L. (1995). Better subset selection using the nonnegative garrotte. *Technometrics* **37**, 373–384.
- Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. *Annals of Statistics* **17** 453–555.
- Chen, Z. (1993). Fitting multivariate regression functions by interaction spline models. *Journal of the Royal Statistical Society, Series B* **55**, 473–491.
- Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407–499.
- Ernst, M. D. (1998). A multivariate generalized Laplace distribution. *Computational Statistics* **13**, 227–232.
- Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics* **19**, 1–141.
- George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association* **95**, 1304–1308.
- George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.
- Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer.

- Gustafson, P. (2000). Bayesian regression modeling with interactions and smooth effects. *Journal of the American Statistical Association* **95**, 795–806.
- Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall.
- Hwang, W., Zhang, H. H. and Ghosal, S. (2009). FIRST: Combining forward selection and shrinkage in high dimensional linear regression. *Statistics and its Interface*, **2**, 341–348.
- Lafferty, J. and Wasserman, L., 2008. Rodeo: sparse, greedy nonparametric regression. *Annals of Statistics* **36**, 28–63.
- Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics* **34**, 2272–2297.
- Linkletter, C., Bingham, D., Hengartner, N. and Higdon, D. (2006). Variable selection for Gaussian process models in computer experiments. *Technometrics* **48**, 478–490.
- Mangold, W. D., Bean, L. and Adams, D. (2003). The impact of intercollegiate athletics on graduation rates among major NCAA division I universities: Implications for college persistence theory and practice. *The Journal of Higher Education* **74**, 540–562.
- Miller, A. (2002). *Subset Selection in Regression, 2nd Edition*. Chapman & Hall/CRC.
- Reich, B. J., Storlie, C. B. and Bondell, H. B. (2008). Bayesian variable selection for nonparametric regression. Unpublished manuscript, http://www4.stat.ncsu.edu/~bondell/Bayes_nonp.pdf
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.
- Shi, P. and Tsai, C. (1999). Semiparametric regression model selections. *Journal of Statistical Planning and Inference* **77**, 119–139.
- Shively, T. S., Kohn, R. and Wood, S. (1999). Variable selection and function estimation in additive nonparametric regression using a data-based prior. *Journal of the American Statistical Association* **94**, 777–794.

- Stone, C. (1985). Additive regression and other nonparametric models. *Annals of Statistics* **13**, 689–705.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58**, 267–288.
- Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics **59**.
- Wood, S., Kohn, R., Shively, T. and Jiang, W. (2002). Model selection in spline nonparametric regression. *Journal of the Royal Statistical Society, Series B* **64**, 119–139.
- Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *Journal of the American Statistical Association* **100**, 1215–1225.
- Yuan, L. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.
- Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* **67**, 301–320.