Computational Statistics and Data Analysis 56 (2012) 1319–1332

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

On efficient calculations for Bayesian variable selection

Eric Ruggieri [a,*], Charles E. Lawrence [b,c,1]

[a] Department of Mathematics and Computer Science, Duquesne University, Pittsburgh, PA 15282, USA

[b] Division of Applied Mathematics, Brown University, Providence, RI 02912, USA

[c] Center for Computational Molecular Biology, Brown University, Providence, RI 02912, USA

ARTICLE INFO

Article history:
Received 15 March 2010
Received in revised form 12 August 2011
Accepted 27 September 2011
Available online 8 October 2011

Keywords:
Bayesian model averaging
Variable selection
Dynamic programming
Inversion of matrix sums
Regression
Spike and slab

ABSTRACT

We describe an efficient, exact Bayesian algorithm applicable to both variable selection and model averaging problems. A fully Bayesian approach provides a more complete characterization of the posterior ensemble of possible sub-models, but presents a computational challenge as the number of candidate variables increases. While several approximation techniques have been developed to deal with problems that contain a large numbers of candidate variables, including BMA, IBMA, MCMC and Gibbs Sampling approaches, here we focus on improving the time complexity of exact inference using a recursive algorithm (Exact Bayesian Inference in Regression, or EBIR) that uses components of one sub-model to rapidly generate another and prove that its time complexity is $O(m^2)$, where m is the number candidate variables. Testing against simulated data shows that EBIR significantly reduces compute time without sacrificing accuracy, while comparisons to the results obtained by MCMC approaches on the Crime and Punishment data set show that model averaging yields improved predictive performance over two model selection approaches. In addition, we show that finite mixtures of centroid solutions provide a means to better characterize the shape of multimodal posterior spaces than any individual model. Finally, we describe how the BIC approximations employed in the BMA and IBMA algorithms can be replaced by an EBIR calculation of equal time complexity and illustrate the departure of the BIC approximation from the exact Bayesian inference of EBIR.

1. Introduction: a Bayesian approach to variable selection

Given an unknown dependent variable, y, and m known predictor variables $x_1, \ldots, x_m$, linear regression methods are based upon the statistical model

$$y = \sum_{j=1}^{m} \beta_j x_j + \varepsilon$$

where β_j is the jth regression coefficient and ε is a random error term. When a large number of predictors are available, interest often focuses on the selection of a subset of these variables.

The number of possible sub-models grows exponentially with the number of predictors, thus traditional methods of variable selection quickly become intractable for high dimensional data. In ultra-high dimensional problems, such as those encountered in genomics (for example, gene expression or Genome Wide Association Studies), a screening procedure aimed

* Corresponding author. Tel.: +1 412 396 4851; fax: +1 412 396 1937.
E-mail addresses: ruggierie@duq.edu (E. Ruggieri), Charles_Lawrence@brown.edu (C.E. Lawrence).
1 Tel.: +1 401 863 1479; fax: +1 401 863 1355.

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.09.026

at dimensional reduction is first implemented to remove many of the predictors from further consideration. For example, the Sure Independent Screening (SIS) procedure (Fan and Lv, 2008) uses the correlation of predictors to the dependent variable to screen, while the greedy Forward Regression algorithm of Wang (2009) uses the traditional forward stepwise regression method combined with the BIC criterion of Chen and Chen (2008) to scan a high dimensional space. Typically these screening procedures are employed to reduce the number of predictors down to a point where more rigorous variable selection techniques such as SCAD (Fan and Li, 2001), Lasso (Tibshirani, 1996), Adaptive Lasso (Zou, 2006; Zhang and Lu, 2007) or Bayesian approaches such as BMA (Raftery, 1995) and the Bayesian Lasso (Park and Casella, 2008; Hans, 2009, 2010) become feasible, especially for the case where the number of predictors exceeds the number of observations (Wang, 2009; Fan and Lv, 2008). Here we focus on Bayesian inference from the posterior space across the ensemble of candidate sub-models. To this end we describe an exact Bayesian algorithm with improved computational efficiency.

In the frequentist setting, common approaches to finding 'good' sets of predictors involves minimizing the sum of squared errors subject to some constraint or selection criteria, such as AIC, BIC or C_p. Examples include greedy forward and backward eliminations (Miller, 2002, and references therein) and the Leaps and Bounds technique (Furnival and Wilson, 1974). A number of penalized regression approaches have been developed including the nonnegative garrote (Breiman, 1995; Miller, 2002), Ridge Regression (Hoerl and Kennard, 1970a,b), which seeks to minimize squared error together with the square of the regression coefficients, the Lasso technique, which minimizes squared error together with a constraint on the sum of the absolute values of the coefficients (Tibshirani, 1996), and SCAD (Fan and Li, 2001), which is similar to the Lasso but uses a smoothly clipped absolute deviation penalty. Least Angle Regression (LARS) (Efron et al., 2004) is a less greedy version of traditional forward selection that with a slight modification can be transformed into the Lasso. In the Adaptive Lasso (Zou, 2006; Zhang and Lu, 2007), adaptive weights are used to penalize different coefficients. Additionally, Zou and Hastie (2005) developed the elastic net, which is a combination of Ridge Regression and the Lasso technique.

Whereas frequentist approaches look for the 'optimal' set of variables, a Bayesian approach to variable selection focuses on finding the posterior distribution across the ensemble of candidate sub-models. The Bayesian Model Averaging (BMA) algorithm introduced by Raftery (1995) can be viewed as a semi-Bayesian approach to selecting a specific subset of variables. BMA uses the branch and bound technique known as Leaps and Bounds (Furnival and Wilson, 1974) to initially screen the solution space. To expedite this process, Leaps and Bounds utilizes an efficient 'matrix sweep' procedure and stores partial results from this step for later use in BMA's evaluation step. The posterior probability of each sub-model that survives the screening process is approximated by and then ranked using the Bayesian Information Criteria (BIC). The space is further truncated by removing any sub-model that does not exceed an arbitrary bound, typically 1/20th of the maximal probability for a sub-model. Iterative BMA (IBMA) (Yeung et al., 2005) was developed in the context of problems where the number of predictors greatly exceeds the number of observations, specifically microarray data. IBMA first ranks variables in order of their individual predictive ability and then successively applies the BMA algorithm to groups of the ordered variables in an attempt to screen out unimportant sets of predictors.

In a Bayesian approach, Mitchell and Beauchamp (1988) found maximum a posteriori (MAP) estimates through an exhaustive search because of the small number of regression terms under consideration; Branch and Bound methods (Land and Doig, 1960; Mitchell and Beauchamp, 1986) were suggested for larger sets of regression terms. As Fernandez et al. (2001b) points out, the 'largest computational burden lies in the evaluation of the marginal likelihood of each model'. Therefore, the ability to make these calculations efficiently is of the utmost importance. Bayesian algorithms that go beyond identification of the mode in high-dimensional spaces include the Bayesian Lasso (Park and Casella, 2008; Hans, 2009, 2010), Markov Chain Monte Carlo (MCMC) approaches (Madigan and York, 1995; Smith and Kohn, 1996; Raftery et al., 1997; Hoeting et al., 1997; Fernandez et al., 2001a) and a Gibbs Sampling algorithm named Stochastic Search Variable Selection (SSVS) (George and McCulloch, 1993; Diebolt and Robert, 1994). MCMC and SSVS methods appear to have an advantage over MAP estimates for large numbers of predictors because they do not have to evaluate all models and also may not require a screening procedure. While of practical value, they provide only an approximate representation of the posterior space and leave open the difficult problem of assessing convergence since not all states will be visited.

Two classes of exact Bayesian approaches have been described. In the 'spike and slab' model of Mitchell and Beauchamp (1988), regression coefficients (β) are modeled by a finite mixture; a point mass at zero was considered the 'spike' while a uniform alternative distribution was considered the 'slab'. In this model, variables assigned to the 'spike' are removed. Alternatively, George and McCulloch (1993) use a mixture of Normals as their prior distribution for β. Both of the Normal distributions are centered at zero, but each has a different variance ($\beta \sim N(0, \sigma^2/k_i)$ or $\beta \sim N(0, \sigma^2/k_e)$). One of the variances is chosen to be sufficiently small so that β's drawn from this distribution can 'safely' be replaced by zero with little loss; the other variance parameter is chosen large enough to give support to the set of predictor variables whose coefficients may differ greatly from zero.

These two approaches highlight the distinction between different schools of thought on variable selection: subset selection and 'shrinkage' procedures. In subset selection, regressors are either retained or dropped completely from the model. The results are easily interpretable, but as discrete processes, they can be highly variable as small changes in the data can cause very different models to be selected (Tibshirani, 1996). Included in the class of subset selection are BMA (Raftery, 1995), IBMA (Yeung et al., 2005), MCMC coupled with a Bayesian graphical model (Madigan and York, 1995; Raftery et al., 1997; Hoeting et al., 1997; Fernandez et al., 2001a), 'spike and slab' (Mitchell and Beauchamp, 1988), and stepwise regression techniques (Miller, 2002). The 'shrinkage' procedures are continuous processes that shrink the regression coefficients of 'uninteresting' regressors towards zero but do not necessarily set them to exactly zero. Included in the 'shrinkage' procedures

are Ridge Regression (Hoerl and Kennard, 1970a,b), nonnegative garrote (Breiman, 1995; Miller, 2002), and Stochastic Search Variable Selection (SSVS) (George and McCulloch, 1993). The Lasso technique (Tibshirani, 1996; Efron et al., 2004) and its Bayesian counterpart (Park and Casella, 2008; Hans, 2009, 2010) are shrinkage procedures in which some of the coefficients are shrunk all the way to zero. Both Ridge Regression and the Lasso technique are equivalent to MAP estimators for Bayesian variable selection procedures with a Normal and Double Exponential prior distribution on the regression coefficients, respectively (Tibshirani, 1996).

Here, we describe Exact Bayesian Inference in Regression (EBIR), a procedure designed to explore the posterior ensemble of candidate predictors. Like SSVS, EBIR falls into the category of 'shrinkage' procedures as we employ a mixture of Normal priors for the regression coefficients, but more efficiently compute the quantities necessary to calculate the posterior probability of sub-models. Using the Crime and Punishment data set (Ehrlich, 1973, 1975; Vandaele, 1978), we will compare the algorithm with three other Bayesian variable selection methods that are also capable of model averaging, the Bayesian Lasso (Hans, 2010), MCMC (Raftery et al., 1997; Fernandez et al., 2001a), and BMA (Raftery, 1995). EBIR can be used not only for variable selection and model averaging, but it also permits direct sampling of subsets from the posterior. Below, we illustrate how these samples can be employed to characterize the posterior ensemble through the use of clustering and centroid estimation. We also show how EBIR employs a binary tree that is identical to the tree used in the 'sweep' procedure of BMA. As a result, EBIR provides a means to replace approximations to the posterior probability (such as BIC) with exact posterior probabilities using an algorithm whose complexity is equivalent to methods that have already been shown to handle problems with a very large number of predictors, specifically, the BMA and IBMA algorithms.

The rest of the paper is organized as follows. Section 2 describes our variable selection model. In Section 3, we describe the EBIR algorithm and prove that its time complexity is $O(m^2)$. Section 4 uses simulations and the Crime and Punishment data set (Ehrlich, 1973, 1975; Vandaele, 1978) to illustrate inferences from EBIR and for comparison with other methods. In order to assess the potential limitations of the BIC approximation used in BMA, we also use the Crime and Punishment data set to compare the results of EBIR to BMA. Section 5 consists of discussions and conclusions.

2. Calculating the probability density of the data $f(y)$

Beginning with the usual linear model assumptions that error terms, ε, are independent, mean zero, normally distributed random variables, we employ the likelihood function $f(y|\beta, \sigma^2, A_m) \sim N(X\beta, \sigma^2 I)$, where X is the matrix of regressors, I is the identity matrix, and A_m is a vector that indicates the included and excluded variables of the model being considered. Conjugate priors are chosen for the prior distributions on the vector of amplitudes, β, and the variance, σ^2. Specifically, β is a mixture of multivariate normal ($\beta \sim N(0, \sigma^2/k_i)$ or $\beta \sim N(0, \sigma^2/k_e)$, depending on whether a specific predictor is included or excluded from the final model, respectively) and $\sigma^2 \sim$ *Scaled-Inverse*$\chi^2\left(v_0, \sigma_0^2\right)$. The marginal probability of the data given a specific model, A_m, is then:

$$f(y|A_m) = \iint f\left(y|\beta, \sigma^2, A_m\right) f(\beta|\sigma^2, A_m) f\left(\sigma^2\right) d\beta \, d\sigma^2.$$

Denote N as the number of data points and m as the total number of possible predictors. Let m_i be the number of included variables in sub-model A_m and let m_e be the number of excluded variables [$m_i + m_e = m$]. Associated with the included variables is a 'wide' prior variance parameter k_i and associated with the excluded variables is a 'narrow' prior variance parameter k_e. Let I_{A_m} be a diagonal matrix with either k_i or k_e on the diagonal corresponding to whether or not a specific predictor is included. Furthermore, define $v_N = v_0 + N$, $\beta^* = \left(X^T X + I_{A_m}\right)^{-1} X^T y$, and $s_N = (y - X\beta^*)^T (y - X\beta^*) + \beta^{*T} I_{A_m} \beta^* + v_0 \sigma_0^2$. Integration yields:

$$f(y|A_m) = \frac{(v_0\sigma_0^2/2)^{v_0/2} \Gamma(v_N/2)(k_i)^{m_i/2}(k_e)^{m_e/2}}{\Gamma(v_0/2)(s_N/2)^{v_n/2}(2\pi)^{N/2}|X^T X + I_{A_m}|^{1/2}}.$$

The marginal probability of the data can then be obtained as follows:

$$f(y) = \sum_{\text{all } A_m} f(y|A_m) f(A_m).$$

Let p_i be the probability of including a predictor ('slab') and let p_e be the probability of excluding a predictor ('spike') [$p_i + p_e = 1$]. Define the prior probability of a sub-model as a Bernoulli product: $f(A_m) = p_i^{m_i} p_e^{m_e}$. Then:

$$f(y) = \frac{(v_0\sigma_0^2/2)^{v_0/2} \Gamma(v_N/2)}{\Gamma(v_0/2)(2\pi)^{N/2}} \sum_{\text{All } A_m} \frac{\left(p_i^2 k_i\right)^{m_i/2} (p_e^2 k_e)^{m_e/2}}{(s_N/2)^{v_n/2}|X^T X + I_{A_m}|^{1/2}}.$$

Notice that only a matrix determinant, matrix inverse (involved in the calculation of s_N), and a count of the number of included (or excluded) model components needs to be completed for each sub-model A_m; the rest of the terms in the density function are independent of A_m. Each of the matrix operations is naively $O(m^3)$, but can be proven to be on the order of matrix multiplication (Bunch and Hopcroft, 1974, Cormen et al., 2001, Ch. 28, Villard, 2003). To date, matrix multiplication is at best $O(m^{2.376})$ by the Coppersmith–Winograd algorithm (Coppersmith and Winograd, 1990). The complexity of s_N

depends only on the matrix inverse if we pre-compute the quantities $y^T y$, $X^T X$, and $X^T y$, and therefore has a time complexity $O(m^{2.376})$. Together with the determinant and the count on the number of included variables, the overall time complexity for an exact representation of the posterior space is $O(Lm^{2.376})$, where L is the total number of sub-models to be examined. Space complexity is upper bounded by the space required for the calculations on an individual sub-model (matrix storage—$O(m^2)$, creation of vector β^*—$O(m)$) and any overhead associated with the data set (pre-computed values such as $X^T y - O(m)$, $X^T X - O(m^2)$, etc.), and is therefore $O(m^2)$ per sub-model. The high time complexity of this exhaustive calculation points to the need for efficient matrix calculations if the regression is to be feasible for a large number of predictors.

3. Efficient calculations

Miller (1981) provided an efficient way to calculate the inverse of a sum of matrices. Suppose that G and H are arbitrary nonsingular square matrices of the same dimension and that we seek the inverse of the matrix $G + H$, should it too be nonsingular. Miller (1981) developed a recursive algorithm to calculate this inverse based upon a fundamental lemma stating that if H is a matrix of rank one, then $(G+H)^{-1} = G^{-1} - \frac{1}{1+g} G^{-1} H G^{-1}$, where $g = \mathrm{tr}(HG^{-1})$. If the matrix H is of rank $r > 1$, then we can write $H = E_1 + E_2 + \cdots + E_r$, and therefore $G + H = G + E_1 + E_2 + \cdots + E_r$ where each E_k, $1 \leq k \leq r$, has rank one (Halmos, 1958), and iteratively apply the lemma (this decomposition is not unique). Finally, Miller (1981) shows that there does exist a decomposition such that each of the 'partial sums' $C_{k+1} = G + E_1 + E_2 + \cdots + E_k$ is nonsingular for $k = 1, \ldots, r$, ensuring that the lemma can be iteratively applied. His theorem is reproduced below:

Theorem. *Let G and $G + H$ be nonsingular matrices and let H have positive rank r. Let $H = E1 + E2 + \cdots + E_r$ where each E_k has rank one and $C_{k+1} = G + E_1 + E_2 + \cdots + E_k$ is nonsingular for $k = 1, \ldots, r$. Then if $C_1 = G$,*

$$C_{k+1}^{-1} = C_k^{-1} - v_k C_k^{-1} E_k C_k^{-1}, \quad k = 1, \ldots, r$$

where

$$v_k = \frac{1}{1 + \mathrm{tr} C_k^{-1} E_k}.$$

In particular

$$(G + H)^{-1} = C_r^{-1} - v_r C_r^{-1} E_r C_r^{-1}.$$

Because of the special structure of the EBIR model, our 'H' matrix $[I_{A_m}]$is a diagonal matrix whose entries specify whether a specific variable will be included or excluded in the final model. Therefore, the rank one matrix, E_k, specifies a variable that is added to (or subtracted from) the current model. To visualize this, think of a full binary tree where one branch signifies 'inclusion' or '1' while the other branch signifies 'exclusion' or '0' and whose depth equals the number of possible predictors. Fig. 1 provides a visual aid with three possible predictors where the root of the tree consists of the null model. Each leaf on the tree holds the probability of one sub-model. The EBIR algorithm uses this lower complexity calculation of matrix inverses and determinants to efficiently work through the tree, performing a calculation only when the 'inclusion' branch is traversed (or the 'exclusion' branch if you start with the full model). Consequently, there is only a single matrix inverse and matrix determinant to calculate (rather than one for each possible sub-model) with the iterative procedure providing all additional matrix inversions and determinants. This binary tree is identical to the one utilized by the Leaps and Bounds procedure, differing only in the matrix operations ('Sweep' versus EBIR) performed at each node. Since summing over all the leaves in the tree marginalizes out each of the indicator random variables to yield the grand normalizing constant, the posterior distribution of the ensemble of all candidate sub-models and the marginal probabilities for inclusion of each variable can be readily obtained.

Theorem. *Let L be the number of sub-models under consideration, N be the length of the data set, and m be the total number of possible predictors. The EBIR algorithm has time complexity $O(m^2 L)$ and space complexity $O(m^3)$.*

In the most general setting, $L = 2^m$. However, restrictions on allowable sub-models, such as a cap on the number of included components, can reduce this number.

Proof. *Time complexity*: For examples of the complexities of recursion trees, see Cormen et al. (2001, Ch. 4). The algorithm can be broken into two parts: Initialization and Recursion. The Recursion can be further partitioned into three stages: Divide (into a number of sub-problems), Conquer (each of the sub-problems), and Combine (solutions to the sub-problems)

(i) Initialization: This step includes all calculations that are independent of sub-model choice (i.e. $(2\pi)^{N/2}$ or $\Gamma(v_N/2)$ — $O(N)$) or which need to be done only once (i.e. $X^T y$ used in the calculation of $\beta^* - O(Nm)$, $(X^T X)^{-1}$ and $|X^T X|^{1/2}$ — $O(m^{2.376})$). All operations are of polynomial order and thus will be dominated by the recursion step.

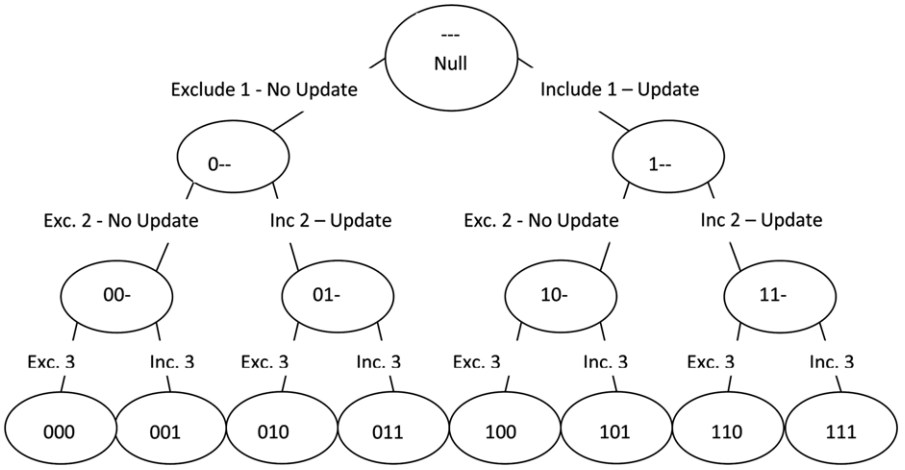

Fig. 1. A Visualization of the Recursive EBIR Procedure. Given three predictors, the binary tree illustrates how to use the initial matrix, represented as '— Null', to iteratively generate each of the eight possible sub-models. A '0' represents a specific predictor being excluded, while a '1' represents inclusion by variable selection. Each level of the tree corresponds to a specific predictor and calculations are performed only when the 'Include' branch is traversed. Since each branch of the tree is independent of its sibling, EBIR is easily parallelizable.

(ii) Recursion.

a. Divide: Each internal node of our recursion tree with two children divides the set of sub-models in half based on whether or not the current predictor is included or excluded. If excluded, no computation needs to be done and the values are simply passed to the next level of the tree—$O(1)$. If included, we need to update our matrix determinant and matrix inverse. In our case, the iterative calculations of Miller (1981) can be further simplified because each of the rank one matrices, E_k, has only one nonzero entry (denoted $e_k =$ the kth position on the diagonal). Therefore, $v_k = \frac{1}{1+C_k^{-1}(k,k)e_k}$, where $C_k^{-1}(k,k)$ is the specified position in the matrix C_k^{-1}; $C_k^{-1}E_kC_k^{-1}$ reduces to vector multiplication ($e_k *$ kth column of ${C_k}^{-1} *$ kth row of ${C_k}^{-1}$). Thus, after the initial matrix inversion and matrix determinant calculation at the root of the tree, each subsequent matrix inverse is an $O(m^2)$ multiplication and an $O(m^2)$ matrix addition, while the matrix determinant is simply an $O(1)$ multiplication of constants $\left[\det C_{r+1} = \frac{1}{v_1 v_2 \cdots v_r} \det C_1\right]$. In total, the divide step is $O(m^2)$.

b. Conquer: The set of sub-models under consideration is continually divided in half until we reach the leaves of the recursion tree. Here, we need to compute $f(y|A_m) \propto \frac{\left(p_i^2 k_i\right)^{m_i/2} (p_e^2 k_e)^{me/2}}{(s_N/2)^{\nu n/2} |X^T X + I_{A_m}|^{1/2}}$. The numerator is $O(m)$, the determinant has already been calculated, and so it remains only to compute $\beta^* = \left(X^T X + I_{A_m}\right)^{-1} X^T y$ and then $s_N = (y - X\beta^*)^T (y - X\beta^*) + \beta^{*T} I_{A_m} \beta^* + v_0 \sigma_0^2$. The matrix inverse and $X^T y$ have already been computed so β^* is $O(m^2)$, while s_N is $O(m^2) + O(m) + O(1) = O(m^2)$. In total, the conquer step is $O(m^2)$.

c. Combine: The cost of combining the sub-problems is $O(1)$ if once calculated, the probabilities of each-sub-model are stored in memory.

To find the overall time complexity of the Recursion step, we can add the time spent at each level of the tree. Each internal node is both a Divide and a Combine step. For a tree with L leaves, there are $L - 1$ internal nodes with two children, a total of $O(m^2 L)$. Each leaf on the tree is a Conquer step. Thus, computation of all leaves costs $O(m^2 L)$. In total, the recursion costs $O(m^2 L) + O(m^2 L) = O(m^2 L)$. Since the Initialization is of polynomial order, the entire algorithm has time complexity $O(m^2 L)$. Alternatively, if $L = 2^m$, define $T(n)$ to be the time complexity of a problem of size n and define the recurrence:

$$T\left(2^n\right) = \begin{cases} 2T\left(2^{n-1}\right) + O(m^2) & n > 0 \\ O(m^2) & n = 0, \text{ i.e. } T(1). \end{cases}$$

Solving the recurrence for $T(2^m)$ gives the desired result.

Space complexity: At Initialization, we create several $O(1)$ variables, the vector $X^T y[O(m)]$, as well as the initial matrix $[O(m^2)]$ whose inverse $[O(m^2)]$ and determinant $[O(1)]$ are calculated. In total, $O(m^2)$. Performing a depth-first traversal of the tree requires the storage of one matrix $O(m^2)$, the value of the determinant $O(1)$, and a vector recording the inclusion/exclusion status of the variables $O(m)$ at each level. Temporary variables involved in the updating of the matrix inverse and determinant are maximally $O(m)$. Since the depth of the tree is m, the space requirement of 'divide' is $O(m^3) + O(m) + O(m^2) + O(m) = O(m^3)$. At each leaf, the calculation of $f(y|A_m)$ requires the temporary creation of the $O(m)$ vector β^*. By writing to disk $f(y|A_m)$ once computed for each sub-model, the total space complexity for 'conquer' is $O(m)$. 'Combine' requires no space. Thus, the overall space complexity for initialization and recursion is $O(m^2) + O(m^3) + O(m) = O(m^3)$. □

Table 1
The Improvement in Speed of EBIR over Brute Force Enumeration. The time (in seconds) required to directly calculate the posterior probability of all possible sub-models through brute force enumeration (Column 1), or via the recursive EBIR algorithm (Column 2). The fold reduction in time (Column 3) compares the timing of EBIR to brute force regression. The posterior distributions for the two methods are identical.

# Variables	Brute force	EBIR	Fold reduction
12	0.64	0.31	2.06
14	3.11	1.22	2.55
16	14.54	5.00	2.91
18	70.44	20.60	3.42
20	325.94	84.99	3.84
22	1531.05	347.85	4.40
24	6918.71	1419.45	4.87
26	34260.74	5921.75	5.79

4. Simulation study and comparison to existing methods

To illustrate the EBIR algorithm, we have created a simulation that displays the improvement in speed of the recursion versus a direct (brute force) calculation of the posterior distribution. In addition, we examine the well-studied Crime and Punishment data set (Ehrlich, 1973, 1975; Vandaele, 1978) and use it to compare EBIR with other Bayesian methods. Bayesian approaches permit an exploration of the full posterior space, rather than returning a single solution that is optimal under a specified loss function. Thus, these approaches permit a greater characterization of the posterior spaces as we will illustrate below. In all cases, EBIR was run on Matlab 2008b on a laptop with an Intel® Core™ 2 Duo CPU 2.26 GHz processor with 2 GB of RAM. For this first example, we choose the parameters of the variable selection algorithm as: $p_i = p_e$ (all sub-models equally likely) and $k_i = 0.01$, $k_e = 100$ which is equivalent to the parameter setting $\left(\frac{\sigma_{\beta_i}}{\tau_i}, c_i\right) = (10, 100)$ in George and McCulloch (1993).

4.1. Improvements in speed

To evaluate the efficiency of the recursion over a brute force approach, we compare the amount of time required to enumerate and calculate the marginal probability of all possible sub-models using EBIR to a direct calculation of the posterior distribution. The simulation involves a progression from $m = 12$ to 26 potential predictors using $N = 200$ observations, yielding $L = 2^{12}$–2^{26} possible sub-models. Predictors are obtained as independent standard normal vectors $X_1, \ldots, X_{26}$ i.i.d. $\sim N_{200}(0, 1)$ so that they are essentially uncorrelated. In all cases, the dependent variable, Y, was generated according to the model

$$Y = 10X_1 - 12X_2 - 7X_3 + 5X_4 + 2X_5 - X_6 + \varepsilon$$

where $\varepsilon \sim N_{200}(0, \sigma^2 I)$ with $\sigma = 2$. As shown in Table 1, the advantage of EBIR grows with the number of variables, resulting in an almost 6-fold reduction in time when 26 variables are considered. This implies that while EBIR's reduction in time complexity may only be $O(m^{0.376})$, the difference in computation time can be significant even for small values of m. Both types of calculations yield an identical posterior distribution on the sub-models, thus the predictive performance is unchanged.

Suppose that you have a very large number of predictors (>100), but suspect that a smaller number of these predictors are significant. Taking advantage of the recursion, we can eliminate from consideration any sub-model whose number of included variables exceeds some threshold ($k_{\max}$). In doing so, entire branches of the binary tree are pruned, reducing L from exponential to $\sum_{k=1}^{k_{\max}} \binom{m}{k} \sim O(m^{k_{\max}})$. To simulate this type of analysis, we will assume that we have $N = 250$ observations and a progression of $m = 50$–250 predictors. We 'expect' that 3 or fewer predictors are significant. Again, predictors are represented as independent standard normal vectors $X_1, \ldots, X_{250}$ i.i.d. $\sim N_{250}(0, 1)$ so that they are essentially uncorrelated. The dependent variable, Y, is generated according to the model

$$Y = 5X_{17} - 6X_{29} + 3X_{41} + \varepsilon$$

where $\varepsilon \sim N_{250}(0, \sigma^2 I)$ with $\sigma = 2$. As Fig. 2 shows, the advantage of EBIR increases rapidly with the number of candidate variables, m, and shows an almost 9-fold reduction in time over a brute force approach with 250 candidate variables. Exact times can be found in the Supplemental Material.

A second approach to handling a very large number of predictor variables (>100) is to screen the variables using the iterative approach employed in the IBMA algorithm (Yeung et al., 2005). Here, the predictors are first ranked according to their individual predictive ability and then groups of variables are swapped in and out of the algorithm. For example, suppose you have 100 predictors and use an iterative approach that considers 20 variables at a time, the top 10 variables in terms of predictive ability and 10 of the remaining 90 variables. The latter group of 10 variables is swapped into and out of the program until all of the variables have been considered. Given that the EBIR algorithm can exhaustively search through 20 variables in roughly 85 s (Table 1), this iterative approach would be able to search 100 variables in about 765 s,

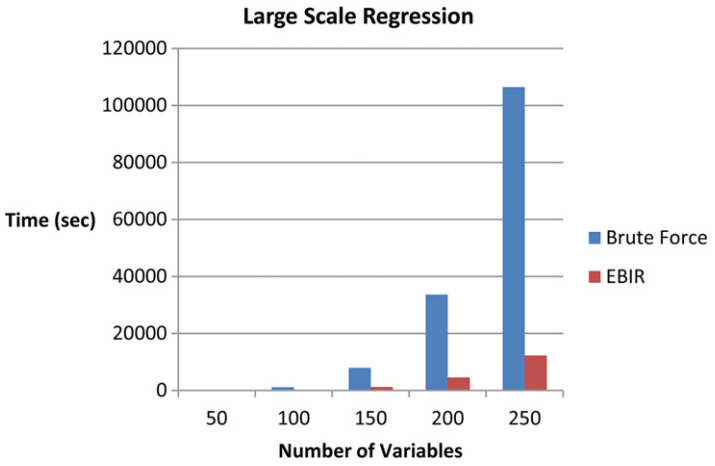

Fig. 2. Simulation of a Large-Scale Regression. Given a large number of candidate variables (m), the simulation assumes that only a small number of variables are significant. Shown is the amount of time (in seconds) required to directly compute the probability of all possible sub-models that have three or fewer interacting predictors by brute force enumeration and through the EBIR recursion given a data set of size $N = 250$.

or just under 13 min. Enlarging the problem further, 250 predictors would able to be analyzed in roughly 34 min. The time required for this approach will obviously change depending on the number of variables considered at any one time and on the number of variables swapped at each iteration. The downside of this approach is that it no longer evaluates the entire posterior space.

4.2. The crime and punishment data set

4.2.1. Comparison to MCMC approaches

Traditionally, crime was characterized by deviant behavior that was linked to the offender's presumed unique motivation, albeit psychological, social, or family circumstances (for an overview, see Taft and England, 1964). In the late 1960s and early 1970s, this paradigm changed and investigators instead sought to examine the relationship between crime and various measurable quantities such as juvenile delinquency, variations in income and unemployment conditions (Fleisher, 1966) and the probability and severity of punishment (Ehrlich, 1973, and references therein). Becker (1968) and Stigler (1970) argued for an 'economics of crime'—the decision to engage in criminal activity was a rational choice determined by the costs and benefits relative to other (legitimate) opportunities. If criminal activity was the outcome of a rational economic decision, then the probability of punishment should act as a deterrent. Ehrlich (1973) developed a theory of the participation in illegitimate activities, specified it mathematically, and tested it against empirical data obtained in 1960 from 47 states (excluding Hawaii, Alaska, and New Jersey). Errors in Ehrlich's (1973) empirical analysis were corrected by Vandaele (1978) whose data is used here. No attempt is made to consider the merits of Ehrlich's theory. This data set is used to illustrate the posterior inferences of EBIR and to compare the results of our variable selection algorithm to the Bayesian Lasso (Hans, 2010) as well as to that of Raftery et al. (1997) and Fernandez et al. (2001a) who also used this data set for illustrative purposes.

In a Metropolis–Hastings MCMC approach to solving a variable selection problem, the Markov chain is constructed by defining a neighborhood $\text{nbd}(A_m)$ for each model that consists of the model A_m itself, and the set of models with either one more or one less variable than A_m (Madigan and York, 1995; Raftery et al., 1997; Fernandez et al., 2001a). The transition matrix $\mathbf{q}$ is then defined as $\mathbf{q}(A_m \to A_m^*)$ equal to a constant for all $A_m^* \in \text{nbd}(A_m)$ and $\mathbf{q}(A_m \to A_m^*) = 0$ otherwise (Raftery et al., 1997). If the chain is currently in state A_m, then the next state is drawn from the transition matrix $\mathbf{q}$ and accepted with probability

$$\min\left\{1, \frac{\Pr(A_m^*|Y)}{\Pr(A_m|Y)}\right\}.$$

To calculate this quantity, the likelihood of the data given the proposed model, $f(y|A_m^*)$, must first be calculated, which as in our variable selection model involves a matrix determinant and a matrix inverse, both of which are $O(m^{2.376})$. Computationally, two differences exist between the MCMC procedures of Raftery et al. (1997) and Fernandez et al. (2001a) and the EBIR procedure. First, the MCMC procedures calculate the likelihood function only if it is the first time that a state is visited. Not all of the states will necessarily be visited by the Markov Chain, while other states may be visited repeatedly. In the EBIR procedure, each and every state is visited exactly once. It is this feature of EBIR that allows for exact inferences to be made about the posterior space. Second, the MCMC procedures of Raftery et al. (1997) and Fernandez et al. (2001a) are 'subset selection' procedures as described earlier. This means that the matrix of predictors is not the full matrix, but includes only

the predictors selected for inclusion in the model. Thus, the matrix that has to be inverted and whose determinant calculated in the likelihood function is usually smaller than the full matrix, reducing the computational burden. However, these matrix operations remain $O(m^{2.376})$ as the speed-up for the matrix calculations we propose (which reduces the complexity to $O(m^2)$) in its current form is not applicable to 'subset selection' procedures and is therefore not utilized by these MCMC algorithms.

As to the Bayesian Lasso, the main difference between the models utilized in the EBIR algorithm and the Bayesian Lasso is the prior distribution on β. EBIR utilizes a Normal prior on β whereas the Bayesian Lasso has a Double Exponential prior on β. Because the Double Exponential prior is not conjugate, computing the marginal likelihood of the data requires the numerical integration of 2^k k-dimensional multivariate Normal integrals for a model of size k (Hans, 2010). Increasing the size of the model increases both the number and dimension of the integrals that need to be evaluated, resulting in longer compute times and possible inaccuracies due to the numerical integration. To enumerate a model space with m possible predictors requires the calculation of 3^m integrals, which becomes intractable for $m > 12$ (Hans, 2010). To circumvent these computational issues, a Gibbs Sampling approach similar to SVSS was developed in which updating each regression coefficient requires only the numerical computation of two one-dimensional Normal probabilities (which can be done quickly and accurately) instead of 2^k k-dimensional multivariate Normal probabilities (Hans, 2010). The Bayesian Lasso has not previously been used to analyze the Crime and Punishment data set. After a burn-in period of 1000, a chain of length 200,000 was generated using the default priors, saving every 5th iteration.

To model the Crime and Punishment data set, we use the basic regression model given above, with the dependent variable, y, representing the per capita crime rate. The fifteen possible predictors are shown in Table 2. As in the original analysis, all data were log transformed except for the indicator variable for southern states. Given the 15 predictors, there are a total of $2^{15} = 32{,}768$ possible sub-models to explore.

The results of the Crime and Punishment data analysis using EBIR are shown in Tables 2 and 3. Table 2 displays the marginal probability of a given regressor being selected for inclusion while Table 3 shows the top performing sub-models. As George and McCulloch (1993) demonstrate, altering the parameters of the prior distribution associated with the inclusion and exclusion of variables places more or less focus on the variables most strongly associated with the response. When comparing these results to previous analyses on the Crime and Punishment data set (Raftery et al., 1997; Fernandez et al., 2001a), we find that setting the parameters for the prior distribution on β to values used by George and McCulloch (1993) $\left(k_i = 0.01,\, k_e = 100 \text{ equivalent to } \left(\frac{\sigma_{\beta_i}}{\tau_i},\, c_i\right) = (10, 100)\right)$ push the algorithm to be conservative in its selection for inclusion. Changing these parameters to, for example, $k_i = 0.09$, $k_e = 100$ $\left(\text{equivalent to } \left(\frac{\sigma_{\beta_i}}{\tau_i}, c_i\right) = (10, 100/3)\right)$ makes the algorithm less conservative in its selections and provides results more akin to those obtained by both Raftery et al. (1997) and Fernandez et al. (2001a) (Table 2). In the latter case, the top sub-model of Raftery et al. (1997) and Fernandez et al. (2001a) has the second largest posterior probability while both of Ehrlich's (1973) models fail to garner much support (Table 3). On the other hand, the Bayesian Lasso was found to be more inclusive of variables than both the EBIR and the MCMC procedures (Table 2). This is reflected not only in the marginal probability of inclusion for each of the individual variables, but in the p_i parameter for this model (represented by ϕ), which the Gibbs Sampler found to be ∼0.80. Perhaps the most interesting feature is that the entire EBIR algorithm takes a mere ∼1.4 s to compute all 2^{15} possible sub-models, far outperforming the Bayesian Lasso (∼90 s) and the MCMC approach of Fernandez et al. (2001a) (80 s), even after consideration for advances made in computing over the last several years. However, one must keep in mind that the time required for MCMC approaches depends on a threshold of tolerance set for the algorithm, which will affect the length of the chain.

4.2.2. Comparison to BMA

The Leaps and Bounds procedure (Furnival and Wilson, 1974) employed by Raftery's (1995) BMA algorithm uses a tree structure that is identical to the binary tree described in Section 3, differing only in the way Leaps and Bounds organizes the variables, which will alter the traversal order. The Leaps and Bounds algorithm creates two trees which are traversed in tandem. The root of one tree is the full model while the root of the other tree is the null model. In an attempt to cluster 'good' models near the roots of the trees and find the optimal sub-models of each size as quickly as possible, the variables are ordered by their individual fit to the data. The 'good' models provide sharper bounds for the algorithm and the quicker that they are found, the more of the sample space that can be truncated by the branch and bound technique.

The move from one sub-model to the next in the Leaps and Bounds technique is accomplished via a matrix 'sweep' operation akin to Gaussian elimination. The 'sweep' operation, like EBIR, uses one sub-model to quickly derive another and like EBIR, is $O(m^2)$. Additionally, since the 'sweep' operation is completed once for each possible sub-model, it will be done an equivalent number of times as the EBIR posterior probability calculation if the same truncation rules are applied to both procedures. 'Sweeping' can be used to produce a matrix inverse $(X^TX)^{-1}$, as well as the quantities β^* and $(y - X\beta^*)^T (y - X\beta^*)$ needed for *least squares* regression (quantities utilized by the BMA procedure). Replacing the 'sweep' operation with the $O(m^2)$ EBIR algorithm described above provides all of the necessary components of the *probabilistic* calculation, including the matrix inverse $\left(X^TX + I_{A_m}\right)^{-1}$ and the matrix determinant $|X^TX + I_{A_m}|^{1/2}$ without increasing the time complexity.

Table 2
The Crime and Punishment Data Set. The marginal probability (×100) of each of the 15 possible predictors of the 1960 crime rate being selected for inclusion using two different parameter settings for EBIR, $k_i = 0.01$, $k_e = 100$ and $k_i = 0.09$, $k_e = 100$. Results are compared to the MCMC analysis of Raftery et al. (1997) (denoted R97) and Fernandez et al. (2001a) (denoted F01), the Bayesian Lasso (Hans, 2010) (denoted H10), and the models hypothesized by Ehrlich (1973) (denoted E1 and E2).

Predictor number	Predictor	$k_i = 0.01$, $k_e = 100$	$k_i = 0.09$, $k_e = 100$	R97	F01	H10	E1	E2
1	Percentage of males age 14–24	41	73	79	86	78		*
2	Indicator variable for southern state	7	17	17	22	43		
3	Mean years of schooling	73	95	98	99	86		
4	Police Expenditure in 1960	66	72	72	67	86		
5	Police Expenditure in 1959	42	49	50	42	72		
6	Labor Force Participation Rate	3	8	6	15	62		*
7	Number of Males per 1000 Females	4	8	7	15	70		
8	State Population	7	19	23	33	58		
9	Number of Nonwhites per 1000 People	21	56	62	69	49	*	*
10	Unemployment rate Urban Males, age 14–24	3	9	11	20	50		*
11	Unemployment rate Urban Males, age 35–39	10	35	45	60	57		
12	Wealth	13	29	30	31	57	*	*
13	Income Inequality	99	100	100	100	99	*	*
14	Probability of Imprisonment	40	78	83	91	80	*	*
15	Average Time Served in State Prisons	4	18	22	33	48	*	*

Table 3
MAP estimates of the Crime and Punishment Data Set. The top 10 sub-models in terms of their posterior probability (×100) as determined by EBIR under two different parameter settings. For comparison purposes, the top sub-model of Raftery et al. (1997) (denoted R97) and Fernandez et al. (2001a) (denoted F01), as well as the models hypothesized by Ehrlich (1973) (denoted E1, E2), have also been included.

Rank:	Posterior probability (%), $k_i = 0.01$, $k_e = 100$	Model												
1	6.79		3	4								13		
2	6.43	1	3	4								13		
3	5.33			4								13		
4	3.76		3		5							13		
5	3.49				5							13		
6	3.35		3	4								13	14	
7	3.22	1	3	4								13	14	
8	2.72		3		5							13	14	
9	2.32		3	4				9				13	14	
10	2.20	1	3		5							13		
R97, F01	0.39	1	3	4				9		11		13	14	
E1	2.41E−10	1				6		9	10		12	13	14	15
E2	1.73E−06							9			12	13	14	15
Rank:	Posterior probability (%) $k_i = 0.09$, $k_e = 100$	Model												
1	2.71	1	3	4				9				13	14	
2	2.01	1	3	4				9		11		13	14	
3	1.91	1	3	4								13	14	
4	1.55	1	3	4				9		11		13	14	
5	1.51		3	4								13	14	
6	1.39	1	3		5			9				13	14	
7	1.32	1	3	4								13		
8	1.31	1	3	4				9				13	14	15
9	1.13	1	3		5			9		11		13	14	
10	1.13	1	3	4				9			12	13	14	
R97, F01	2.01	1	3	4				9		11		13	14	
E1	4.37E−09	1				6		9	10		12	13	14	15
E2	1.21E−06							9			12	13	14	15

Leaps and Bounds returns the sum of squared error for each sub-model that it analyzes. BMA converts this squared error to a BIC statistic which approximates the probability of the remaining sub-models given the data. Given the squared error, calculation of the BIC statistic requires a constant number of operations (two logarithms and three $O(1)$ additions and multiplications). On the other hand, once the EBIR matrix operations are complete, two $O(m)$ vector multiplications and a constant number of other operations (two logarithms, and fewer than a dozen $O(1)$ additions and multiplications, depending on how parameter values are stored in memory) need to be computed to obtain the exact posterior probability. Thus, when same elimination rules are employed, only a small difference in the constants exists between the two procedures.

Table 4
Comparing BMA and EBIR. Each row of the table represents one of the 14 (corrected) sub-models selected by *Strict Occam's Window* with its corresponding BIC Score and BIC Probability, normalized to only the 14 sub-models shown below (Raftery, 1995). The Exact Probability column gives the posterior probability of these 14 sub-models using EBIR, again normalized to only the 14 models shown below (which account for only ∼19% of the posterior space). The True Rank column shows the overall probabilistic rank of the given sub-model from the posterior distribution when the exact probabilities are calculated on the entire sample space. This column can be compared to the first, which gives the rank of the sub-models according to the BMA procedure. The last three rows of the table give the marginal probability of each variable being selected for inclusion (PMP = Posterior Model Probability). The 'True PMP' is taken from the analysis done for Table 3. All probabilities have been multiplied by 100.

BMA rank	Models															BIC score	BIC prob.	Exact prob.	True rank
1	1		3	4					9		11		13	14	15	−55.9	24.0	5.1	**17**
2	1		3	4					9		11		13	14		−55.4	18.3	11.4	**2**
3	1		3		5				9		11		13	14		−54.4	11.3	6.4	**9**
4	1		3	4					9				13	14	15	−53.8	8.3	7.4	**8**
5	1		3	4							11		13	14		−53.6	7.7	8.8	**4**
6	1		3	4				8	9				13	14		−53.1	5.8	4.4	**21**
7	1		3		5						11		13	14		−52.7	4.9	5.1	**16**
8			3	4				8	9				13	14		−52.4	4.2	5.7	**14**
9	1		3	4					9				13	14		−52.4	4.2	15.3	**1**
10	1		3		5				9				13	14	15	−51.5	2.7	2.5	**47**
11			3		5			8	9				13	14		−51.3	2.4	3.1	**34**
12	1		3		5				9				13	14		−51.2	2.3	7.9	**6**
13	1		3	4							11		13			−50.9	2.0	6.1	**12**
14	1		3	4									13	14		−50.9	1.9	10.8	**3**
	93	0	100	76	24	0	0	12	84	0	68	0	100	98	35	BIC PMP			
	91	0	100	75	25	0	0	13	69	0	43	0	100	94	15	Exact PMP			
	73	17	95	72	49	8	8	19	56	9	35	29	100	78	18	True PMP			

For example, in a direct comparison on the Crime and Punishment data set, the ratio of the compute time for EBIR to BMA was ∼1.5:1, implying that the cost of an exact answer is minimal. Furthermore, these calculations were less than 25% of the total run time of Leaps and Bounds. These specific results are of course dependent on the computer and programming language used, and the efficiency of the written code. Since our exact probabilistic calculation can be done in nearly the same time as the BIC approximation, an approximation of the posterior probability is no longer necessary for a method that has been shown to work for very large problems (Yeung et al., 2005; Eicher et al., 2007).

Given the set of sub-models produced by the Leaps and Bounds algorithm on the Crime and Punishment data set (Raftery, 1995), a further truncation of the sample space is done by BMA based on the BIC statistic. *Symmetric Occam's Window* eliminates all sub-models that are much less likely than the most likely sub-model, by default 20 times less likely. *Strict Occam's Window* further reduces the set of models by eliminating all sub-models with more likely sub-models nested within them. The *Symmetric Occam's Window* leaves 51 possible sub-models covering ∼31% of the posterior space, while the *Strict Occam's Window* leaves 14 possible sub-models covering ∼18% of the posterior space for the Crime and Punishment data set (Raftery, 1995). Table 4 compares the BIC approximation of the probability of each sub-model left by *Strict Occam's Window* to its true posterior probability, normalized by the sum of these 14 sub-models. Table 4 also provides the probabilistic ranking of each of the sub-models according to the BMA procedure and to the posterior probabilities from EBIR. As shown in the table, the model ranks based on BMA do not correspond well with ranks based on exact posterior probabilities as only two have the same rank. Furthermore, the model with the highest posterior probability has BMA rank #9, while the model ranked highest by BMA is ranked #17 based on its EBIR posterior probability.

4.2.3. Model averaging and predictive performance

Using the MAP model as a point estimate is only one of the possible ways to describe the posterior distribution over the set of included variables. Since the EBIR algorithm calculates the posterior probability of every possible sub-model, these probabilities can be combined to form the 'average' solution through a process known as model averaging (Raftery, 1995). For example, suppose that Δ is a quantity of interest. The Bayesian inference about Δ is based on its posterior distribution, which is:

$$f(\Delta|Y) = \sum_{\text{All } A_m} f(\Delta|Y, A_m) f(A_m|Y)$$

i.e. The posterior distribution of the quantity of interest, Δ, given the data, Y, is the sum of the posterior distribution of Δ given each of the possible models, weighted according to their posterior probability. For the Crime and Punishment data set, the quantity of interest is the prediction of the level of crime for a given set of variables; the average solution would be the one where each variable is included in proportion to its marginal probability.

However, the 'average' solution is typically not among the set of feasible solutions to a variable selection problem because marginal probabilities are almost certainly not integer. The solution which falls closest to the average solution under squared error loss is the centroid, which is the solution that minimizes the expected squared distance to the posterior mean (Carvalho and Lawrence, 2008). Furthermore, when inferences on binary or nominal variables are of interest, centroid estimators yield solutions that minimize Hamming distance and all pth power loss functions. To find the centroid solution in a variable

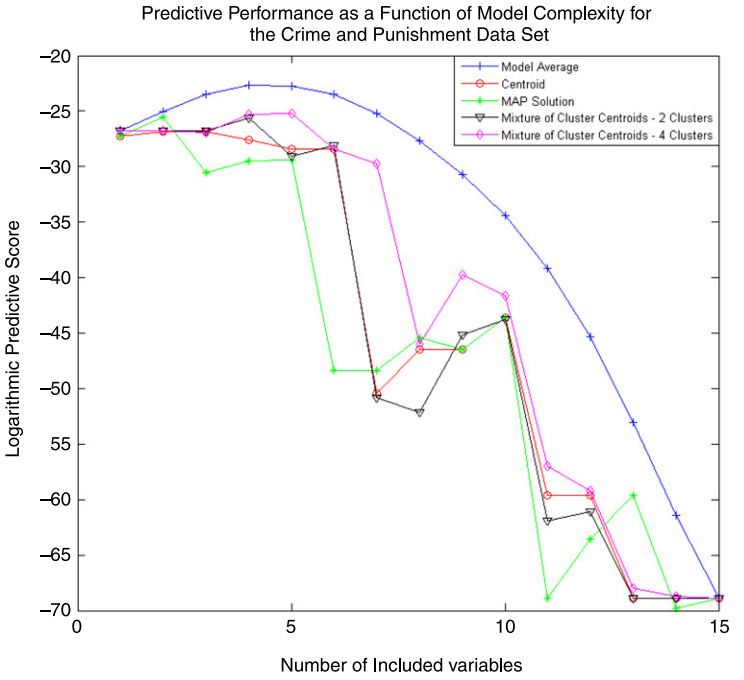

Fig. 3. Predictive performance as a function of model complexity for the Crime and Punishment data set. Using the logarithmic scoring rule of Good (1952), the predictive performance of the model averaged, the ensemble centroid, a mixture of cluster centroids, and the MAP solution are plotted for a given number of included variables on the Crime and Punishment data set.

selection problem, include any predictor whose marginal probability is above 50% and exclude any predictor whose marginal probability is below 50%. For example, in the Crime and Punishment data set, the centroid solution would be to include variables 1, 3, 4, 9, 13 and 14, which coincidentally is also the MAP estimate. By choosing a set of variables for inclusion, we expect to give up some of the predictive ability of the model averaged result. But how much do we give up by choosing a single solution? To find out, we compare the predictive performance of these solutions.

To measure predictive performance, we randomly split the Crime and Punishment data set into two halves, one, Y_{Train}, which will be used for training and the other, Y_{Test}, which will be used to test a model's predictive ability. For an individual model, A_m, the logarithmic scoring rule of Good (1952) takes the form:

$$\sum_{y \in Y_{\text{Test}}} \log f(y|A_m, Y_{\text{Train}}).$$

Predictive performance for the model averaged solution is therefore:

$$\sum_{y \in Y_{\text{Test}}} \log \left\{ \sum_{\text{All } A_m} f(y|A_m, Y_{\text{Train}}) f(A_m|Y_{\text{Train}}) \right\}.$$

Fig. 3 shows the score of the MAP, Centroid, and Model Averaged solution for a given number of included predictors. Built into a Bayesian model is the trade-off between model complexity and a better fitting model. Thus, we expect a logarithmic scoring rule (which is similar to Shannon Entropy) to peak, after which an improvement in fit is counterbalanced by an increase in the model complexity. Fig. 3 shows that this peak occurs for models with 4–5 included variables. As you would expect, the model averaged solution provides a better prediction than does any individual model using this logarithmic scoring rule (Madigan and Raftery, 1994). Note that the MAP solution is not always the best model in terms of its predictive ability. The centroid solution appears to fare somewhat better where the curve for the model averaged solution peaks.

However, there is no assurance that the posterior distribution in a model selection problem is convex. Instead, it may contain multiple modes. Clustering provides a useful way to ascertain multimodality in posterior spaces (Ding et al., 2006). When the posterior contains distinct clusters of solutions, no single model can reasonably characterize the posterior space, since such a point estimate must either fit one of the clusters or lie in the space outside of each. In this setting, the mean will lie in the 'desert' between the distinct clusters, while the mode can be anywhere in the space, even far from the bulk of the posterior mass (Carvalho and Lawrence, 2008). Thus, the characterization of a multimodal posterior space will require a least one point estimate for each distinct cluster.

To explore the posterior space for evidence of multimodality, we clustered the posterior weighted ensemble of possible sub-models using the K-Means clustering algorithm built into Matlab. The centroids for two clusters contain variables 1,

3, 9, 13, 14 and either variable #4 or #5. With four clusters, we have the same two cluster centroids as before, both with and without variable #11. The predictive performance of the posterior weighted mixture of the centroids of the clusters is shown in Fig. 3. As you can see, with two clusters, the weighted mixture of centroids tracks the predictive performance of the ensemble centroid. With four clusters, the weighted mixture of centroids outperforms both the MAP and the ensemble centroid solution, and has a predictive performance closer to the model averaged solution. Using a larger number of clusters would allow a weighted mixture of centroids to more closely approximate the full model averaged solution and therefore reduce the difference in predictive performance between these two solutions.

5. Discussion and conclusions

EBIR efficiently provides an exact representation of the posterior space of all possible sub-models and consequently, the marginal probability of including each of the predictor variables when the number of variables is not too large. Thus, this fully Bayesian model can be used for variable selection, model averaging applications, and examination of the shape of the posterior space. For a posterior distribution that is multi-modal, no single solution can accurately characterize this complex space. Clusters of solutions can help to explain why the MAP or ensemble centroid solutions do not perform as well, in terms of predictive performance, as the model averaged result. However, while best in terms of predictive performance, a model averaged result does not return a feasible solution to the model selection problem that end line users often want. A mixture of cluster centroids provides a compromise between these two extremes by returning a small set of feasible solutions weighted according to their posterior probabilities.

There are three circumstances in which EBIR may be used to advantage. (1) Any application in which BMA or IBMA has been or could be used to advantage; (2) Problems involving a large number of regression analyses. For example, in change point analysis (Auger and Lawrence, 1989; Liu and Lawrence, 1999; Ruggieri et al., 2009), variable selection is required on every possible substring of the data. In this case, the reduction in compute time can be realized for each of the $N(N+1)/2$ substring calculations that need to be performed on a data set of length N; (3) Problems involving a very large number of predictors, in which there is interest in examining the posterior distribution of lower order tuples of variables, such as in genome studies where epitasis plays a role in phenotype or disease risk. To address the practicality of this application, we simulated a moderate sized problem (Fig. 2) and project this to the very large size problem that could be run on current parallel computing systems. Our simulation studies showed a significant and increasing reduction in time as the number of predictors grew (Table 1, Fig. 2).

Because any internal node of the binary tree (Fig. 1) can have its independent 'include' and 'exclude' branches computed separately, this algorithm is readily amenable to parallelization. Thus, when parallel systems are available, the number of models that can be included in exact posterior probability calculations increases nearly linearly with the number of available processors. Setting a hard limit on the number of variables selected for inclusion in an exhaustive search can reduce the time complexity from exponential to polynomial, and thus the posterior distribution can be obtained for all permitted models. We encountered no numerical difficulties as our results consistently matched brute force methods. It is possible that numerical difficulties specific to this approach may arise in other applications, but we expect that these are unlikely. While we believe that our projection from a moderate to a very large sized problem is reasonable, practical problems could arise in the implementation of this approach on parallel computing systems.

This procedure, like all others, becomes overwhelmed by the exponentially increasing number of sub-models when all combinations of candidate predictors are considered. MCMC approaches (Smith and Kohn, 1996; Raftery et al., 1997; Fernandez et al., 2001a) can handle a larger number of predictors before running into this computational 'wall' because they do not evaluate all possible sub-models. While in such cases their solutions can be of practical value, convergence of the Markov Chain must be addressed. Although not discussed in this context, EBIR could also be utilized by an MCMC approach that traverses the probability space through models that are 'neighbors' as defined in Raftery et al. (1997). When the number of predictors becomes very large, a preliminary screening procedure, parallel processing, or a restriction on the number of included predictors is required to reduce the dimensionality of the problem.

EBIR was tested not only with simulations, but also on the publicly available Crime and Punishment data set (Vandaele, 1978). Our analysis of the Crime and Punishment data set showed that BIC based solutions differed substantially from exact Bayesian inferences (Table 4). One possible explanation is that the BIC approximation (Raftery, 1995) is based upon a convergence result—an approximation that improves as the number of observations increases relative to the number of predictors.

Alternatively, as with all Bayesian procedures, the posterior distribution is dependent on the prior models and their parameter settings. Thus, discrepancies in the BMA and EBIR probabilistic rankings may stem from differences in their prior models. The prior model assumptions of BIC are implicit, and as pointed out by Chen and Chen (2008) can be quite unreasonable. Specifically, the constant prior behind BIC amounts to assigning probabilities to classes of sub-models that are proportional to their size. Accordingly, the prior increases almost exponentially in class size. "This is obviously unreasonable, being strongly against the principle of parsimony (Chen and Chen, 2008)".

As indicated by George and McCulloch (1993), altering the 'spike' and 'slab' parameters can change the set of sub-models with high posterior probability, a result echoed for EBIR through the Crime and Punishment example. Specifically, since EBIR falls under the shrinkage family of variable selection models, the investigator can tune the degree of shrinkage. Thus, the use of this approach requires an investigator to select a range of parameter values near zero that are too small to be of

"interest", just as an investigator is required to set the level of Type I error that merits further study. For EBIR, the larger the difference between the values of k_e and k_i, the larger the 'penalty' will be for including an additional variable in the model. Further study is required to fully understand the impact of these settings. A second limitation to the EBIR procedure is that the probability density function is designed for homoscedastic Gaussian errors. Thus, the procedure should be able to generalize to mixtures of Gaussians, but may not be easy to generalize to non-Gaussian distributions.

One important argument for the 'shrinkage' approach is based on the reasoning of Efron (2004) who argues for the use of the terms 'interesting' and 'uninteresting' rather than describing variables in terms of their statistical significance. An investigator may be uninterested either because the discovery of such small differences would be of no practical value or because such small differences could too easily be the result of artifacts such as unobserved confounding or unexpected correlation (Efron, 2004). Variables that fall in the 'uninteresting' category have their coefficients close to zero, so that even if they are statistically significant, their effect is too small to be of interest.

EBIR can be used to replace BIC approximations with an exact probabilistic calculation in model selection procedures without a change in time complexity. Since EBIR employs the same tree configuration as the 'sweep' procedure of Leaps and Bounds and because it has the same time complexity, $O(m^2)$, exact posteriors can be employed in a Leaps and Bounds based screening procedure with little addition to the computational load. In this setting, our procedure provides a way to efficiently calculate the posterior probabilities of the models selected by the screening step up to the unknown normalizing constant and thus a means to rank the selected models according to their posterior probabilities. Additionally, since BMA (specifically, IBMA which repeatedly uses BMA) has been shown to work for very large problems, EBIR can capitalize on this existing iterative method to become applicable to problems of this size without a substantial difference in time. Of course, since the posterior space is truncated, EBIR will give posterior probabilities only for the remaining portion of the posterior space.

We capitalize on Miller's (1981) recursive procedure to circumvent the most computationally intensive calculations for a variable selection procedure, matrix inversions and matrix determinants, as addition or deletion of a variable is equivalent to adding a matrix of rank one. This reduction in time complexity to $O(m^2)$ extends the number of models amenable to exact posterior calculations.

Availability: The Matlab implementation of this algorithm is available under the GNU Public License by contacting Eric Ruggieri at ruggierie@duq.edu.

Acknowledgments

The authors would like to thank T. Herbert and K. Lawrence for bringing a variable selection question to our attention. The authors would also like to thank the anonymous reviewers and associate editor whose comments helped to greatly improve this manuscript.

Appendix. Supplementary data

Supplementary material related to this article can be found online at doi:10.1016/j.csda.2011.09.026.

References

Auger, I.E., Lawrence, C.E., 1989. Algorithms for the optimal identification of segment neighborhoods. Bull. Math. Biol. 51, 39–54.
Becker, G.S., 1968. Crime and punishment: an economic approach. J. Polit. Econ. 76, 169–217.
Breiman, L., 1995. Better subset regression using the nonnegative garrote. Technometrics 37, 373–384.
Bunch, J.R., Hopcroft, J.E., 1974. Triangular factorization and inversion by fast matrix multiplication. Math. Comp. 28, 231–236.
Carvalho, L., Lawrence, C., 2008. Centroid estimation in discrete high-dimensional spaces with applications in biology. PNAS 105 (9), 3209–3214. doi:10.1073/pnas.0712329105.
Chen, J., Chen, Z., 2008. Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95 (3), 759–771. doi:10.1093/biomet/asn034.
Coppersmith, D., Winograd, S., 1990. Matrix multiplication via arithmetic progressions. J. Symbolic. Comput. 9, 251–280.
Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., 2001. Introduction to Algorithms, 2nd ed. MIT Press, New York.
Diebolt, J., Robert, C.P., 1994. Estimation of finite mixture distributions through Bayesian sampling. J. R. Stat. Soc. Ser. B 56, 363–375.
Ding, Y., Chan, C.Y., Lawrence, C.E., 2006. Clustering of RNA secondary structures with application to messenger RNAs. J. Mol. Biol. 359, 554–571.
Efron, B., 2004. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J. Amer. Statist. Assoc. 99 (465), 96–104. doi:10.1198/016214504000000089.
Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32, 407–499.
Ehrlich, I., 1973. Participation in illegitimate activities: a theoretical and empirical investigation. J. Polit. Econ. 81, 521–565.
Ehrlich, I., 1975. The deterrent effect of capital punishment: a question of life and death. J. Polit. Econ. 81, 521–567.
Eicher, T.S., Papageorgiou, C., Roehn, O., 2007. Unraveling the fortunes of the fortunate: an iterative Bayesian model averaging (IBMA) approach. J. Macroecon. 29, 494–514.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360. doi:10.1198/016214501753382273.
Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser. B 70, 849–911.
Fernandez, C., Ley, E., Steel, M.F.J., 2001a. Benchmark priors for Bayesian model averaging. J. Econometrics 100, 381–427.
Fernandez, C., Ley, E., Steel, M.F.J., 2001b. Model uncertainty in cross-country growth regression. J. Appl. Econometrics 16, 563–576.
Fleisher, B.M., 1966. The Economics of Delinquency. Quadrangle, Chicago.
Furnival, G.M., Wilson, R.W., 1974. Regression by leaps and bounds. Technometrics 16, 499–511.
George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. J. Amer. Statist. Assoc. 88, 881–890.

Good, I.J., 1952. Rational decisions. J. R. Stat. Soc. Ser. B 14, 107–114.
Halmos, P.R., 1958. Finite Dimensional Vector Spaces. VanNostrand, Princeton.
Hans, C., 2009. Bayesian lasso regression. Biometrika 96, 835–845. doi:10.1093/biomet/asp047.
Hans, C., 2010. Model uncertainty and variable selection in Bayesian lasso regression. Stat. Comput. 20, 221–229. doi:10.1007/s11222-009-9160-9.
Hoerl, A.E., Kennard, R.W., 1970a. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67.
Hoerl, A.E., Kennard, R.W., 1970b. Ridge regression: applications to nonorthogonal problems. Technometrics 12, 69–82.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1997. Bayesian model averaging: a tutorial. Statist. Sci. 14, 382–417. Corrected version from: http://www.stat.washington.edu/www/research/online/hoeting1999.pdf.
Land, A.H., Doig, A.G., 1960. An automatic method of solving discrete programming problems. Econometrica 28 (3), 497–520.
Liu, J.S., Lawrence, C.E., 1999. Bayesian inference on biopolymer models. Bioinformatics 15 (1), 38–52.
Madigan, D., Raftery, A.E., 1994. Model selection and accounting for model uncertainty in graphical models using Occam's window. J. Amer. Statist. Assoc. 89, 1535–1546.
Madigan, D., York, J., 1995. Bayesian graphical models for discrete data. Internat. Statist. Rev. 63, 215–232.
Miller, K.S., 1981. On the inverse of the sum of matrices. Math. Mag. 54 (2), 67–72.
Miller, A.J., 2002. Subset Selection in Regression, 2nd ed. Chapman and Hall, New York.
Mitchell, T.J., Beauchamp, J.J., 1986. Algorithms for Bayesian variable selection in regression. In: Boardman, T.J. (Ed.), Computer Science and Statistics: Proceedings of the 18th Symposium on the Interface. American Statistical Association, Washington, DC, pp. 181–182.
Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression (with discussion). J. Amer. Statist. Assoc. 83, 1023–1036.
Park, T., Casella, G., 2008. The Bayesian lasso. J. Amer. Statist. Assoc. 103 (482), 681–686. doi:10.1198/016214508000000337.
Raftery, A.E., 1995. Bayesian model selection in social research. Sociol. Methodol. 25, 111–163.
Raftery, A.E., Madigan, D., Hoeting, J.A., 1997. Bayesian model averaging for linear regression models. J. Amer. Statist. Assoc. 92, 179–191.
Ruggieri, E., Herbert, T., Lawrence, K.T., Lawrence, C.E., 2009. Change point method for detecting regime shifts in paleoclimatic time series: application to d18O time series of the Plio–Pleistocene. Paleoceanography 24, PA1204. doi:10.1029/2007PA001568.
Smith, M., Kohn, J., 1996. Nonparametric regression using Bayesian variable selection. J. Econometrics 75, 317–343.
Stigler, G.J., 1970. The optimum enforcement of law. Political Economy 78, 526–536.
Taft, D.R., England Jr., R.W., 1964. Criminology, 4th ed. MacMillan, New York.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B 58, 267–288.
Vandaele, W., 1978. Participation in illegitimate activities: Ehrlich revisited. In: Blumstein, A., Cohen, J., Nagin, D. (Eds.), Deterrence and Incapacitation. National Academy of Sciences Press, Washington, DC, pp. 270–335.
Villard, G., 2003. Computation of the inverse and determinant of a matrix. In: Chyzak, F. (Ed.), Algorithms Seminar. INRIA Rocquencourt, France, pp. 29–32.
Wang, H., 2009. Forward regression for ultra-high dimensional variable screening. J. Amer. Statist. Assoc. 104 (488), 1512–1524. doi:10.1198/jasa.2008.tm08516.
Yeung, K.A., Bumgarner, R.E., Raftery, A.E., 2005. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics 21, 2394–2402.
Zhang, H.H., Lu, W., 2007. Adaptive lasso for Cox's proportional hazard model. Biometrika 94 (3), 691–703. doi:10.1093/biomet/asm037.
Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429. doi:10.1198/016214506000000735.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320.