

Posterior Model Probabilities via Path-based Pairwise Priors

James O. Berger¹

*Duke University and Statistical and Applied Mathematical Sciences Institute,
P.O. Box 14006, RTP, Durham, NC 27709, U.S.A.*

German Molina²

*Credit Suisse First Boston, Fixed Income Research,
One Cabot Square, London, E14 4QJ, United Kingdom.*

We focus on Bayesian model selection for the variable selection problem in large model spaces. The challenge is to adequately search the huge model space, while accurately approximating model posterior probabilities for the visited models. The issue of choice of prior distributions for the visited models is also important.

Key words and phrases. Pairwise Model Comparisons, Model Selection, Inclusion Probabilities, Search in Model Space.

1. Introduction

1.1. Model Selection in Normal Linear Regression

Consider the usual normal linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where \mathbf{Y} is the $n \times 1$ vector of observed values of the response variable, \mathbf{X} is the $n \times p$ ($p < n$) full rank design matrix of covariates, and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients. We assume that the coordinates of the random error vector $\boldsymbol{\varepsilon}$ are independent, each with a normal distribution with mean 0 and common variance σ^2 that is unknown. The least squares estimator for this model is thus $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.

Equation (1.1) will be called the *full* model \mathcal{M}_F , and we consider selection from among submodels of the form

$$\mathcal{M}_i : \mathbf{Y} = \mathbf{X}_i \boldsymbol{\beta}_i + \boldsymbol{\varepsilon}, \tag{1.2}$$

¹berger@stat.duke.edu

²german@alumni.duke.edu

where β_i is a $d(i)$ -dimensional subvector of β and \mathbf{X}_i is the corresponding $n \times d(i)$ matrix of covariates. Denote the density of \mathbf{Y} , under model \mathcal{M}_i , as $f_i(\mathbf{Y} | \beta_i, \sigma^2)$.

In Bayesian model selection, the model is treated as another unknown parameter (see for instance RAFTERY *et al*, 1997, CLYDE, 1999, and DELLAPORTAS *et al*, 2002), and one seeks to obtain the posterior probabilities (or other features of interest) of the models under consideration. It is convenient to approach determination of these probabilities by first computing Bayes factors. Defining the marginal density of the data under model \mathcal{M}_i as

$$m_i(\mathbf{Y}) = \int f_i(\mathbf{Y} | \beta_i, \sigma^2) \pi_i(\beta_i, \sigma^2) d\beta_i d\sigma^2,$$

where $\pi_i(\beta_i, \sigma^2)$ denotes the prior density of the parameter vector under \mathcal{M}_i , the Bayes factor of \mathcal{M}_i to \mathcal{M}_j is given by

$$BF_{ij} = \frac{m_i(\mathbf{Y})}{m_j(\mathbf{Y})}. \quad (1.3)$$

The Bayes factor can be interpreted as the posterior odds ratio of the two models under consideration, and, as such, can be used directly as an instrument to compare (any) two models, once we observe the data. However, it can be difficult to use, in that it involves integration of the joint distribution (likelihood times prior) over the parameter space.

A second difficulty is that of appropriate choice of the priors. This is not generally a major problem when we are simply interested in parameter inference within a given model, because objective priors are available, but, for comparisons between models, objective priors cannot typically be used.

A third difficulty that arises is related to the size of the model space. In the linear regression problem for instance, the 2^p -element model space can be too large to allow computation of the marginal density of the data under each of the possible models, regardless of how fast we are able to perform that computation. In this situation, we need to be able to search through model space (see GEORGE and MCCULLOCH, 1993, CHIPMAN *et al*, 1998 and CHIPMAN *et al*, 2001) to find the important models, without visiting every model.

From Bayes factors, it is possible to compute a wide variety of quantities relevant to model uncertainty. Foremost among these are model posterior probabilities, which, under a uniform prior on the model space ($P(\mathcal{M}_i) = P(\mathcal{M}_j)$, $\forall i, j$), are given by the normalized Bayes factors,

$$P(\mathcal{M}_i | \mathbf{Y}) = \frac{m_i(\mathbf{Y})}{\sum_j m_j(\mathbf{Y})} = \frac{BF_{iB}}{\sum_j BF_{jB}}, \quad (1.4)$$

where B can refer to any of the models; it is often convenient to single out a particular model \mathcal{M}_B , which we will call the base model, compute all pairwise Bayes factors with respect to this model, and then use (1.4) to determine posterior probabilities.

Further quantities of great interest are the *inclusion probabilities*, q_i , which are defined as the probabilities of a variable being in a model (in the model space under consideration) given the data. Thus, q_i is the sum of the model posterior probabilities of all models containing the variable β_i ,

$$q_i = P(\beta_i \in \text{model} | \mathbf{Y}) = \sum_{j: \beta_i \in \mathcal{M}_j} P(\mathcal{M}_j | \mathbf{Y}). \quad (1.5)$$

These are useful in defining the the *median probability model*, which is the model consisting of those variables whose posterior inclusion probability is at least 1/2. Surprisingly, as is shown in BARBIERI and BERGER (2004), the median probability model is often better for prediction than is the highest posterior probability model.

1.2. Conventional Prior Approaches

Two of the traditional approaches to prior elicitation in the model selection context in multiple linear regression for the regression parameters are the use of *g-priors* and the *Zellner-Siow*-type priors (ZELLNER and SIOW, 1980). They both utilize the full design matrix to construct the prior for the different models.

1.2.1. *g*-Priors

Perhaps the most commonly used conventional prior for testing is the *g*-prior, because of its computational advantages. However, *g*-priors have a number of disadvantages (BERGER and PERICCHI, 2001), and indeed can perform quite poorly. For instance, in regression, serious problems can occur if the the intercept term is not treated separately. Since all our examples will be regression problems, we thus follow the usual convention with *g*-priors of only considering models that have the intercept β_1 and assigning this parameter a constant prior density. Also assume that the covariates have been centered, so that the design matrix $\mathbf{X} = (\mathbf{1} \ \mathbf{X}^*)$, with the columns of \mathbf{X}^* being orthogonal to $\mathbf{1}$ (the vector of ones), and write $\boldsymbol{\beta}^* = (\beta_2, \dots, \beta_p)$. The same notational devices will be used for submodels: $\mathbf{X}_i = (\mathbf{1} \ \mathbf{X}_i^*)$ and $\boldsymbol{\beta}_i^*$ is those

coordinates of β_i other than β_1 . The g -prior for M_i is then defined as

$$\pi_i^g(\beta_1, \beta_i^*, \sigma^2) = \frac{1}{\sigma^2} \times \text{Normal}_{(d(i)-1)}(\beta_i^* \mid \mathbf{0}, cn\sigma^2(\mathbf{X}_i^* \mathbf{X}_i^*)^{-1}),$$

where c is fixed, and is typically set equal to 1 or estimated in an empirical Bayesian fashion (see e.g., CHIPMAN, GEORGE and MCCULLOCH, 2001). For given c , the Bayes factor for comparison of any two submodels has the closed-form expression

$$BF_{iB} = (1 + cn)^{\frac{d(B)-d(i)}{2}} \left(\frac{|\mathbf{Y} - \hat{\beta}_1 \mathbf{1}|^2 + cn|\mathbf{Y} - \mathbf{X}_B \hat{\beta}_B|^2}{|\mathbf{Y} - \hat{\beta}_1 \mathbf{1}|^2 + cn|\mathbf{Y} - \mathbf{X}_i \hat{\beta}_i|^2} \right)^{\frac{n-1}{2}},$$

in terms of the residual sums of squares for the models (with $\hat{\beta}_i$ being the least square estimate for M_i). Having such a simple closed form expression has made these priors widely used.

1.2.2. Zellner-Siow priors

The Zellner-Siow priors have several possible variants, depending on the base model, M_B , with respect to which the Bayes factors are computed. We consider here only the base model consisting of the linear model with just the intercept β_1 . The resulting prior, to be denoted ZSN, is given by

$$\begin{aligned} \pi_B^{ZSN}(\beta_1, \sigma^2) &= \frac{1}{\sigma^2}; \\ \pi_i^{ZSN}(\beta_1, \beta_i^*, \sigma^2) &= \frac{1}{\sigma^2} \times \text{Normal}_{(d(i)-1)}(\beta_i^* \mid \mathbf{0}, cn\sigma^2(\mathbf{X}_i^* \mathbf{X}_i^*)^{-1}), \\ \pi_i^{ZSN}(c) &\sim \text{InverseGamma}(c \mid 0.5, 0.5), \end{aligned}$$

i.e., is simply a scale mixture of the g -Priors. The resulting expression for BF_{iB} is

$$BF_{iB} = \frac{\int_0^\infty (1 + cn)^{(n-d(i))/2} (|\mathbf{Y} - \hat{\beta}_1 \mathbf{1}|^2 + cn|\mathbf{Y} - \mathbf{X}_i \hat{\beta}_i|^2)^{-(n-1)/2} c^{-3/2} e^{-1/(2c)} dc}{\int_0^\infty (1 + cn)^{(n-d(B))/2} (|\mathbf{Y} - \hat{\beta}_1 \mathbf{1}|^2 + cn|\mathbf{Y} - \mathbf{X}_B \hat{\beta}_B|^2)^{-(n-1)/2} c^{-3/2} e^{-1/(2c)} dc}.$$

While the ZSN prior does not yield closed-form Bayes factors, they can be computed by one-dimensional numerical integration. Also, a quite accurate (essentially) closed form approximation is given in PAULO (2003).

1.3. The choice of Priors

One major concern regarding the use of the g -Priors or the Zellner-Siow type priors is that, when the two models under comparison have a large difference in dimension,

the difference in dimension in the conventional priors is also large. We would thus be assigning a large-dimensional proper prior for the parameters that are not common. Since these are proper priors, one might worry that this will have an excessive influence on model selection results. The alternatives we introduce in the following section offer possible solutions to this (potential) problem: model comparisons will only be done pairwise, and with models differing by one dimension.

2. Search in Large Model Space

In situations where enumeration and computation with all models is not possible, we need a strategy to obtain reliable results while only doing computation over part of model space. For that, we need to find a search strategy that locates enough models with high posterior probability to allow for reasonable inference. We consider situations in which marginal densities can be computed exactly or accurately approximated for a visited model, in which case it is desirable to avoid visiting a model more than once. This suggests the possibility of a sampling without replacement search algorithm in model spaces. What is required is a search strategy that efficiently leads to the high probability models.

Alternative approaches, such as reversible jump-type algorithms (GREEN, 1995, DELLAPORTAS *et al*, 2002), are theoretically appealing, but not as useful in large model spaces since it is very unlikely that we will visit any model frequently enough to obtain a good estimate of its probability by Monte Carlo frequency. (See also CLYDE *et al*, 1996, RAFTERY *et al*, 1997.)

2.1. A search algorithm

The search algorithm we propose makes use of previous information to provide orientation for the search. The algorithm utilizes not only information about the models previously visited, but also information about the variables visited (estimates of the inclusion probabilities) to specify in which direction to make (local) moves.

Recall that the marginal posterior probability that a variable β_i , arises in a model (the posterior inclusion probability) is given by equation (1.5). Define the estimate of the inclusion probability, \hat{q}_i , of the variable β_i , to be the sum of the estimated posterior probabilities of the visited models that contain β_i , i.e.,

$$\hat{q}_i = \sum_{j:\beta_i \in \mathcal{M}_j} \hat{P}(\mathcal{M}_j | \mathbf{Y});$$

the model posterior probabilities, $\hat{P}(\mathcal{M}_j \mid \mathbf{Y})$, could be computed by any method for the visited models, but we propose a specific method – that meshes well with the search – in the next section.

One difficulty is that all or none of the models visited could include a certain variable and, then, the estimate of that inclusion probability would be one or zero. To avoid this degeneracy, a tuning constant \mathbb{C} will be introduced in the algorithm to keep the \hat{q}_i bounded away from zero or one. The value chosen for implementation was $\mathbb{C} = 0.01$. Larger values of \mathbb{C} allow for bigger moves in Model Space, as well as reducing the susceptibility of the algorithm to getting stuck in local modes. Smaller values allow for more concentration on local modes.

The stochastic search algorithm consists of the following steps:

1. Start at any model, that we will denote as \mathcal{M}_B . This could be the full model, the null model, any model chosen at random, or an estimate of the median probability model based on adhoc estimates of the inclusion probabilities.
2. At iteration k , compute the current model probability estimates, $\hat{P}(\mathcal{M}_j \mid \mathbf{Y})$, and current estimates of the variable inclusion probabilities, \hat{q}_i .
3. Return to one of the $k - 1$ distinct models already visited, in proportion to their estimated probabilities $\hat{P}(\mathcal{M}_j \mid \mathbf{Y})$, $j = 1, \dots, k - 1$.
4. If the chosen model is the full model, choose to remove a variable.
If the chosen model is the null model, choose to add a variable.
Otherwise choose to add/remove a variable with probability 1/2.
If adding a variable, choose variable i with probability $\propto \frac{\hat{q}_i + \mathbb{C}}{1 - \hat{q}_i + \mathbb{C}}$.
If removing a variable, choose variable i with probability $\propto \frac{1 - \hat{q}_i + \mathbb{C}}{\hat{q}_i + \mathbb{C}}$.
5. If the model obtained in step 4 has already been visited, return to Step 3.
If it is a new model, perform the computations necessary to update the $\hat{P}(\mathcal{M}_j \mid \mathbf{Y})$ and \hat{q}_i , update k to $k + 1$, and go to Step 2.

A number of efficiencies could be effected (e.g., making sure never to return in Step 3 to a model whose immediate neighbors have all been visited), but a few explorations of such possible efficiencies suggested that they offer little actual improvement in large model spaces.

The starting model for the algorithm does have an effect on its initial speed, but seems to be essentially irrelevant after some time. If initial estimates of the posterior probabilities are available, using these to estimate the inclusion probabilities and starting at the median probability model seems to beat all other options.

3. Path-based priors for the model parameters

3.1. Introduction

In this section we propose a method for computation of Bayes factors that is based on only computing Bayes factors for adjacent models, that is, models that differ by only one parameter. There are several motivations for considering this:

1. Proper priors are used in model selection, and these can have a huge effect on the answers for models of significantly varying dimension. By computing Bayes factors only for adjacent models, this difficulty is potentially mitigated.
2. Computation is an overriding issue in exploration of large model spaces, and computations of Bayes factors for adjacent models can be done extremely quickly.
3. The search strategy in the previous section, operates on adjacent pairs of models, and so it is natural to focus on Bayes factor computations for adjacent pairs.

Bayes factors will again be defined relative to a base model M_B , that is, a reference model under which we can compute the Bayes factor BF_{kB} for any other model, and then equation (1.4) will be used to formally define the model posterior probabilities.

BF_{kB} will be determined by connecting M_k to M_B through a path of other models, $\{M_k, M_{k-1}, M_{k-2}, \dots, M_1, M_B\}$, with any consecutive pair of models in the path differing by only one parameter. The Bayes factor $BF_{k,k-1}$ between two adjacent models in the path is computed in ways to be described shortly, and then BF_{kB} is computed through the ‘chain rule’ of Bayes factors:

$$BF_{kB} = BF_{k,k-1} \times BF_{k-1,k-2} \times BF_{k-2,k-3} \times \dots \times BF_{1,B}. \quad (3.1)$$

To construct the paths joining M_B with the other visited models, we will use the search strategy from the previous section, starting at M_B . A moments reflection will make clear that the search strategy yields a connected graph of models, so that any model that has been visited is connected by a path to M_B , and furthermore that this path is unique. This last is an important property because the methods we use to compute the $BF_{k,k-1}$ are not themselves ‘chain rule coherent’ and so two different paths between M_k and M_B could result in different answers. Unique paths avoid this complication but note that we are, in a sense, then using (3.1) as a definition of the Bayes factor, rather than as an inherent equality. Note that the product chain computation is not necessarily burdensome since, in the process of determining the estimated inclusion probabilities to drive the search, one is keeping a running product of all the chains.

The posterior probabilities that result from this approach are random, in two ways. First, the search strategy often only visits a modest number of models, compared to the total number of models (no search strategy can visit most of 2^p models for large p), and posterior probabilities are only computed for the visited models (effectively assuming that the non-visited models have probability zero). Diagnostics then become important in assessing whether the search has succeeded in visiting most of the important models.

The second source of randomness in the posterior probabilities obtained through this search is that the path that is formed between two models is random and, again, different paths can give different Bayes factors. This seems less severe than the above problem and the problem of having to choose a base starting model.

3.2. Pairwise expected arithmetic intrinsic Bayes factors

The path-based prior we consider is a variant of the Intrinsic Bayes factor introduced in BERGER and PERICCHI (1996). It leads, after a simplification, to the desired pairwise Bayes factor between model \mathcal{M}_j and \mathcal{M}_i (\mathcal{M}_j having one additional variable),

$$BF_{ji} = \left(\frac{|\mathbf{Y} - \mathbf{X}_i \hat{\boldsymbol{\beta}}_i|}{|\mathbf{Y} - \mathbf{X}_j \hat{\boldsymbol{\beta}}_j|} \right)^{n-d(i)} \left(\frac{d(i) + 2}{n} \right)^{1/2} \frac{(1 - e^{-\lambda/2})}{\lambda}, \quad (3.2)$$

$$\lambda = (d(i) + 2) \left(\frac{|\mathbf{Y} - \mathbf{X}_i \hat{\boldsymbol{\beta}}_i|^2}{|\mathbf{Y} - \mathbf{X}_j \hat{\boldsymbol{\beta}}_j|^2} - 1 \right),$$

where $\hat{\boldsymbol{\beta}}_k = (\mathbf{X}'_k \mathbf{X}_k)^{-1} \mathbf{X}'_k \mathbf{Y}$. Quite efficient schemes exist for computing the ratio of the residual sums of squares of adjacent nested models, making numerical implementation of (3.2) extremely fast.

3.2.1. Derivation of (3.2)

To derive this Bayes factor, we start with the Modified Jeffreys prior version of the Expected Arithmetic Intrinsic Bayes factor (EAIBF) for the linear model. To define the EAIBF, we need to introduce the notion of an m -dimensional *training design matrix* of the n -dimensional data. Say that, for model \mathcal{M}_k , we have the $n \times m$ design matrix \mathbf{X}_k . Then there are $L = \binom{n}{m}$ possible training design matrices, to be denoted $X_k(l)$, each consisting of m rows of X_k . For any model i nested in j with a difference of dimension 1, and denoting l as the index for the design rows,

the Expected Arithmetic Intrinsic Bayes factor was then given by

$$BF_{ji} = \left(\frac{|\mathbf{X}_i' \mathbf{X}_i|}{|\mathbf{X}_j' \mathbf{X}_j|} \right)^{1/2} \left(\frac{|\mathbf{Y} - \mathbf{X}_i \hat{\boldsymbol{\beta}}_i|}{|\mathbf{Y} - \mathbf{X}_j \hat{\boldsymbol{\beta}}_j|} \right)^{n-d(i)} \frac{1}{L} \sum_{l=1}^L \frac{|X_j'(l) X_j(l)|^{1/2} (1 - e^{-\lambda(l)/2})}{|X_i'(l) X_i(l)|^{1/2} \lambda(l)}, \quad (3.3)$$

where $\lambda(l) = n |\mathbf{Y} - \mathbf{X}_j \hat{\boldsymbol{\beta}}_j|^{-2} \hat{\boldsymbol{\beta}}_j' X_j'(l) [\mathbf{I} - X_i(l) (X_i'(l) X_i(l))^{-1} X_i'(l)] X_j(l) \hat{\boldsymbol{\beta}}_j$.

A technique that requires averaging over all training design matrices is not appealing, especially when a large number of such computations is needed (large sample size or model space). Therefore, one needs a shortcut that avoids the computational burden of summing/enumerating over training design matrices. One possibility is the use of a *representative training sample design matrix*, that will hopefully give results close to those that would be obtained by averaging over the $X_j(l)$.

Luckily, a standard result for matrices is that

$$\sum_{l=1}^L X_k(l)' X_k(l) = \begin{pmatrix} n-1 \\ m-1 \end{pmatrix} \mathbf{X}_k' \mathbf{X}_k,$$

from which it follows directly that a representative training sample design matrix could be defined as a matrix $\tilde{\mathbf{X}}_k$ satisfying

$$\tilde{\mathbf{X}}_k' \tilde{\mathbf{X}}_k = \frac{1}{L} \sum_{l=1}^L X_k(l)' X_k(l) = \frac{m}{n} \mathbf{X}_k' \mathbf{X}_k.$$

Again focusing on the case where \mathcal{M}_i is nested in \mathcal{M}_j and contains one fewer parameter, it is natural to assume that the representative training sample design matrices for models i and j are related by $\tilde{\mathbf{X}}_j = (\tilde{\mathbf{X}}_i \tilde{\mathbf{v}})$, where the columns have been rearranged if necessary so that the additional parameter in model j is the last coordinate, with $\tilde{\mathbf{v}}$ denoting the corresponding covariates. Noting that the training sample size is $m = d(i) + 2$, algebra shows that any such representative training sample design matrices will yield the Bayes factor in (3.2). Related arguments are given in DE VOS (1993) and CASELLA and MORENO (2002).

4. An application: Ozone data

4.1. Ozone 10 data.

To illustrate the comparisons of the different priors on a real example, we chose a subset of covariates used in CASELLA and MORENO (2002), consisting of 178 observations of 10 variables of an Ozone data set previously used by BREIMAN and

Table 1: For the Ozone 10 data, the variable inclusion probabilities (i) for the conventional priors, and (ii) for the path-based prior with different [starting models], run until 512 models had been visited; numbers in parentheses are the run-to-run standard deviation of the inclusion probabilities.

	g-Prior	ZSN	EAIBF[Null]	EAIBF[Full]	EAIBF[Random]
x1	0.393	0.458	0.449 (.0016)	0.449 (.0014)	0.449 (.0012)
x2	0.086	0.111	0.107 (.0003)	0.107 (.0003)	0.107 (.0003)
x3	0.106	0.135	0.131 (.0004)	0.131 (.0004)	0.131 (.0004)
x4	0.093	0.117	0.114 (.0003)	0.114 (.0004)	0.114 (.0004)
x5	0.075	0.095	0.092 (.0003)	0.092 (.0003)	0.092 (.0003)
x6	0.995	0.995	0.994 (.0000)	0.994 (.0000)	0.994 (.0000)
x7	1.000	1.000	1.000 (.0000)	1.000 (.0000)	1.000 (.0000)
x8	0.915	0.905	0.905 (.0010)	0.905 (.0011)	0.905 (.0010)
x9	0.194	0.236	0.231 (.0012)	0.231 (.0012)	0.231 (.0010)
x10	0.353	0.408	0.401 (.0016)	0.400 (.0011)	0.400 (.0010)

FRIEDMAN (1985). Details on the data can be found in CASELLA and MORENO (2002). We will denote this data set as *Ozone 10*.

4.1.1. Results under conventional priors

For the Ozone 10 data we can enumerate the model space, which is comprised of 1024 possible models, and compute the posterior probabilities of each of the models under conventional priors. The first two columns of Table 1 report the inclusion probabilities for each of the 10 variables under each of the conventional priors.

4.1.2. Results under the Path-based Pairwise prior

Path-based Pairwise priors provide results that will depend both on the starting point of the path (Base Model) and on the path itself. We use the *Ozone 10* data to conduct several Monte Carlo experiments to study this dependence, and to compare the pairwise procedure with the conventional prior approaches.

We ran 1000 different (independent) paths for the Path-based Pairwise prior, using the search algorithm detailed in Section 3, under three differing conditions: (i) Initializing the algorithm by including all models with only 1 variable and using the Null Model as Base Model; (ii) Initializing the algorithm with the Full Model as the Base Model; (iii) Initializing the algorithm at a model chosen at random with equal probability in the model space, and choosing it as the Base Model. Each path was stopped after exploration of 512 different models.

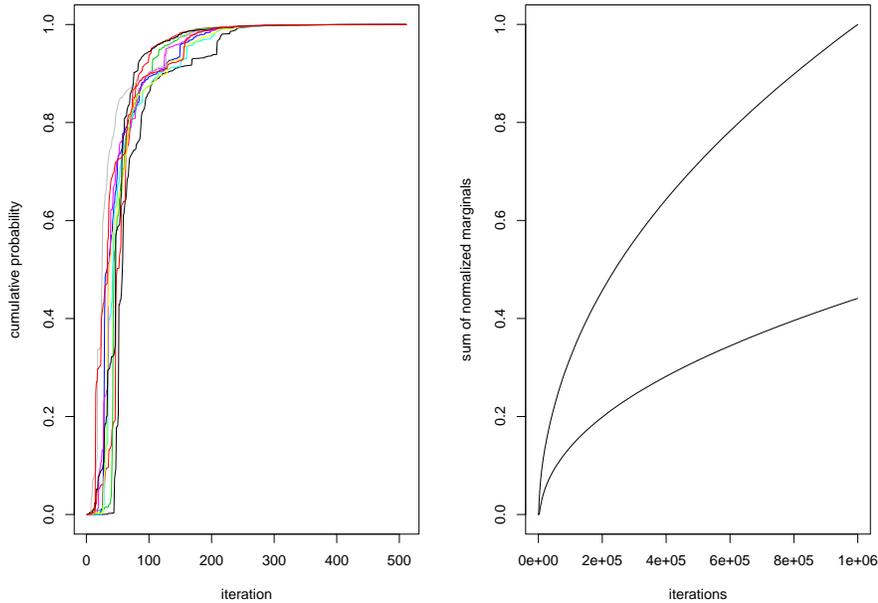


Figure 1: Cumulative posterior probability of models visited for (left) ten independent paths for the path-based EAIBF with random starting models for the Ozone 10 data; (right) the two paths for the path-based EAIBF starting at the Full Model for the Ozone 65 data.

The last three columns of Table 1 contain the mean and standard deviation (over the 1000 paths run) of the resulting inclusion probabilities. The small standard deviation indicates that these probabilities are very stable over the differing paths. Also, recalling that the *median probability model* is that model which includes the variables having inclusion probability greater than 0.5, it is immediate that any of the approaches results here in the median probability model being that which has variables x_6, x_7 , and x_8 (along with the intercept, of course). Interestingly, this also turned out to be the highest probability model in all cases.

The left part of Figure 1 shows the cumulative posterior probability of the models visited for the path-based EAIBF, for each of ten independent paths with random starting points in the model space. We can see that, although the starting point and the path taken have an influence, most of the high probability models are explored in the first few iterations. Another indication that the starting point and random path taken are not overly influential is Figure 2, which gives a histogram (over the 1000 paths) of the posterior probability (under a uniform prior probability on the model space) of the Highest Posterior Probability (HPM) model. The variation in this posterior probability is slight.

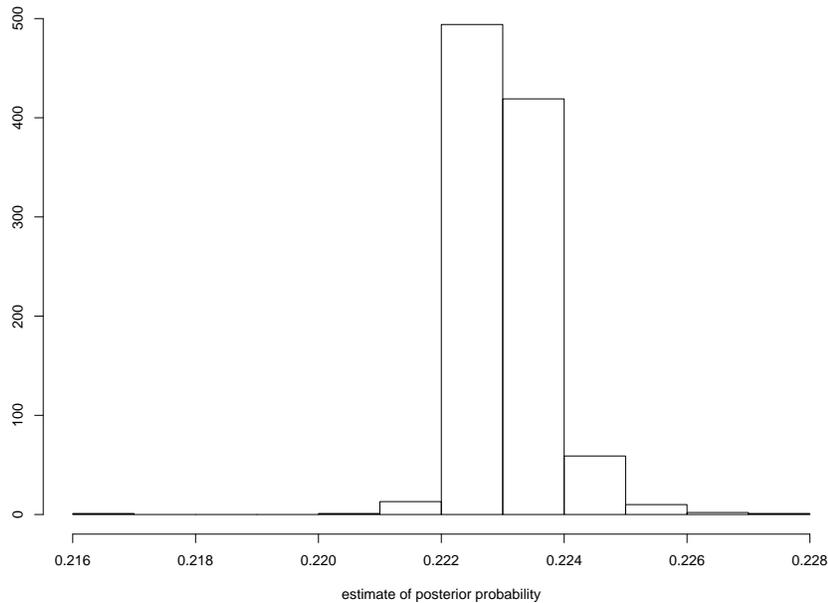


Figure 2: Posterior probabilities of the HPM for the Ozone 10 data over 1000 independent runs with randomly chosen starting models for the path-based EAIBF.

4.2. Ozone 65 data

We compare the different methods under consideration for the Ozone dataset, but this time we include possible second-order interactions. We will denote this situation as *Ozone 65*. There are then a total of 65 possible covariates that could be in or out of any model, so that there are 2^{65} possible models. Enumeration is clearly impossible and the search algorithm will be used to move in the model space. We ran two separate paths of one million distinct models for each of the priors under consideration, to get some idea as to the influence of the path.

All approaches yielded roughly similar expected model dimension (computed as the weighted sum of dimensions of models visited, with the weights being the model posterior probabilities). The g-Prior and the Zellner-Siow prior resulted in expected model dimension of 11.8 and 13.2, respectively. The expected dimension for the path-based EAIBF models was 13.6 and 13.7, when starting at the null and full models, respectively. The right part of Figure 1 shows the cumulative sum of (estimated) posterior probabilities of the models visited, for the two path searches. Note that these do not seem to have flattened out, as did the path searches for the Ozone 10 situation; indeed, one of the paths seems to have found only about half the probability mass as the other path. This suggests that even visiting a million models is not enough, if the purpose is to find most of the probability mass.

Table 2: Estimates of some inclusion probabilities in a run (1,000,000 models explored) of the path-based search algorithm for the Ozone 65 situation for the different priors [base models] (with two independent runs for each EAIBF pair).

	g-Prior	ZSN	EAIBF[Null]	EAIBF[Null]	EAIBF[Full]	EAIBF[Full]
x1	0.9048	0.9830	0.9902	0.9843	0.9889	0.9898
x2	0.0672	0.0864	0.0806	0.0902	0.0905	0.0705
x3	0.0385	0.0454	0.0441	0.0472	0.0465	0.0392
x4	0.9921	0.9986	0.9993	0.9982	0.9993	0.9993
x5	0.2312	0.3464	0.3911	0.3872	0.3871	0.3995
x6	0.2258	0.4041	0.4933	0.4787	0.4796	0.5345
x7	0.2176	0.2373	0.2293	0.2390	0.2232	0.2230
x8	0.9732	0.9968	0.9971	0.9979	0.9977	0.9968
x9	0.0391	0.0577	0.0833	0.0694	0.0692	0.1077
x10	0.9999	0.9999	0.9999	0.9999	0.9999	0.9997
x1.x1	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
x9.x9	0.9998	0.9999	0.9998	0.9999	0.9999	0.9999
x1.x2	0.5909	0.7773	0.8394	0.8515	0.8557	0.8350
x1.x7	0.1013	0.1938	0.2101	0.2285	0.2390	0.1899
x3.x7	0.7532	0.3309	0.0118	0.0343	0.0333	0.0280
x4.x7	0.3454	0.5108	0.5485	0.5690	0.6034	0.5142
x6.x8	0.8025	0.9623	0.9320	0.9831	0.9821	0.8664
x7.x10	0.9768	0.9977	0.9687	0.9986	0.9975	0.9048

Table 2 gives the resulting estimates of the marginal inclusion probabilities, for all main effects and for the quadratic effects and interactions that had any inclusion probabilities above 0.1. There does not appear to be any systematic difference between starting the EAIBF at the full model or the null model. There is indeed variation between runs, indicating that more than 1 million iterations is necessary to completely pin down the inclusion probabilities. The trend for the different priors is that the EAIBF seems to have somewhat higher inclusion probabilities than ZSN, which in turn has somewhat higher inclusion probabilities than g -priors. (Higher inclusion probabilities means support for bigger models.) The exception is the x_3x_7 interaction; we do not yet understand why there is such a difference between the analyses involving this variable.

Table 3 gives the Highest Probability Model from among those visited in the 1,000,000 iterations, and the Median Probability Model. Here slight differences occur but, except for the x_6x_7 variable, all approaches are reasonably consistent.

Table 3: Variables included in the Highest Probability Model (H) and Median Probability Model (M), after 1,000,000 models have been visited in the path-based search, for the Ozone 65 situation for the different priors [Base Model] (with two independent runs included for each EAIBF pair).

	g-Prior	ZSN	EAIBF[Null]	EAIBF[Null]	EAIBF[Full]	EAIBF[Full]
x1	H M	H M	H M	H M	H M	H M
x4	H M	H M	H M	H M	H M	H M
x6	- -	- -	- -	- -	- -	- M
x8	H M	H M	H M	H M	H M	H M
x10	H M	H M	H M	H M	H M	H M
x1.x1	H M	H M	H M	H M	H M	H M
x9.x9	H M	H M	H M	H M	H M	H M
x1.x2	H M	H M	H M	H M	H M	H M
x3.x7	- M	- -	- -	- -	- -	- -
x4.x7	H -	H M	H M	H M	H M	H M
x6.x8	H M	H M	H M	H M	H M	H M
x7.x10	H M	H M	H M	H M	H M	H M

5. Conclusion

The main purpose of the paper was to introduce the path-based search algorithm of Section 3 for the linear model, as well as the corresponding Bayes factors induced from pairwise Bayes factors and the ‘chain rule.’ Some concluding remarks:

1. The path-based search algorithm appears to be an effective way of searching model space, although even visiting a million models may not be enough for very large model spaces, as in the Ozone 65 situation.
2. With large model spaces, the ‘usual’ summary of listing models and their posterior probabilities is not possible; instead, information should be conveyed through tables of inclusion probabilities (and possibly multivariable generalizations).
3. Somewhat surprisingly, there did not appear to be a major difference between use of the path-based priors and the conventional g -priors and Zellner-Siow priors (except for one of the interaction terms in the Ozone 65 situation). Because the path-based priors utilized only one-dimensional proper Bayes factors, a larger difference was expected. Study of more examples is needed here.
4. While the path-based prior is not formally coherent, in that it depends on the path, the effect of this ‘incoherence’ seems minor, and probably results in less variation than does the basic fact that any chosen path can only cover a small amount of a large model space.

References

- BARBIERI, M. and J. BERGER (2004), Optimal predictive model selection. *Annals of Statistics*, **32**, 870–897.
- BERGER, J.O. and L.R. PERICCHI (1996), The Intrinsic Bayes Factor for Model Selection and Prediction, *Journal of the American Statistical Association*, **91**, 109–122.
- BERGER, J.O. and L.R. PERICCHI (2001), Objective Bayesian methods for model selection: introduction and comparison (with Discussion). In *Model Selection*, P. Lahiri, ed., Institute of Mathematical Statistics Lecture Notes, volume 38, Beachwood Ohio, 135–207.
- BREIMAN, L. and J. FRIEDMAN (1985), Estimating Optimal Transformations for Multiple Regression and Correlation, *Journal of the American Statistical Association*, **80**, 580–619.
- CASELLA, R. and E. MORENO (2002), Objective Bayesian Variable Selection. Technical Report 2002-026, Statistics Department, University of Florida.
- CHIPMAN, H., E. GEORGE, and R. MCCULLOCH (1998), Bayesian CART Model Search. *Journal of the American Statistical Association*, **93**, 935–960.
- CHIPMAN, H., E. GEORGE, and R. MCCULLOCH (2001), The Practical Implementation of Bayesian Model Selection with discussion. In *Model Selection*, P. Lahiri, ed., Institute of Mathematical Statistics Lecture Notes, volume 38, Beachwood Ohio, 67–134.
- CLYDE, M. (1999). Bayesian Model Averaging and Model Search Strategies. *Bayesian Statistics 6*, J.M. Bernardo, A.P. Dawid, J.O. Berger, and A.F.M. Smith eds. Oxford University Press, pp. 157–185.
- DELLAPORTAS, P., J. FORSTER, and I. NTZOUFRAS (2002), On Bayesian Model Selection and Variable Selection using MCMC. *Statistics and Computing*, **12**, 27–36.
- DE VOS, A. F. (1993), A Fair Comparison Between Regression Models of Different Dimension. *Technical report*, The Free University, Amsterdam.
- GEORGE, E., and R. MCCULLOCH (1993), Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association*, **88**, 881–889.
- GREEN, P. (1995), Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, **82**, 711–732.
- PAULO, R. (2003), Notes on Model Selection for the Normal Multiple Regression Model. *Technical Report*, Statistical and Applied Mathematical Sciences Institute.
- RAFTERY, A., E. MADIGAN, and J. HOETING (1997), Bayesian Model Averaging for Linear Regression Models. *Journal of the American Statistical Association*, **92**, 179–191.
- ZELLNER, A. and A. SIOW (1980), Posterior Odds Ratios for Selected Regression Hypothesis. In *Bayesian Statistics*, J. M. Bernardo et. al. (eds.), Valencia Univ. Press, 585–603.