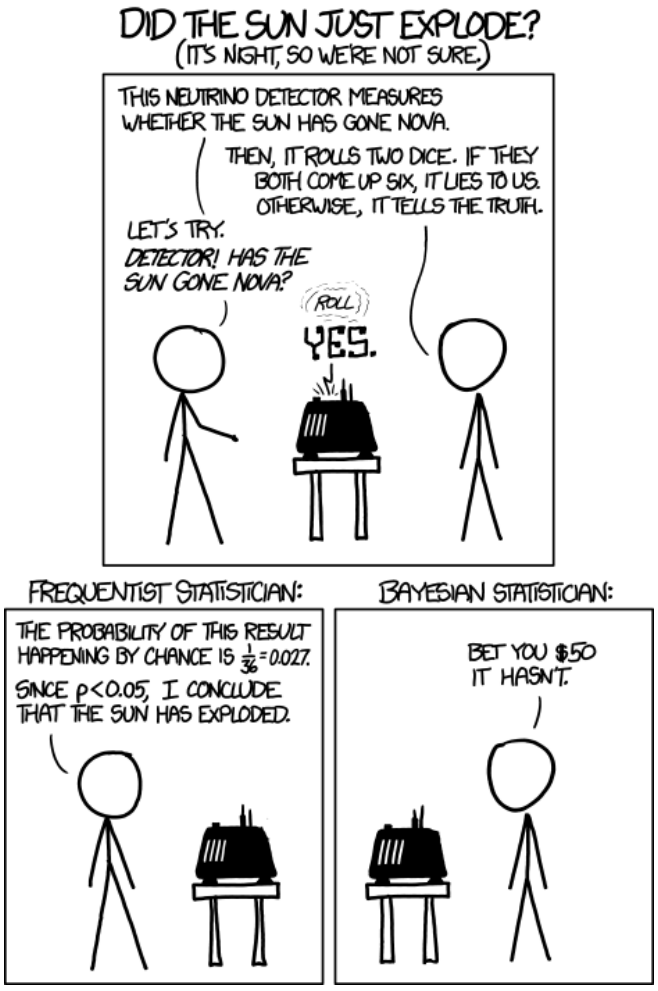
École nationale
de la statistique
et de l'administration
économique

université
PARIS-SACLAY

BAYESIAN STATISTICS

# Objective Bayesian variable selection




Lina CRISTANCHO
Charlotte PASQUIER
Jean-Michel ROUFOSSE

January 2018

# I Introduction

For this project we were interested in applying Bayesian variable selection techniques to normal regression models, as explained in Casella and Moreno's paper[1] published in 2006.

When performing regression modeling, a question that sometimes arises is whether to keep all the database's initial variables or not. It is indeed true that there might be a subset of the variables that could perform better, rather than using all of them.

The problem however is the following. Consider a dependant random variable $Y$ and a set $\{X_1, X_2, \ldots, X_k\}$ of $k$ variables of a given dataset. The number of combinations of all the $k$ variables that could possibly explain $Y$ is $2^k$. When $k$ is small, for example 10, it is not time consuming for a computer to test all the possible combinations. However, if for example an econometrician wishes to add the squares and the interaction terms, then the number of variables increases drastically and it becomes quickly impossible to test all the models.

We have written in R the methods explained in the mentioned paper [1]. The selection mechanism is completely automatic and criterion-based. This means the models are selected according to a chosen criterion, based on their Bayes Factor, and automatic because there are no hyper-parameters or parameters to tune.

At first, in the paper it is considered using default priors for the regression coefficients and the error variance. However, as argued by the authors these priors are improper and thus are not suited to be used for model choice or prediction. As such, they inspired their method by Berger and Pericchi's use of an empirical measure for model comparison : the intrinsic priors[2], which approximate the Bayes Factor, and are roughly calculated thanks to a training sample of the data. Plus, with intrinsic priors, there are no parameters to adjust, as they are derived from the model's structure. Then posteriors are easily computed and comparison between models can be made either by these probabilities, by the Bayes factors or by some other statistical criterion.

However, when dealing with models of high dimension, that is with databases having too many variables, an optimized search between the models is needed. This search should be able to find models with "high" posterior probability. Some work had been done by performing this search thanks to Gibbs Sampling [3], novel idea at the time since it gave a ranking of good models instead of the best one found, with in addition giving some assurance that the algorithm would not be stuck in local modes. Yet enough, Casella and Moreno's paper suggests a stochastic search based on a Markov Chain with a stationary distribution proportional to the posterior probabilities. Hence the use of a Metropolis-Hastings algorithm.

In our project report, we first illustrate the use of the Bayes factor criterion and how we used it in our R code. Then, we explain in depth how the article performs the stochastic search by explaining how it is based on the Metropolis-Hastings algorithm. Finally, every example from the article has been reimplemented, results have been compared with the article's ones, and found to be very satisfying in terms of accuracy as well as for the computing time.

[1] George Casella and Elias Moreno. Objective Bayesian variable selection. J. Amer. Statist. Assoc., 101(473):157-167, 2006.

# II Using Bayes Factor for variable selection

The approach proposed in the article focuses on two aspects:

- First, as it is natural for a Bayesian approach, the selection is based on the posterior probabilities of the models. The authors propose this computation to be done via the so-called "intrinsic priors".

  However, we have to out-stand the fact that on this aspect we followed the approach of Liang, Paulo, Molina, Clyde, and Berger (2008) [4] based on mixtures of g- priors, that is used in the R package "BayesFactor". Indeed, as a more recent approach, the former takes into consideration the remarks made in Castella and Moreno (2006) about having a desirable default prior and keeps the computational facilities of the original g-prior formulation. We will see in the applications that the use of these priors gives approximately the same results for the posterior probabilities as the intrinsic priors used in the article. We remain therefore with a fully automatic Bayesian approach for the computation of these posterior probabilities.

- In order to tackle the problem of calculating the posterior probabilities for all the models when there are to many of them (when $2^k$ is too big), the authors propose a stochastic search approach based on a Metropolis-Hastings algorithm whose stationary distribution is proportional to the posterior probabilities of the model.

## II.1 Posterior probabilities computation

The authors consider the normal regression model

$$y = X\alpha + \epsilon$$

where $y$ is the vector of $n$ observations, $X$ is the $n \times k$ design matrix, $\alpha$ is the vector of the $k$ regression coefficients, and $\epsilon \sim N_n(0, \sigma^2 I_n)$. This full model is denoted $N_n(y|X\alpha, \sigma^2 I_n)$.

The idea here is to rank the models using their posterior probability as the ranking criterion. The general procedure was the following:

1. For each model, we calculate the Bayes factor with respect to the full model:

$$B_{\gamma 1}(y, X) = \frac{m_\gamma(y, X)}{m_1(y, X)} = \frac{\int \mathcal{N}_n(y|X\beta_\gamma, \sigma_\gamma^2 I_n)\pi(\beta_\gamma, \sigma_\gamma)d\beta_\gamma d\sigma_\gamma}{\int \mathcal{N}_n(y|X\alpha, \sigma_\gamma^2 I_n)\pi(\alpha, \sigma)d_\alpha d\sigma}$$

   where $\gamma$ can be thought of representing a specific model, $\gamma = 1$ corresponding to the full model. We recall that this Bayes factor is computed via the R package "BayesFactor" which follows the mixture of g-priors approach.

2. Compute the posterior probability of each model as:

$$Pr(M_\gamma|y, X) = \frac{B_{\gamma 1}(y, X)}{1 + \sum_{\gamma \in \Gamma, \gamma \neq 1} B_{\gamma 1}(y, X)} \quad \gamma \in \Gamma$$

   where $\Gamma$ denotes the set of $2^{k-1}$ models to be compared.

3. Arrange the models in decreasing order according the their posterior probabilities.

Therefore, the higher the Bayes factor, the higher the probability that the model $M_\gamma$ will be a good model.

We may note that for each model $\gamma$, the hypothesis to be tested is

$$H_0 : \textit{True model is Model } \gamma \qquad vs \qquad H_1 : \textit{True model is the Full Model}$$

Then, for each model the test is centered at its null and the full model is the overall reference to which they are all directly compared to.

## II.2 Stochastic search

As mentioned before, for high dimension problems, it becomes impossible to evaluate the posterior probability of each one of the models. In this context, a stochastic search approached becomes necessary. The idea is to use a MCMC algorithm such that the stationary distribution is proportional to the posterior probabilities.

The stochastic-search algorithm does the following:

- Initialization:
  1. Split the set of models as $B = \bigcup_{i=1} B_i$, where $B_i$ is the set of all the models $M_\gamma$ having strictly $i$ variables.
  2. Select a small percentage of models from $B$, that gives us roughly the same amount of models in each $B_i$.
  3. For the selected models, calculate the posterior probabilities $p_{ij}$ of model $j$ in $B_i$. Then $\hat{P}(B_i) = \frac{\sum_{j \in B_i} p_{ij}}{\sum_{ij} p_{ij}}$

  $\hat{P}(B_i)$ is what helps the algorithm chose how many variables to select. The higher it is, the higher the chance of getting models drawn from that subset $B_i$. This is good since it means the Markov Chain will not visit models having a number of variables that does not give good results (does not have a high posterior probability).

- At iteration $t$:
  1. Compute the distribution $G_t = (\hat{P}_0, ..., \hat{P}_k)$ of subsets of the partition, where:

$$\hat{P}_i \propto \frac{1}{k+1}\frac{1}{log(t+1)} + \frac{\sum_{j \in \mathcal{B}_i} p_{ij}}{\sum_{ij} p_{ij}}$$

  2. According to the distribution $G_t$, select the subset $B_i$
  3. Metropolis-Hasting:
     - At random, select a proposed model $M_{\gamma_{t'}}$ among the models in the selected subset $B_i$
     - Accept the proposed model with probability:

$$min\left(1, \frac{Pr(M_{\gamma_{t'}}|y, X)\mathcal{G}(M_{\gamma_t})}{Pr(M_{\gamma_t}|y, X)\mathcal{G}(M_{\gamma_{t'}})}\right)$$

We note that indeed, the proposed algorithm covers the whole space of models visiting them a number of times proportional to their posterior probabilities. Plus, since $B_i$ is chosen independently, the constructed process is thus a reversible ergodic Markov Chain.

## II.3 Examples

We implemented the procedure described in sections II.1 and II.2 to the same simulated and real examples as the ones used in the original article.

**Example 1.**

For the first example, we simulated 1000 datasets, each one of $n = 10$ observations, according to the true model $y = \beta_0 + \beta_1 x_1 + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$ We considered the full model of $k = 4$ variables to be: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \varepsilon$. We then had $2^4 = 16$ models. Since we did not have a high dimension problem, there was not need for a stochastic search approach. The Bayes factor and therefore the posterior probabilities of each model were then calculated with respect to this model. The results for the model ranking are displayed in Table 1.

We found that the best model is the true model where only $x_1$ is included as regressor. Our results are very similar to the ones in the article. The true model has an average posterior probability equal to 0.39, compared to 0.41 in the article, with a slightly lower standard deviation (0.16 instead of 0.21). Just like Casella and Moreno, we also identify the model with $x_1^2$ as the second best model, but our average posterior probability is lower than theirs (0.12 against 0.259). The rank of the other models differs from their result. Though, it is not an issue as all other models were found to have very low and close average posterior probabilities, compared to the true model.

| | Average posterior probability | Standard deviation | % maximum BF |
|---|---|---|---|
| $x_1$ | 0.3981 | 0.1649 | 79.9 |
| $x_1^2$ | 0.1285 | 0.1319 | 12.7 |
| $x_1 + x_2^2$ | 0.0979 | 0.0536 | 2.6 |
| $x_1 + x_2$ | 0.0963 | 0.0487 | 2.1 |
| $x_1 + x_1^2$ | 0.0915 | 0.0547 | 1.5 |
| $x_2 + x_1^2$ | 0.0341 | 0.0339 | 0.7 |
| $x_1^2 + x_2^2$ | 0.0328 | 0.0308 | 0.0 |
| $x_1 + x_2 + x_2^2$ | 0.0327 | 0.0358 | 0.2 |
| $x_1 + x_2 + x_1^2$ | 0.0300 | 0.0237 | 0.0 |

Table 1: For example 1 a. with n=10, $\sigma = 2$ and 1000 simulations

Later, we considered a different true model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. With $\varepsilon$ as before.

Again, based on this true model we simulated 1000 datasets, each one of $n = 10$ observations. We took as the full model the same as before. So here we also had $2^4 = 16$ models, for which we calculated the posterior probabilities with respect to the full model. The results for the model ranking in this case are displayed in Table 2.

Once again, our two best models are also the ones as the ones in the article. We find that the true model (with regressors $x_1$ and $x_2$), which is the one we select, has an average posterior probability equals to 0.5. That is, we obtained even better results than ones of the paper (their average posterior probability for the true model was equal to 0.309). The model including only $x_2$ and $x_1^2$ was ranked second with an average posterior probability of 0.16, which is similar to what the authors found (0.215). The remaining models have very low posterior probability (less than 0.1) and their ranking order does not correspond exactly to the article's one.

| | Average posterior probability | Standard deviation | % maximum BF |
|---|---|---|---|
| $x_1 + x_2$ | 0.5003 | 0.2157 | 77.8000 |
| $x_2 + x_1^2$ | 0.1603 | 0.1685 | 12.7000 |
| $x_1 + x_2 + x_1^2$ | 0.0766 | 0.0652 | 1.2000 |
| $x_1 + x_2 + x_2^2$ | 0.0742 | 0.0499 | 0.5000 |
| $x_1^2 + x_2^2$ | 0.0702 | 0.1028 | 4.7000 |
| $x_1 + x_2^2$ | 0.0597 | 0.0932 | 3.0000 |
| $x_2 + x_1^2 + x_2^2$ | 0.0272 | 0.0251 | 0.0000 |
| $x_1 + x_2 + x_1^2 + x_2^2$ | 0.0150 | 0.0199 | 0.1000 |
| $x_1 + x_1^2 + x_2^2$ | 0.0138 | 0.0166 | 0.0000 |

Table 2: For example 1.b with n=10, $\sigma = 2$ and 1000 simulations

**Example 2.**

In this example as in the first one, we simulated 1000 datasets of $n = 10$ observations, according to the same true model of example 1.b:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

But here we considered as the full model:

$$y = \beta_0 + \sum_{i=1}^{3} \beta_i x_i + \sum_{i=1}^{3} \tau_i x_i + \sum_{i>j}^{3} \eta_{ij} x_i x_j + \eta_{123} x_1 x_2 x_3 + \epsilon$$

where the $x_i$ were generated uniformly on $[0, 10]$ and $\epsilon \sim N(0, \sigma^2)$. As we have 10-predictors in the full model, there are $2^{10} = 1024$ models to be compared. Given the fact that the dimension of the problem was not too small, this example was useful to test the stochastic search algorithm. Results for when $\sigma = 2$ , and $\sigma = 5$ are displayed in Table 3 and Table 4.

For both values of $\sigma$ used, we see that with respect to the results of the article, our implementation of the stochastic search not only selects as best model the true model, but visits it a much higher proportion of times (10% more for the case $\sigma = 2$ and around 40% more for the case $\sigma = 5$). We then remark that even when we have a higher variance with $\sigma = 5$, our stochastic search algorithm had great results. We only display the best three models, since the rest of them had an almost null visit proportion.

| Model : $\sigma = 2$ | Proportion of visits |
|---|---|
| $x_1 + x_2$ | 0.96 |
| $x_1 + x_2 + x_1 x_3$ | 0.02 |
| $x_1 x_2$ | 0.01 |

Table 3: For example 1.b with n=10, $\sigma = 2$ and 1000 simulations. Using our Stochastic Search Algorithm

Perhaps we should nuance the results of the visit proportion being of 0.96. It might be because the stochastic search found the model at the beginning of its search and the Metropolis-Hastings algorithm would keep that model and never change to another one. The 0.96 might be sheer luck, yet even a lesser proportion would still be very satisfying.

| Model : $\sigma = 5$ | Proportion of visits |
|---:|:---:|
| $x_1 + x_2$ | 0.63 |
| $x_2 + x_1^2 + x_1x_2 + x_2x_3 + x_1x_2x_3$ | 0.10 |
| $x_2 + x_1^2$ | 0.10 |
| $x_1 + x_2 + x_3^2 + x_1x_2$ | 0.03 |
| $x_2 + x_1x_2 + x_2x_3$ | 0.02 |

Table 4: For example 1.b with n=10, $\sigma = 5$ and 1000 simulations. Using our Stochastic Search Algorithm

**Example 3. Hald Data**

This was the first of our real data examples. It consists on 13 observations, where the response variable $y$ is the heat evolved in a cement mix, and the four explanatory variables are ingredients of the mix. Since we have a reduced number of variables, the stochastic search is of no use. We followed the same procedure as in example 1, and decreasingly sorted the models according to their posterior probability. Table 5 refers to the models whose posterior probability was found to be at least 0.001.

We notice that our first two best models are the same as in the article, and their posterior probabilities are also similar to the ones found by the authors (0.5224 and 0.1295 respectively). Models classed in the third and fourth position are in inverse order with respect to the authors findings, although their posterior probabilities are very similar. For the rest of the models the ranking is exactly the same, even if evidently the posterior probabilities differ a little from the articles results. To sum up, like the authors, we reinforced the conclusion made by Berger and Pericchi (1996) [2], by finding that the model $\{x_1, x_2\}$ is strongly preferred over the models $\{x_1, x_4\}$ and $\{x_3, x_4\}$.

| | Posterior probability |
|---:|:---:|
| $x_1 + x_2$ | 0.5308 |
| $x_1 + x_4$ | 0.1700 |
| $x_1 + x_2 + x_4$ | 0.0962 |
| $x_1 + x_2 + x_3$ | 0.0952 |
| $x_1 + x_3 + x_4$ | 0.0765 |
| $x_2 + x_3 + x_4$ | 0.0176 |
| $x_1 + x_2 + x_3 + x_4$ | 0.0096 |
| $x_3 + x_4$ | 0.0039 |

Table 5: For example 3 Hald Data

**Example 4. Ozone Data**

The final example to which we applied the model selection method was the Ozone data. The problem was to predict the daily maximum one-hour-average ozone reading $y$. There were 12 additional variables (3 categorical and 7 numerical ones),all described in Figure 1.

| *Variable* | *Description* |
|---|---|
| $y$ | Response = Daily maximum 1-hour-average ozone reading (ppm) at Upland, CA |
| $x_1$ | Month: 1 = January, ..., 12 = December |
| $x_2$ | Day of month |
| $x_3$ | Day of week: 1 = Monday, ..., 7 = Sunday |
| $x_4$ | 500-millibar pressure height (m) measured at Vandenberg AFB |
| $x_5$ | Wind speed (mph) at Los Angeles International Airport (LAX) |
| $x_6$ | Humidity (%) at LAX |
| $x_7$ | Temperature (°F) measured at Sandburg, CA |
| $x_8$ | Inversion base height (feet) at LAX |
| $x_9$ | Pressure gradient (mm Hg) from LAX to Daggett, CA |
| $x_{10}$ | Visibility (miles) measured at LAX |

Figure 1: All variables used for the Ozone Database

As mentioned by the authors, in the correlation plot for the numerical variables (Figure 2.) we can see that there are two variables ($x_{11}$ and $x_{12}$, which correspond to temperature at El Monte, CA and temperature at LAX) highly correlated with other variables of the dataset. To avoid this multicollinearity issue, we deleted them as it was done in the original article. At first sight, we observe a strong correlation between $y$ and most of the numerical predictors, except with $x_5$ and $x_9$, and especially with $x_7$.

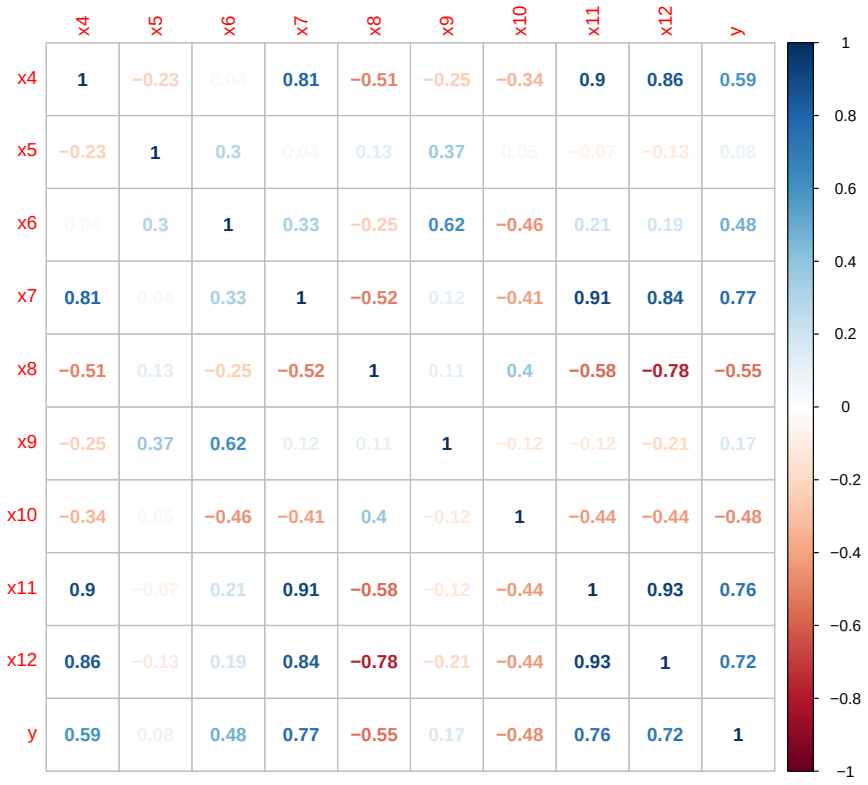



Figure 2: Correlation plot between numerical variables of Ozone Data

Regarding the categorical variables $x_1$, $x_2$ and $x_3$ (corresponding to the month, the day of the month and the day of the week), we see in figure 2. that the response variable $y$ takes in general higher values and also higher dispersion for months around the middle of the year. The other two variables do not seem to have such a strong relation with the response variable.

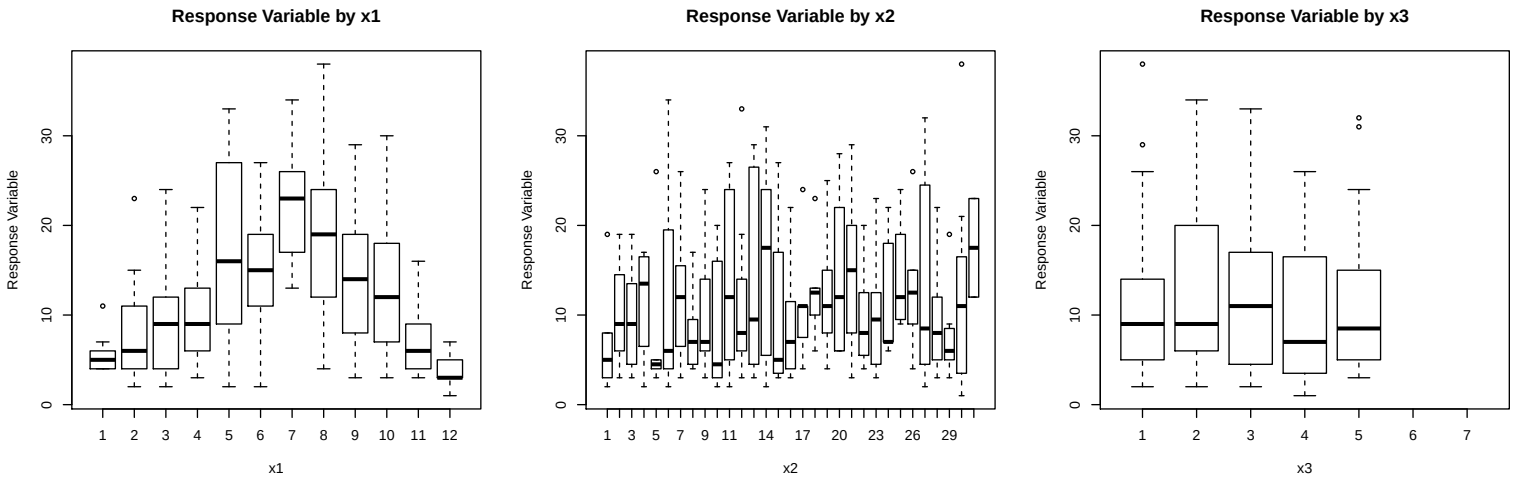



Figure 3: Box plots between the categorical variables and the response variable of Ozone Data

We first evaluated the models with respect to the full model containing all the simple terms (no quadratic terms nor interactions). We then had $2^{10} = 1024$ models. In this case we did not make use of the stochastic search. In section III we will consider the full model as the one containing all the quadratic terms and interactions, in which case we will need the stochastic search. As the authors, we constructed the models using 178 observations picked randomly, and saved the other 25 to evaluate the $R^2$ and $SME$ (the squared root of the mean of the squared prediction errors).

In Table 6 we have the results for this application. Our first remark is that contrary to the results of the authors, all of our best models contain the variable $x_1$, month of the year. However, we noticed that when evaluating the $R^2$ and the SME of the models chosen by the authors we did not obtained the same conclusion about their accuracy and performance.

We may ask ourselves whether this difference in results is due to the coding or manipulation of the categorical variables $x_1$, $x_2$, and $x_3$ that the authors may have used. Indeed, our results approach a lot more to theirs when we only consider the numerical explanatory variables in the full model. By doing so, we find the same best model as them: $x_6 + x_7 + x_8$. Another possible explanation to the difference in the results could be the randomness introduced by the division of the dataset in training and test data. Finally, our dataset itself could not correspond perfectly to the one that was used by the authors, as they mention they did not used exactly the same one as Breiman and Friedman (1985)[5].

| Model | Posterior probability | $R^2$ | Adjusted $R^2$ | SME |
|---|---|---|---|---|
| $x_1 + x_6 + x_7 + x_8 + x_9 + x_{10}$ | 0.1796 | 0.7580 | 0.7340 | 4.7537 |
| $x_1 + x_6 + x_7 + x_9 + x_{10}$ | 0.1126 | 0.7513 | 0.7283 | 4.8632 |
| $x_1 + x_6 + x_7 + x_8 + x_9$ | 0.1119 | 0.7464 | 0.7229 | 4.8114 |
| $x_1 + x_4 + x_6 + x_7 + x_9 + x_{10}$ | 0.1061 | 0.7615 | 0.7377 | 4.9764 |
| $x_1 + x_4 + x_6 + x_7 + x_8 + x_9 + x_{10}$ | 0.0851 | 0.7653 | 0.7403 | 4.8565 |
| $x_1 + x_4 + x_6 + x_7 + x_8 + x_9$ | 0.0485 | 0.7544 | 0.7300 | 4.8796 |
| $x_1 + x_4 + x_6 + x_7 + x_9$ | 0.0368 | 0.7490 | 0.7257 | 5.0493 |
| $x_1 + x_6 + x_7 + x_9$ | 0.0364 | 0.7372 | 0.7146 | 4.9874 |
| $x_1 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$ | 0.0254 | 0.7593 | 0.7338 | 4.7287 |

Table 6: For example 4 Ozone Data using as full model the model without quadratic terms or interactions

# III Ozone Data Stochastic Search

Finally for the Ozone data we consider the full model containing all the simple terms, quadratic terms and interactions. But we had in addition to add all the dummy variables. Since there are a lot of categories, 31 days, 7 weekdays, 12 months, there are a lot of crossed variables. As readability was a cause of concern, we decided to write dummy variable as such: $x_{1_7}$ is the dummy variable of variable $x_1$ with the value 7. When it equals to one it means that the time of interest is in July.

We then have 720 variables, instead of 64 for the article. That number is greater than the number of lines in our dataset. Therefore, we can only test the stochastic search with a maximum number of the number of lines in our dataset. We therefore have a number of models with almost 53 zeros. Calculating a model per second would mean having to wait $3.10^{45}$ years! Given the high dimension, in this case we are forced to make use of the stochastic search.

After calculating posterior probabilities of 500 000 models from different $B_i$, with $i$ lesser than the maximum number of variables we can choose, we obtained the following ranking in Table 7 of the groups having the best models. Table 7 shows the models with the highest Bayes factor in the 500 000 simulations come from 7 variables models. Then comes models with 1 variable and ranking third is models with 5 variables. Not seeing variables with more than 20 variables is normal since the probability of choosing them are all under $10^{-4}$. Model Simplicity is therefore preferred.

| Number of Variables | Estimation of Posterior Probability of $B_i$ |
|---:|---:|
| 7 | 0.39 |
| 1 | 0.31 |
| 5 | 0.08 |
| 2 | 0.08 |
| 10 | 0.05 |
| 9 | 0.04 |

Table 7: Top probabilities used in the Search algorithm to chose a number of variables

Table 8 below displays the obtained results of the Metropolis-Hastings algorithm :

| Model | Proportion of visits | $R^2$ | Adjusted $R^2$ |
|---|---|---:|---:|
| $x_6x_7 + x_6x_8$ | 0.71 | 0.69 | 0.69 |
| $x_{1_2}x_{2_{27}}+x_{1_2}x_{3_2}+x_{1_2}x_{3_4}+x_{1_{11}}x_{2_{24}}+x_{1_{11}}x_5+$ $x_{2_{19}}x_9 + x_{2_{22}}x_5 + x_6x_7 + x_6x_8$ | 0.11 | 0.73 | 0.71 |
| $x_5^2 + x_{2_7}x_7 + x_{2_{15}}x_{3_2} + x_{2_{17}}x_9 + x_{2_{30}}x_{3_1} +$ $x_6x_7 + x_6x_8$ | 0.06 | 0.72 | 0.71 |
| $x_6+x_8+x_{1_2}x_8+x_{1_9}x_{2_{13}}+x_{2_{13}}x_7+x_{2_{17}}x_{3_2}+$ $x_6x_7$ | 0.06 | 0.72 | 0.71 |
| $x_6^2 + x_{1_3}x_5 + x_{1_9}x_{2_8} + x_{2_6}x_7 + x_{2_{12}}x_{3_3} +$ $x_{2_{16}}x_6 + x_6x_7$ | 0.02 | 0.71 | 0.70 |
| $x_7^2$ | 0.01 | 0.63 | 0.63 |

Table 8: For example 4 Ozone Data using as full model the model with quadratic terms and interactions and using dummy variables ran on 1 000 000 simulations

Those numbers should be way higher if perfect results are the goal, since currently with them we are only analyzing $10^{-47}$% of the database. However, we decided not to spend too much time on the computation.

From Table 8, we notice that our $R^2$ are close from the article's results. We also notice that every type of model was chosen at least once, complex ones or simple ones. However, as our posterior probability distribution is more centered towards the lower end of the $B_i$, models with more than 15 variables were rarely considered.

In addition, some variables come back often, as we can see that the square of $x_7$ and $x_6$ appears in several models, as well as the dummy variable of the month of July, and end of the month variables like $x_{2_{30}}$. Using these variables is not such a far-fetched idea, since in a way it illustrates the same effect perceived in the box-plots earlier.

Overall, it is very satisfying that with more than 720 variables possible and despite running few simulations to find $G$, we still find reasonable models which have very close $R^2$ with the article's results.

# IV Conclusion

As argued by the authors in the article, their proposed method is composed of two independent parts: the first one based on an objective Bayesian prior to construct the criteria for selecting models and the second one consisting of a search algorithm that allows to avoid the calculation of this criterion for every model when there is a large number of them. We implemented the papers[1] proposition by using for the first part a mixture of g priors, and keeping the second part exactly the same.

Our algorithms evidence very good results on simulated data as in real data. First, we observed that for the low dimensional problems, which only required to apply the first part of the methodology, the selection using the posterior probabilities of the models based on the g priors allowed to choose the true model in the simulated examples, and the same model selected by the authors for the Hald Data. Although it was not the case for the Ozone Data, there are some considerations that might explain the differences.

As for the high dimensional problems, which made use of the stochastic search part, we not only chose the true model as the best one for the simulated examples, but also our proportion of visits to it was much higher that the one presented in the original paper, even when having a more variable noise in the response variable. Finally for the real application on the high dimensional version of the Ozone Data example, we managed to recover a satisfying model, considering we had 720 variables and not 65 like in the article. Increasing simulations could help bring forth even better results.

In conclusion, we saw that the proposed method can be efficiently applied to simulated or real data, obtaining very good results: selecting the true model with a high posterior probability for the simulated data, and selecting a model with a high posterior probability and good performance (amount of variability explained by the model, and prediction accuracy) for real data. Undoubtedly, it is a method that will be useful for us in further applications.

# References

[1] George Casella and Elias Moreno. Objective Bayesian variable selection. J. Amer. Statist. Assoc., 101(473):157-167, 2006.

[2] Berger, J. O., and Pericchi, L. R. (1996), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.

[3] George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.

[4] Liang, F., Paulo, R., Molina, G., Clyde, M. A.,and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. Journal of the American Statistical Association, 103(481), 410-423.

[5] Breiman, L., and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. Journal of the American statistical Association, 80(391), 580-598.

[6] Jean-Luc Jannink,1 and Rohan L. Fernando (2004). "On the MetropolisHastings acceptance probability to add or drop a quantitative trait locus in Markov Chain Monte Carlo-based Bayesian analyses"." *Genetics* Jan;166(1):641-3.

[7] Green, P. J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82: 711–732.

[8] Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer-Verlag, 2nd edition

[9] `https://imgs.xkcd.com/comics/frequentists_vs_bayesians.png`