ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Bayesian variable selection using hyper-g prior and Adaptive sampling

By

Fivos G. Anastasakis

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfilment of the requirements for

the degree of Master of Science in Statistics

Athens, Greece

June 2015

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ

Μπεϋζιανή επιλογή μεταβλητών με χρήση g-prior και προσαρμοστικής δειγματοληψίας

Φοίβος Γ. Αναστασάκης

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής

του Οικονομικού Πανεπιστημίου Αθηνών

ως μέρος των απαιτήσεων για την απόκτηση

Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Ιούνιος 2015

DEDICATION

To my grandmother.

ACKNOWLEDGEMENTS

I would like to thank Dr. Ioannis Ntzoufras firstly for his guidance but mainly for his patience.

VITA

I was born in January 15th, 1985 where I got my High School Diploma. In 2008 I graduated from the Department of Statistics of Athens University of Economics and Business and applied for the Full Time Master of Science in Statistics. I am currently working in a market research company as a statistical analyst.

ABSTRACT

Fivos Anastasakis

Bayesian variable selection using hyper-g prior and Adaptive sampling

June 2015

Bayesian variable selection has become an area of extensive research through the last decades. The two main challenges that a researcher confronts, is the specification of the prior distribution on model parameters and the calculation of the posterior model probability which makes the evaluation of a candidate model feasible. In linear models, popular prior choices are based on conjugate analysis of Normal-Gamma family. Among them, alternatives based on Zellner's g-prior are mainly preferred, as they lead to tractable marginal likelihoods. On the other hand, since posterior inference is related to high dimensional integrals, Bayesian model selection became popular only after the adoption of advanced simulation algorithms, that are used to overcome demanding computational issues.

In the current thesis, we will attempt a review of the existing methodologies that deal with the Bayesian model selection problem. Different ways of estimating Bayes Factors will be covered and major MCMC based algorithms that deal with the exploration of model space and estimation of posterior will be presented. Emphasis will be given on Bayesian adaptive sampling algorithm of Clyde et al. (2011) that exploits the idea of adaptive sampling algorithms and adopts Zellner's g-prior to perform sampling over model space. Its performance will be explored both using small and large simulated data.

ΠΕΡΙΛΗΨΗ

Φοίβος Αναστασάκης

Μπεϋζιανή επιλογή μεταβλητών με χρήση g-prior κατανομής και προσαρμοστικής δειγματοληψίας

Ιούνιος 2015

Οι μέθοδοι Μπεϋζιανής επιλογής μεταβλητών αποτελούν τις τελευταίες δεκαετίες ένα τομέα ενδελεχούς έρευνας. Οι δύο σημαντικότερες προκλήσεις που οι ερευνητές έχουν να αντιμετωπίσουν είναι η επιλογή της εκ των προτέρων κατανομής των παραμέτρων του μοντέλου και ο υπολογισμός της εκ των υστέρων κατανομής του μοντέλου, η οποία καθιστά ικανή την αξιολόγησή του. Στα γραμμικά υποδείγματα, δημοφιλείς επιλογές εκ των προτέρων κατανομών, βασίζονται στην συζυγή ανάλυση μέσω της οικογένειας Κανονικής και Γάμα κατανομής. Εξ αυτών, προτιμιτέες εναλλακτικές, βασίζονται στην εκ των προτέρων κατανομή g του Zellner, δεομένου ότι καθιστούν τον υπολογισμό της περιθώριας κατανομής εφικτή. Από την άλλη, δεδομένου ότι η εκ των υστέρων συμπερασματολογία, σχετίζεται με τον υπολογισμό ολοκληρωμάτων υψηλών διαστάσεων, οι Μπεϋζιανές μέθοδοι επιλογής μεταβλητών απέκτησαν δημοτικότητα με την υιοθέτηση εξειδικευμένων αλγορίθμων προσομοίωσης, που χρησιμοποιούνται για να ξεπεραστούν απαιτητικά υπολογιστικά προβλήματα.

Στην παρούσα εργασία, θα επιχειρηθεί μία ανασκόπηση των υπαρχουσών μεθοδολογιών που σχετίζονται με την Μπεϋζιανή επιλογή μεταβλητών. Θα καλυφθούν διαφορετικές μέθοδοι εκτίμησης του παράγοντα Bayes καθώς και βασικοί αλγόριθμοι βασισμένοι στην MCMC μεθοδολογία, που σχετίζονται με την δειγματοληψία στο χώρο του μοντέλου και στην εκτίμηση της εκ των υστέρων κατανομής. Έμφαση θα δοθεί στον αλγόριθμο Μπεϋζιανής προσαρμοστικής δειγματοληψίας των Clyde et al (2011), ο οποίος χρησιμοποιεί την ιδέα της προσαρμοστικής δειγματοληψίας και υιοθετεί την εκ των προτέρων κατανομή g του Zellner για να πραγματοποιήσει

δειγματοληψία στον χώρο του μοντέλου. Η επίδοσή του θα μελετηθεί σε προσομοιωμένα δεδομένα.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1 : Introduction

1.1 Purpose of the thesis

The Bayesian approach for model selection problems, unlike classical methods, attempts to control for model and parameter uncertainty simultaneously. Implementation of model selection from a Bayesian perspective, entails two challenging problems: prior specification and posterior calculations. In practice, applying Bayesian methods, is associated with computationally demanding integrals, necessary in order to evaluate candidate models. These quantities can be evaluated only under specific cases, strongly related to the selection of the prior. Moreover, in cases of large model spaces, where computational complexity increases, makes the evaluation of all candidates prohibitive.

MCMC algorithms, offer a powerful tool, that helps surpass both demanding posterior calculation difficulties through approximations and facilitates the evaluation of candidate models through sampling over model space. An optional algorithm, namely Bayesian adaptive sampling (BAS), that has been introduced to perform Bayesian model selection, is provided by Clyde et al (2011). In contrast to MCMC methods, BAS performs sampling without replacement over model space. The main argument that makes the algorithm applicable, is that when model probabilities are tractable, visiting past sampled models is not necessary. In order to evaluate posterior probabilities analytically, BAS is based on conjugate Bayesian analysis and adopts Zellner's g prior over parameters. The purpose of the current thesis is to present BAS and explore its performance on simulated data.

1.2 Structure of the thesis

Chapter 2 focuses on classical methods for model selection. Traditional hypothesis testing for comparing models and basic model selection criteria that are commonly used in classical approach are presented. We review the family of information criteria and distinguish their performance based on their asymptotic properties. Basic model selection algorithms are also presented.

The third chapter deals with the Bayesian approach. Inference based on Bayes Factor and posterior odds is presented, focusing on computational difficulties and paradoxes related to it. Alternatives such as Bayes Factor variants and mathematical ways of approximating it, are reviewed. In the last section of the chapter, Monte Carlo methods of approximating Bayes Factor

are presented. We review Monte Carlo integration, Importance sampling and MCMC algorithms introduced to deal with the specific problem.

In the fourth chapter we present existing MCMC based algorithms for model determination that avoid the extensive enumeration of all candidate models. The first part describes methods for variable selection, namely SSVS, KM Sampler and GVS, while the second part discusses algorithms that directly sample from model space. Some latest advances are additionally reviewed in the final part.

Chapter 5 focuses on the Bayesian Adaptive Sampling algorithm for variable selection and is divided in two parts. After introducing the basic concepts of adaptivity in MCMC algorithms, we fully review Zellner's g prior for conjugate analysis and present different variants that have been introduced. The second part focuses on BAS algorithm. We present a full review of the algorithm, including its sampling strategy, how is adaptive and the way is implemented in detail.

The sixth chapter provides with a review of the BAS package and its functions, that Clyde et al (2011) developed in R, in order to implement the algorithm. The performance of the package is also examined in small and large sample simulated data, focusing on the comparison of results under different prior distributions that are provided by BAS package.

Chapter 2 : Classical Methods

2.1 Introduction

In classical statistics model selection issues have been studied thoroughly. In brief, there are two major ways used for model determination : hypothesis tests and model selection criteria. In hypothesis testing a model's performance is mainly evaluated in terms of error sums of squares (SSError). SSError is a measure of discrepancy between the data and the model's estimation and provides a statistic that measures the adequacy of the model (Dobson, 2002, sec 2.3.4). By comparing the model's fit with and without a vector of variables, it is actually tested whether the reduction in SSError is statistically significant. In order to conduct such a test, the models that are compared must be nested; Model A is nested within model B if it derives from model B by deleting a number of terms; in other words models A is a special case of model B (Agresti, 2002, sec 4.5.4) . Alternatively, model comparison, even between non-nested models, can be based on model selection criteria. They are mainly functions of the likelihood followed by an extra term, which is used to penalize for the addition of any extra term in the equation. Especially information criteria have become very popular and tend to replace the hypothesis tests in model comparison, due to the fact that the latter are occasionally misused. Both significance tests and selection criteria are used as stopping rules in stepwise algorithms. As it is implied by their name, stepwise algorithms, sequentially fit models by adding or deleting terms in the equation. Depending on the stopping rule, at each step, models are either sequentially tested until there is no significant improvement in fit, or evaluated by a selected criterion until it reaches an optimal value. With the development of computer science, these procedures have become the most widespread tool in model determination since they can be computed automatically, easily and rapidly.

2.2 Hypothesis Testing

The general form of hypothesis testing in model selection for normal regression is

$$\begin{aligned} H_0 &: Y = X_q\beta_q + \varepsilon \\ H_1 &: Y = X_p\beta_p + \varepsilon \end{aligned} \quad , \; q < p \qquad (2.1).$$

As it can be seen, the model corresponding to the null hypothesis is a special case of the one in the alternative and derives by setting a regressor coefficient or a vector of coefficients equal to zero $(\beta_i = 0 \;,\; i = q+1,\ldots, p)$. Using goodness of fit statistics that are based either on maximized likelihood or minimized SSError, it is actually tested whether the improvement in fit due to added

regressors in the full model is statistically significant. By accepting the null hypothesis, there is no evidence of statistically significant improvement in fit by adding extra terms in the equation and therefore the simpler model is preferable.

In linear regression applications, under the hypothesis that the reduced model fits the data adequately and assuming that the error term is identically and independently distributed, the hypothesis of reduced model against the full model is tested through an F test of the following form

$$F = \frac{(SSE_{Reduced} - SSE_{General})/(p - q)}{SSE_{General}/(n-p)} \quad (2.2),$$

where n is the number of observations, p is the number of parameters of the general model and q is the number of parameters of the reduced model. It is a log-likelihood ratio (LRT) typed test and the sampling distribution of the statistic is F with (p-q) and (n-p) degrees of freedom, which derives as the ratio of two chi-square distributions. In particular, the numerator is chi-squared distributed with p-q degrees of freedom, while the denominator is chi-squared distributed with n-p degrees of freedom (Dobson, 2002, sec 6.2.4). Further details on LRT tests can be found in Dobson, 2002, sec 5.5.

Two special cases of the F test is the F to enter statistic and the lack of fit test. The F to enter statistic tests for the statistical significance of only one regressor coefficient and it is computed by the ratio

$$F_{enter} = \frac{SSE_p - SSE_{p+1}}{SSE_{p+1}/(n-p-1)}. \quad (2.3)$$

It is F distributed with 1 and n-p-1 degrees of freedom and although it has been criticized as an inappropriate tool for model selection (Miller 1984) it is widely used, especially in stepwise algorithms which will be discussed later in this chapter.

The lack of fit statistic, tests the adequacy of a fitted model containing q out p existing variables against the full model which contains all p variables (Full model). Under the assumption that the full model fits the data well, this statistic can be used to determine whether any significant variables are missing or misspecified in the equation of the reduced model.

In generalized linear models (GLM) applications the log likelihood ratio test is computed through a different goodness of fit measure, called deviance. It is defined as

$$D_{0s} = 2(l_s - l_0), \quad (2.4)$$

where l_s is the maximized log likelihood of the saturated model; the model that fits the data exactly and its number of parameters coincides with the number of observations, and l_0 is the maximized log likelihood of a reduced model. The saturated model is the one that fits the data exactly and will always have the greater value of maximized likelihood in contrast to other models

with less number of parameters. Hence, the difference between a reduced model and the saturated one provides a distance measure through which can be tested whether a simpler model fits the data adequately.

Under the null hypothesis D_{0s} is asymtotically chi-squared distributed with degrees of freedom equal to the difference between the size of the two models and by accepting the null hypothesis the simple model can be used to describe the data sufficiently. Assuming, now, that there are two models that have been accepted through the LRT test, a simple one with number of parameters q and deviance D_{0s} and a more general with number of paramaters p, so that q<p and deviance D_{1s}, the difference in deviances can be used in order to compare the fit of the two models. Again, under the assumption that the two models fit the data well, $D_{0s}-D_{1s}$ is asymptotically chi squared distributed with p-q degrees of freedom. The above results are asymptotic; further details and examples on GLM can be found in Dobson, 2002, chapter 3, while asymptotic properties of the sampling distribution of deviance and examples can be found in Dobson, 2002, sec. 5.6.

2.3 Model Selection Criteria

Apart from significance tests, the decision on the number of variables that should be included in a model can be based on model selection criteria. They are either functions of SSError or of the likelihood and are used to evaluate the performance of a model. Through this approach, some candidate models are fitted, a selected criterion corresponding to each model is calculated and the calculated values are then compared. The model that produces the best value according to the selected criterion provides the most adequate description of the data. It follows a brief description of the most popular model selection criteria.

The main goal in regression is the understanding and reducing of the observed and unexplained variance of the dependent variable, through some explanatory ones. The overall variance can be analyzed and expressed as the sum of two different quantities; the sum of squares due to regression (SSReg)

$$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \tag{2.5}$$

and the sum of squares due to error (SSError)

$$\sum_{i=1}^{n}(y_i-\hat{y})^2. \tag{2.6}$$

Since the SSError is a distance measure between the observed and fitted values, among a set of

candidate models, the best one is expected to minimize the error's sum of squares. Consequently, it would be reasonable to use this quantity as a selection criterion. However, it is known that as variables are added in the equation there will always be a decrease in SSError, independently of the variables' importance. This property makes SSError itself inappropriate for use as a model selection criterion, due to the fact that the full model will be always producing the smallest SSError. It is presented though, as it is the basic goodness of fit measure, especially in linear regression.

A useful exploratory tool for model determination is the unbiased estimator of σ^2 :

$$MSE = \frac{SSError}{n-p} \tag{2.7},$$

where n is the number observations and p is a candidate model's size, including constant. In a data set where n is essentially greater than p (Drapper and Smith, 1981), and given that all the necessary-important variables are considered as possible covariates, MSE tends to approximate the real value of variance as variables are added in the equation. Thus, even when overfitted models are used, MSE will provide a good estimate of σ^2, which can be also determined by models of smaller complexity.

Another statistic that is used for model evaluation is the coefficient of determination

$$R^2 = 1 - \frac{SSError}{SSTotal} \quad , \quad R^2 \in [0,1]. \tag{2.8}.$$

It measures the proportion of the error variance explained by the regressors and a value equal to 1 implies that the variance due to prediction is fully explained by the fitted model. However, since it is a function of SSError, it behaves in a similar way as discussed above. A modified form of R^2, is the adjusted coefficient of determination

$$R^2_{adj} = \frac{MSE}{\hat{\sigma}^2_y} = 1 - (1 - R^2)\frac{(n-1)}{(n-p)}, \tag{2.9},$$

where s^2_y is the sample variance of y. R^2_{adj} adjusts R^2 for the number of explanatory variables p in the model. Unlike R^2, it increases only if the reduction in MSError is appreciable and it can possibly be used for comparison among models of different size. Finally, three generalizations of R^2 have been proposed for the evaluation of GLMs :

- McFadden's (1974) pseudo $R^2_M = 1 - \frac{logL(\hat{\theta})}{logL(0)}$, which tends to be smaller than R^2 ; values between 0.2 and 0.4 can be considered satisfatory,
- Cox and Snell's (1989) pseudo $R^2_{CS} = 1 - \exp\left\{-\frac{2}{n}[logL(\hat{\theta}) - logL(0)]\right\}$, which takes the number of observations into account but it also does not reach the value of one and

- Negelkerke's (1991) pseudo $R_N^2 = \frac{R_{CS}^2}{R_{max}^2}$, $R_{max}^2 = 1 - \exp\left\{\frac{2}{n} logL(0)\right\}$, which modifies Cox and Snell's R_{CS}^2, such that its upper bound is equal to one; where n is the number of observations, $L(\hat{\theta})$ is the likelihood of the fitted model and $L(0)$ is the likelihood of the constant or null model.

The last criterion that will be discussed in this section (before presenting the family of information criteria) is the Mallows' C_p statistic. Mallows (1973) introdused the following statistic

$$C_p = \frac{SSError}{\hat{\sigma}^2} - n + 2\mathrm{p}, \qquad (2.10),$$

where p is the number of candidates in a model and $\hat{\sigma}^2$ is the estimate of σ^2 using the full model. Then, by assuming that the full model provides with an unbiased estimate of σ^2, it follows that $SSError = (n-p)\hat{\sigma}^2$ and hence

$$C_p = p \qquad (2.11).$$

According to this criterion, a well-performing model is expected to produce a low C_p value that approximates p. Due to randomness, C_p may take value lower than p (Mallows 1975). On the contrary, models that produce C_p greater than p , should be considered biased and excluded from the analysis.

2.4 Information Criteria

The information criteria (IC) includes several model evaluation tools, that are based on the maximised likelihood, and take the following general form

$$IC = -2\mathrm{logL}(\hat{\theta}_p / y) + C(p, n). \qquad (2.12)$$

The likelihood's negative logarithm is used as a goodness of fit measure which, however, decreases every time a variable is added in the model. The second term, is mainly a function of the number of parameters p and possibly of the sample size n, which, in contrast to, increases as the number of covariates increases. In other words, information criteria are trying to balance between the goodness of fit and the model complexity by penalising the likelihood for each variable is in the model. The IC values are meaningless by their own, they are used only for comparison between models of different size even if they are non-nested. The model with the lowest IC, is selected. Akaike (1973), used Kullback-Leibler's Information (K-L direct divergence) to produce the first IC defined as

$$AIC = -2\log f(\hat{\theta}_p / y) + 2p \quad (2.13),$$

which is called Akaike's information criterion or simply AIC.

2.4.1 K-L Based Information Criteria

Let us assume that the observed data $S:(y_1, y_2, \ldots, y_n)'$ generated from a true but unknown density g(y) which cannot be determined exactly due to its complexity. Let us further assume a parametric family of models $f(y/\theta_m)$, $m = 1, \ldots, M$, that is used to approximate the true data generating mechanism g(y). K-L Information is defined as

$$I(\theta_m) = E_g\left\{\log \frac{g(y)}{f(y|\theta_m)}\right\} \quad (2.14),$$

and quantifies the distance between g(y) and f(y). In order to determine which model best approximates g without important loss of information, the K-L discrepancy (or relative K-L information)

$$d(\theta_m) = E[-2\log f(y|\theta_m)] \quad (2.15)$$

can be used as a distance measure. In practice, θ must be estimated from the data and the above quantity could be estimated by

$$d(\hat{\theta}_m) = E_g[-2\log f(y|\theta_m)]_{\theta=\hat{\theta}_m}. \quad (2.16)$$

However, since the lack of knowledge concerning g, makes this computation impossible, Akaike, proposed $-2\log f(y|\hat{\theta}_m)$ as a biased estimator of $d(\hat{\theta}_p)$ and estimated the bias

$$E[d(\hat{\theta}_p)] - E[-2\log f(y|\hat{\theta}_m)] \quad (2.18)$$

to be, asymptotically equal to 2p. Therefore, AIC provided an asymptotically unbiased estimator of the average distance between a fitted model and the true but unknown density that generated the observed data. In other words, the model that minimises AIC, is expected *to provide the closest approximation to g(y)*. This is the main difference between criteria that derived based on K-L Information and consistent criteria. The main goal of the latter is to asymptotically identify the exact true model that generated the observed data instead of minimizing the distance between them.

The introduction of AIC in the statistical inference generated a new research on the topic, introducing new and variants of it. Some of these variants derived from taking different starting points (f.i. Bayesian analysis, predictive risk, etc), while others derived by relaxing the initial assumptions of AIC's construction, and others focused on the improvement its properties.

Regarding the assumptions, the asymptotic unbiasedness of AIC holds only when (i) the true density that generated the data S is a member of the models under consideration and (ii) a large sample relative to p is available.

Takeuchi (1976), by assuming that the true density is not included in the set of models under consideration, introduced a generalised large sample estimate of K-L discrepancy defined as

$$TIC = -2\log L(\hat{\theta}_m \, divided \, y) + 2\mathrm{tr}[J(\hat{\theta}_m) I(\hat{\theta}_m)^{-1}] \qquad (2.19),$$

where

$$J(\hat{\theta}_m) = -\frac{\partial^2 logf(x|\hat{\theta}_m)}{\partial \theta_m^2} \qquad (2.20)$$

and

$$I(\hat{\theta}_p) = \sum_{i=1}^{n} \left[\frac{\partial}{\partial \theta} logf(y|\hat{\theta}_m)\right]\left[\frac{\partial}{\partial \theta} logf(y_i|\hat{\theta}_m)\right]' \qquad (2.21).$$

When (I) is true, then TIC's penalty coincides with AIC's penalty. Eventhough, TIC is a useful generalization of AIC, accurate estimation of I and J requires large samples, which practically is difficult to be obtained.

Hurvich and Tsai (1989) introduced

$$AIC_c = AIC + 2m\frac{(m+1)}{n-m-1} \qquad (2.22),$$

as a corrected form of AIC for small samples. In their simulation studies, they showed that in cases when n is small in comparison to the number of regressors (rule of thumb n/m < 40), AIC tends to select over-fitted models, while AIC_c selects the correct one.

Fujikoshi and Satoh (1997) introduced modified AIC by relaxing both the true model and the large sample assumptions, which in linear regression applications takes the form of

$$MAIC = AIC_c + \left[2m\left(\frac{(n-m)\hat{\sigma_m}^2}{(n-M)\hat{\sigma}^2_{full}}\right) - 2\left(\frac{(n-m)\hat{\sigma_m}^2}{(n-M)\hat{\sigma}^2_{full}} - 1\right)^2\right] \qquad (2.23),$$

where m is the size of the candidate model, M the size of the full model, $\hat{\sigma_m}^2$ the maximum likelihood (ML) estimator of error variance associated to the candidate model and $\hat{\sigma}^2_{full}$ the ML estimator of error variance associated to the full model) .

Leberton et al. (1992) introduced the Quasi-AIC family which is appropriate in GLM applications where overdispersion is detected. In the exponential family of distributions, if μ represents the mean of the dependent variable, then the variance of the dependent variable Y is computed from $Var(Y) = a(\varphi)Var(\mu)$. Usually, $a(\varphi)$ has the form of $\frac{\varphi}{w}$, where w is a

known weight for each observation (for instance the number of observations n) and φ is the dispersion parameter. Theoretically, for binomial or Poisson distributed data, φ is equal to one and under the assumption of independence and homogeneity among the observations, it can be estimated from the global model, by dividing the models deviance by the number of residual degrees of freedom. When $\hat{\varphi}>1$, the observed variance is greater than the expected (overdispersion) and this could lead in selecting over-fitted models. If this is the case, the extra variability should be taken into account and the selection should be then based on the following modified criteria

$$QAIC = -\frac{2\log f(\hat{\theta}_m|y)}{\hat{\varphi}} + 2m \quad \text{and} \quad QAIC_c = QAIC + 2m\frac{(m+1)}{n-m-1} \qquad (2.24).$$

Finally, Cavanaugh (1999, 2004) proposed the use of the K-L symmetric divergence as an alternative basis for information criteria derivation. The K-L symmetric divergence, defined as the sum of two directed as

$$J(\theta_m) = E_g\left[\frac{g(y)}{f(y|\theta_m)}\right] + E_f\left[\frac{f(y|\theta_m)}{g(y)}\right] \qquad (2.25),$$

was suggested as a more sensitive distance measure that would reflect more accurately the separation between two densities. Depending on that, Cavanaugh, introduced

$$KIC = -2\log f(\hat{\theta}_m|y) + 3m \qquad (2.26),$$

as a large sample estimator of K-L symmetric divergence and a corresponding correction for small samples

$$KICc = KIC + \frac{2m(m+1)}{n-m-1} \qquad (2.27).$$

2.4.2 Consistent Criteria

A distinction between model selection criteria are made according to their asymptotic properties. In particular, they are divided in two categories; the first one includes AIC, AIC_c, C_p and *adjusted* R^2 all of which are asymptotically efficient with respect to MSE. This means that as the number of observations increases, the above criteria tend to select the model that minimizes MSE.

The second group of information criteria, is characterized by property of consistency. Assuming that the true model is among the list of candidate models, then an asymptotically consistent criterion, will choose the true model with probability tending to one (weak consistency) or almost surely (strong consistency) for large samples.

Obviously, the above assumption is not realistic in most applications and, thus, consistency has been also examined in terms of K-L distance. Then, a consistent criterion would select the model that minimized the K-L distance. Furthermore, due to the parsimony, a consistent criterion would select the simplest model when two or more models are equivalent in terms of K-L distance. AIC, TIC, AIC_c and, all criteria with penalty that does not depend on n, do not achieve consistency. As a consequence, there is a possibility of selecting, unnecessarily, over-fitted models. (Claeskens G. and Hjort N. L. , 2008, sec. 4.1)

The most popular consistent criterion was introduced by Schwarz (1978) and derived using purely Bayesian arguments. A Bayesian rule of selection would choose the candidate model M with the highest posterior probability

$$P(m|y) = \frac{\pi(m)\int L(\theta_m|y)g(\theta_m|m)\partial\theta_m}{h(y)} \quad (2.28),$$

where $\pi(m)$ is a discrete prior distribution over the model m, $L(\theta_m|y)$ is the likelihood function, $g(\theta_m|m)$ is the prior distribution over parameter vector given the model and $h(y)$ is the marginal distribution of the data vector.

By minimizing $-2\log P(m|y)$, Schwarz proposed the Schwarz Information Criterion

$$BIC = -2\log L(\hat{\theta}_m|y) + p\log(n) \quad (2.29),$$

as a large sample approximation of the log-transformed posterior distribution of model m. BIC, is closely related to AIC, yet, it penalizes model complexity more stringently.

Hannan and Quinn (1979), in order to achieve strong consistency in the selection of time series model, proposed

$$HQC = -2\log L(\hat{\theta}_m|y) + c\log[\log(n)],\ c \geq 2 \quad (2.30),$$

and Bozdogan (1987), based on Akaike's work, extended AIC to make it consistent. He introduced consistent AIC, defined as

$$CAIC = -2\log L(\hat{\theta}_p/y) + p[\log(n)+1] \quad (2.31)$$

and consistent AIC with Fisher Information, defined as

$$CAICF = AIC + p\log(n) + \log\left|J(\hat{\theta}_p)\right| \quad (2.32).$$

2.4.3 Minimum Description Length

Rissanen (1978) introduced the minimum description length principle, which again is derived from the field of information theory. According to MDL principle, one would choose the model that achieves the shortest description of the data. MDL is based on Kolmogorov's theory of algorithmic complexity which is simply the length of the shortest computer program that describes a sequence. In a similar way to the K-L Distance, the algorithmic complexity cannot be computed (theorem of incomputability of Kolmogorov's complexity). Thus, Rissanen suggested that the encoding of the data could be achieved using probability distributions. This is strongly related to Shannon's source coding theorem, which provides a lower bound for iid variable compression, without crucial loss of information. In that sense, a probability distribution is just used as a description measure of complexity in order to achieve the shortest data compression.

Depending on the strategy used for data encoding, there have been proposed several functions providing a lower bound of compression (valid description length). The first strategy applied to produce such a compression is called two-stage coding scheme and its corresponding lower bound coincides with BIC. Other coding schemes include : Mixture MDL (a coding scheme that resembles to bayesian analysis), Normalized maximum likelihood MDL and predictive MDL all of which lead to different MDL based criteria for linear regression applications while other MDL criteria focus on GLMs are also available in the literature (Hansen and Yu, (2003).

Rissanen's approach differs the traditional derivation of IC, since there are no assumptions regarding the random process of the data. Moreover, probability distributions are only used just as description tools of the data. This allows comparisons between models of different type.

2.5 Model Selection Procedures

In the current section the most popular techniques for model construction will be discussed.

2.5.1 All Subsets Regression

A reasonable, yet computationally exhaustive method, is to evaluate all possible models using a model selection criterion. This is often called full enumeration or exhaustive search. For instance, in a data set of p=5 explanatory variables, an all subsets algorithm would consist of the following steps :

- Compute the intercept model $Y_i = \beta_0 + e_i \ , \ i = 1,\ldots,n$ and evaluate it.
- Compute all models including one explanatory variable
  $Y_{i,j} = \beta_0 + \beta_j \ X_{i,j} + e_{i,j} \ , \ i = 1,\ldots,n \ , \ j = 1,\ldots,5$ and evaluate them.
- Compute all models including two explanatory variables
  $Y_{i,j} = \beta_0 + \beta_j \ X_{i,j} + \beta_k \ X_{i,k} + e_{i,j} \ , \ i = 1,\ldots,n \ , \ j,k = 1,\ldots,5 \ , \ j \neq k$
  and evaluate them.

  ⋮
- Compute full model and evaluate it.
- Choose the best model according to a selected criterion.

In practice, such a thorough search is not useful when dealing with real data sets, especially large ones. A huge number of possible models needs to be evaluated even for moderate p since the number of all possible models will be equal 2^p. Moreover, the researcher's aim of constructing simple and small-sized models with satisfactory fit and possibly the prior knowledge concerning the relationship between dependent and independent variables (in the sense that some regressors should or should not be excluded from the equation), make the use of all subset regression unnecessary in practical problems. Yet, in cases of small or moderate-sized data sets, it could still be a useful tool.

2.5.2 Stepwise Algorithms

An effective, timesaving and computationally simple alternative compared to all possible regressions, is offered by techniques. There are three main approaches that are widely used in model construction via stepwise search: forward selection, backward elimination and stepwise selection. Depending on the algorithm, the particular methodology attempts to construct a model by sequentially adding or deleting one variable at each step, taking into account the presence of the other regressors that are in the model. In order to achieve this, the algorithm selects a candidate variable and evaluates its contribution, measuring whether the reduction in the total SSError is large enough or not. In other words, the model that includes and the one that does not include the candidate variable are compared and is decided weather the model's fit is improved. The evaluation can be performed in two ways; either by using an F or chi-squared test and their corresponding p-value, or a model selection criterion.

(a) Forward Selection & Backward Elimination

The forward selection algorithm starts with the intercept model, adding at each step the independent variable X that is most significant, according to a significance test or a selection criterion. The algorithm stops when no further improvement is achieved by adding an explanatory variable. A general form of forward addition using an F test would be:

- Begin with the intercept model
- Choose the k^{th} regressor that produces the maximum F test value $max_{k=1...p} F_k ; k = argmax(F_k)$
- If $max\, F_k > F_{enter}$, where F_{enter} is an arbitrary significance value, (usually the 95th percentile of the F distribution) add X_k in the model and go to step 2
- Stop if for all remaining candidates $max\, F_k < F_{enter}$

Conversely, the Backward Elimination Algorithm, would have as a starting point the model that contains all available variables, and would decide, which of the regressors are not statistically significant, so as to be removed. An example of such an algorithm would be:

- Begin with the full model
- Choose the k^{th} regressor that produces the minimum F test value

$min_{k=1...p} F_k ; k = argmin(F_k)$

- If $min F_k < F_{remove}$, where F_{enter} is similarly an arbitrary significance value, (usually the 5th percentile of the F distribution) remove X_k from the model and go to step 2
- otherwise STOP

(b) Stepwise Search

As it can be seen, the above algorithms do not re-examine the significance of a variable that is added or deleted from the model. In forward selection, once a variable is added, it will never be deleted. Similarly, in backward elimination, once a variable is excluded, it will never be added in any further step. Stepwise Algorithms, combining Forward and Backward selection, provide a solution to this problem

There are two alternatives when using stepwise search: Stepwise Forward Algorithm, begins with the null model, adding at each step the most significant independent variable. However, before proceeding to a new addition, re-examines whether any of the previously added variables have become insignificant in the presence of the new variable. On the other hand, Stepwise Backward Algorithm begins with the full model, testing for variables to exclude. Since a variable is excluded, previously deleted variables are re-examined for potential re-addition. Obviously, despite the fact that the computations in Forward and Backward Selection are quicker, stepwise search is preferable since it conducts double tests at each step.

Stepwise procedures, have become very popular and are widely used by non-profesional statistician too, as they can be easily performed automatically in all statistical packages. However, there are disadvantages that have been discussed in several papers, and are briefly summarized in the next paragraph.

Firstly, not only there is no guarantee that the different algorithms will select the same model, but also, there is no guarantee that the selected model will be the correct one. The order of variable addition or deletion and the selection criterion, affects the final construction of the model. As a result, there might exist a different one, performing equally well which will not be examined.. An additional issue that weakens the performance of the algorithms, is the multiple testing of the null hypothesis $H_0 : \beta_k = 0$. That leads to an increase of the tests Type I and Type II Error (include not significant regressors that should be deleted and delete important regressors that should be added in the model). Furthermore, the distribution of the F statistic is also affected as the selection of variables is decided with respect to the existing observations (Pope and Webster, 1972).

Finally, since the problem of multi-collinearity is not taken into consideration as the algorithm proceeds, the final model will require further examination and in most cases corrections.

Chapter 3 : Bayesian Methods

3.1 Introduction

Inference using classical statistical arguments, is based on probability density function $f(y|\hat{\theta}_p)$. The observed data vector y is considered to be an outcome from a random variable Y, which is characterized by the population parameter vector θ. The parameter vector is assumed to be a constant quantity, that has to be estimated properly. In order to achieve that, likelihood function $L(\theta_p|y)$ is used. By maximising the likelihood function with respect to θ, the appropriate estimates are obtained.

Bayesian statistics use an alternative approach. Taking into account the uncertainty that is produced due to the ignorance concerning the parameter vector, the latter is treated as a random variable and the statistical inference is based on the posterior probability function of the parameter vector, given the observed data $p(\theta_p|y)$. Posterior probability is defined through Bayes theorem as the joint distribution of the data and the parameter vector, divided by the marginal distribution of the data.

$$p(\theta_p|y) = \frac{f(y,\theta_p)}{f(y)} = \frac{f(y|\theta_p)\pi(\theta_p)}{f(y)} = \frac{f(y|\theta_p)\pi(\theta_p)}{\int f(y|\theta_p)\pi(\theta_p)d\theta_p} \qquad (3.1).$$

The integral in the denominator is called normalizing constant and and is used in case of continuous priors. When discrete priors are used, the integral is replaced by the corresponding sum $\sum_i f(y|\theta_i)\pi(\theta_i)$.

The joint probability of θ and y can be calculated as the product of the likelihood function and $\pi(\theta_p)$, which is called prior distribution of the parameter vector. The existence of the prior distribution distinguishes the Bayesian from the classical analysis and represents the prior knowledge that someone may have concerning the parameter of interest. Depending on the prior knowledge, it could be chosen an informative distribution that would favor certain values for the parameter vector instead of others, or it could be chosen an uninformative one, such as the uniform distribution, assigning equal probabilities irrespective of the value.

3.2 Model Comparison & Hypothesis Testing

In model selection problems using Bayesian arguments, the comparison between candidate models and the decision on which model best describes the observed data, is based on the comparison of models' posterior probabilities. Assume that the observed data $y = \{y_{1,} y_{2,\ldots}, y_n\}'$ have been generated from one of the two following models, $\{M_0 \ , \ M_1\},$ according to a density $f(y|M_0)$ or $f(y|M_1).$ The first step in testing a hypothesis of the following form

$$\begin{aligned} H_0 &: M = M_0 \\ H_1 &: M = M_1 \end{aligned} \qquad (3.2),$$

is to assign prior probabilities on each model. Note that the models do not need to be nested as in classical model comparison. Let $\pi(M_0)$ denote the prior probability over model M_0 and $\pi(M_1) = 1-\pi(M_0)$ the prior probability over model $M_1.$ Then, the posterior probability of model $M_i \ , \ i=\{0,1\}$ is defined as

$$p(M_i|y) = \frac{f(y|M_i)\pi(M_i)}{\sum_i f(y|M_i)\pi(M_i)} \qquad (3.3).$$

The decision on which of the two models is preferable can be based simply on the comparison of their posterior probabilities. Then, for example, the model in the null hypothesis would be accepted if

$$p(M_0|y) > p(M_1|y) \qquad (3.4).$$

Alternatively, the posterior model odds can be used, defined as

$$PO_{01} = \frac{p(M_0|y)}{p(M_1|y)} = \frac{p(M_0|y)}{1-p(M_0|y)} = \frac{f(y|M_0)}{f(y|M_1)} * \frac{\pi(M_0)}{\pi(M_1)} = \qquad (3.5).$$
$$Bayes\ Factor \ \times \ prior\ model\ odds$$

It can be easily seen that there is no need to compute the normalizing constant appearing in the denominator of eq (3.3) since it appears both in the numerator and the denominator and it cancels out. Similarly, in cases of uninformative priors over candidate models, the posterior model odds equals to the Bayes factor.

Bayes factor, which was introduced in 1948 by Jeffreys, is of major importance in Bayesian inference and is defined as the ratio of posterior odds over prior odds. In model comparison, each model is fully specified by its parameters, yielding a likelihood of the form $f(y|M_p, \theta_{M_p}),$ where p is the size of the parameter vector. Hence, the computation of Bayes factors needs the integration of $f(y|M_p, \theta_{M_p})$ (rather than its maximization) over the parameter vector. Let BF_{01} denote the Bayes factor of model M_0 over model $M_1.$ Then, BF_{01} is computed by the following

equation

$$BF_{01} = \frac{f(y|M_0)}{f(y|M_1)} = \frac{\int f(y|M_{0,}\theta_{M_0})\pi(\theta_{M_0}|M_0)d\theta_{M_0}}{\int f(y|M_{1,}\theta_{M_1})\pi(\theta_{M_1}|M_1)d\theta_{M_1}} \quad (3.6).$$

The quantity $f(y|M_p)$ is called marginal likelihood of the data vector given a model of size p, or the predictive probability of the observed data under a model M of size p and represents the probability of obtaining the actually observed data, before any data are available, under the assumption that the model M is the real stochastic mechanism that generated the observed data (Ntzoufras 1999) ; $f(y|M_p,\theta_{M_p})$ is the likelihood function of the data vector and $\pi(\theta_{M_p}|M_p)$ is the prior distribution of the parameter vector of size p under the model M. As mentioned before, in cases of no prior knowledge concerning the candidate models, inference is based only on BF_{01}, expressing the evidence in favor of the model corresponding to the null hypothesis. Conversely, BF_{10} provide evidence against the model corresponding to the null hypothesis. (see tables provided by Kass & Raftery (1995) with numerical values of BF_{10} in logarithmic scale, on which the inference can be based).

3.2.1 Problems using Bayes factor

There are two significant problems when using Bayes factors for Bayesian inference, both of which concern the specification of the prior distribution $\pi(\theta_{M_p}|M_p)$. The first one, is related to the calculation of the integral $\int f(y|M_p,\theta_{M_p})\pi(\theta_{M_p}|M_p)d\theta_{M_p}$. In many cases the use of complex informative priors may lead in various computational problems that are impossible to overcome. This results in incapability of evaluating the above integral, unless numerical methods are used. The exact derivation can be achieved when modeling a likelihood of the exponential family with conjugate prior distribution for the parameter vector (see for instance Zellner 1971)

The second problem occurs due to the dependency and sensitivity of Bayes factors on the choice of the prior distribution. It was studied firstly by Lindley (1957) and then by Bartlett (1957) and is called Lindley's (or Bartlett) paradox, while others refer to it as Jeffrey's paradox (Lindley, 1980). The term is used to describe any situation in which classical and Bayesian analysis provide contradicting results in hypothesis testing problems and occurs in the presence of uninformative prior distributions.

In brief, it can be shown that the prior variance affects the value of Bayes factor in a way that as the variance increases, Bayes factor also increases, working always in favor of the simplest

model. This implies that flat proper priors cannot be used. Moreover, in cases of improper priors, Bayes factor can be defined only to an undetermined constant and it cannot be fully specified. Further details on Lindley's paradox are provided by Shafer (1982).

However, the use of uninformative prior distributions is essential in Bayesian analysis and therefore several marginal likelihood's estimators and Bayes factors variants that try to cope with the discussed problems, have been proposed in the literature.

3.3 Bayes factors' variants

In order to avoid BF indeterminacy when using improper priors, three variants of Bayes factors have been proposed the following

- Posterior Bayes Factor (Aitkin 1991)
- Fractional Bayes Factor (O' Hagan 1995) and
- Intrinsic Bayes Factor (Berger & Pericchi 1996)

Posterior Bayes Factor uses the ratio of the likelihood's posterior means instead of the likelihood's prior means ratio. Defining the likelihood's posterior means under model M_0 and model M_1 as

$$PM_0 = \int f(y|\theta_{M_0}, M_0)\, p(\theta_{M_0}|y, M_0)\, d\theta_{M_0}$$
$$PM_1 = \int f(y|\theta_{M_1}, M_1)\, p(\theta_{M_1}|y, M_1)\, d\theta_{M_1} \quad \text{and}$$

respectively, then the posterior Bayes factor is given by the ratio $PBF_{01} = \frac{PM_0}{PM_1}$, providing evidence in favor of the model M_0. More specifically, Aitkin claimed that $PBF_{01}'s$ values that are less than $\frac{1}{20}, \frac{1}{100}$ and $\frac{1}{1000}$, provide strong, very strong and overwhelming evidence against M_0.

PBF_{01} avoids Lindley's paradox and is not affected by normalizing constants. However, it uses the data twice and hence it does not go along with Bayesian rationale. Furthermore, as pointed out in Berger and Pericchi (1995) and O' Hagan (1995), it is not consistent in a sense that it does not tend to infinity as the sample size increases.

Intrinsic Bayes Factor, derived by using the idea of Partial Bayes Factors, which was introduced by Spiegelhalter & Smith (1982). According to the Partial Bayes Factor approach, when there is weak prior information the observed data y could be divided into two parts (y_0, y_1),

one of quite small length l and one of length $n-l$. Then, the first subsample y_0 could be used as a training sample to update the prior and the rest of the sample y_1 to obtain Bayes Factor. For two models M_0 and M_1, the partial Bayes factor based on subsample y_0 is defined as

$$PBF_{01}^{y_0} = \frac{f(y_1|y_{0,}M_0)}{f(y_1|y_{0,}M_1)} = \frac{\int f(y_1|\theta_{M_0},M_0)\pi(\theta_{M_0}|y_{0,}M_0)d\theta_{M_0}}{\int f(y_1|\theta_{M_1},M_1)\pi(\theta_{M_1}|y_{0,}M_1)d\theta_{M_1}} \quad (3.7),$$

where $\pi(\theta_{M_p}|y_{0,}M_p)$ is the updated prior based on y_0, computed as

$$\pi(\theta_M|y_{0,}M_p) \propto f(y_0|\theta_M,M)\pi(\theta_M|M) \quad (3.8).$$

$PBF_{01}^{y_0}$ is less sensitive to the prior distribution of the parameter vector and do not face the problem of unknown constants in cases of improper priors. This happens since it can be expressed as the ratio of the overall Bayes factor, that is based on the full data vector, over the Bayes factor based on the subsample y_0. The overall marginal likelihood of the data given the model M, can be written in the following form

$$f(y|M)=f(y_{0,}y_1|M)=f(y_1|y_{0,}M)f(y_0|M). \quad (3.9)$$

It follows that

$$f(y_1|y_{0,}M) = \frac{f(y|M)}{f(y_0|M)} \Leftrightarrow PBF_{01}^{y_0} = \frac{BF_{01}}{BF_{01}(y_0)}. \quad (3.10)$$

However, the disadvantage of Partial Bayes Factor is its high dependence on the subsample y_0 and its corresponding size. In order to decrease this dependency, Berger & Pericchi (1996) suggested to replace $BF_{01}(y_0)$ in (2.10) by its average $\overline{BF_{01}}(y_0)$ computed over all L samples for which all the parameters corresponding to all models are identifiable (minimal training samples; Berger & Pericchi 1996) and derived the Intrinsic Bayes Factor

$$IBF_{01} = \frac{BF_{01}}{B\bar{F}_{01}(y_0)} = BF_{01} * B\bar{F}_{10}(y_0) \quad (3.11).$$

Depending on which way the average is calculated, the following variants of IBF_{01} derive

- the Arithmetic IBF_{01} for which $\overline{BF_{10}}(y_0) = \frac{1}{L}\sum_{l=1}^{L} BF_{10}[y_0(l)]$
- the Geometric IBF_{01} for which $\overline{GBF_{10}}(y_0) = \left[\prod_{l=1}^{L} BF_{10}[y_0(l)]\right]^{(\frac{1}{l})}$

  (DeSantis & Spezzaferri, 1997)
- the Median IBF_{01} for which $\overline{MBF_{10}}(y_0) = \underset{l\in L}{med}\, BF_{10}[y_0(l)]$

  (Berger & Pericchi, 1998)

Finally, when the size of the subsample l and the size of the sample n is essentially

large, the following approximation is obtained

$$f(y_0|\theta_{M_p}, M_p) \approx f(y|\theta_{M_p}, M_p)^b \quad , \quad b = \frac{l}{n} \; < \; 1 \tag{3.12},$$

where b is called fractional parameter. Then, if in equation (2.11), $B\bar{F}_{01}(y_0)$ is substituted by the following quantity

$$BF_{01}^b \; = \; \frac{\int f(y|\theta_{M_0}, M_0)^b \pi(\theta_{M_0}|M_0)\, d\theta_{M_0}}{\int f(y|\theta_{M_1}, M_1)^b \pi(\theta_{M_1}|M_1)\, d\theta_{M_1}} \tag{3.13},$$

the Fractional Bayes Factor is obtained

$$FBF_{01} = \; \frac{BF_{01}}{BF_{01}^b} \tag{3.14}.$$

Further details concerning the derivation and properties of Bayes factors' variants are provided in the original papers of Spiegelhalter & Smith (1982), Aitkin (1991), O' Hagan (1995) and Berger & Pericchi (1996, 1998), while discussion on the use and comparisons between the different variants, are provided by DeSantis & Spezzaferri (1997) and Berger & Pericchi (1998).

3.4 Approximating Bayes factors

3.4.1 The Bayesian Information Criterion

As discussed in the first chapter, the Schwarz information criterion defined as

$$SIC = -2\log L(\hat{\theta}_{M_p}|y) + p\log(n) \tag{3.15},$$

derived as a large sample approximation of the log transformed posterior density of a candidate model , given the observed data. Hence, the following quantity (based on the Schwarz criterion)

$$S_{01} = logL(\hat{\theta}_{M_0}|y, M_0) - logL(\hat{\theta}_{M_1}|y, M_1) - \frac{1}{2}(p_{M_0} - p_{M_1})\log(n) \tag{3.16},$$

provides a rough approximation of the log transformed Bayes factor of model M_0 over model M_1 , since its main property is

$$\frac{S_{01} - logBF_{01}}{logBF_{01}} \to 0, \; for \; n \to \infty \tag{3.17}.$$

Kass & Wasserman (1995), studied under which conditions (3.17 holds and provided a corrected form of the approximation where needed. They showed that a wide range of prior distributions exists, under which S_{01} provide a useful approximation and they suggested it as preferable in contrast to BF variants, since it, in addition, does not require very large samples to provide adequate results. In general, even though S_{01} is the simplest and not always the best

approximation of Bayes Factor, it can be used as an explanatory tool, to evaluate the evidence in favor of the null hypothesis, in cases where prior distributions are hard to determine

3.4.2 The Laplace approximation

Under certain circumstances (see Kass & Raftery, 1995 and references therein), one accurate way to approximate the marginal likelihood, is the Laplace approximation (De Bruijn 1970 and Tierney & Kadane 1986). According to this method, if a real valued function h(.) of a p dimensional vector x, is expanded quadratically using the Taylor series about the value $\tilde{x}$ (the value that h(x) attains its maximum) and then is exponentiated, its integral can be approximated using the following formula

$$\int \exp\{h(x)\}dx \;=\; (2\pi)^{\frac{p}{2}}|A|^{\frac{1}{2}}\exp\{h(\tilde{x})\}. \tag{3.18}$$

The quantity A equals to minus the inverse Hessian matrix $H(h)$ of $h(x)$, evaluated at $\tilde{x}$. The Hessian matrix is simply the square matrix of second order partial derivatives of h, describing the local curvature of the function and is computed providing that all second partial derivatives exist.

$$H(h)=\left[\begin{array}{cccc} \frac{\partial^2 h}{\partial x_1^2} & \frac{\partial^2 h}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 h}{\partial x_1 \partial x_p} \\ \frac{\partial^2 h}{\partial x_2 \partial x_1} & \frac{\partial^2 h}{\partial x_2^2} & \cdots & \frac{\partial^2 h}{\partial x_2 \partial x_p} \\ \vdots & & & \vdots \\ \frac{\partial^2 h}{\partial x_p \partial x_1} & \frac{\partial^2 h}{\partial x_p \partial x_2} & \cdots & \frac{\partial^2 h}{\partial x_p^2} \end{array}\right]_{x=\tilde{x}} . \tag{3.19}$$

In order to apply the method for approximating the marginal likelihood $f(y|M_p) \;=\; \int f(y|M_p,\theta_{M_p})\pi(\theta_{M_p}|M_p)d\theta_{M_p}$, the posterior density of the parameter vector $p(\theta_{M_p}|y,M_p)$ should be unimodal, *or at least dominated by one single mode. This occurs for large samples, when the likelihood function* $f(y|\theta_{M_p},M_p)$ *is highly peaked near its maximum* θ_{max} (Gelfand & Dey, 1994 ; Kass & Raftery, 1995).

Then, if $h(x)$ is substituted in (2.9), by the logarithm of $f(y|M_p,\theta_{M_p})\pi(\theta_{M_p}|M_p)$, the following estimate is obtained

$$\hat{f}_{Laplace}(y|M_p) \;=\; (2\pi)^{\frac{p}{2}}|A|^{\frac{1}{2}}_{\theta_p=\tilde{\theta}_{M_p}} f(y|\tilde{\theta}_{M_p},M_p)\pi(\tilde{\theta}_{M_p}|M_p) \tag{3.20},$$

where p is the dimension of the parameter vector $\tilde{\theta}_{M_p}$ is the posterior mode of θ_{M_p} and A is minus the inverse Hessian matrix of the function $f(y|M_p,\theta_{M_p})\pi(\theta_{M_p}|M_p)$ evaluated at $\theta_{M_p}=\tilde{\theta}_{M_p}$.

The approximation error (also relative error), $\left|\frac{f-\hat{f}}{f}\right|$, as proved in the appendix of Tiernay & Kadane (1986) is $O(n^{-1})$ and when the approximation is applied for estimating both marginal densities for the computation of Bayes factor, the relative error remains the same.

3.4.3 Variants of Laplace

Kass and Vaidyanathan (1992) derived two variants of Laplace's method that are easier to compute, but are less accurate compared to the first approximation. Yet, they provide useful alternatives that remain reliable.

The first one is of the following form

$$\hat{f}_{MLE}(y|M_p) = (2\pi)^{\frac{p}{2}}\left|\hat{\Sigma}\right|^{\frac{1}{2}} f(y|\hat{\theta}_{MLE},M)\pi(\hat{\theta}_{MLE}|M) \qquad (3.21),$$

where $\hat{\theta}_{MLE}$ is the maximum likelihood estimator of the log-likelihood and $\hat{\Sigma}^{-1}$ is the observed information matrix; the Hessian matrix of the log-likelihood evaluated at $\hat{\theta}_{MLE}$.

The second variant is obtained simply by substituting the observed information matrix with the expected information matrix (Fisher Information) $I(\theta)=E\left[\frac{\partial}{\partial\theta}logf(y;\theta)\right]$, as the asymptotic covariance matrix. Their corresponding approximation error is $O(n^{-1})$ for the first estimate, while for the second is larger, equal to $O\left(n^{-1/2}\right)$.

Another variant of Laplace' s method was introduced by Raftery (1996a) and is useful in cases of difficulties concerning the computation of the posterior mode and the inverse of the Hessian matrix. Then, by generating a sample of size T, from the posterior $p(\theta_{M_p}|y,M_p)$ using an MCMC algorithm, the above quantities could be substituted by the simulated estimates of the posterior mean and the posterior covariance matrix. Such an algorithm would consist of the following steps :

1. Generate a sample $\{\theta^{(1)},\theta^{(2)},\theta^{(3)},\cdots,\theta^{(T)}\}$ from the posterior density $p(\theta_{M_p}|y,M_p)$
2. Calculate $\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T}\theta^{(t)}$ and $S = \frac{1}{T-1}\sum_{t=1}^{T}(\theta^{(t)}-\bar{\theta})(\theta^{(t)}-\bar{\theta})'$

3. Calculate $\hat{f}_{Metropolis}(y|M_p) = (2\pi)^{\frac{p}{2}}|S|^{\frac{1}{2}} f(y|\bar{\theta}_{M_p}, M_p)\pi(\bar{\theta}_{M_p}|M_p)$

3.4.4 Monte Carlo integration and importance sampling estimators

An optional way to estimate the value of the integral is provided through Monte Carlo integration. Assume that we want to calculate the value of the integral $I = \int g(x)f(x)dx$ and assume that $f(x)$ is a probability density function. Then, from probability theory is known that the above integral equals to the expected value of $g(x)$ with respect to the density $f: I = \int g(x)f(x)dx = E_f[g(x)]$. Then, by using the law of large numbers, for an adequately large random sample $\{x_1, x_2, x_3, \cdots, x_n\}$, $n \to \infty$, an estimator of the expected value is obtained by the sample mean as $\hat{I} = \frac{1}{n}\sum_{i-1}^{n} g(x_i)$. The variance of the estimator can be easily shown that is proportional to $\frac{1}{n}$ and hence, asymptotically, $\hat{I}$ is expected to be near the real value of I. Proper choice of g and f results in minimizing the estimator's variance.

Monte Carlo integration can be applied directly for estimating the marginal density $f(y|M_p)$, simply by generating a random sample from the prior distribution $\pi(\theta_{M_p}|M_p)$ and calculating the sample mean of the likelihood. In particular, since the following equation holds

$$f(y|M_p) = \int f(y|M_p, \theta_{M_p})\pi(\theta_{M_p}|M_p)d\theta_{M_p} = E_{\pi(\theta_{M_p}|M_p)}[f(y|M_p, \theta_{M_p})] \quad (3.22)$$

then, by generating a random sample $\{\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \cdots, \theta^{(T)}\}$ from the prior distribution $\pi(\theta_{M_p}|M_p)$ an estimate of the marginal density is obtained by

$$\hat{f}(y|M_p) = \frac{1}{T}\sum_{T=1}^{T} f(y|M_p, \theta_{M_p}^{(t)}) \quad (3.23).$$

As mentioned before, the variance of the estimator is affected by the functions involved in the integral. Hence, the choice of the prior distribution affects the efficiency of the estimator. More precisely, the choice of uninformative priors leads in increasing the variance of the estimator and the convergence of the algorithm will be slow.

3.4.5 Importance sampling

One way to obtain smaller values of the estimator's standard error is provided by importance sampling. Assuming a density function $h(x)$, the integral I can be rewritten in the following form

$$I = \int g(x) f(x) dx = \int g(x) \frac{f(x)}{h(x)} h(x) dx = \int w(x) h(x) dx = E_h[w(x)] \quad (3.24).$$

Then, by generating a random sample $\{x_1, x_2, x_3, \cdots, x_n\}$ from $h(x)$, the estimator that is obtained is

$$\hat{I} = \frac{1}{n} \sum_{i-1}^{n} w(x_i) = \sum_{i=1}^{n} g(x_i) \frac{f(x_i)}{h(x_i)} \quad (3.25).$$

The advantage of importance sampling is that it can obtain estimators of zero variance as long as the importance sampling density $h(x)$ is shaped in a similar way to the function of g.

A marginal likelihood approximation through importance sampling method, is obtained by considering the following formula

$$f(y|M_p) = \int f(y|M_p, \theta_{M_p}) \frac{\pi(\theta_{M_p}|M_p)}{h(\theta_{M_p})} h(\theta_{M_p}) d\theta_{M_p} = E_{h(\theta_{M_p})}[\frac{f(y|M_p, \theta_{M_p}) \pi(\theta_{M_p}|M_p)}{h(\theta_{M_p})}] \quad (3.26).$$

By generating a random sample $\{\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \cdots, \theta^{(T)}\}$ from the importance sampling density $h(\theta_{M_p})$, the marginal likelihood estimate has the following form

$$\hat{f}(y|M_p) = \frac{1}{T} \sum_{t=1}^{T} \frac{f(y|M_p, \theta_{M_p}^{(t)}) \pi(\theta_{M_p}^{(t)}|M_p)}{h(\theta_{M_p}^{(t)})} \quad (3.27).$$

In some cases the importance sampling density is known up to a constant C, such that $h(\theta_p) = C\varphi(\theta_p)$. When this is the case, the constant C could also be expressed as an expectation, using the following formula

$$C = \int C\pi(\theta_{M_p}|M_p) = \int \frac{h(\theta_{M_p})}{\varphi(\theta_{M_p})} \pi(\theta_{M_p}|M_p) = E_h\left[\frac{\pi(\theta_{M_p}|M_p)}{\varphi(\theta_{M_p})}\right] \quad (3.28).$$

and the marginal likelihod's estimate is obtained by generating a random sample $\{\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \cdots, \theta^{(T)}\}$ from $h(\theta_{M_p})$ and calculating the quantity

$$\hat{f}(y|M_p) = \frac{\sum_{t=1}^{T} f(y|M_p, \theta_{M_p}^{(t)}) w(\theta_{M_p}^{(t)})}{\sum_{t=1}^{T} w(\theta_{M_p}(t))}, \quad w(\theta_{M_p}^{(t)}) = \frac{\pi(\theta_{M_p}^{(t)})}{\varphi(\theta_{M_p}^{(t)})} \quad (3.29).$$

3.4.6 Sampling from the posterior

Another alternative that derived using Monte Carlo method of estimating the marginal likelihood, is the harmonic mean estimator (Newton & Raftery ,1994). By considering Bayes' theorem,

$$p(\theta_{M_p}|y,M_p) = \frac{f(y|M_p,\theta_{M_p})\pi(\theta_{M_p}|M_p)}{f(y|M_p)} \Leftrightarrow \frac{p(\theta_{M_p}|y,M_p)}{f(y|\theta_{M_p},M_p)} = \frac{\pi(\theta_{M_p}|M_p)}{f(y|M_p)} \quad (3.30)$$

the marginal likelihood can be expressed as an expectation with respect to the posterior distribution $p(\theta_{M_p}|y,M_p)$, since

$$[f(y|M_p)]^{-1} = \int [f(y|M_p)]^{-1}\pi(\theta_{M_p}|M_p)d\theta_{M_p} = \int p(\theta_{M_p}|y,M_p)[f(y|\theta_{M_p},M_p)]^{-1}d\theta_{M_p} = E_{p(\theta_{M_p}|y,M_p)}[(f(y|\theta_{M_p},M_p))^{-1}] \quad (3.31).$$

By generating a random sample $\{\theta^{(1)},\theta^{(2)},\theta^{(3)},\cdots,\theta^{(T)}\}$ from the posterior probability $p(\theta_{M_p}|y,M_p)$, the harmonic mean estimator is defined as

$$\hat{f}(y|M_p) = \left[\frac{1}{T}\sum_{T=1}^{T}[f(y|M_p,\theta_{M_p}^{(t)})]^{-1}\right]^{-1} \quad (3.32).$$

Despite that the estimator is consistent and simple to compute, is proved to be unstable and often its variance appears to be infinite.

In order to overcome harmonic mean's instability, Newton & Raftery (94) proposed the use of a mixture of prior and posterior distribution as an importance sampling density, defined as

$$h(\theta_{M_p}) = w\pi(\theta_{M_p}|M_p)+(1-w)p(\theta_{M_p}|y,M_p) \ , \ 0<w<1 \quad (3.33).$$

The corresponding estimator of the marginal density is

$$\hat{f}(y|M_p) = \frac{\sum_{T=1}^{T}[f(y|M_p,\theta_{M_p}^{(t)})]/h(\theta_{M_p}^{(t)})}{\sum_{t=1}^{T}h(\theta_{M_p}^{(t)})} \quad (3.34).$$

Finally Gelfand & Dey (1994) derived the generalized harmonic mean estimator, an unbiased and consistent estimator of the marginal likelihood (Kass & Raftery 95 , Chib 95). Again, by using the Bayes theorem, the following formula is obtained

$$p(\theta_{M_p}|y,M_p) = \frac{f(y|M_p,\theta_{M_p})\pi(\theta_{M_p}|M_p)}{f(y|M_p)} \Leftrightarrow \frac{p(\theta_{M_p}|y,M_p)}{f(y|\theta_{M_p},M_p)\pi(\theta_{M_p}|M_p)} = \frac{1}{f(y|M_p)} \quad (3.35).$$

By choosing an importance sampling density $h(\theta_{M_p})$, such that

$$\frac{1}{f(y|M_p)} = \int \frac{1}{f(y|M_p)} h(\theta_{M_p}) d\theta_{M_p} = \int \frac{p(\theta_{M_p}|y, M_p)}{f(y|\theta_{M_p}, M_p)\pi(\theta_{M_p}|M_p)} h(\theta_{M_p}) d\theta_{M_p} = E_{p(\theta_{M_p}|y, M_p)}\left[\frac{h(\theta_{M_p})}{f(y|\theta_{M_p}, M_p)\pi(\theta_{M_p}|M_p)}\right] \tag{3.36}$$

the generalized harmonic mean estimator is defined as

$$\hat{f}(y|M_p) = \left[\frac{1}{T}\sum_{t=1}^{T} \frac{h(\theta_{M_p}^{(t)})}{f(y|\theta_{M_p}^{(t)})\pi(\theta_{M_p}^{(t)}|M_p)}\right]^{-1} \tag{3.37}.$$

Again the performance of the method depends on the proper choice of sampling density. As pointed out in Kass & Raftery (1995) and Chib (1995), for high dimensional problems, $h(\theta_{M_p})$ is difficult to determine, while in low-dimensional problems a proper choice of $h(\theta_{M_p})$ provide satisfying results. Especially in cases where $h(\theta_{M_p})$ is proportional to the likelihood, the method provides also efficient estimators.

3.4.7 The Chib's estimator

Finally, in high-dimensional problems the use of MCMC methods is claimed to be the most promising. MCMC algorithms provide useful tools for simulating from a multivariate density. The most popular are the Gibbs sampler (Geman & Geman 1984) and the Metropolis-Hastings algorithm (Metropolis et al. 1953, Hastings 1970).

Using the Gibbs' sampler, one can simulate from a p-dimensional distribution $f(\theta_p|y)$, by simply generating values from its p conditional distributions, $f(\theta_i|\theta_j, y)$, $i \neq j$. The algorithm is of the following form :

- Set a vector of initial values $\theta_p^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \cdots, \theta_p^{(0)})$
- Generate $\theta_1^{(1)} \sim f(\theta_1|y, \theta_2^{(0)}, \theta_3^{(0)}, \cdots, \theta_p^{(0)})$

  $\theta_2^{(1)} \sim f(\theta_2|y, \theta_1^{(1)}, \theta_3^{(0)}, \cdots, \theta_p^{(0)})$

  $\vdots$

  $\theta_p^{(1)} \sim f(\theta_p|y, \theta_1^{(1)}, \theta_2^{(1)}, \cdots, \theta_{p-1}^{(1)})$
- at the k^{th} iteration

  Generate $\theta_j^{(k)} \sim f(\theta_j|y, \theta_1^{(k)}, \theta_2^{(k)}, \cdots, \theta_{j-1}^{(k)}, \theta_{j+1}^{(k-1)}, \cdots, \theta_p^{(k-1)})$ etc.

Then, after a few iterations, the simulated vectors $\theta^{(t)}, \theta^{(t+1)}, \cdots, \theta^{(T)}$, will be a sample from the multivariate density $f(\theta_p|y)$. At a chosen point θ', the multivariate density is approximated by

$$\hat{f}(\theta'|y) = \frac{1}{T}\sum_{t=1}^{T} f(\theta_1'|y, \theta_2^{(t)}, \cdots, \theta_p^{(t)}) f(\theta_2'|y, \theta_1', \theta_3^{(t)}, \cdots, \theta_p^{(t)}), \cdots, f(\theta_p'|y, \theta_1', \theta_2', \cdots, \theta_{p-1}') \tag{2.38}.$$

In order to apply the method for calculating the marginal likelihood at the point of interest $\theta^{'}$, one should generate a sample from the posterior density $p(\theta_{M_p}|y, M_p)$ and evaluate it at the following formula, using (2.38)

$$\hat{f}(y|M_P) = \frac{f(y|\theta^{'}_{M_p})\pi(\theta^{'}_{M_p}|M_p)}{\hat{p}(\theta^{'}_{M_p}|y, M_p)} \tag{2.39}$$

A slightly different approach is proposed by Chib (1995). The general idea remains the same, yet, in order to approximate posterior density at a chosen point $\theta^{'}$, , he uses the following equation

$$p(\theta^{'}_{M_p}|y, M_p) = p(\theta^{'}_1|y, M_p)\, p(\theta^{'}_2|\theta^{'}_1, y, M_p), \cdots, p(\theta^{'}_p|\theta^{'}_1, \theta^{'}_2, \cdots, \theta^{'}_{p-1}, y, M_p) \tag{2.40}$$

Each conditional density can be estimated using the following updating scheme

$$\begin{aligned}
\hat{p}(\theta^{'}_1|y, M_p) &= \frac{1}{T}\sum_{t=1}^{T} p(\theta^{'}_1/\theta^{(t)}_2, \theta^{(t)}_3, \cdots, \theta^{(t)}_p, y, M_p) \\
\hat{p}(\theta^{'}_2|\theta^{'}_1, y, M_p) &= \frac{1}{T}\sum_{t=1}^{T} p(\theta^{'}_2|\theta^{'}_1, \theta^{(t)}_3, \cdots, \theta^{(t)}_p, y, M_p) \\
\hat{p}(\theta^{'}_3|\theta^{'}_1, \theta^{'}_2, y, M_p) &= \frac{1}{T}\sum_{t=1}^{T} p(\theta^{'}_3|\theta^{'}_1, \theta^{'}_2, \theta^{(t)}_4, \cdots, \theta^{(t)}_p, y, M_p) \\
&\vdots \\
\hat{p}(\theta^{'}_p|\theta^{'}_1, \theta^{'}_2, \cdots, \theta^{'}_{p-1}, y, M_p) &= \frac{1}{T}\sum_{t=1}^{T} p(\theta^{'}_1, \theta^{'}_2, \theta^{'}_2, \cdots, \theta^{'}_{p-1}, y, M_p)
\end{aligned} \tag{2.41}$$

A presentation of the above two methodologies, discussion on their properties and applications on real data sets are provided by Yu & Tunner (1999), while Chib & Jeliazkov (2001) extended the same method, using the Metropolis-Hastings algorithm. There have been also introduced estimators that are based on different sampling schemes, namely the bridge sampling estimator (Meng & Wong, 1996) and the path sampling estimator (Gelman & Meng, 1998).

3.5 Discussion

Except from BIC, the methods presented above for approximating the marginal likelihood, are a few alternatives that have been developed in order to deal with the so called integration problem. The adequate approximation of such integrals, as the one needed for the calculation of the marginal likelihood, is of major importance, especially when Bayesian analysis is applied. Evans & Swartz (1995) have summarized and categorized the existing methods in the five following groups :

1. Asymptotic methods
2. Importance sampling
3. Adaptive importance sampling

4. Multiple quadrature and
5. Markov Chain methods

In their paper each method is presented and evaluated through several examples.

Both Evans & Swartz (1995) and Kass & Raftery (1995) suggest the use of asymptotic methods as more accurate and efficient, but only in cases where the integrated function is obviously unimodal. Otherwise, importance sampling should be preferred, even though it appears to be computationally more demanding and less accurate. Yet, in order to be reliable, the importance sampling density must be chosen carefully. An accurate, effective and strongly recommended alternative, is provided by quadrature methods and especially by the subregion adaptive integration (Genz & Kass, 1993), which was not discussed here. However, it is preferable only for low-diamentional problems $p \leq 8$ (Kass & Raftery, 1995). In high dimensional problems, MCMC based methods as Chib's estimator Chib & Jeliaskov's extension (2001) or Chen's extension (2005), that deals with the case when latent variables are present, retain their usefulness and popularity. However, all methods discussed in this Chapter require the estimation of all marginal likelihoods of models under comparison, which in real life problems can be computationally prohibitive. In following Chapter, we will deal with the Bayesian algorithms that have been developed, in order to explore model space and try to uncover efficiently candidate models of higher posterior probabilities Further details and review of the methods of approximating BF can be found in Gamerman & Lopes (2006) and in Ntzoufras (2011, pg. 392-397)

Chapter 4 : Bayesian Variable Selection

4.1 Introduction

The third chapter of this dissertation dealt with the problem of Bayesian model comparison. Inference based on posterior odds and Bayes Factors, requires the computation of the the posterior model probabilities $p(M_p|y)$ of all candidate models $M_p \in M$ and their corresponding marginal likelihood $f(y|M_p)$. The main problems that one may face when using the above tools, concerns the intractability of the integrals needed for the computation of $f(y|M_p)$ or the behavior of posterior odds under the use of certain kind of priors (Lindley's Paradox). Despite the fact that several methods have been developed to overcome the above mentioned problems, the choice of the most promising model requires the evaluation of all candidates. As mentioned in the first chapter, a thorough search might be time consuming especially in cases where many regressors are present. Hence, similarly to stepwise algorithms, several Bayesian algorithms have been proposed which efficiently explore large model spaces, focusing on the most probable posterior models (Ntzoufras 2009, p. 405).

Bayesian variable selection algorithms were, initially introduced by George and McCullagh (1993). Using an indicator variable to identify the candidate subsets and a hierarchical structure of the regression model, which will be presented in the next paragraph, they developed , the Stochastic Search Variable Selection (SSVS), a general Gibbs based algorithm, to sample from models with highest a posteriori probabilities, avoiding the exhaustive evaluation of all 2^p candidates (George and McCullach, 1993). Apart from SSVS, other methods that have been developed and will be discussed in the first part of chapter 3, exploiting the idea of George and McCullogh, is the Kuo and Mallick sampler (1998) and the Gibbs Variable Selection (GVS), introduced by Dellaportas et al (2002).

The second group of algorithms that will be described, were proposed as an extension and generalization of Gibbs - based algorithms, in order to overcome convergence issues arose when using Gibbs - based algorithms. Main representatives of this group are the Carlin Chib method (1995) and the Reversible Jump MCMC algorithm (Green, 1995).

Finally, some latest advances will be briefly reviewed, including Population-based Reversible Jump MCMC (Jasra et al, 2007), Shotgun Stochastic Search (Hans et al, 2007) and Subspace Carlin and Chib algorithm (Petralias and Dellaportas, 2012).

4.2 Bayesian Variable Selection For Normal and GLMs : Initial Concepts

4.2.1 Model structure

Consider the linear regression formula

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + e_i \quad , \quad e_i \sim N_n(0, \sigma^2 I) \quad , \tag{4.1}$$

It assumes a linear relationship between a random variable y, with y_i , $i = 1,\dots,N$ outcomes and p independent variables, taking values $x_{i,j} = 1,\dots,p$, given a vector of parameters $\beta_j = 1,\dots,p$. Parameter β_0 is the constant of the model and e_i is the error term. By assuming that y is generated from a distribution that is a member of the exponential family, we obtain the generalized linear model, introduced by McCullach and Nelder (1989) and is given by the following formula

$$E[g(y_i)] = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} \quad , \tag{4.2}$$

The extension of the linear regression formula, enables us to model both discrete and continuous data, including Poisson or Binomial data. In order to do so, the link function $g(.)$ is introduced in the model and is used to combine the stochastic part of (4.2) with the the systematic part $\beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}$. For instance, in case of Poisson data, the logarithm of the expected value of y, is used as a link function, while when modeling binomial data the logarithm of odds is used.

An alternative way to represent formulas (3.1) and (3.2), is by introducing a binary indicator variable γ_j, that takes two possible values $\{0,1\}$. By doing so, the presence, when $\gamma_j = 1$, or absence, when $\gamma_j = 0$, of a candidate variable could be controlled. Depending on how γ_j is treated, the above formulas can be formed in two different ways.

George and McCullogh (1993) introduced γ_j without embedding it in the linear predictor (4.2). The parameter vector consists only of the effects of the covariates, i.e $\theta = (\theta_1, \theta_{2,\dots}, \theta_p)' = (\beta_1, \beta_{2,\dots}, \beta_p)'$ and the regression formula remains as in (3.2). In that case, the indicator gets involved in the model through the following hierarchical structure

$$\begin{aligned} y|\beta &\sim f(y|\beta) \\ \beta|\gamma &\sim \pi(\beta|\gamma) \\ \gamma &\sim \pi(\gamma) \end{aligned} \tag{4.3}$$

where $f(y|\beta)$ is the likelihood function, while $\pi(\beta|\gamma)$ and $\pi(\gamma)$ indicate the priors of

parameter vector and indicator variable respectively.

Conversely, by defining as β_j the effect of the j^{th} covariate, the indicator could be embedded in 3.2 (or 3.1), yielding the expanded regression formula (Kuo and Mallick, 1998), used for the Kuo and Mallick sampler and GVS. Then, the parameter vector takes the following form $\theta = (\theta_1, \theta_2, \ldots, \theta_p)' = (\gamma_1\beta_1, \gamma_2\beta_2, \ldots, \gamma_p\beta_p)'$ and equation 3.2 can be rewritten as

$$E[g(y_i)] = \eta_i = \sum_{j=0}^{p} \gamma_j \beta_j x_{i,j} \tag{4.4}$$

4.2.2 The Gibbs algorithm

Regardless of the model structure, both type of algorithms proceed by using a Gibbs sampler to obtain values from the joint posterior distribution $p(\gamma, \beta|y)$. The main interest lies in the simulation of the posterior $p(\gamma_j=1|y)$, , which is called marginal inclusion probability. Full specification of marginal inclusion probabilities, enables us to identify those variables with higher posterior probabilities (Kuo and Mallick, 1998). The produced sequence $\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(T)}$, as pointed out in George and McCullogh (1993), converges rapidly in the target distribution and contains all the information needed for variable selection. This occurs due to the fact that candidate models with higher posterior probabilities (those for which $\gamma_j=1)$, appear in the simulated sample with high frequencies. On the contrary, those candidate models that hardly appear, are simply not of interest and can be excluded from the analysis. Hence, the method requires to generate a sequence

$$\gamma^{(1)}, \beta^{(1)}, \gamma^{(2)}, \beta^{(2)}, \ldots, \gamma^{(T)}, \beta^{(T)}$$

from the full conditional posterior distributions of γ and β iteratively and then identify the most promising candidate models, by counting the frequency of their appearance (Kuo and Mallick, 1998). By denoting as β_{-j} all the effects except the one associated to covariate $x_{i,j} = 1,\ldots, p$, the general steps of Gibbs algorithm that are used to produce values from the full posterior $p(\beta, \gamma|y)$ are :

- Update β_j for $j=1,\ldots, p$ from full conditional posterior $p(\beta_j|y, \beta_{-j}, \gamma)$, where $\beta_{-j}=(\beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$
- Update γ_j for $j=1,\ldots, p$, sequentially or in random order, from full conditional posterior $p(\gamma_j|y, \gamma_{-j}, \beta)$, where $\gamma_{-j}=(\gamma_1, \gamma_2, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p)$, with probability .

4.2.3 Posterior Inference

Once the algorithm is terminated, the results can be used in order to draw inference on the model with the highest estimated posterior probability. In terms of Bayesian analysis, this is referred to, as the maximum a posteriori probability (MAP) model. Following Ntzoufras (2009, sec. 11.6), a common way to estimate a MAP model is to use the indicator variable γ, in order to index existing candidate models. Let m be an indicator variable that is to be used to index all candidate models. A formula that provides with a one-to-one transformation between γ and m is given by

$$m(\gamma) = \sum_{j=k}^{p} \gamma_j 2^{j-k}, \tag{4.5}$$

where $k=1$ in case of including the constant in the model and $k=0$ otherwise. Assuming that the algorithm has been implemented for T iterations, let B be the number of burn-in iterations and denote as $m^{(t)}$ the indicator of model m at iteration t. Then, the corresponding posterior probability of each model can estimated in a straightforward manner, by

$$\hat{p}(m|y) = \frac{1}{T-B} \sum_{t=B+1}^{T} I(m^{(t)}=m) \tag{4.6}.$$

As described in chapter 3, posterior model probabilities, can also be used to estimate posterior model odds or Bayes Factors and draw inference through them, by comparing for instance, each model to the one with the highest estimated model posterior probability.

It is also possible to estimate the marginal inclusion posterior probabilities for every candidate by

$$\hat{p}(\gamma_j=1|y) = \frac{1}{T-B} \sum_{t=B+1}^{T} I(\gamma_j^{(t)}=1) \tag{4.7}.$$

Apart from tracing the desired MAP model using (4.6), the MCMC output enables us to estimate the Median Probability (MP) model via (4.7). The latter, was introduced by Barbieri and Berger (2004) as an optimal choice of model with better predictive performance under certain conditions. According to this approach, MP model is defined as the model which consists of those candidates whose marginal inclusion probability is greater than $\frac{1}{2}$, that is $\gamma^{MP}=\gamma_j : p(\gamma_j=1|y) > \frac{1}{2}$.

Before proceeding in a more detailed description of the developed algorithms, there is one estimation issue that should be further pay attention. As described in section 3.2.2, the number of visits of candidate regressors models differs, according to the target posterior model probabilities. Consequently, the estimates' accuracy of posterior model probabilities, posterior model odds and Bayes Factors, increases with the number of times a variable is sampled, during the algorithm. In

order to ensure that all candidates are sampled in a sufficient number of times, Ntzoufras (2009, p. 413) provides with two alternative strategies. The first approach, requires a proper choice of prior distribution over γ, favoring candidates with lower probabilities of sampling; for details and illustration in a simple setup see Ntzoufras et al, (2005). The second one, introduced by Fouskakis et al. (2009), can also be implemented in data sets consisting with large p, compared to the sample size. When this is the case, he suggests a reduction in the number of candidate models by excluding from the analysis those ones with low estimated marginal inclusion probabilities.

4.3 Variable Selection Methods

Despite the fact that the general idea when using Gibbs sampler remains the same, there are differences between the methods mentioned above. Apart from the model scheme, the assumptions made regarding the relationship between γ and β and the prior specification of (γ, β) has led in the development of different methods. Following O Hara and Sillanpaa (2009), Bayesian variable selection methods can be categorized in four groups :

- Stochastic Search Variable Selection (George and McCullogh, 1993)
- Indicator Model Selection
- Adaptive Shrinkage
- Model Space Search

4.3.1 Stochastic Search Variable Selection (SSVS)

To begin with, SSVS was introduced by George and McCullogh (1993), as a Bayesian variable selection procedure, used to identify the most promising models between all 2^p candidates. As mentioned above, in order to avoid the exhaustive evaluation of models' complete posterior distribution, they developed a Gibbs based algorithm to sample from models with higher posterior probabilities. To do so, they considered the hierarchical structure of the regression equation (4.4), introduced the auxiliary variable γ, to indicate whether a variable is included or excluded from the likelihood and assigned such a prior distribution on model effects β, to keep the parameter space for all models constant. The latter is the main advantage of the particular method, since it ensures the convergence of the algorithm.

4.3.1.1 Prior specification

To complete the full specification of Bayesian variable selection via SSVS, the specification of the priors $\pi(\beta|\gamma)$ and $\pi(\gamma)$, is required. For the first prior, the method assumes a 'slab and spike'-type prior on each parameter β_j. In brief, the 'slab and spike' prior, which was introduced by Mitchel and Beauchamp (1988), is a mixture of priors on each β_j. which puts mass on zero with probability h_{0j} (spike), while is uniformly distributed over the range $(-f_j, f_j)$, for some large f_j, with density $\frac{1-h_{0j}}{2f_j}$ (slab) (Miller, 2002).

Exploiting Mitchel and Beauchamp's (1988) idea, George and McCullogh (1993) proposed the following mixture of Normal distributions

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(0,\tau_j^2)+\gamma_j N(0,g_j^2\tau_j^2) \quad , \tag{4.8}$$

which does not set unimportant regressors exactly equal to zero, but forces them to be close to it. This can be achieved by setting $\tau_j>0$ small, so that if $\gamma_j=0$, the effect of the candidate variable β_j to be safely estimated by zero and $g_j>1$ large, so that if $\gamma_j=1, \beta_j$ is non zero and therefore, the regressor is candidate for inclusion (Dellaportas et al, 2002). The use of such a prior does not actually drop variables out of the model, but shrinks them toward zero and, by this way, keeps the size of the equation constant in every step of the algorithm, in order to ensure the convergence of Gibbs algorithm. (Carlin and Chib, 1995, p. 475)

As far as $\pi(\gamma)$ is concerned, George and McCullogh propose alternatives to depict the prior knowledge on the indicator γ For instance, by assuming each γ_j as independent Bernoulli trials with probability p_j, then

$$\pi(\gamma)=\prod_j p_j^{\gamma_j}(1-p_j)^{1-\gamma_j} \quad , \tag{4.9}$$

implying that the addition of each variable does not depend on the inclusion of another. A special case of (4.9), is obtained if each γ_j is $Bernoulli\left(\frac{1}{2}\right)$ distributed. Then, the corresponding prior is

$$\pi(\gamma)=\frac{1}{2^p} \quad , \tag{4.10}$$

implying prior ignorance concerning the inclusion of each variable.

4.3.1.2 The SSVS algorithm and derivation of conditional posterior distributions

After defining the priors involved in the model's construction, the conditional posterior distributions must be calculated, in order to implement the Gibbs sampler in the SSVS approach. The full conditional posterior of the parameters can be simply derived from

$$p(\beta_j|y,\gamma,\beta_{-j}) \propto f(y|\beta,\gamma)\pi(\beta_j|\gamma_j). \tag{4.11}$$

Taking into account mixture of Normals as specified in (3.8), then (3.11) can be rewritten as

$$\begin{array}{ll} p(\beta_j|y,\gamma,\beta_{-j}) \propto f(y|\beta)N(0,g_j^2\tau_j^2), & for \quad \gamma=1 \\ p(\beta_j|y,\gamma,\beta_{-j}) \propto f(y|\beta)N(0,\tau_j^2), & for \quad \gamma=0 \end{array} \tag{4.12}$$

By considering, now, the full conditional posterior $p(\gamma_j|y,\gamma_{-j},\beta)$, the hierarchical structure of the model given in (3.4), implies independence between γ and y and therefore, the likelihood $f(y|\beta,\gamma)$ does not get involved in the computation (George and McCullogh, 1993). The corresponding posterior depends only on priors and is again Bernoulli, with success probability

$$p(y;\gamma,\beta) \sim Bernoulli(p_{post}), \quad with$$

$$p_{post} = p(\gamma_j=1|\beta,\gamma_{-j}) = \frac{a_j}{a_j+b_j} = \frac{a_j/b_j}{(a_j/b_j)+1} = \frac{O_j}{O_j+1}, \tag{4.13}$$

$$O_j = \frac{\pi(\beta|\gamma_j=1,\gamma_{-j})}{\pi(\beta|\gamma_j=0,\gamma_{-j})} \frac{\pi(\gamma_j=1|\gamma_{-j})}{\pi(\gamma_j=0|\gamma_{-j})}$$

(Ntzoufras, 2009,p. 411)

Then, the Gibbs sampler can be applied as described in paragraph 4.2.2 by iteratively producing values from (4.9) and (4.10).

4.3.1.3 Discussion

The effectiveness of the algorithm strongly depends on the parameters of $\pi(\beta|\gamma)$ that must be specified and this, can be considered as the main disadvantage of the method (O' Hara and Sillanpaa, 2009). It must be noted that in case of linear regression, the method gets further complicated, since it requires the specification of an additional prior on σ^2. In that case the Gibbs algorithm requires an intermediate step of updating from the corresponding posterior $p(\sigma^2|y,\beta,\gamma)$; for details see George and McCullogh (1993), where it can be found an extensive discussion on algorithm's convergence issues and suggestions through which the method gets simplified. They also provide details on how the parameters should be tuned to obtain a sufficiently well behaved SSVS algorithm.

However, the algorithm can be considered rather simple and hence, the method has been adopted in many applications. Extensions of the method have been proposed for GLM models (George and McCullogh, 1996 ; 1997), Poisson log-linear models (Ntzoufras et al, 2000), Multivariate regression (Brown et al, 1997), genetics applications (Oh et al, 2003; Yi et al, 2003), implementions using BUGS; see Ntzoufras (2009, sec. 11.7) and factor analytics models (Mavridis & Nzoufras, 2014).

4.3.2 Indicator variable selection

As discussed in section 4.2.1, an alternative way to use the indicator variable in Bayesian variable selection, is to embed γ directly in the likelihood equation as in (4.3). The two methods developed by this approach (KM sampler and GVS), do not only differ from SSVS in model structure (the likelihood depends on the indicator), but also in prior specification of $\beta_j|\gamma_j$. The spike part of the prior is centered exactly on zero, while the slab part is Normally distributed around a pre-specified value β_{0j}, representing the prior belief on each variable; for discussion on the choice of priors see for example Ntzoufras (2009). The difference between KM sampler and GVS, lies in how $\pi(\beta,\gamma)$ is specified.

4.3.2.1 Kuo and Mallick sampler

The simplest way to define the prior $\pi(\beta,\gamma)$, was suggested by Kuo and Mallick (1998) and assumes independence between β and γ. Then it follows that

$$\pi(\beta,\gamma)\propto\pi(\beta)\pi(\gamma) \tag{4.14}$$

By also assuming the partition of β in $(\beta_\gamma, \beta_{-\gamma})$ as in Ntzoufras (2009), then, the parameter vector is divided in two parts: the active part, which consists of those β that are included in the equation, noted as β_γ (those for which $\gamma=1$) and the remaining part of the vector, which consists of the variables excluded from the model, noted as $\beta_{-\gamma}$ (those for which $\gamma=0$). Then, the prior $\pi(\beta,\gamma)$ is defined

$$\pi(\beta,\gamma)\propto\pi(\beta_\gamma|\beta_{-\gamma})\pi(\beta_{-\gamma})\pi(\gamma) \tag{4.15}$$

and the corresponding posterior required for the Gibbs algorithm is of the following form

$$\begin{array}{ll} p(\beta_j|y,\gamma,\beta_{-j})\propto f(y|\gamma,\beta)\pi(\beta_j|\beta_{-j}), & for\ \ \gamma=1 \\ p(\beta_j|y,\gamma,\beta_{-j})\propto\pi(\beta_j|\beta_{-j}), & for\ \ \gamma=0. \end{array} \tag{4.16}$$

The presence of the above equation implies that when updating the parameter vector β in

the Gibbs algorithm, in case of the presence of a variable, the produced values depend on the likelihood and the posterior derives as usual. Conversely, when the effect of a variable is constrained to zero, the algorithm proposes values from a linking density, which depends only on the conditional prior $\pi(\beta_j|\beta_{-j})$. As pointed out in Kuo and Mallick (1998), this is reasonable, because in the absence of the variable, all information needed can be provided only from the prior and not from the data. This prior can be characterized as a pseudoprior, a term introduced by Carlin and Chib (1995) and was used for this purpose by Dellaportas et al (2002).

As far as $p(\gamma_j|y,\gamma_{-j},\beta)$ is concerned, Kuo and Mallick, using similar arguments for the prior $\pi(\gamma)$ specification as in George and McCullogh (1993), derived a Bernoulli full conditional posterior, with success probability given by

$$p_{post}=p(\gamma_j=1|\beta,\gamma_{-j})=\frac{O_j}{O_j+1},$$
$$O_j=\frac{f(y|\gamma_j=1,\gamma_{-j},\beta)}{f(y|\gamma_j=0,\gamma_{-j},\beta)}\frac{\pi(\gamma_j=1,\gamma_{-j})}{\pi(\gamma_j=0,\gamma_{-j})} \quad (4.17).$$

As it can be seen, the posterior probability of the indicator, depends on the likelihood, since it is embedded in the model, however it is independent of the parameters' prior due to prior independence assumption.

KM sampler is simple to apply, avoids the exhaustive evaluation of all candidate models and, unlike SSVS, requires only the specification of the prior on the parameter vector and the indicator variable (Kuo and Mallick, 1998). However, as stated in Dellaportas et al (2000, 2002), the fact that the conditional prior $\pi(\beta_j|\beta_{-j})$ derives directly from the prior of β, may be considered as a disadvantage, since this restriction may cause inefficiency of the method, due to 'bad' behaved pseudopriors.

4.3.2.2 Gibbs Variable Selection (GVS)

GVS is the second method that uses the indicator variable as part of the model equation and was introduced by Ntzoufras (1999) and Dellaportas et (2000, 2002), extending the idea of Carlin and Chib (1995). In GVS the prior is formed in the following way

$$\pi(\beta,\gamma)\propto\pi(\beta_\gamma|\gamma)\pi(\beta_{-\gamma}|\beta_\gamma,\gamma)\pi(\gamma) \quad , \quad (4.18)$$

where the intermediate term $\pi(\beta_{-\gamma}|\beta_\gamma,\gamma)$ is a pseudoprior which does not affect the posterior of $p(\beta_\gamma|y,\gamma)$, since $\beta_{-\gamma}$ is independent of the likelihood. This independence allows the user to specify the pseudoprior 'freely', and unlike KM sampler make the method work more efficiently.

(Ntzoufras, 2009, sec. 11.5.3)

The full conditional posterior is obtained by

$$\begin{gathered} p(\beta_\gamma|y,\beta_{-\gamma},\gamma)\propto f(y|\beta,\gamma)\pi(\beta_\gamma|\gamma)\pi(\beta_{-\gamma}|\beta_\gamma,\gamma) \\ p(\beta_{-\gamma}|y,\beta_\gamma,\gamma)\propto \pi(\beta_{-\gamma}|\beta_\gamma,\gamma) \end{gathered} \tag{4.19}$$

and the full conditional posterior of the indicator variable is obtained by

$$\begin{gathered} p_{post}=p(\gamma_j=1|\beta,\gamma_{-j})=\frac{O_j}{O_j+1}, \\ O_j=\frac{f(y|\gamma_j=1,\gamma_{-j},\beta)}{f(y|\gamma_j=0,\gamma_{-j},\beta)}\frac{\pi(\beta|\gamma=1,\gamma_{-j})}{\pi(\beta|\gamma=0,\gamma_{-j})}\frac{\pi(\gamma_j=1,\gamma_{-j})}{\pi(\gamma_j=0,\gamma_{-j})} \end{gathered} \tag{4.20}.$$

As pointed out in Ntzoufras (2009, p.409), the dependence of the full conditional posterior $p(\beta_\gamma|y,\beta_{-\gamma},\gamma)$ on the pseudoprior $\pi(\beta_{-\gamma}|\beta_\gamma,\gamma)$, can be useful when collinearity is detected between candidate variables. However, in cases of orthogonal candidates, the dependence between β_γ and $\beta_{-\gamma}$ is useless. Then, it follows that

$$\pi(\beta_{-\gamma}|\beta_\gamma,\gamma)=\pi(\beta_{-\gamma}|\gamma) \tag{4.21}$$

and the computation of the full conditional posterior gets simplified in

$$\begin{gathered} p(\beta_\gamma|y,\gamma,\beta_{-\gamma})\propto f(y|\gamma,\beta)\pi(\beta_\gamma|\gamma) \\ p(\beta_{-\gamma}|y,\gamma,\beta_\gamma)\propto \pi(\beta_{-\gamma}|\gamma) \end{gathered} \tag{4.22}$$

Ways to further simplify the method are presented in Dellaportas et al (2000, 2002) and in Ntzoufras (2009, p. 409, 410), by assuming prior conditional independence for all parameters given the model γ. As stated in Ntzoufras (2009) they are rather restrictive, however might be reasonable, for instance, when the candidate variables are centered, or standardized, or orthogonal.

4.3.2.3 Discussion

The methods described above, provide smart and efficient Gibbs based algorithms, that are used as faster alternatives for Bayesian model specification. Their difference among them, lies in the model formulation as described in section 4.2.1 and the assumptions regarding the relationship between the parameter vector and the indicator variable. Dellaportas et al., (2002) summarize these differences by commenting on how the initial assumptions affect the posterior conditional probability $p(\gamma_j|y,\gamma_{-j},\beta)$.

SSVS assumes a hierarchical structure for the model equation and therefore, $p(\gamma_j|y,\gamma_{-j},\beta)$ does not depend on the likelihood, but only on priors. The method requires the careful treatment of several tuning parameters, the specification of which strongly affects the efficiency of the algorithm. On the other hand, GVS and KM sampler, embed the indicator variable

in the model equation and $p(\gamma_j|y,\gamma_{-j},\beta)$ depends on the likelihood. For the KM sampler, the independence assumption on priors of the parameter vector and the indicator variable, implies that in the computation of $p(\gamma_j|y,\gamma_{-j},\beta)$, only the prior of γ gets involved. It is considered to be the simplest of the Gibbs based methods and the efficiency of the algorithm strongly depends on the specification of the parameter's prior. Finally, in GVS, the conditional posterior is not only affected by the likelihood and the prior of γ but also by the pseudoprior $\pi(\beta_{-\gamma}|\beta_{\gamma},\gamma)$. The use of the latter, despite the fact that improves the efficiency of the algorithm can be also considered as a drawback, since it requires careful treatment.

As stated in Dellaportas et al (2000), all methods can be easily applied using the Gibbs sampler algorithm, however they require a careful specification of priors and as pointed out in O' Hara and Sillanpaa (2009) they should not be used unwisely. Review of the methods, examples and applications using BUGS on different kinds of data are provided by Dellaportas et al (2000) and Ntzoufras (2009, chapter 11, 11.5.2, 11.5.3, 11.7)

4.3.3 Model space search

The second group of algorithms are more general and have been developed to cope with the model determination problem. They use MCMC techniques to sample directly from the joint posterior distribution $p(m,\beta_m|y)$. Under this notation, $m=1,\ldots,M$ is used to index the candidate models and each β_m represents its corresponding parameter vector. Therefore, the parameter space consists of all $(\beta_1,\ldots,\beta_M)$ vectors. By applying an MCMC algorithm, the interest lies in sampling directly from models of high posterior probability (Han and Carlin, 2001). The methods that are mainly used, are the Carlin Chib method (Carlin and Chib, 1995) and the reversible jump MCMC (Green, 1995).

4.*3.3.1 The Carlin Chib method*

The introduction of the integer valued parameter m, $m=1,2,\ldots,M$ in Bayesian model selection, was proposed by Carlin and Chib (1995) in order to overcome convergence problems that arose in Gibbs algorithm, when sampling from models of different size. Unlike SSVS which forces the dimension of the model to be fixed throughout the sampling procedure, Carlin and Chib worked with the product space of all parameter vectors and the model indicator $(m,\beta)\in M\times\prod_{m\in M}B_m$. Their algorithm samples over the defined product space, which is now constant, independently of

the size of the parameter vector. (Godsill, 2001 ; Han and Carlin, 2001)

To derive the Carlin Chib method, each model m is associated to the likelihood $f(y|\beta_m, m)$ and the corresponding prior $\pi(\beta_m|m)$. Given a model m, the data vector is allowed to depend only on its corresponding parameter vector β_m and thus, the likelihood is of the following form

$$f(y|\beta, m) = f(y|\beta_m, m) \tag{4.23}$$

By also assuming conditional independence among the data vectors for simplicity, the marginal likelihood is obtained by

$$f(y|m) = \int f(y|\beta, m)\pi(\beta|m)\,d\beta = \int f(y|m, m)\pi(\beta_m|m)\,d\beta_m \tag{4.24}$$

To completely specify the model, a pseudoprior $\pi(\beta_{m'}|m' \neq m)$ is required (Dellaportas et al, 2002); it can be formed, though, independently from the usual prior, since it does not get involved in the above computation and works as a linking density to improve the efficiency of the algorithm (Ntzoufras, 2009). Under the prementioned assumptions, the full conditional posterior of the parameter vector required for the first step of Gibbs sampler, is given by

$$\begin{array}{ll} p(\beta_{m'}|\beta_m, y, m) \propto f(y|\beta_m, m)\pi(\beta_m|m), & m' = m \\ p(m'|\beta_m, y, m) \propto \pi(\beta_{m'}|m), & m' \neq m \end{array} \tag{4.25}$$

To derive the conditional posterior of m, the usual discrete prior π_m on each model and the joint probability of y and β under the model m is required. Given the independence assumptions, the latter is obtained over the product space as

$$f(y, \beta, m) = f(y|\beta_m, m)\left\{\prod_{m\in M}\pi(\beta_m|m)\right\}\pi_m \tag{4.26}$$

Then, the posterior distribution can be generated as a discrete random variable (Dellaportas et al, 2002) using the following formula

$$p(m|\beta, y) = \frac{f(y|\beta_m, m)\left\{\prod_{m\in M}\pi(\beta_m|m)\right\}\pi_m}{\sum_{m\in M} f(y|\beta_m, m)\left\{\prod_{m\in M}\pi(\beta_m|m)\right\}\pi_m}, \quad \forall m \in M \quad . \tag{4.27}$$

Han and Carlin (2001), suggest that the prior probabilities on the models π_m, should be chosen in such a way that would facilitate the algorithm to sample from each model equally and consequently obtain more accurate estimators. By also commenting on the use of pseudopriors, they argue that the efficiency of the method depends on their proper specification and as stated in Dellaportas et al (2002) they should resemble the scheme of their corresponding posterior distribution $p(\beta_{m'}|y, m' \neq m)$. As in the indicator selection algorithms, the use of pseudopriors can be considered as the main drawback of the method. However, unlike GVS and KM sampler which require only one prior at each step of the algorithm, for the Carlin-Chib method, in order to

sample from the full condition posterior $p(m|\beta, y)$, the specification of all $\pi(\beta_m|m)$, $m \in M$ is needed. In practice, when the number of candidate models is large, the latter can be time consuming and restricts the performance and the implementation of the method (Ntzoufras, 2009).

4.3.3.2 The Metropolised Carlin Chib

In order to overcome the exhaustive use of too many pseudopriors, the Metropolised Carlin Chib method, or independence sampler, was introduced by Dellaportas et al (2002). Instead of updating from the full conditional posterior $p(m|\beta, y)$, the second step of the Gibbs algorithm is substituted by a Metropolised move from model m to m'. Analytically, the steps of the algorithm are the following

- Let the current state of the algorithm be (m, β_m), where $\beta_m \sim p(\beta_m|y, m)$.
- Propose a new model with probability $h(m, m')$.
- Generate $\beta'_{m'} \sim \pi(\beta'_{m'}|y, m \neq m')$, where $\pi(\beta_{m'}|y, m \neq m')$ a pseudoprior.
- Accept model m' with probability

$$a_{m \to m'} = min\left\{1, \frac{f(y|\beta'_{m'}, m')\pi(\beta'_{m'}|m')\pi(\beta_m|m')\pi(m')h(m', m)}{f(y|\beta_m, m)\pi(\beta_m|m)\pi(\beta_{m'}|m)\pi(m)h(m, m')}\right\}$$

Obviously the method gets simplified since it requires the use of only one pseudoprior at each run of the algorithm.

4.3.3.3 Reversible Jump MCMC (RJMCMC)

An alternative Metropolis-based algorithm that has been developed in order to generate values from the joint posterior $p(m, \beta_m|y)$ is the reversible jump MCMC (Green, 1995), which explores the parameter and model space, by allowing sampling from models of varying dimensions (Han and Carlin, 2001).

Supposing that the current state of the algorithm is (m, β_m), where β_m is the parameter vector associated to model m, of dimension $dim(\beta_m)$, the algorithm proceeds by proposing a new value $(m', \beta'_{m'})$, where $dim(\beta'_{m'})$ can be of different length from β_m. Due to this change in the length of the chain, the algorithm's convergence is ensured under the condition of reversibility and dimension matching (Hartman and Hart, 2009). In order to satisfy the above conditions, an auxiliary random variable $u \sim q(u|\beta_m, m, m')$ is introduced. The latter is associated to each candidate model m, $m \in M$, so that the dimension of (β_m, u) remains constant for all

models (dimension matching). In other words, when proposing a move from model m to m' where $dim(\beta_m) \neq dim(\beta'_{m'})$, the following equality should hold

$$dim(\beta_m) + dim(u) = dim(\beta'_{m'}) + dim(u') \quad , \tag{4.28}$$

where u, actually, consists of the random elements that needs to be added in β_m so as to match the dimension of $\beta'_{m'}$.

In addition each (β_m, u) is associated to (β'_m, u') through an invertible function g, so that

$$(\beta'_m, u') = g_{m\ m'}(\beta_m, u) \tag{4.29}$$

The latter, satisfies reversibility, which implies that the algorithm can move backwards from the proposed values to the current state.

Finally, the proposed move is accepted or not, by calculating the acceptance probability. Its calculation is similar to the usual acceptance probability of Metropolis algorithm, but is adjusted for the change in dimension by multiplying it with the Jacobian $J = \left| \frac{\partial g(\beta_m, u)}{\partial (\beta_m, u)} \right|$ (Godsill, 2001).

The algorithm as described in Han and Carlin (2001) and Dellaportas et al (2002) uses the following steps

- Let the current state be (m, β_m)
- Propose a new model m' with probability $h(m, m')$
- Generate $u \sim q(u|\beta_m, m, m')$
- Set $(\beta'_{m'}, u') = g_{m\ m'}(\beta_m, u)$
- Accept the proposed move with probability

$$a_{m \to m'} = min\left\{1, \frac{f(y|\beta_{m'}, m')\pi(\beta_{m'}|m')\pi(\beta_m|m')\pi(m')h(m', m)}{f(y|\beta_m, m)\pi(\beta_m|m)\pi(\beta_{m'}|m)\pi(m)h(m, m')} \times \left| \frac{\partial g(\beta_m, u)}{\partial (\beta_m, u)} \right| \right\}$$

RJMCMC provides a useful tool for model determination and has become one of the most widely applicable algorithms as it allows moves between models of different size in a flexible way (Dellaportas et al., 2002). More details about the algorithm are provided by Han and Carlin (2001) and variations of the method can be found in Dellaportas et al. (2002) and Ntzoufras (2009). A rather comprehensive mathematical derivation with applications on genetics is provided by Waagepetersen and Sorensen (2001), while Hartman and Hart (2009) offer a nice review with applications on econometrics. Finally, further details and the relationship between RJMCMC and Carlin Chib method can be found in Godsill (2001), who also records lots of references of the algorithm's applications.

4.3.3.4 Model Composition (MC^3)

One of the earliest and easiest Metropolis based model search algorithms, is the Markov Chain Monte Carlo Model Composition (MC^3), which requires posterior model probabilitie $p(M|y)$ (Fernandez et al., 2001). MC^3 was introduced by Madigan and York (1995) for Bayesian analysis of graphical models for discrete data and was then adopted by Raftery et al. (1997) and Fernandez et al. (2001) for linear regression (Miller, 2002). The algorithm operates over model space and searches for the most probable a posteriori candidate models by comparing them through posterior model odds. In its general form, if the current state of the chain is in model M a new model in the neighborhood of M is proposed with probability $h(M, M')$. The neighborhood of the current model includes the current model M and those candidates that are formed after adding one more variable or removing one of the existing ones (Raftery et al., 1997). Then the proposed move is accepted with probability

$$a_{M \to M'} = min\left\{1, \frac{p(M'|y)\,h(M, M')}{p(M|y)\,h(M', M)}\right\}. \tag{4.30}$$

MC^3 algorithm, as described in Miller (2002), uses an integer variable $j = 0, \ldots, p$ to index the candidate variables. For a randomly selected model M, a number j is sampled. Zero value corresponds to the existing model and when is picked, the chain remains in the current state. If number $j \in (1, \ldots, p)$ corresponds to an absent candidate, the latter is added in M Otherwise, if j corresponds to an existing variable, it is removed from the model. In both cases posterior odds between the produced model M' and the current one is calculated and is decided weather the first will be accepted.

Dellaportas et al (2002) proved that MC^3 is a special case of Metropolised Carlin & Chib algorithm, while Miller (2002) characterizes MC^3 as an extension of Efroymson's (1960) stepwise algorithm. As he states, the results that both methods produce are similar. However, occasionally, MC^3 tends to provide with more accurate results due to its stochastic nature. The main disadvantage of the method is that the posterior distribution on the model space is not always tractable. When this is the case, Fernandez et al (2001) suggests the use of Green's RJMCMC.

4.4 Latest Advances

4.4.1 Population-based Reversible Jump MCMC (Pop-RJMCMC)

As described in previous paragraphs, trying to deal with convergence issues that appear in MCMC algorithms when sampling from models of different dimensions, led to the development of several alternatives in Bayesian variable selection. RJMCMC (Green, 1995) is one of the most efficient choices. However, a second issue that arises when applying MCMC methods in variable selection problems, concerns the nature of the joint posterior distribution $p(m, \beta_m|y)$. As Brooks et al. (2002) states, in case of multi-modal target distributions, traditional MCMC methods, fail to explore adequately both within and between distribution's local maxima. Motivated by this problem, Jasra et al. (2007), adopted population based sampling methods and developed the population based Reversible Jump Markov chain Monte Carlo algorithm (Pop-RJMCMC). The key difference between traditional MCMC and population based algorithms, is the ability of the latter to sample from various number of chains - say for instance N - in parallel (Liu, 2001 chapter 11).

In brief, following Fouskakis et al (2009), the algorithm is constructed to generate $l=1,\ldots,N$ parallel auxiliary chains, in order to achieve a thorough search over model space. Practically, apart from sampling over set of parameters (β, γ), using traditional RJMCMC steps, this leads to an additional sampling step over candidate chains. This is carried out by raising each candidate chain to a power $t_l > 0$ referred to us, as the *temperature*. Assuming that at iteration (t), the algorithm's state is in chain l, sampling from chains powered on lower values of t_l, will result in larger jumps over model space and thus in visits over regions of lower probability. On the other hand, when sampling from chains powered on greater values, the algorithm will be limited to current state's neighborhood.

The efficiency of Pop - RJMCMC depends mainly on two aspects. The first one concerns the conditions of reversibility and dimension matching, as described in section 3.3.3.3. Since these conditions are covered, convergence of the algorithm is ensured and the results are valid (Jasra et al., 2007). The second one, concerns the ability of the algorithm to explore the whole model space efficiently. To achieve this, running a large number of candidate chains, is required. One way to deal with this, is by specifying a sufficiently large number of different t_l temperatures. However, as Fouskakis et al. (2009) claims, this could be computationally exhaustive. To avoid the latter, they propose the use of only two auxiliary chains. The first one is suggested to be raised to a power such that $0 < t_1 < 1$, while the other to be raised to a power of $t_2 > 1$. Then, by assigning a distribution

over t_l, $l=1,2$ temperatures are considered as random variables and at each step different values of t_l can be sampled, resulting in an efficient exploration of the model space.

Details on theoretical and practical aspects concerning Pop-RJMCM and extended discussion on every aspect of Pop-RJMCM can be found in Jasra et al. (2007), while Fouskakis et al (2009) apply the method in a health evaluation study indicating its efficiency over original RJMCMC algorithm.

4.4.2 Shotgun Stochastic Search (SSS)

An MCMC motivated algorithm, which is developed to explore model space more rapidly and aggressively in contrast to other MCMC algorithms and therefore can be more effective in problems of higher dimensions, is the Shotgun Stochastic Search (SSS), introduced by Hans et al (2007).

Model selection based on MCMC methods discussed above, aims to simulate the posterior distribution of the model space, by seeking for individual models at each step of the algorithm. One candidate is randomly selected, is evaluated and is accepted if it is of higher posterior probability in contrast to the model in the current state of the chain. Finally, by estimating the posterior probability of each model, depending on how many times each candidate was visited, one can trace the maximum a posteriori model, as described in Ntzoufras (2009, sec. 11.6). Conversely, SSS, is designed to search for regions with models of higher probability, by running multiple parallel chains at each iteration. By exploring regions and evaluating more than one models at each step, SSS, accelerates the models space search and is likely to reach the best model faster.

In brief, let p denote the number of all candidate models and γ a $p\times 1$ vector indicating the presence or the absence of the j^{th} variable if $\gamma=1$ or $\gamma=0$ respectively. Supposing that the current state of the algorithm is in model $\gamma_k, 1\le k\le p$, a neighborhood $nbd(\gamma_\kappa)$ of proposal models is defined, based on the current candidate. The neighborhood of proposals consists of three possible model sets : $\{\gamma_k^+, \gamma_k^o, \gamma_k^-\}$, where γ_k^+ is produced after adding one more variable in the current model, γ_k^o after replacing one of the existing variables and γ_k^- after removing one of the existing ones. Each proposal in this neighborhood is then evaluated in parallel, using the models' posterior probability $p(\gamma|y)$ or other models' fit criterion. Depending on the models' score a new candidate is chosen and the algorithm repeats the above steps. This procedure, actually, results in ranking a large number of candidates, until the algorithm reaches the one that sufficiently describes the data.

Hans et al (2007) discus extensively the aspects of the method, describing the steps of the algorithm, suggesting alternatives on evaluating the proposal models and comparing it with traditional MCMC methods through examples in linear and binary regression.

4.4.3 Subspace Carlin and Chib (SCC)

The last algorithm that will be briefly discussed, is the Subspace Carlin Chib algorithm (SCC), which has been developed by Petralias and Dellaportas (2012) as a combination of the Carlin Chib, the Metropolised Carlin Chib and the Shotgun Stochastic Search algorithm. As discussed in paragraph 3.3.3.2, in order to avoid the exhaustive calculation of all possible pseudopriors $\pi(\beta_m|m)$, $m\in M$, MCC replaces Gibbs sampling over candidate models with a metropolised step, by proposing a new model with probability $h(m,m')$. SCC adopts the idea of SSS's sampling on neighborhoods and the proposed move form model m to model m' is restricted, in the sense that it allows jumps between models that are formed by deleting, replacing or adding one of the existing variables in the current model. Analytically, considering that the current state of the algorithm is in model m, then neighborhood of candidate models for evaluation in the next step of the algorithm is defined to be $S_{m'}=\{S^{-}_{m'}, S^{o}_{m'}, S^{+}_{m'}\}$. The algorithm samples at random a neighborhood of models with probability $Q_{m'}=\{q^{-}_{m'}, q^{o}_{m'}, q^{+}_{m'}\}$ and proceeds by sampling a model in the sampled neighborhood with probability $h(m,m')$.

4.5 Discussion

In this chapter the basic algorithms used for Bayesian model determination have been presented. The first part describes methods for variable selection, namely SSVS, KM Sampler and GVS, while the second part discusses algorithms that directly sample from model space. A review with applications on variable selection strategies using Gibbs sampler is provided by Dellaportas et al (2000). A comparative review on model selection algorithms is provided by Han and Carlin (2001) and their relationship is examined in Godsill (2001). Ntzoufras (2009) reviews all discussed methods, apart from those reviewed in paragraph 3.4, and provides examples using BUGS. Details on Pop RJMCMC can be found in Jasra et al (2007) and Fouskakis et al (2009). SSS, is extensively discussed in Hans et al (2007) and a combination of SSS and Carlin Chib method is provided by Petralias and Dellaportas (2012).

Chapter 5: Bayesian Adaptive Sampling for Variable Selection and Model Averaging

5.1 Introduction

As discussed in Chapter 3 MCMC techniques provide an important and easy to use computational tool, especially in complicated statistical problems, i.e in high dimensional problems or in computations of analytically intractable posterior distributions. In order to efficiently sample from large model spaces and draw inference regarding higher posterior models, the Gibbs sampler and the Metropolis Hastings algorithm were adopted, based on which, several algorithms have been developed.

Despite the fact that MCMC based model selection algorithms resulted in overcoming major computational problems in Bayesian variable selection, their efficiency is not guaranteed, since it strongly depends on the careful specification of their related proposal distributions and the specification of their corresponding tuning parameters. Mistreating them may result in slow mixing of the chains or worse, failure of algorithms' convergence and hence inefficient estimation of the desired posterior distributions. Gibbs based algorithms can be highly affected either by careless specification of the prior distributions or by highly correlated data. In the first case, improper specification of a prior distribution over candidate models, may prevent the algorithm from sampling equally from all models. Furthermore, a bad choice of pseudopriors may restrict the algorithm to 'local' moves, meaning that the algorithm can be 'trapped' in areas of lower dimensions compared to the proposed ones (Dellaportas et al. 2002). The shape and size of the proposal distribution of the Metropolis-Hastings algorithm, also plays a key role on the efficiency of such algorithms. Heavy tailed proposals, decreases the number of the accepted points forcing the algorithm to stand still in specific areas of the target distribution. On the other hand, using a proposal that increases the acceptance rate of the algorithm leads in small jumps and hence full exploration of the target distribution is very slow (Haario et al, 1999, Pasarica and Gelman 2003).

Selecting an efficient proposal distribution within an MCMC method can be proved a hard and time consuming task for the researcher. During the last decade, adaptive modifications of traditional Markov Chain Monte Carlo (AMCMC) schemes have been proposed, as a way to accelerate and improve the efficiency of the Gibbs sampler and the M-H algorithm (Gilks et al., 1998, Haario et al., 2001, Atchade and Rosenthal, 2003, Pasarica and Gelman, 2003, Andrieu and Thoms, 2008, Roberts and Rosenthal, 2009). The main idea behind AMCMC alternatives is to take advantage of the algorithm's history and let the proposal distribution learn from it. In other words, as the algorithm proceeds, past sampled values are used in order to modify, mainly, the proposal distribution and hence automatically tune it during the simulation, aiming in faster convergence and more efficient estimation of posterior quantities (Ji and Schmidler, 2009).

Applications of adaptive methods have also been presented in Bayesian inference (Tierney and Mira, 1999), in variable selection and model averaging (Nott and Kohn, 2005, Clyde et al., 2011, Lamnisos et al., 2012). This chapter focuses on the Bayesian Adaptive Sampling (BAS) algorithm. The algorithm was recently developed algorithm by Clyde et al (2011) and exploits the idea of adaptation. It can be applied in linear regression where marginal likelihood is analytically available or easy to estimate. As described in their paper, BAS algorithm, samples models without replacement from the model space. In cases of moderate number of candidate models, $p \approx 30$, BAS fully explores model space in 2^p iterations, while it provides perfect samples without replacement when the number of candidate models is large to handle. Sampling is based on marginal inclusion probabilities which are adaptively calculated as the algorithm proceeds, in order to avoid re-sampling a model that was visited in a previous step. Clyde et al (2011) focus on variable selection problems in linear regression, adopting the Normal-Gamma prior family. For its computational simplicity, they adopt Zellner's g-prior (Zellner, 1986) and their extensions; the Zellner-Siow Cauchy prior (Zellner and Siow, 1980) and the hyper g-prior (Liang et al., 2008). Hence, before presenting BAS algorithm, there will be a brief representation regarding main results for the Normal-Gamma formulation in Bayesian variable selection problem and a more detailed review of the Zellner's family of priors.

5.2 Conjugate Analysis for Linear Regression models

Recall the linear regression formula, as described in paragraph 3.2.1, equation 3.1.

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + e_i \quad , \quad e_i \sim N_n(0, \sigma^2 I)$$

and the p-diamensional vector $\gamma \in \{0,1\}^p$ used to index the candidate variables that are included in the model. In matrix form, the linear regression formula under model M_γ can be rewritten as

$$Y|\beta_\gamma, \sigma^2, M_\gamma \sim N_n(X_\gamma \beta_\gamma, \sigma^2 I_n) \tag{5.1}$$

Under this notation, the likelihood takes the following form:

$$f(Y|\beta_\gamma, \sigma^2, M_\gamma) = \exp\left\{-\frac{1}{2\sigma^2}(X_\gamma \beta_\gamma)'(X_\gamma \beta_\gamma)\right\} \tag{5.2}$$

The maximum likelihood estimators of a model's unknown parameters can be computed as

$$\begin{aligned} \hat{\beta}_\gamma &= (X_\gamma' X_\gamma)^{-1} X_\gamma y, \\ \hat{\sigma}^2 &= \frac{1}{n}(y - X_\gamma \hat{\beta}_\gamma)'(y - X_\gamma \hat{\beta}_\gamma), \\ \hat{Cov}(\hat{\beta}_\gamma | \sigma^2, X_\gamma) &= (X_\gamma' X_\gamma)^{-1} \hat{\sigma}^2 \end{aligned} \tag{5.3}$$

providing that $(X_\gamma' X_\gamma)^{-1}$ exists.

Instead of using σ^2, when modeling the error term, (4.1) can be parametrized in terms of

precision $\tau=\sigma^{-2}$. By doing so, a more straightforward interpretation of the parameter is achieved, since τ quantifies the accuracy of the estimated quantity that is to be used to summarize Y .

Proper prior formulation for unknown parameters in linear model is of major concern since it strongly affects the posterior results. There are two main alternatives when choosing a prior distribution over (β_γ, τ). On the one hand, the need of representing prior ignorance concerning the parameter vector, is supported by using a non-informative prior. This approach eliminates subjectivity during the analysis and maximizes data contribution in posterior results.

In brief, there are three basic methods in prior formulation when dealing with objective Bayesian analysis. Laplace's rule is based on the principle of insufficient reason and states that if there is no reason to a priori favor specific values of the parameter vector, then each one should be treated and weighted equally. This can be achieved by assigning a uniform prior over parameter vector. However, such a selection is sensitive in terms of invariance principle, meaning that any transformation upon the parameters, affects the prior distribution, making it potentially informative. Motivated by that, Jeffrey's (1946) proposed a widely accepted prior distribution, proportional to the square root of Fisher information matrix, which remains invariant, independently of the parametrization. The third option for constructing a non-informative prior is the reference prior, introduced by Bernardo (1979). By maximizing Kullback-Liebler divergence a reference prior, attempts to maximize the distance between the prior and the posterior function and hence maximize the contribution of the data in posterior estimation (Kass and Wasserman, 1996).

In case of subjective Bayesian analysis, researcher's prior beliefs can be represented by assigning an informative prior distribution over the parameters. Formulation of an informative prior for linear regression has been mainly based on conjugate analysis with Normal-Gamma distribution (Smith and Kohn, 1996; George and McCullagh, 1997; Raftery et al, 1997), since it facilitates computations regarding the posterior quantity of interest and avoids intractability of models' marginal likelihood. The Normal-Gamma prior scheme is presented by

$$\left(\beta_\gamma, \tau\right) \sim NG\left(\beta_{\pi\gamma}, \frac{V_\pi}{\tau}, a_\pi, b_\pi\right) \tag{5.4},$$

with

$$\begin{aligned} \beta_\gamma|\tau, \gamma &\sim N_p\left(\beta_{\pi\gamma}, \frac{V_\pi}{\tau}\right) \\ \tau|\gamma &\sim Gamma(a_\pi, b_\pi) \end{aligned} \tag{5.5},$$

where π is used to index the prior distribution parameters.

The resulting posterior is again a multivariate Normal-Gamma distribution, $\beta_\gamma, \tau \sim NG(\tilde{\beta}_\gamma, \tilde{T}, \tilde{a}, \tilde{b})$ with updated parameters given by

$$\begin{aligned}\tilde{\beta}_\gamma &= \tilde{T}\,(X_\gamma' y + V_\pi^{-1}\beta_{\pi\gamma}) \\ \tilde{T} &= (V_\pi^{-1} + X_\gamma' X_\gamma)^{-1} \\ \tilde{a} &= \frac{1}{2}n + a_\pi \\ \tilde{b} &= \frac{1}{2}SS + b_\pi \\ SS &= y' y - \tilde{\beta}_\gamma' \tilde{T}^{-1}\tilde{\beta}_\gamma + \beta_{\pi\gamma}' V_\pi^{-1}\beta_{\pi\gamma}\end{aligned} \tag{5.6}$$

Considering the marginal posterior of parameter vector β_γ, by integrating out the precision τ, we obtain the following p-dimensional Student's t distribution with $n+2\mathrm{a}$ degrees of freedom

$$\beta_\gamma | y, M_\gamma \sim MSt_p\left(\tilde{\beta}_\gamma, \frac{SS+2\mathrm{b}}{n+2\mathrm{a}}\tilde{T}, n+2\mathrm{a}\right). \tag{5.7}$$

Similarly, by integrating out β_γ we obtain the marginal posterior of the precision as a Gamma distribution

$$\tau | y, M_\gamma \sim G(\tilde{a}, \tilde{b}) \tag{5.8}.$$

A detailed review concerning conjugate analysis for Normal data and Bayesian linear regression can be found in Ntzoufras (2009, ch 1.5, p 9 – 13), while computational details on posterior derivation can be found in Bernardo and Smith (1994). Detailed review concerning non-informative prior formulation can be found in Kass and Wasserman (1996).

5.3 Zellner's g prior

5.3.1 Introduction

Even under the conjugate Normal-Gamma structure, prior formulation is not straightforward (Zellner, 1983). Specification of prior parameters involved in equation (4.4), has been an area of extended research in literature, especially focusing on prior covariance matrix formulation. Zellner (1986), proposed a specific prior scheme based on conjugate Normal-Gamma family. The so called g prior for a model M_γ, assumes a Jeffrey's prior over the precision and a p-dimensional Normal distribution over coefficient vector β_γ, with prior covariance matrix proportional to the inverse of Fisher information matrix;

$$\begin{aligned}\beta_\gamma | \tau, \gamma &\sim N_p(\beta_{\pi\gamma}, g\,\tau^{-1}(X_\gamma' X_\gamma)^{-1}). \\ f(\tau|\gamma) &\propto \tau^{-1}\end{aligned} \tag{5.9}$$

In contrast to the scheme presented in (4.5), Zellner's prior simplifies prior covariance set-up by reducing the number of unknown parameters to one. The unspecified parameter g, plays a key role in the analysis, since it controls prior weight and quantifies the prior contribution in posterior results (Liang et al., 2008). The influence of g can be measured, in terms of additional observational

units added by the prior, in conditional posterior of $\beta_\gamma|y,\tau,\gamma$. A choice of $g=1$ corresponds to adding n observations in the analysis; in other words the posterior result, depends on the prior on a 50%. Similarly, a choice of $g=10$ implies a prior weight equal to 10% contribution. Larger values of g, reflect prior ignorance regarding β_γ. Detailed choices concerning g, will be discussed in section 5.3.3.

5.3.2 Model comparison via Zellner's g-prior

Apart from simplifying prior set-up, Zellner's prior became popular due to the fact that it leads to closed form expressions of marginal likelihoods. Consequently, Bayes factors, which can now be expressed as a function of the coefficient of determination R^2_γ, facilitates and accelerates computations in model comparison (Liang et al., 2008).

Since for any base model M_b, we can compute the Bayes factor of model M_γ over M_b, by

$$BF_{\gamma,b}=\frac{f(y|M_\gamma)}{f(y|M_b)}, \tag{5.10}$$

we can compare any two models M_γ and $M_{\gamma'}$ by

$$BF_{\gamma,\gamma'}=\frac{BF_{\gamma b}}{BF_{\gamma' b}}. \tag{5.11}$$

A common strategy for model comparison, is to compare nested models. In such cases Zellner and Siow (1980), proposed to to assign a flat prior over the parameters that appear in both models and a g-prior over the remaining parameters of the more complex model. In this context, the use of the null or the full model as a base model, makes each pair under consideration nested (Liang et al., 2008, Guo and Speckman, 2009).

If we chose as the base the model null; M_{null}, under equation (4.1), σ^2 or τ is the only common parameter between all models. A simplification adopted by Fernandez et al., 2001, Liang et al., 2008, Bottolo and Richardson, 2010 and others, occurs by assuming a centering of the covariates, so that $1'X_\gamma=0_\gamma$. Then, the intercept β_0 can be considered as a common parameter between any model M_γ and M_{null} and can be treated in the same way as τ. The above scheme leads to the following optional form of Zellner's g-prior

$$\begin{aligned}\beta_\gamma|\tau,\gamma &\sim N_p(0_\gamma, g\,\tau^{-1}(X_\gamma' X_\gamma)^{-1})\\ \beta_0,\tau|\gamma &\sim \tau^{-1}\end{aligned} \tag{5.12}.$$

Under (4.12) and centered covariates the marginal likelihood can be analytically computed as

$$f(y|\gamma, g)=\frac{\Gamma((v-1)/2)}{\sqrt{(\pi)}^{n-1}\sqrt{(n)}}\|y-\bar{y}\|\frac{(1+g)^{(n-1-p_\gamma)/2}}{[1+g(1-R_\gamma^2)]^{(n-1)/2}}, \quad (5.13)$$

where R_γ^2 is the coefficient of determination of candidate model M_γ.

The resulting Bayes factor, used for comparisons between M_γ and M_{null} can be obtained as a function of R_γ^2 and g

$$BF_{\gamma, null}=\frac{(1+g)^{(n-p_\gamma-1)/2}}{[1+g(1-R_\gamma^2)]^{(n-1)/2}}. \quad (5.14)$$

Using similar arguments, see Liang et al., (2008) for details, if we consider the full model M_{full} as a base for comparisons, Bayes factors can be obtained, as

$$BF_{\gamma, full}=(1+g)^{-(n-p-1)/2}\left[1+g\frac{1-R_{full}^2}{1-R_\gamma^2}\right]^{(n-p_\gamma-1)/2}. \quad (5.15)$$

5.3.3 Selecting g

5.3.3.1 Fixed values

As discussed in section 5.3.1, the choice of g quantifies the amount of subjectivity in the analysis and in case of uninformative prior over models, it actually controls model selection. In general, as pointed out in George and Foster (2000), a choice of larger values of g, leads to models with fewer parameters and large coefficients. On the other hand, smaller values are associated to a selection of saturated models with small values of coefficients. Moreover, as stated in Liang et al., (2008), g acts as a dimensionality penalty and specific fixed choices of g, have been studied and introduced in relation to information criteria, such as BIC.

A popular choice in case of prior ignorance, is setting $g=n$. Such a selection, retain the spirit of unit information priors of Kass and Wassermann (1995) and corresponds to prior distribution that adds information equal to one observation in the posterior analysis (Fernandez et al., 2001). Besides that, by setting $g=n$, a strong connection between BIC and the log posterior $f(\gamma|y)$ is obtained. In Chapter 2 we discussed that BIC derives as a large sample approximation of the log transformed posterior of model M_γ and penalizes model complexity by adding in deviance measure a penalty term equal to $p_\gamma \log(n)$

$$BIC_\gamma=C+nlog(RSS_\gamma)+p_\gamma\log(n) \quad (5.16),$$

where C is a common constant over all candidate models.

A similar expression of the log transformed posterior of model M_γ is obtained, under the unit information prior and a uniform distribution over model space.

$$-2\log f(\gamma|y)=const.+nlog(RSS_\gamma)+p_\gamma\log(n+1) \quad (5.17).$$

The penalty term in (5.16) is replaced by $p_\gamma \log(n+1)$, depicting the influence of one additional informational unit added by the prior (Ntzoufras, 2009 p. 96-98).

Other recommendations for a fixed-valued g, include

- $g = p_\gamma^2$, introduced by Foster and George (1994), connecting the g-prior with the risk inflation criterion (RIC)
- $g = max(n, p_\gamma^2)$, namely the benchmark prior, introduced by Fernandez et al., (2001), combining BIC and RIC (BRICK)
- $g = \log(n)^3$, which asymptotically mimics the Hanna-Quinn criterion.

5.3.3.2 Empirical Bayes methods

Under the null baseline model approach, Liang et al., (2008) examine the influence of selecting a fixed value for g, and focus on two undesirable issues that arise. In case of prior ignorance, selecting large values for g imply an uninformative prior over parameter space. Such a choice, though, activates the Lindley's paradox. Supposing a fixed value for n and p_γ, Bayes factor, as derived in (4.14), always favors the null model, irrespective of the evidence provided by the data; i.e when $g \to \infty$, $BF_{\gamma, null} \to 0$. In addition, in case of a perfectly fitted model, so that $R_\gamma^2 \to 1$, a fixed choice of g does not allow Bayes factor go to infinity, activating the information paradox. In other words, for a constant value of g , n and p_γ , as $R_\gamma^2 \to 1$, Bayes factor converges to a constant $BF_{\gamma, null} \to (1+g)^{(n-p_\gamma-1)/2}$.

Obviously, a preselected value of g, is related to some undesirable issues and as stated in Celeux et al., (2010), although they rely on asymptotic properties, they heavily depend on sample size, involve a degree of arbitrariness and, thus, could be characterized as unsatisfactory choices. In an attempt to provide with more objective approaches, George and Foster (2000) and Clyde and George (2000), exploited empirical Bayes methods to develop a common or global data dependent estimate of g .

The global empirical Bayes approach assumes a common value of g over all models, which is estimated by maximizing the marginal likelihood $p(y|\gamma, g)$, as an average over all models.

$$\hat{g}^{EBG} = \underset{g>0}{argmax} = \sum_\gamma p(\gamma) \frac{(1+g)^{(1-p_\gamma-1)/2}}{[1+g(1-R_\gamma^2)]^{(n-1)/2}}, \tag{5.18}$$

A related approach is the local empirical Bayes estimate, introduced by Hansen and Yu (2001). It assumes varying data-based estimates of g for each model, obtained by

$$\hat{g}_\gamma^{EBL} = max\{F_\gamma - 1, 0\} \tag{5.19},$$

where F_γ is the common F statistic for testing the hypothesis $H_0: \beta_\gamma = 0$;

$$F_\gamma = \frac{R_\gamma^2 / p_\gamma}{(1-R_\gamma^2)/(n-1-p_\gamma)} \tag{5.20}$$

The resulting Bayes factor is obtained below

$$BF_{\gamma,null} = \frac{[1+R_\gamma^2(n-1-p_\gamma)/p_\gamma(-R_\gamma^2)]^{(n-p_\gamma-1)/2}}{[1+(n-1-p_\gamma)R_\gamma^2/p_\gamma(1-R_\gamma^2)]^{(n-1)/2}} \tag{5.21}$$

5.3.3.3 Full Bayes approach

The information paradox is resolved through empirical Bayes methods; see Liang et al., (2008). However, model selection consistency, as considered in Fernandez et al., (2001), is not guaranteed. In other words, the selection of the true model is not asymptotically certain by EB approaches. Besides that, empirical methods stand in contrast to fully Bayesian approaches and are often criticized for using data-based estimates for prior quantities. Instead of fixing the unknown parameter, the most natural alternative to deal with the uncertainty of g, is to assign a weakly informative hyperprior over g (Zellner, 1986). The most popular choices that have been developed in literature, as fully Bayesian approaches, are the Zellner-Siow prior (Zellner and Siow, 1980) and the hyper-g family of priors, introduced by Liang et al., (2008).

The Zellner-Siow prior, has been developed, exploiting Jeffreys' (1961) work on hypothesis testing of univariate normal means. Jeffreys proposed using a Cauchy prior instead of a Normal, to avoid inconsistency related to Bayes factors. Following Jeffreys, Zellner and Siow (1980), proposed a hierarchical scheme for comparing nested models in linear regression, based on the multivariate Cauchy distribution. Utilizing the same strategy described in 4.3.2, they proposed a flat prior on parameters that appear in both models under comparison and a Cauchy prior on the remaining ones. Representing the Zellner-Siow priors as a scale mixture of Gaussian random variables, the recommended prior on β_γ can be expressed as

$$\pi(\beta_\gamma|\tau) \propto \int N(\beta_\gamma|0, g\tau^{-1}(X_\gamma' X_\gamma)^{-1}\pi(g)\,dg) \tag{5.22}$$

with an Inverse Gamma assigned on g, so that

$$g \sim IG\left(\frac{1}{2}, \frac{n}{2}\right) \tag{5.23}$$

The main drawback of the Zellner-Siow priors is that marginal likelihoods are intractable and can be evaluated using for instance a Laplace approximation (Liang et al., 2008).

Liang et al., (2008), introduced a broader class of hyper-g priors as an alternative to Zellner-Siow prior, given by

$$\pi(g) = \frac{\beta_0 - 2}{2}(1+g)^{-\beta_0/2} \tag{5.24}$$

The prior (5.24) is a proper distribution for $\beta_0 > 2$. For $\beta_0 = 2$ can be considered both as a reference and a Jeffreys prior, while a reasonable choice include a range of $2 < \beta_0 \leq 4$. Further details and arguments on hyper parameter specification can be found in Liang et al. (2008).

In contrast to Zellner-Siow priors, the hyperprior of Liang et al., allows a closed form expression for marginal likelihood and posterior quantities of interest, even though it requires evaluation of the Gaussian hypergeometric function, which is proved to be problematic under certain circumstances. However, the Zellner-Siow prior results in a consistent model selection procedure under the null model, as considered in Fernandez et al. (2001). The property does not hold for the hyper-g family of priors even though Liang et al. (2008) proved that the null model remains the highest probability model. This occurs due to the fact that Zellener-Siow prior allows to depend on the sample size n . Motivated by that, Liang et al., proposed a sample-size dependent modification, namely the hyper-g/n prior

$$\pi(g) = \frac{\beta_0 - 2}{2\mathrm{n}} \left(1 + \frac{g}{n}\right)^{-\beta_0/2} \tag{5.25}.$$

The hyper-g/n prior is model selection consistent, but it does not allow analytical evaluation of posterior quantities. Computation of these quantities require a Laplace approximation, provided by Liang et al., (2008).

5.3.3.4 Discussion and Further Extensions of g-priors.

Uncertainty over g has been encountered through the three different ways described in this section. Liang et al., (2008), either by considering their theoretical properties in terms of consistency and through simulation studies, showed that fully Bayesian approaches outperform selecting a fixed value of g or estimating from the data using empirical Bayes approaches. Choices of a fixed value, are strongly related to the information paradox and are not model selection consistent, except from those that correspond to the BIC or BRIC approach. Empirical methods, apart from being partially Bayesian although they do not activate the information paradox, they do not guarantee model selection consistency. Setting a prior over g resolve both the information paradox and model selection inconsistency. Besides that they retain prediction consistency, a property which is occasionally preferred.

Apart from Liang et al., (2008), several authors have presented work based on g-priors, either proposing modifications of traditional Zellner's prior or studying theoretical aspects on existing extensions. Some of this work include Marin and Robert (2007), who propose a continuous improper prior on hyperparameter g, without considering consistency properties, Krishna et al., (2009), who provide with a modification of the prior covariance set up in Zellner's g-prior, Guo and Speckman (2009), who examine consistency issues of several g-prior settings, including improper

prior of Marin and Robert (2009), Celeux et al., (2010) who compare g-prior modifications to frequentist approaches in case of $p=n$, Bottolo and Richardson (2010), Maruyama and George (2010), Yang and Song (2010) and Baragatti and Pommeret (2012) who adopt g-prior to cope with 'large p small n' problem, and Fouskakis et al (2009, 2015, 2015). As stated in the introduction of the current chapter, Zellner's g-prior was also adopted by Clyde et al., (2011) to develop their Bayesian adaptive sampling scheme (BAS), which will be discussed in section 5.4 that follows, exploiting the computational simplifications that it provides.

5.4 Bayesian Adaptive Sampling

5.4.1 Introduction

BAS was introduced as an alternative to traditional MCMC algorithms, making use of an innovative model search algorithm which performs sampling without replacement from the posterior distribution. The key idea is described in Clyde et al., (2009); MCMC algorithms are designed to sample with replacement from finite model spaces. Then, by counting the visits on each model; i.e. by counting the MCMC model frequencies, each model is a posteriori ranked or selected as the highest probability one. In case of conjugate analysis or in general, in cases where marginal likelihoods are analytically tractable, the latter can replace MCMC model frequencies to produce marginal likelihoods and provide comparisons through models under consideration. In that sense, re-sampling over model space, does not actually provide with any additional information and sampling over model spaces without replacement could provide with a more efficient strategy for model search.

The algorithm's main features can be summarized in five bullets:

- BAS samples models without replacement.
- It fully enumerates model space, for a moderate number of covariates $p \approx 30$.
- It provides perfect samples, under the condition of orthogonality or of limiting dependence, when the number of covariates is larger and sampling is unavoidable.
- It samples near the the median probability model, providing that the sampling probabilities are the marginal inclusion probabilities.
- Estimates marginal inclusion probabilities adaptively, as they are not known beforehand.

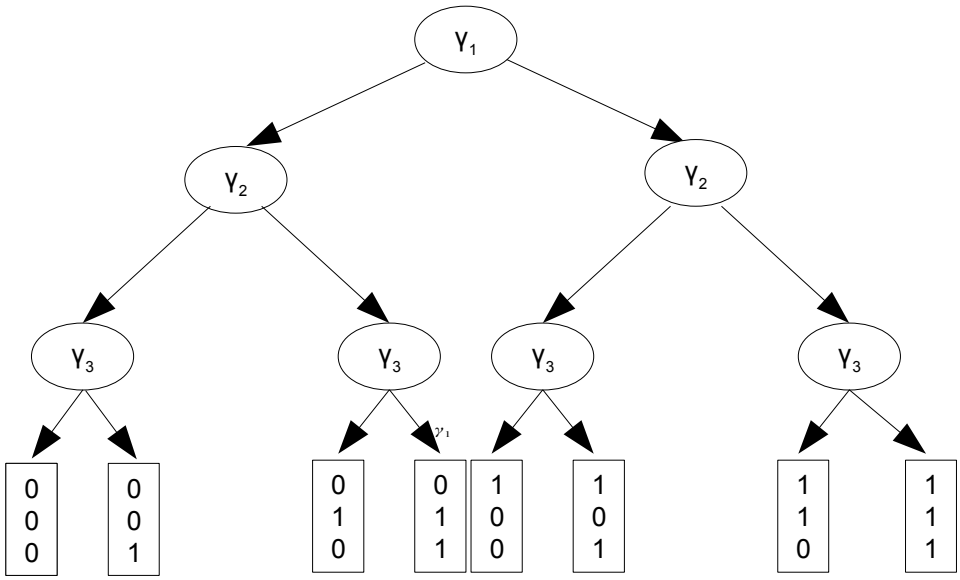
5.4.2 Sampling strategy

According to Clyde et al., (2011) the simplest way to perform sampling without replacement over the model space Γ , is by assigning equal probabilities to all models and draw a simple

random sample of size T . However, such an approach, does not account for the magnitude of each model, leading to samples that may exclude models of high posterior probabilities, especially in cases of carelessly selected sample size T .

In contrast to simple random sampling without replacement, they suggest a probability proportional to size sampling scheme (PPS). The idea behind PPS sampling, is to use an auxiliary variable as a size measure for each model, in order to improve sampling accuracy and efficiency. The efficiency of such a sampling scheme, strongly depends on the constructed size variables, which should be correlated to the variable of interest. In this context, Clyde et al., (2011) propose that sampling variables should be correlated to the product of the marginal likelihood and the prior probability of each model. Once a model is sampled, to ensure that the latter will not be re-sampled, its probability is set to zero and the remaining sampling probabilities are re-caclulated, under the restricted set of unselected models.

In order to perform PPS sampling, one needs to fully specify the initial sampling probabilities over all models in advance and at each step re-normalize them under the restricted sampling frame. Such an exhaustive enumeration is avoided by representing the model space by a binary tree. The following scheme represents the model space in case of 3 candidate variables.



Each node represents a candidate variable. Beginning from top, the first node represents γ_1. followed by the remaining 2 candidates. Each branch corresponds to the inclusion or exclusion of each candidate. Specifically, the left branch corresponds to the exclusion of a variable $\gamma_j=0.$ while the right one represents the inclusion of a variable $\gamma_j=1.$ Each path leads to a unique model, with $2^3=8$ endpoints, which represent all candidate models. Under this scheme, after a model is selected, its contribution is removed and as the algorithm proceeds the sampling distributions are re-calculated only once a node is sampled. In other words, the update of sampling

probabilities, is limited to the ones getting involved in the path being sampled at each step. The sampling probabilities that do not include the sampled model remain the same as in the previous iteration.

5.4.3 BAS notation and implementation

Following Clyde et al., (2011), assume that under BAS, a model is sampled without replacement using probabilities proportional to a probability mass function $f(\gamma)$. Since γ_j is binary, we can consider that the pmf used to sample models, is a product Bernoulli distribution with probability ρ_j , so that

$$f(\gamma)=\prod_{j=1}^{p}\rho_j^{\gamma_j}(1-\rho_j)^{(1-\gamma_j)} \tag{5.26}$$

Let, $\gamma_{<j}$ denote the subset of indicators $\{\gamma_k\}$ for $k<j$ and $\gamma_{\geq j}$ the corresponding subset for $k\geq j$. Then, the pmf $f(\gamma)$ can be expressed as a product of univariate conditional distributions as above:

$$f(\gamma)=\prod_{j=1}^{p}f(\gamma_j|\gamma_{<j}) \tag{5.27}$$

Using (4.26), $f(\gamma)$ becomes

$$f(\gamma|\rho)=\prod_{j=1}^{p}\rho_{j|<j}^{\gamma_j}(1-\rho_{j|<j})^{(1-\gamma_j)} \tag{5.28}$$

where ρ is the sequence of all $\{\rho_{j|<j}\}$ formed as $\rho_{j|<j}=f(\gamma_j=1|\gamma_{<j})$. The latter corresponds to the partial conditional distribution of inclusion of variable $j.$ given the inclusion history of past sampled variables $1,2,\ldots,j-1.$

Then the algorithm consists of the following steps:

- At time $t=0$, initialize with starting sampling probabilities $\rho=\rho^{(0)}$.
- For $t=1,2,\ldots,T$, consider the partition of model space Γ between $\Gamma_t^{(s)}$, the set of all previously selected models and $\Gamma_t^{(u)}$ the set of the remaining unselected models, so that $\Gamma=\{\Gamma_t^{(s)},\Gamma_t^{(u)}\}$
- Sample a model $\gamma^{(t)}$ with $\gamma_j^{(t)}|\gamma_{<j}^{(t)}\sim Ber\left(\rho_{j|<j}^{(t-1)}\right)$.
- Set $\Gamma_t^{(s)}=\Gamma_{t-1}^{(s)}\cup\{\gamma^{(t)}\}$.
- For $j=1,2,\ldots,p$ update the conditional probabilities $\rho_{j|<j}^{(t-1)}$ over the path of the binary tree of model $\gamma^{(t)}$ with

$$\rho_{j|<j}^{(t)}=\frac{\rho_{j|<j}^{(t-1)}-f(\gamma_{\geq j}^{(j)}|\gamma_{<j}^{(t)},\rho^{(t-1)})\gamma_j^{(t)}}{1-f(\gamma_{\geq j}^{(t)}|\gamma_{<j}^{(t)},\rho^{(t-1)})} \tag{5.29}$$

where

$$f(\gamma_{\geq j}^{(t)}|\gamma_{<j}^{(t)}, \rho^{(t-1)}) = \prod_{k=j}^{p} \left(\rho_{k|<k}^{(t-1)}\right)^{\gamma_k^{(t)}} \left(1-\rho_{k|<k}^{(t-1)}\right)^{1-\gamma_k^{(t)}} \quad (5.30).$$

- For all other paths let

$$\rho_{j|<j}^{(t)} = \rho_{j|<j}^{(t-1)} \quad (5.31).$$

Under equation (4.29), BAS ensures two things; either that all models will be sampled in $T=2^p$ iterations and also that at each step, $f(\gamma|\rho^{(t)})$ assigns zero probability to past sampled models $\gamma \in \Gamma_t^{(s)}$ while for unselected models $\Gamma_t^{(s)} = \Gamma_{t-1}^{(s)} - \{\gamma^{(t)}\}$ assigns probability equal to one. The proof is provided in the supplement of Clyde et al., (2011) in the Supplemental Materials

5.4.4 Approximation and adaptivity

A main problem in the algorithm described above, is setting the starting values $\rho^{(0)}$ that initialize model sampling. In practice, the partial conditional posterior distribution $\rho_{j|<j} = f(\gamma_j = 1|\gamma_{<j})$ is unknown and evaluating them in advance, is computationally exhaustive. Clyde et al., (2011) suggest using marginal inclusion probabilities $p(\gamma_j = 1|y)$ instead, as an approximation of the posterior model probabilities. In particular a first order approximation of posterior model probabilities in terms of Kullback-Leibler divergence can be obtained, using the *current* estimates of the marginal inclusion probabilities, at each step of the algorithm, tas described in equation (4.26); see proof 2 at the Supplemental Material of Clyde et al (2011). So, ideally, the rationale of the algorithm, is to utilize past sampled models, in an adaptive way, to update marginal inclusion probabilities of each candidate at each step, through

$$\hat{\rho}_{j|<j}^{(t)} = \frac{\sum_{\gamma \in \Gamma_t^{(s)}} p(y|\gamma)\, p(\gamma)\gamma_j}{\sum_{\gamma \in \Gamma_t^{(s)}} p(y|\gamma)\, p(\gamma)} \quad (5.32),$$

and decide weather a candidate should be included in the model or not.

Updating the sampling inclusion probabilities at each step of the algorithm, is computationally expensive, since it also requires re-normalization over the sampling probability sequence, to avoid duplications over past sampled models. Clyde et al., (2011) suggest a compromise, estimating the marginal inclusion probabilities periodically, every U iterations, so that there is a significant change and the update is meaningful. They also claim that the update of $\rho^{(0)}$ should not be implemented too early, so that estimates do not receive zero probability. So,

they propose shrinking $\rho_{j|<j}^{(t)}$ away from zero or one, so that all models receive a possitive probability. Taking into account the above, the proposed algorithm, proceeds using the following steps:

- Set ε and $\delta=\sqrt{\varepsilon}$. Set $T\leq 2^p$.
- At time $t=0$, initialize with starting sampling probabilities $\rho=\rho^{(0)}$.
- For $t=1,2,\dots,T$, consider the partition of model space Γ between $\Gamma_t^{(s)}$, the set of all previously selected models and $\Gamma_t^{(u)}$ the set of the remaining unselected models, so that $\Gamma=\{\Gamma_t^{(s)},\Gamma_t^{(u)}\}$
- Sample a model $\gamma^{(t)}$ with $\gamma_j^{(t)}|\gamma_{<j}^{(t)}\sim Ber\left(\rho_{j|<j}^{(t-1)}\right)$.
- Set $\Gamma_t^{(s)}=\Gamma_{t-1}^{(s)}\cup\{\gamma^{(t)}\}$.
- For $j=1,2,\dots,p$ update the conditional probabilities $\rho_{j|<j}^{(t-1)}$ over the path of the binary tree of model $\gamma^{(t)}$ with

$$\rho_{j|<j}^{(t)}=\frac{\rho_{j|<j}^{(t-1)}-f(\gamma_{\geq j}^{t})\gamma_j^t}{1-f(\gamma_{\geq j}^{t})} \tag{5.33},$$

where

$$f(\gamma_{\geq j}^{(t)})=\prod_{k=j}^{p}\left(\rho_{k|<k}^{(t-1)}\right)^{\gamma_k^{(t)}}\left(1-\rho_{k|<k}^{(t-1)}\right)^{1-\gamma_k^{(t)}} \tag{5.34}.$$

- For all other paths let

$$\rho_{j|<j}^{(t)}=\rho_{j|<j}^{(t-1)} \tag{5.35}.$$

- If $t\ mod\ U=0$, estimate marginal inclusion probabilities

$$\hat{p}_j^{(t)}=\frac{\sum_{i-1}^{t}p\left(y|\gamma^{(i)}\right)p\left(\gamma^{(i)}\right)\gamma_j^{(i)}}{p\left(y|\gamma^{(i)}\right)p\left(\gamma^{(i)}\right)} \tag{5.36}.$$

- If $\|\rho_{j|<j}^{(t)}-\hat{p}_j^{(t)}\|^2/p>\delta$
  - Set $\rho_{j|<j}^{(0)}=min\left(max\left(\varepsilon,\hat{p}_j^{(t)}\right),1-\varepsilon\right)$
  - Re-normalize probabilities using equations (4.33) and (4.34) and sample with a new $\rho_{j|<j}^{(0)}$.

5.4.5 Estimation of initial values

The last step to complete the algorithm is to set the initial sampling probabilities. As discussed in Clyde et al., (2011), under the assumption of orthogonality, marginal inclusion probabilities, can be evaluated prior to sampling. In the general case, these quantities, must be estimated. The authors provide with three different choices on how this could be achieved. Initially, they propose using $\rho_{j|<j}^{(0)}=1/2$ corresponding to simple random sampling without replacement.

The second approach suggests estimating $\rho_{j|<j}^{(0)}$ through p-values, based on the work of Selke et al., (2001), on p-value calibration for testing precise hypothesis. The methodology proceeds as above:

- Fit the full model to the data.
- For $j=1,\dots,p$ test $H_{0j}:\beta_j=0$ versus $H_{1j}:\beta_j\neq 0$, given that the remaining coefficients β_j , $i=1,\dots,j-1,j+1,\dots,p$, are not zero.
- Calculate p , p-values, namely $p_j^{(v)}$.
- Set $\hat{\rho}_{j|<j}^{(0)}=1/\{1-ep_j^{(v)}logp_j^{(v)}\}$ $\quad if\ p_j^{(v)}<1/e\approx 0.37$
- Otherwise set $\hat{\rho}_{j|<j}^{(0)}=1/2$.

The third option that is highly recommended in case of highly correlated data, is to estimate initial values of marginal inclusion probabilities through MCMC frequencies. The suggested estimate is calculated as

$$\hat{\rho}_{j|<j}^{(0)}=\sum_{\gamma\in A}\gamma_j\,\hat{p}^{MC}(\gamma|y) \qquad (5.37),$$

where A corresponds to the unique sampled models and $\hat{p}^{MC}(\gamma|y)$ is estimated as in equation (3.7), Section 3.2.3.

Chapter 6: Illustration and examples of the BAS package

6.1 Introduction

In the final chapter of the current dissertation, the performance of the BAS algorithm will be tested. We will explore the ability of the algorithm to uncover the true model, for the case of linear regression. The main area of focus will be the effect of the coefficient vector's prior, on posterior results. Both independent and correlated simulated data sets will be used. At first, we will study the performance of the BAS algorithm in case of a relatively small model space using 10 candidates. Recall that, when this is the case, BAS ensures that model space will be fully explored. The second part will deal with a larger model space of 30 candidates, where sampling is required. Before that, there will be a brief presentation of the main formulas that are included in the BAS package.

6.1 The BAS package in R

The BAS package version 1.0, has been developed by Clyde and Littman (2005) as a tool for Bayesian model selection in R and implements the BAS algorithm that has been described in chapter 4.

The main formula that applies the aforementioned algorithm, is the the *bas.lm()* function. It performs random or deterministic sampling without replacement in model space using a prior distribution on coefficients that belong to Zellner's g-prior family for $p>15$. For $p<15$ it fully enumerates the marginal likelihoods of all models under consideration (equal to 2p) Possible choices include Zellner's g-prior, Zellner-Siow Cauchy prior, hyper-g prior of Liang et al. (2008), Local and Global empirical Bayes estimates of g and AIC or BIC, as model selection criteria. To initialize the algorithm BAS provides with two options on the starting marginal inclusion probabilities. One can assign either equal probabilities on each predictor or can use the p-value calibration of Selke et al. (2001), as described in section 4.4.5. There is also a possibility of preselecting the number of models that the algorithm will sample and the frequency of sampling probabilities updates. Results can be updated using a different prior, without rerunning the algorithm, through the *update.bma()* function.

Considering the results, the *summary.bma()* function, by default, prints the top 5 highest posterior probability models with their corresponding Bayes factor, posterior probability, R square, dimension and logarithm of the marginal likelihood. The marginal posterior summaries of coefficients can be obtained by the *coef.bma()* function, which prints their posterior means, standard deviations and marginal inclusion probabilities, under Bayesian Model Averaging. The Posterior distributions of coefficients can also be graphically displayed using the *plot.coef.bma()* function.

The BAS package includes two more plotting functions, namely *image.bma()* and *plot.bma()*. The first function displays a heat map of the model space sampled under BAS, while *plot.bma()* returns four plots; the residuals vs fitted values plot, the cumulative model probability plot, the models' log marginal likelihood vs model complexity and a graph of marginal inclusion probabilities.

Finally, fitted values and predictions can be calculated through *fitted.bma()* and *predict.bma()* functions. The *fitted.bma()* function returns fitted values under the highest probability model, the median probability model and the posterior means of BMA using the top m sampled models. *predict.bma()* calculates the predicted values using BMA. The last function which deals with predictions, is the *cv.summary.bma()* function. It provides with out of sample predictions, given the output of *predict.bma()* function, returning the average prediction error form the highest probability model and the average prediction error under BMA.

6.2 Examples

6.2.1 Priors used in BAS

Under the data that will be generated and the prior model distribution, we applied the BAS algorithm under the priors choices that can be implemented using BAS package and have been discussed in detail, in chapter 4, section 4.3. In specific, we applied the following prior choices:

- AIC: The Akaike Information Criterion,
- BIC: The Bayesian Information Criterion,
- g-prior: The prior of Zellner with $g=100$ corresponding to the Unit information prior, of Kass and Wassermann (1995)
- ZS-Null: The prior of Zellner & Siow, utilizing the null model as a base for comparison, Zellner and Siow (1980)
- ZS-Full: The prior of Zellner & Siow, utilizing the full model as a base for comparison, Zellner and Siow (1980)
- Hyper-g: The prior of Liang et al. (2008), with $a=3$ as recommended in Clyde et al (2011),
- Hyper-g-Laplace: The prior of Liang et al. (2008), using a Laplace approximation to estimate g,
- EB-Local: The local empirical Bayes estimate of g, of Hansen and Yu (2001)
- EB-Global: The global empirical estimate of g, of George and Foster (2000) and Clyde and George (2000)

6.2.2 Full enumeration – Simulated Data

In our first example, we used simulated data with $p=10$ candidates and $n=150$. observations. The data set was split, so that the first 100 observations were used to apply the algorithm and the last 50 observations were used to perform out of sample predictions. All columns of the design matrix were generated from independent $N(0,1)$ random variables. The parameters were, deliberately, chosen to be relatively high, to ensure that will be included in the model, independently of the prior coefficients setup. In particular, we chose $\alpha=4$, $\beta=(3.2,-1.05,0,0,0,0,0,0,0.5,0)$ and $\varphi=1$. To complete the prior specification, the prior distribution over model space was set to be Uniform, using $p(M_\gamma)=\frac{1}{2^p}$.

Posterior Results

All methods achieved to detect the true model as the maximum a posteriori model (MAP), with the inclusion probabilities of the first 2 candidates to be equal to 1. The corresponding probability of the ninth candidate was estimated to be approximately 0.99. However, as noticed, the 'AIC' and 'ZS-Full' methods, seem less confident in detecting the true model. Their corresponding model's posterior probability appeared to be significantly low; 0.07 and 0.17, in contrast to all other methods, the value of which, fluctuated around 0.5. In addition, the above methods assigned significantly higher marginal inclusion probabilities to the variables excluded from the model. In contrast, the highest model posterior probabilities, corresponded to the Empirical Bayes methods (approximately 0.57), which also computed the lowest values of marginal inclusion probabilities for the excluded candidates.

Regarding the marginal posterior means of each coefficient under BMA, the intercept was estimated to be 3.97 with a marginal posterior standard deviation around 0.11. The closest estimates of the effects were provided by the 'g-prior' method, while 'AIC' and 'BIC', provided the most distant ones. The 'BIC' method, seemed to be related to higher estimates of marginal posterior standard deviations for each coefficient under BMA.

Finally, concerning the Average Prediction Error (APE), the 'AIC', 'BIC' and 'ZS-Full' methods, provided with the lowest values of both in sample and out of sample APE under MAP (1.053645 and 0.9247008). The corresponding highest values were detected under the 'g-prior'. 'AIC' prior, also provided with the smallest in sample APE under BMA (1.046514) but the largest out of sample APE under BMA (0.9348658). The highest ones were computed under the 'hyper-g' and 'BIC' prior, respectively. Summary tables of the results are provided in Appendix A.

Consistency of results

In order to examine the stability of the results, we repeated the experiment by generating 100 samples as described above and applied the algorithm in each sample. The first two candidates, X_1 and X_2 were included in the selected model in all samples with an average marginal inclusion probability 1. The ninth candidate was selected 99 times, with an average marginal inclusion probability 0.97, except from the 'AIC' and 'ZS-Full' case where it was also found in the selected model in all samples.

Under 'AIC' prior, the algorithm failed to detect the true model as the HPM, in 76 samples As it can be seen in table 1, 'AIC' prior and tends to select overfitted models and is related to higher rate of selecting candidates with zero effect, as significant ones. The 'ZS-Full' method also seems inefficient, however the probability of selecting the true model is increased, in contrast to 'AIC' (57%). The most effective methods are the 'ZS-Null', 'Hyper-g', 'Hyper-g Laplace', 'EB-Local' and 'EB-Global', which succeeded to uncover the true model in 89 samples, with average posterior probability of the model selected to be on average 0.5.

Table 6.1: Frequencies of candidate spotted as important for 100 simulations

	AIC	BIC	g-prior (g=n)	ZS-Null	ZS-Full	Hyper-g (a=3)	hyper-g Laplace (a=3)	EB Local	EB Global
Intercept	100	100	100	100	100	100	100	100	100
x1*	100	100	100	100	100	100	100	100	100
x2*	100	100	100	100	100	100	100	100	100
x3	15	3	3	3	5	3	3	3	3
x4	19	4	3	3	10	2	2	2	3
x5	12	2	1	0	4	0	0	0	0
x6	17	4	3	1	9	1	1	1	1
x7	19	5	3	3	9	3	3	3	3
x8	17	4	4	2	11	2	2	2	2
x9*	100	99	99	99	100	99	99	99	99
x10	14	1	0	0	4	0	0	0	0
count	34	78	83	89	57	89	89	89	89
MPP	0,062	0,344	0,358	0,494	0,136	0,501	0,499	0,504	0,51
SD	0,012	0,078	0,076	0,111	0,032	0,111	0,111	0,112	0,116
log marginal likelihood	-230,3	-235,4	115,4	116,5	10,7	115	115	118	118
SD log	7,738	7,513	8,684	9,605	0,98	9,571	9,571	9,68	9,682

count: no of times that the true model was detected as the HPM
MPP: average posterior probability of the true model over 100 samples
SD HPM: average standard deviation of true model's posterior probability over 100 samples
log marginal likelihood: average log marginal likelihood of the true model
SD log: average sd of log maginal likelihood of the true model

Regarding the marginal posterior inclusion probabilities of zero effect candidates, most methods estimate them, on average, at around 0.1. Under 'ZS-Full', their estimates were increased at 0.25, while using the 'AIC' prior the estimates are even higher, reaching 0.4. As it can be seen in

graph 1, the variability of zero effect candidates is considerably increased for the 'AIC' and 'ZS-Full' priors, in contrast to other. However, a closer look in the 9th candidate, depicts an opposite attribute. The distribution of the marginal inclusion probability under the aforementioned priors has less low extreme values (graph 6.1).

Considering the posterior means of the coefficients X_1, X_2 and X_9 , 'AIC' prior provided with the closest estimates, under BMA. Their distribution, though do not seem to alter considerably (graph 6.2). On the other hand, the distribution of the zero effect candidates is found to be of greater variance (graph 6.3).

Finally, as far as APE is concerned, 'AIC' and 'ZS-Full' prior are related to smallest in sample APE and out of sample APE under BMA. However, the corresponding out of sample APE for HPM, can be seen that is higher, in contrast to other priors.

Figure 6.1: Marginal inclusion probabilities. (100 samples)

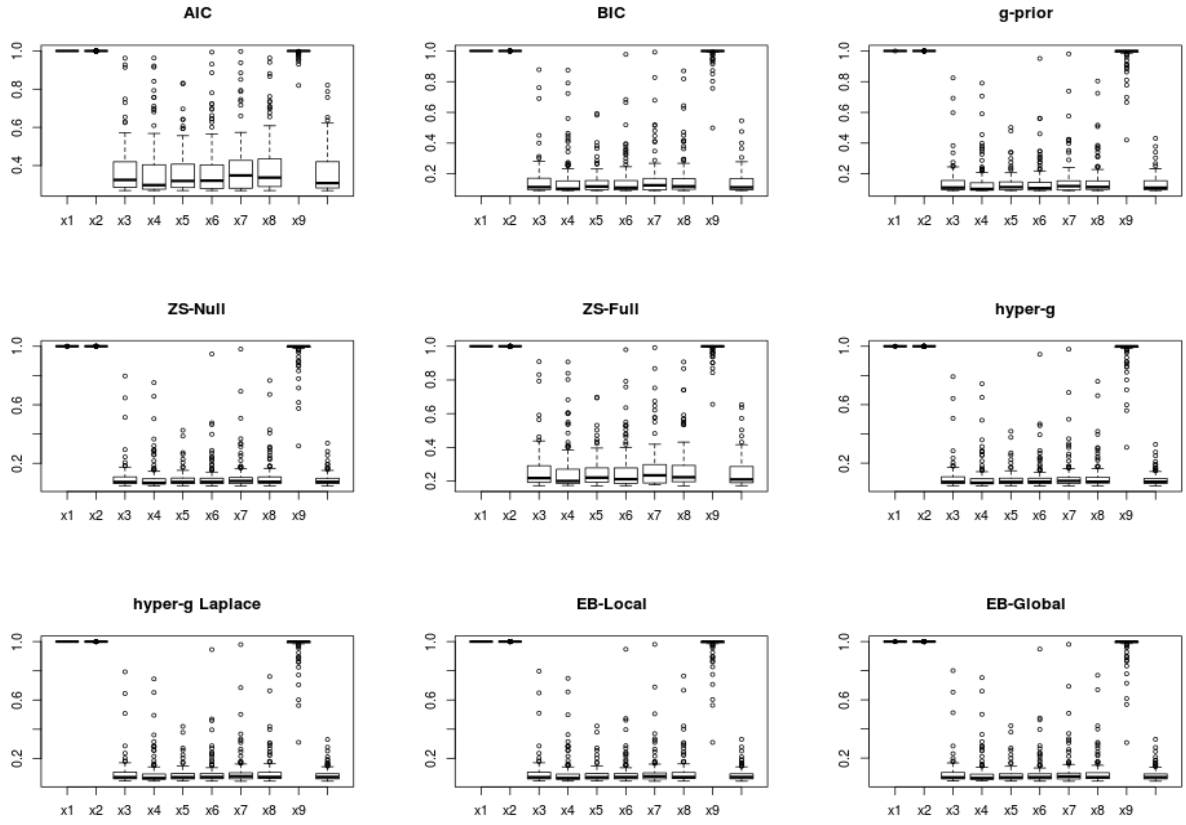


Figure 6.2: Posterior means – Non-zero coefficients. (100 samples)

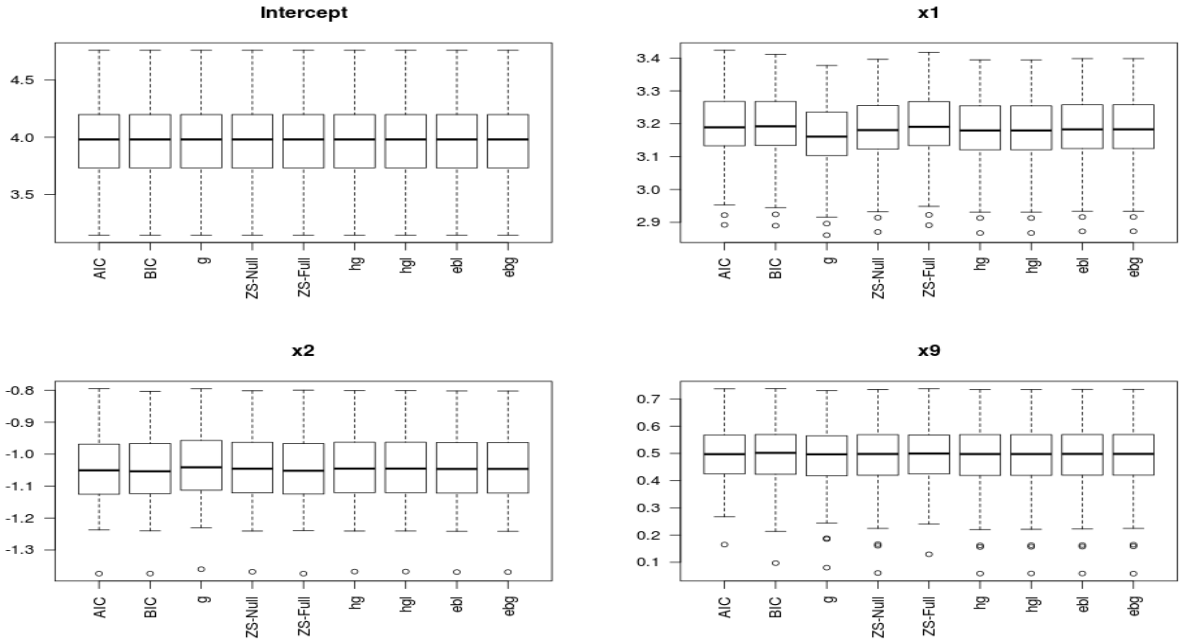


Figure 6.3: Posterior means – Zero coefficients. (100 samples)

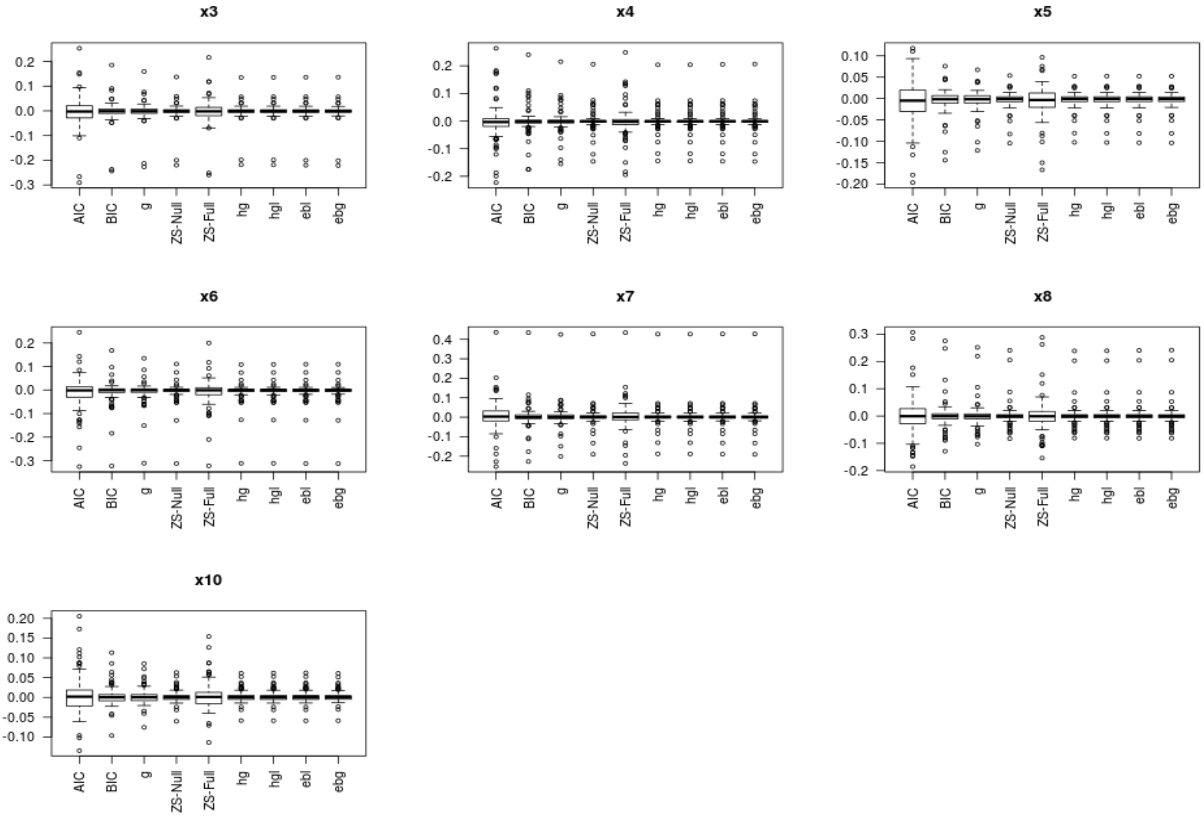
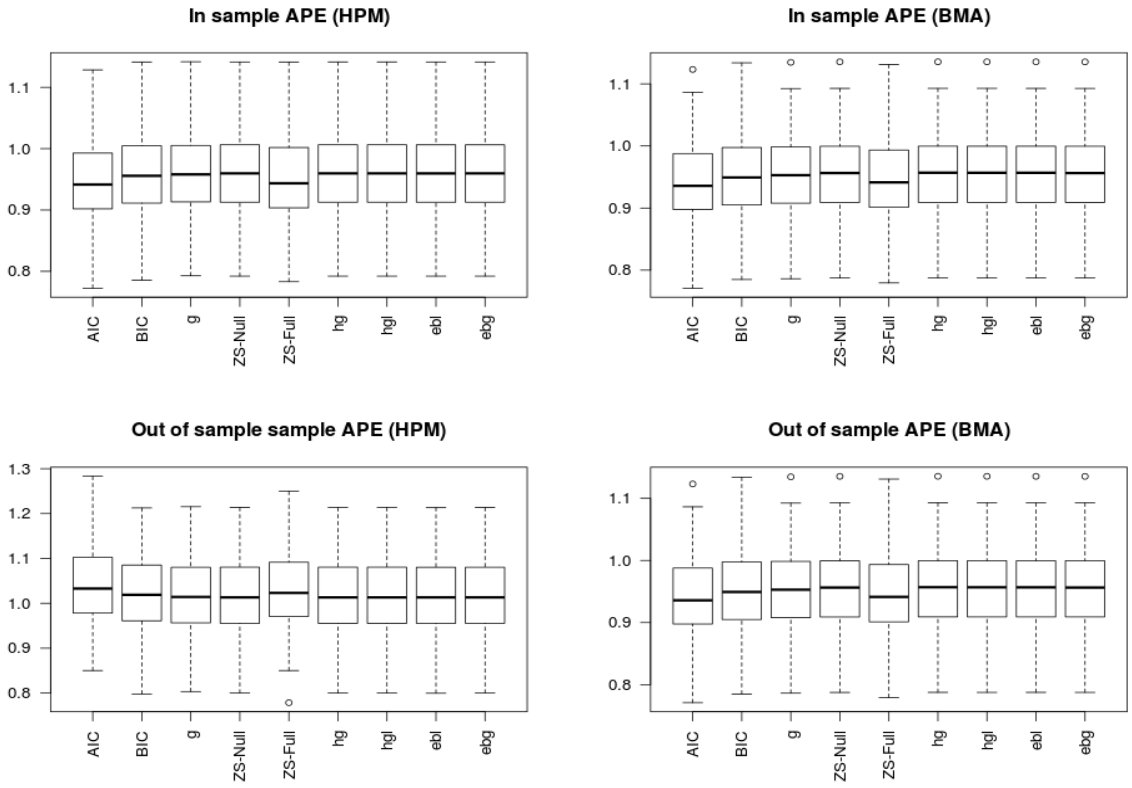


Figure 6.4: In sample & Out of sample Average Prediction Error. (100 samples)



6.2.3 Adaptive Sampling - Simulated Data

In the second example, we explored the performance of the algorithm using simulated data with $p=30$ and $n=15,$ so that all calculated results may be obtained by sampling the model space. The first 26 columns of the model space were generated using independent $N(0,1.)$ The last four candidates were generated under the following correlation matrix

Table 6.2: Correlation matrix

	x27	x28	x29	x30
x1	**0,993**	-0,033	-0,007	0,080
x2	-0,061	**0,794**	0,069	-0,065
x3	0,055	0,059	**0,730**	-0,191
x4	-0,050	-0,132	-0,205	**0,689**

The regression parameters were chosen as

$a=3.2,\ \beta=(2.3,-1.5,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,-0.8,0,0,0,0,0,0,0,0,0,0,0,-0.48)'$

and $\varphi=1.$ We ran the algorithm for 2^{15} iterations and the updating step for sampling inclusion probabilities was chosen to be 500. At first, the performance of the algorithm was explored under Uniform initial probabilities and afterwards the p-value calibration was examined.

Similarly to the first example, the true model was successfully identified as the MAP, under all priors except from the AIC and ZS-Full prior, which tend to select models of considerably higher

dimensions. In particular, for the top 20 ranked models, independently from the initial sampling probabilities choice, while all other priors selected models of 5 or 6 candidates, under AIC and ZS-Full prior, the number of significant candidates, varied between 7 and 11. (Table 6.3)

Table 6.3: Dimension of the top 20 sampled models (constant included)

Model Rank	Initial probabilities: Uniform									Initial probabilities: p-value calibration								
	AIC	BIC	ZS-Null	ZS-Full	g-prior	hyper-g	Laplace	EB-Local	EB-Global	AIC	BIC	ZS-Null	ZS-Full	g-prior	hyper-g	Laplace	EB-Local	EB-Global
1	8	5	5	8	5	5	5	5	5	8	5	5	8	5	5	5	5	5
2	9	6	5	7	5	5	5	5	5	9	6	5	7	5	5	5	5	5
3	10	6	6	7	6	6	6	6	6	10	6	6	7	6	6	6	6	6
4	9	5	6	9	6	6	6	6	6	9	5	6	9	6	6	6	6	6
5	7	6	6	8	6	6	6	6	6	7	6	6	8	6	6	6	6	6
6	9	7	6	9	6	6	6	6	6	9	7	6	9	6	6	6	6	6
7	10	6	6	8	7	6	6	6	6	10	6	6	8	7	6	6	6	6
8	7	7	6	10	6	6	6	6	7	7	7	6	10	6	6	6	6	7
9	9	6	6	9	7	6	6	6	6	9	6	6	9	7	6	6	6	6
10	8	6	6	8	6	6	6	7	6	8	6	6	8	6	6	6	7	6
11	10	6	7	9	6	7	7	6	6	10	6	7	9	6	7	7	6	6
12	9	6	6	8	6	6	6	6	7	9	6	6	8	6	6	6	6	7
13	10	6	6	9	6	6	6	6	6	10	6	6	9	6	6	6	6	6
14	8	6	7	8	6	7	7	7	6	8	6	7	8	6	7	7	7	6
15	9	6	6	8	6	6	6	6	6	9	6	6	8	6	6	6	6	6
16	8	6	6	8	6	6	6	6	6	8	6	6	8	6	6	6	6	6
17	9	6	6	9	6	6	6	6	6	9	6	6	9	6	6	6	6	6
18	10	6	6	9	6	6	6	6	6	10	6	6	9	6	6	6	6	6
19	11	6	6	6	6	6	6	6	6	11	6	6	6	6	6	6	6	6
20	9	6	6	9	6	6	6	6	6	9	6	6	9	6	6	6	6	6

Graphically, as it can be seen from the top 100 models sampled (see Figure 6.5 -6.6), each prior's results, does not seem to alter between the two alternatives of initial sampling probabilities. AIC and ZS-Full, tend to additionally support the selection of the subset of non significant covariates $\{x6, x11, x14\}$. Moreover seem more confident concerning the identification of candidate $\{x30\}$, as significant one (selected in all 100 models).

Figure 6.5: Top 100 models sampled (Initial Probabilities: Uniform)

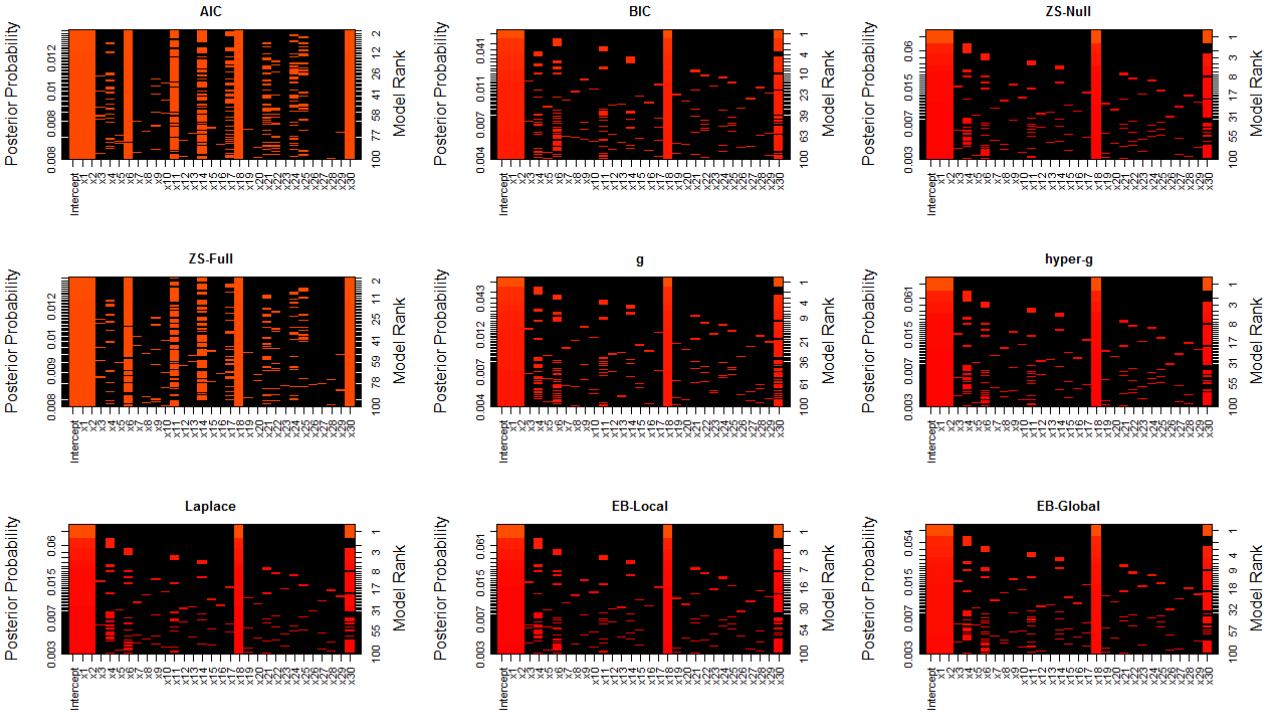
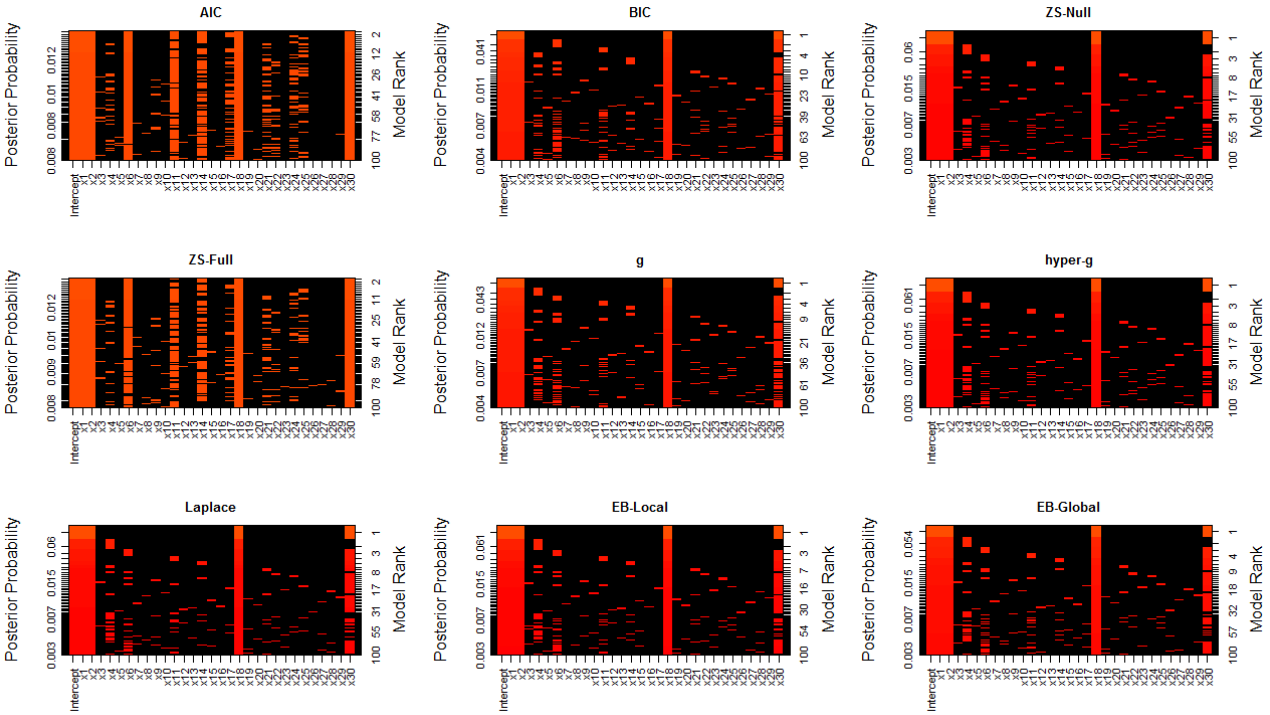


Figure 6.6: Top 100 models sampled (P-Value calibration)



As mentioned in the previous paragraph, the results, are not affected by the two alternative choices of initial sampling probabilities. The main difference between the two methods lies in the reduction of the number of updates that the p-value calibration needs to reach the final model. In particular, in most cases, the latter required almost half updates to select the final model. Interestingly, under AIC, g-prior and EB-Global prior required a larger number of updates.

Table 6.4: Marginal Inclusion Probabilities

	Initial probabilities: Uniform									Initial probailities: p-value calibration								
	AIC	**BIC**	**ZS-Null**	**ZS-Full**	**g-prior**	**hyper-g**	**Laplace**	**EB-Local**	**EB-Global**	**AIC**	**BIC**	**ZS-Null**	**ZS-Full**	**g-prior**	**hyper-g**	**Laplace**	**EB-Local**	**EB-Global**
const	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,000**
x1*	**1,000**	**0,998**	**0,996**	**0,999**	**0,997**	**0,996**	**0,996**	**0,996**	**0,996**	**0,9998**	**0,9978**	**0,9954**	**0,9995**	**0,9961**	**0,9955**	**0,9956**	**0,9954**	**0,9959**
x2*	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**
x3	0,172	0,085	0,076	0,180	0,086	0,075	0,076	0,075	0,076	0,1815	0,0906	0,0752	0,1063	0,0823	0,0767	0,0778	0,0752	0,0760
x4	0,322	0,314	0,305	0,350	0,314	0,315	0,312	0,306	0,312	0,3308	0,2735	0,3041	0,2284	0,3323	0,3122	0,3098	0,3203	0,3090
x5	0,105	0,067	0,056	0,183	0,064	0,060	0,057	0,057	0,059	0,1257	0,0626	0,0573	0,0765	0,0641	0,0584	0,0576	0,0576	0,0572
x6	**0,906**	0,340	0,256	**0,793**	0,288	0,253	0,256	0,256	0,256	**0,9555**	0,3493	0,2541	**0,8296**	0,2875	0,2535	0,2580	0,2586	0,2590
x7	0,094	0,066	0,057	0,158	0,064	0,058	0,058	0,057	0,058	0,1085	0,0647	0,0572	0,0967	0,0639	0,0579	0,0583	0,0577	0,0572
x8	0,083	0,061	0,057	0,062	0,063	0,058	0,057	0,058	0,057	0,1027	0,0652	0,0574	0,0630	0,0641	0,0578	0,0575	0,0568	0,0574
x9	0,170	0,093	0,084	0,165	0,093	0,084	0,084	0,082	0,083	0,1833	0,0939	0,0830	0,1669	0,0938	0,0834	0,0824	0,0830	0,0832
x10	0,151	0,071	0,060	0,081	0,068	0,061	0,060	0,061	0,062	0,1297	0,0684	0,0611	0,1479	0,0680	0,0625	0,0614	0,0619	0,0613
x11	**0,840**	0,262	0,198	**0,604**	0,229	0,198	0,197	0,201	0,201	**0,8732**	0,2653	0,1988	**0,6955**	0,2236	0,2011	0,2017	0,1969	0,2017
x12	0,101	0,068	0,060	0,113	0,069	0,062	0,060	0,061	0,061	0,1043	0,0653	0,0587	0,1188	0,0682	0,0612	0,0620	0,0610	0,0605
x13	0,096	0,059	0,056	0,075	0,061	0,055	0,055	0,056	0,056	0,0652	0,0616	0,0552	0,1103	0,0611	0,0564	0,0558	0,0549	0,0549
x14	**0,554**	0,209	0,164	**0,518**	0,185	0,164	0,165	0,164	0,162	**0,5542**	0,2202	0,1619	**0,5425**	0,1875	0,1612	0,1650	0,1671	0,1638
x15	0,121	0,064	0,059	0,160	0,065	0,058	0,059	0,059	0,058	0,0906	0,0642	0,0583	0,1306	0,0643	0,0591	0,0590	0,0571	0,0582
x16	0,095	0,062	0,055	0,086	0,062	0,055	0,056	0,055	0,055	0,0727	0,0604	0,0546	0,0768	0,0609	0,0550	0,0557	0,0553	0,0548
x17	0,366	0,086	0,072	0,239	0,085	0,072	0,072	0,072	0,073	0,3737	0,0925	0,0721	0,3146	0,0787	0,0722	0,0727	0,0727	0,0736
x18*	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**	**1,0000**
x19	0,063	0,063	0,058	0,069	0,064	0,059	0,058	0,058	0,058	0,0912	0,0654	0,0583	0,0932	0,0631	0,0578	0,0589	0,0572	0,0579
x20	0,085	0,060	0,056	0,063	0,063	0,055	0,057	0,056	0,056	0,1109	0,0612	0,0565	0,0744	0,0616	0,0565	0,0565	0,0559	0,0556
x21	0,501	0,149	0,116	0,281	0,139	0,119	0,117	0,117	0,120	0,5032	0,1374	0,1174	0,3908	0,1398	0,1200	0,1176	0,1164	0,1191
x22	0,293	0,106	0,089	0,170	0,104	0,093	0,092	0,090	0,092	0,2428	0,1066	0,0899	0,2411	0,1028	0,0917	0,0904	0,0936	0,0903
x23	0,074	0,065	0,059	0,068	0,066	0,060	0,061	0,060	0,061	0,0578	0,0697	0,0596	0,1297	0,0686	0,0610	0,0602	0,0592	0,0607
x24	0,428	0,108	0,092	0,207	0,101	0,089	0,092	0,091	0,089	0,4507	0,1149	0,0902	0,3240	0,1015	0,0897	0,0918	0,0874	0,0923
x25	0,351	0,100	0,083	0,231	0,095	0,084	0,085	0,083	0,084	0,3998	0,1011	0,0835	0,2947	0,0988	0,0839	0,0848	0,0829	0,0864
x26	0,058	0,057	0,053	0,055	0,059	0,054	0,056	0,053	0,054	0,1095	0,0590	0,0531	0,0535	0,0594	0,0536	0,0543	0,0536	0,0538
x27	0,076	0,065	0,062	0,098	0,068	0,062	0,062	0,062	0,063	0,1157	0,0662	0,0633	0,0976	0,0694	0,0629	0,0620	0,0622	0,0618
x28	0,110	0,067	0,061	0,131	0,067	0,061	0,062	0,061	0,061	0,0781	0,0682	0,0606	0,1026	0,0674	0,0610	0,0617	0,0594	0,0607
x29	0,065	0,065	0,059	0,117	0,064	0,059	0,058	0,057	0,059	0,0881	0,0636	0,0591	0,0957	0,0656	0,0580	0,0580	0,0572	0,0580
x30*	**0,989**	**0,827**	**0,798**	**0,981**	**0,808**	**0,788**	**0,792**	**0,798**	**0,791**	**0,9892**	**0,8655**	**0,7981**	**0,9901**	**0,7903**	**0,7913**	**0,7942**	**0,7846**	**0,7939**
Dimension	8	5	5	8	5	5	5	5	5	8	5	5	8	5	5	5	5	5
posterior probability	0,001	0,025	0,057	0,001	0,030	0,056	0,056	0,057	0,049	0,001	0,026	0,057	0,001	0,030	0,056	0,056	0,057	0,049
R2	0,920	0,912	0,912	0,920	0,912	0,912	0,912	0,912	0,912	0,920	0,912	0,912	0,920	0,912	0,912	0,912	0,912	0,912
In sample APE (HPM)	0,946	0,990	0,990	0,946	0,990	0,990	0,990	0,990	0,990	0,946	0,990	0,990	0,946	0,990	0,990	0,990	0,990	0,990
In sample APE (BMA)	0,935	0,974	0,976	0,942	0,975	0,976	0,976	0,976	0,975	0,935	0,974	0,976	0,942	0,975	0,976	0,976	0,976	0,975
Out of sample APE (HPM)	1,080	1,039	1,040	1,080	1,043	1,041	1,041	1,040	1,041	1,080	1,039	1,040	1,080	1,043	1,041	1,041	1,040	1,041
Out of sample APE (BMA)	1,058	1,042	1,046	1,050	1,048	1,046	1,046	1,046	1,046	1,058	1,042	1,046	1,050	1,048	1,046	1,046	1,046	1,046

Dimension: Dimension of HPM

posterior probability: Posterior Probability of HPM

R2: R2 of HPM

In sample APE (HPM): In sample average prediction error under HPM

In sample APE (BMA): In sample average prediction error under BMA

Out of sample APE (HPM): Out of sample average prediction error under HPM

Out of sample APE (BMA): Out of sample average prediction error under BMA

Both AIC and ZS-Full method, are related to lower posterior probability of selecting the HPM and the lowest in sample prediction error; both under MAP and BMA. Considering the out of

sample prediction error, the lowest values were calculated under the BIC prior. Finally, regarding the marginal inclusion probabilities the main difference lies in candidate $\{X_{30}\}$, for which both AIC and ZS-Full prior supports higher values and equal to 1, while the rest of the methods support its selection with values approximately 0.8. (Summary results regarding the coefficients estimates can be found in Appendix B.)

Consistency of results

Similarly to the first example we repeated the algorithm for 100 times for each prior, in order to examine the stability of results. Due to lack of computational power and restricted memory capacity, we adopted a different approach regarding the number of sampled models that , however approaches the one applied above. In particular, we performed the algorithm by reducing the number of iterations in 2^{10}, but we increased the number of updates, by choosing a step of updating every 100 iterations.

Each candidate included in the true model was sampled as a significant one for over than 90 samples, with an average marginal inclusion probability to be approximately equal to one, independently of the initial sampling probability choice. As in the full enumeration example, AIC and ZS-Full method, both under Uniform initial sampling probabilities and p-value calibration, tend to include insignificant candidates more times in the MAP model, tend to select models of significantly higher dimension and as it can be seen in graph 6 and 7, the distribution of candidates' marginal posterior probabilities appear to be of greater variance. Regarding the subset of true candidates, all methods, independently from the prior set up, select $\{x_{2,}x_{18}\}$, with marginal inclusion probability equal to one and with zero variance. On the other hand, the marginal inclusion probability distribution of the subset $\{x_{1,}x_{30}\}$, displays a significant number of low extreme values, indicating a slight instability in selecting candidates of higher correlations or lower effects. Especially for the subset $\{x_{1,}x_{27}\}$, which is constructed to be highly correlated, the estimates of their posterior means appear to be of considerably greater variance (approximately 0.5), something that is also depicted in graph 6.9 and 6.10. (Summary posterior tables are provided in Appendix C)

Table 6.5: Number of inclusion for each variable

	Initial probabilities: Uniform									Initial probailities: p-value calibration								
	AIC	BIC	g-prior	ZS-Null	ZS-Full	hyper-g	Laplace	EB-Local	EB-Global	AIC	BIC	g-prior	ZS-Null	ZS-Full	hyper-g	Laplace	EB-Local	EB-Global
x1*	97	95	98	94	97	96	93	94	93	98	95	93	95	96	96	96	95	95
x2*	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
x3	28	4	3	2	20	1	0	1	1	25	5	3	1	22	2	2	2	1
x4	15	8	5	3	15	4	5	3	4	19	8	6	5	15	4	4	4	5
x5	16	4	2	2	12	3	3	3	4	15	5	4	4	12	4	4	3	4
x6	21	6	3	4	18	4	4	3	4	18	6	5	5	15	4	5	5	5
x7	20	7	6	5	17	6	6	6	7	20	9	6	5	17	5	6	7	5
x8	24	7	6	6	19	4	6	4	4	24	9	6	5	17	6	4	4	4
x9	21	3	2	1	14	1	1	2	1	19	5	1	1	15	1	1	1	1
x10	18	3	2	2	17	2	2	1	2	17	5	2	2	12	3	2	3	2
x11	23	4	1	2	14	2	2	2	3	22	6	3	2	16	2	2	3	4
x12	22	5	4	3	18	3	4	3	3	22	6	4	3	17	3	3	3	3
x13	25	2	0	0	21	1	2	1	1	25	3	3	0	18	1	2	1	0
x14	21	7	5	5	15	4	4	3	4	19	8	7	7	16	5	5	5	6
x15	24	7	5	3	23	4	4	5	4	24	9	5	5	17	5	5	4	5
x16	15	6	4	5	14	4	4	4	5	16	6	5	5	9	5	4	4	5
x17	30	6	3	3	29	3	4	3	3	30	8	3	3	26	3	3	3	3
x18*	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
x19	20	2	1	1	16	1	1	2	1	20	4	1	1	14	1	1	1	1
x20	24	4	1	1	13	2	1	1	2	19	7	3	2	13	2	3	2	2
x21	29	2	2	2	23	2	2	2	2	27	3	2	2	20	2	2	2	2
x22	21	5	5	2	14	3	5	4	3	18	6	5	4	12	3	4	4	5
x23	17	2	1	2	11	1	2	1	2	16	2	2	1	12	2	2	2	2
x24	29	3	2	1	20	2	1	2	1	27	3	2	2	23	3	2	2	2
x25	23	3	3	3	19	2	3	4	2	26	5	4	3	17	2	4	3	2
x26	24	3	0	0	16	0	1	1	1	17	4	0	1	19	1	1	0	1
x27	19	10	8	8	17	5	9	8	9	22	9	9	7	15	7	8	9	8
x28	25	5	2	2	14	2	2	2	3	21	4	3	2	17	3	3	3	3
x29	21	3	3	3	16	3	3	3	3	16	3	3	3	14	3	3	3	3
x30*	98	95	94	96	96	95	95	96	95	98	95	95	95	97	95	96	95	96
mpp	0,016	0,060	0,067	0,091	0,016	0,089	0,088	0,089	0,084	0,010	0,055	0,061	0,084	0,009	0,083	0,082	0,085	0,079
mpp sd	0,0049	0,021	0,023	0,037	0,006	0,036	0,035	0,035	0,030	0,002	0,019	0,021	0,033	0,002	0,034	0,034	0,035	0,033
median dim	10	6	5	5	9	5	5	5	5	10	6	5	5	9	5	5	5	5
dim sd	2,275	1,338	1,140	0,925	1,977	1,018	1,065	1,029	1,092	2,273	1,361	1,136	1,037	2,014	1,009	1,056	1,046	1,041

mpp: Average Model's Poterior Probability

mpp sd: Std Deviation of Average Model's Posterior Probability

median dim: HPM Median Dimension

dim sd: Std Deviation of HPM Average Dimension

Figure 6.7: Marginal Inclusion Probabilities (Initial Probabilities: Uniform)

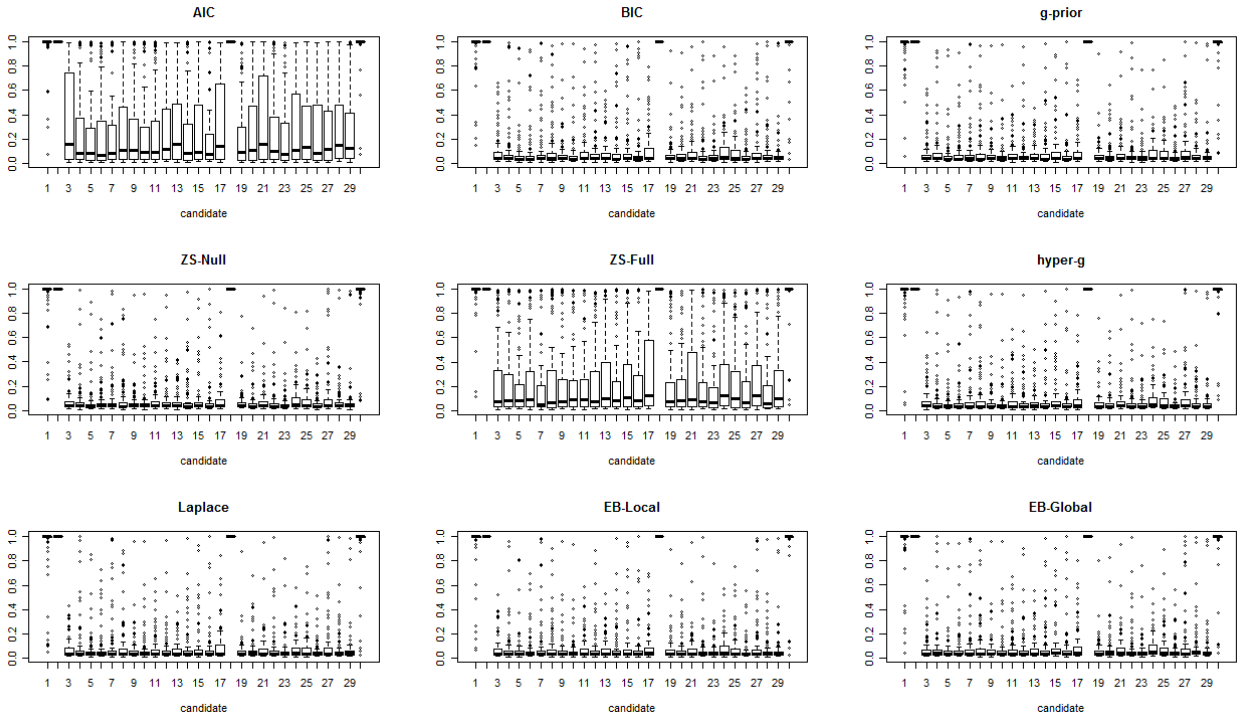


Figure 6.8: Marginal Inclusion Probabilities (P-Value calibration)

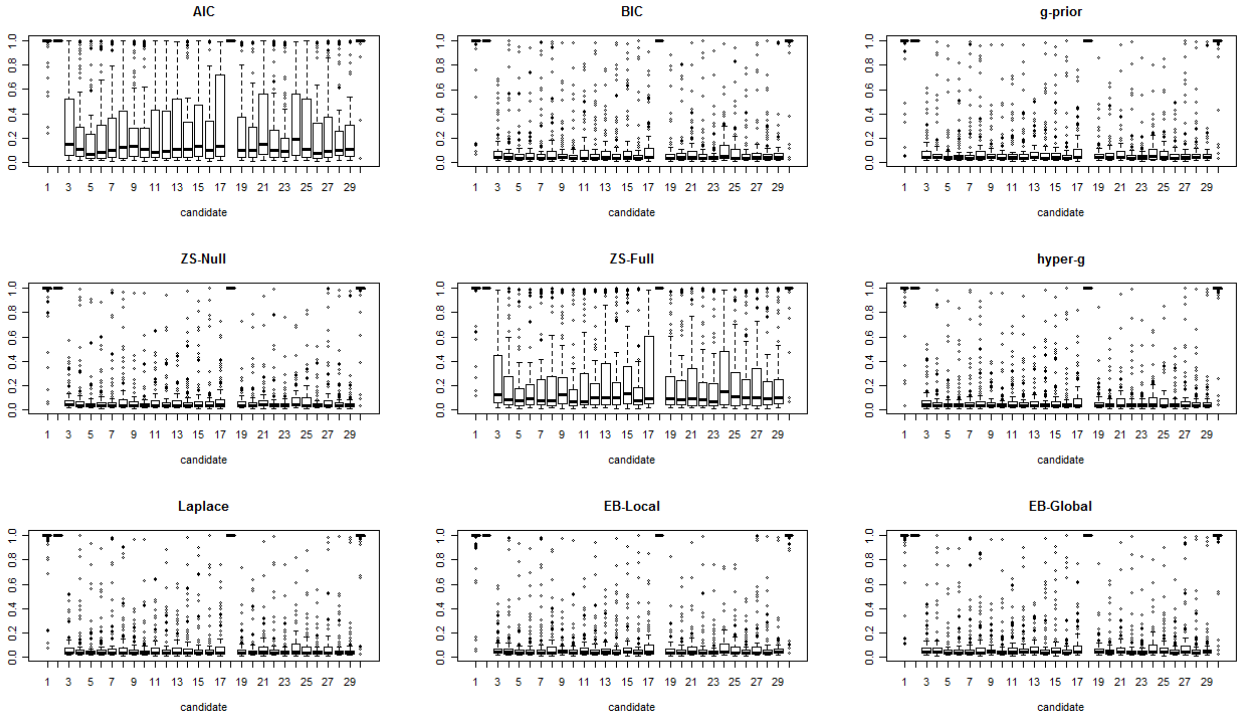


Figure 6.9: BMA Posterior Means of coefficients (Initial Probabilities: Uniform)

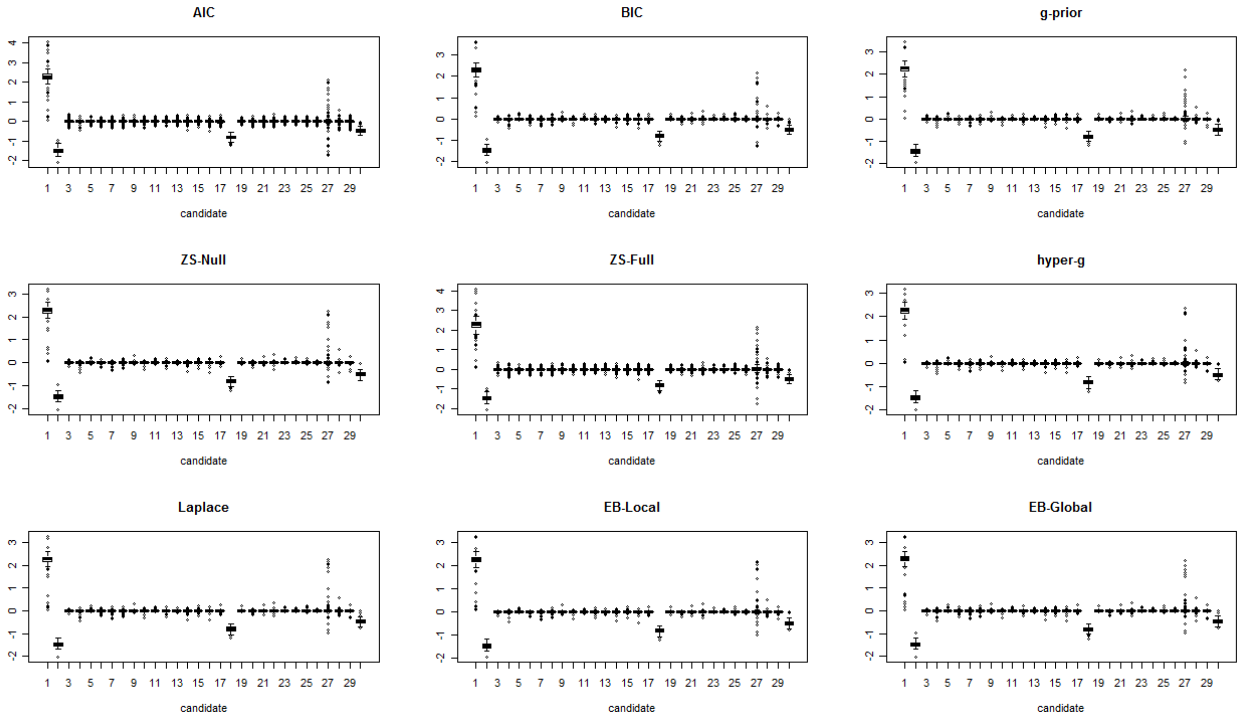
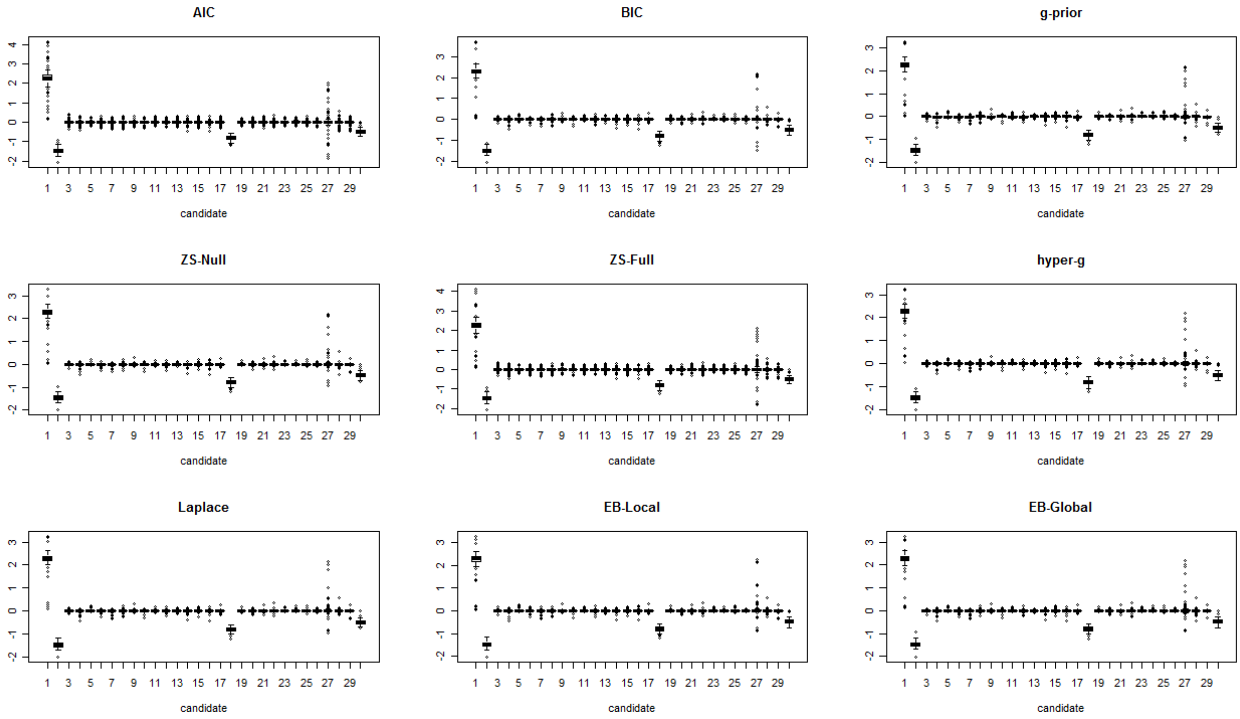


Figure 6.10: BMA Posterior Means of coefficients (P-Value calibration)



Finally, in Table 6.6 and 6.7 we present the in sample and out of sample average prediction error, both under HPM and BMA. A first remark is that there is no obvious difference in APE between the initial sampling probability choice. As it can be noticed, BMA provides lower average APE, both in sample and out of sample. Looking deeper in differences between prior set up, AIC prior and ZS-Full method, provides with the lowest average in sample APE (<0.9), while under all other priors the latter fluctuates around 0.95.

Table 6.6: In sample APE

	HPM				**BMA**			
	Uniform		**P-value Calibration**		**Uniform**		**P-value Calibration**	
	Average	Std Deviation	Average	Std Deviation	Average	Std Deviation	Average	Std Deviation
AIC	0,881	0,075	0,881	0,076	0,875	0,074	0,874	0,075
BIC	0,938	0,080	0,938	0,080	0,923	0,077	0,923	0,077
g-prior	0,946	0,079	0,946	0,079	0,930	0,074	0,929	0,075
ZS-Null	0,951	0,078	0,950	0,078	0,934	0,074	0,933	0,074
ZS-Full	0,894	0,076	0,890	0,075	0,886	0,075	0,883	0,074
hyper-g	0,950	0,078	0,951	0,078	0,934	0,074	0,934	0,074
Laplace	0,950	0,078	0,950	0,078	0,934	0,074	0,933	0,074
EB-L	0,951	0,078	0,950	0,078	0,934	0,074	0,933	0,074
EB-G	0,948	0,078	0,949	0,078	0,932	0,074	0,932	0,074

On the other hand, average out of sample APE seems to follow a completely opposite pattern. AIC and ZS-Full priors are associated to higher out of sample APE (around 1.13). The corresponding average for all other priors does not exceed the value of 1.075, the lowest value of which is provided by Local Empirical Bayes method.

Table 6.7: Out of sample APE

	HPM				**BMA**			
	Uniform		**P-value Calibreation**		**Uniform**		**P-value Calibreation**	
	Average	Std Deviation	Average	Std Deviation	Average	Std Deviation	Average	Std Deviation
AIC	1,145	0,131	1,144	0,133	1,138	0,131	1,139	0,130
BIC	1,075	0,134	1,077	0,135	1,071	0,129	1,072	0,131
g-prior	1,067	0,131	1,067	0,130	1,062	0,128	1,063	0,128
ZS-Null	1,061	0,129	1,062	0,130	1,058	0,128	1,059	0,127
ZS-Full	1,130	0,128	1,138	0,128	1,123	0,131	1,128	0,128
hyper-g	1,063	0,129	1,063	0,131	1,059	0,128	1,059	0,128
Laplace	1,064	0,130	1,063	0,130	1,061	0,129	1,059	0,128
EB-L	1,063	0,128	1,063	0,129	1,060	0,128	1,060	0,128
EB-G	1,064	0,132	1,064	0,130	1,061	0,128	1,060	0,128

Chapter 7: Discussion-Further Research

7.1 Conclusion

In the current thesis we attempted a review of basic concepts and tools for Bayesian model selection, focusing on the Bayesian adaptive sampling algorithm of Clyde et al (2011). Differences between classical and Bayesian approaches were presented, while more focus is given in the Bayesian variable selection methods. Bayesian adaptive sampling (BAS) of Clyde et al. (2011) was fully reviewed explaining the key difference between the sampling strategies of traditional MCMC algorithms and its performance under different kinds of priors was explored.

For the small sample case, where full enumeration is allowed, Bayesian model selection using several priors was applied both in one sample of 100 observations and 100 samples in order to explore the stability of results for each prior. The general picture obtained is that under AIC and ZS-Full prior, the models selected are overfitted. Under AIC, the true model was selected only on 30% of cases, while for ZS-Full the corresponding rate was increased at 60%. The rest of the methods using other prior schemes, identified the true model in 90% of cases. Naturally, AIC and ZS-Full priors are related to higher marginal inclusion probabilities for zero coefficients. For AIC, the average inclusion probability for non zero coefficients was 40%, for ZS-Full prior was 20%, while for other priors was 10%. Following this result, AIC and ZS-Full methods are more confident in selecting non-zero coefficients of lower values. Finally, on average, AIC and ZS-Full prior was related to lower in-sample APE and out-of-sample APE under BMA, but greater out-of-sample APE under MAP

For the large sample case, were sampling is required, we performed the algorithm in a similar way as above. By controlling for the number of iterations and the updating step, the performance of each prior was explored, using both initial sampling probabilities and p-value calibration. A first conclusion reached, is that the algorithm performed almost identically irrespective from the initial probabilities set up. Naturally, the only difference lied in the fact that under p-value calibration, the algorithm converged faster. Regarding the performance of each prior, the results did not differentiate much in comparison to the small sample case. The main difference observed was the incapability of the AIC and ZS-Full method to detect the true model in all cases. The latter consistently selected models of higher dimension as the Highest probability model. Finally, a last remark that observed is related to the APE. In particular, while for AIC and ZS-Full prior in-sample APE was lower, out-of-sample APE for both methods, was greater.

7.2 Further Research

Apart from the cases described in the current dissertation, the performance of BAS needs to be explored further, especially in case of large samples, where sampling is required. The optimal number of iterations that provides with trustworthy results is a field that could easily be identified, while the effect of the updating step could be examined in detail for each prior set up. A second point that worth examining in detail, is the ability of the algorithm to detect non zero coefficient candidates of different values for small effects, for instance below 0.6. In the current thesis we observed that for a value of 0.5 the candidate was supported in the highest probability model as significant one under AIC and ZS-Full prior, however more detailed simulation studies for such cases could provide a deeper insight and a more detailed review for each prior. Similarly, detailed simulation studies could provide with results for the performance of the algorithm in cases of different values in correlated data. By doing so, for each prior, it could be explored under which cases the algorithm is able to distinguish a true candidate from a correlated one. Moreover, the algorithm should be tested in different kinds and of varying complexity real data.

Apart from deepening in its performance, BAS should be examined in comparison to existing Bayesian algorithms that are already widely applicable. Its extension in generalized linear models should be deeply explored both in simulated and real data. Variants of g-priors for GLM that have been introduced lately in the literature (Fouskakis et al., 2009, Gupta and Ibrahim, 2009, Bove & Held, 2009, Hanson et al., 2014) might be explored in conjuction with BAS. The large p small n problem, could also be an area, in which BAS could be examined, adopting for instance the proposed generalization of Zellner's g-prior of Maruyama and George (2011), that allows for p>n. Finally, non-local priors, introduced by Johnson & Rossell (2010) and adopted for Bayesian variable selection in high dimensional problems (Johnson and Rossell, 2012) could be an alternative to the current used family of priors and could be used to extend BAS.

Appendix A: Full Enumeration (One Sample)

Table A1: marginal posterior inclusion probabilities

		AIC	BIC	ZS-Null	ZS-Full	g-prior (g=n)	hyper-g (a=3)	hyper-g Laplace (a=3)	EB Local	EB Global
	Intercept	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	x1*	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	x2*	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
	x3	0,323	0,118	0,085	0,224	0,114	0,084	0,085	0,083	0,079
	x4	0,284	0,097	0,070	0,194	0,096	0,070	0,070	0,069	0,065
	x5	0,403	0,158	0,113	0,283	0,150	0,112	0,113	0,111	0,106
	x6	0,358	0,135	0,096	0,249	0,129	0,096	0,097	0,095	0,091
	x7	0,270	0,091	0,066	0,185	0,091	0,066	0,066	0,065	0,061
	x8	0,270	0,091	0,066	0,185	0,091	0,066	0,066	0,065	0,061
	x9*	1,000	0,998	0,996	0,998	0,996	0,996	0,996	0,996	0,996
	x10	0,280	0,096	0,069	0,191	0,095	0,069	0,069	0,068	0,064
	model's posterior probability	0,070	0,431	0,559	0,167	0,441	0,561	0,559	0,566	0,573
	log marginal likelihood	-239,484	-244,694	101,092	11,716	101,426	99,676	99,634	102,471	102,450
in sample	HPM	1,054	1,054	1,054	1,054	1,054	1,054	1,054	1,054	1,054
APE	BMA	1,047	1,051	1,051	1,049	1,051	1,052	1,051	1,051	1,055
out of sample	HPM	0,925	0,925	0,927	0,925	0,930	0,927	0,927	0,927	0,927
APE	BMA	0,935	0,928	0,930	0,930	0,934	0,930	0,930	0,929	0,929

Table A2: marginal posterior means of coefficients under BMA

	AIC	BIC	ZS-Null	ZS-Full	g-prior (g=n)	hyper-g (a=3)	hyper-g Laplace (a=3)	EB Local	EB Global	true value
Intercept	3,972	3,972	3,972	3,972	3,972	3,972	3,972	3,972	3,972	4,000
x1*	3,310	3,312	3,296	3,311	3,280	3,294	3,294	3,299	3,299	3,200
x2*	-1,093	-1,090	-1,084	-1,092	-1,079	-1,084	-1,084	-1,085	-1,085	-1,050
x3	-0,025	-0,010	-0,007	-0,018	-0,009	-0,007	-0,007	-0,007	-0,006	0,000
x4	0,013	0,004	0,003	0,009	0,004	0,003	0,003	0,003	0,003	0,000
x5	-0,044	-0,017	-0,012	-0,031	-0,016	-0,012	-0,012	-0,012	-0,012	0,000
x6	0,035	0,014	0,010	0,025	0,013	0,010	0,010	0,010	0,009	0,000
x7	0,001	-0,001	-0,001	0,000	-0,001	-0,001	-0,001	-0,001	-0,001	0,000
x8	0,001	0,001	0,000	0,001	0,001	0,000	0,000	0,000	0,000	0,000
x9*	0,541	0,530	0,524	0,535	0,523	0,523	0,523	0,524	0,524	0,500
x10	-0,011	-0,004	-0,003	-0,007	-0,004	-0,003	-0,003	-0,003	-0,002	0,000

Table A3: marginal posterior standard deviations of coefficients under BMA

	AIC	BIC	ZS-Null	ZS-Full	g-prior (g=n)	hyper-g (a=3)	hyper-g Laplace (a=3)	EB Local	EB Global
Intercept	0,108	0,108	0,108	0,108	0,108	0,108	0,108	0,108	0,108
x1*	0,122	0,124	0,123	0,122	0,123	0,122	0,122	0,122	0,122
x2*	0,112	0,114	0,113	0,112	0,113	0,112	0,112	0,112	0,112
x3	0,038	0,073	0,046	0,039	0,062	0,045	0,039	0,039	0,039
x4	0,033	0,069	0,040	0,034	0,056	0,039	0,034	0,034	0,033
x5	0,047	0,084	0,057	0,048	0,073	0,055	0,048	0,048	0,048
x6	0,044	0,082	0,054	0,046	0,071	0,052	0,045	0,046	0,045
x7	0,027	0,057	0,033	0,028	0,047	0,032	0,028	0,028	0,027
x8	0,025	0,054	0,031	0,026	0,044	0,031	0,026	0,026	0,026
x9*	0,128	0,127	0,126	0,128	0,127	0,128	0,128	0,128	0,128
x10	0,032	0,066	0,039	0,033	0,055	0,038	0,033	0,033	0,033

Table B1: Marginal Posterior means

		Initial probabilities: Uniform									Initial probailities: p-value calibration								
	TRUE	**AIC**	**BIC**	**ZS-Null**	**ZS-Full**	**g-prior**	**hyper-g**	**Laplace**	**EB-L**	**EB-G**	**AIC**	**BIC**	**ZS-Null**	**ZS-Full**	**g-prior**	**hyper-g**	**Laplace**	**EB-L**	**EB-G**
const	3.2	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6	3,6
x1*	2.3	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2
x2*	-1,5	-1,6	-1,6	-1,6	-1,6	-1,6	-1,6	-1,6	-1,6	-1,6	-1,6	-1,6	-1,6	-1,6	-1,6	-1,6	-1,6	-1,6	-1,6
x3	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x4	0	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	-0,1	0,0	-0,1	-0,1	-0,1	-0,1	-0,1
x5	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x6	0	-0,2	-0,1	-0,1	-0,2	-0,1	-0,1	-0,1	-0,1	-0,1	-0,2	-0,1	-0,1	-0,2	-0,1	-0,1	-0,1	-0,1	-0,1
x7	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x8	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x9	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x10	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x11	0	0,2	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,2	0,1	0,0	0,1	0,0	0,0	0,0	0,0	0,0
x12	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x13	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x14	0	0,1	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0
x15	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x16	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x17	0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x18*	-0,8	-0,9	-0,8	-0,8	-0,9	-0,8	-0,8	-0,8	-0,8	-0,8	-0,9	-0,8	-0,8	-0,9	-0,8	-0,8	-0,8	-0,8	-0,8
x19	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x20	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x21	0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0
x22	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x23	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x24	0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x25	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,1	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x26	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x27	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x28	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x29	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x30*	-0,48	-0,3	-0,3	-0,3	-0,3	-0,3	-0,3	-0,3	-0,3	-0,3	-0,3	-0,3	-0,3	-0,4	-0,3	-0,3	-0,3	-0,3	-0,3

Table B2:Standard Deviation of Marginal Posterior means

	Initial probabilities: Uniform									Initial probailities: p-value calibration								
	AIC	BIC	ZS-Null	ZS-Full	g-prior	hyper-g	Laplace	EB-L	EB-G	AIC	BIC	ZS-Null	ZS-Full	g-prior	hyper-g	Laplace	EB-L	EB-G
const	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10	0,10
x1*	**0,28**	**0,26**	**0,28**	**0,30**	**0,28**	**0,28**	**0,28**	**0,27**	**0,28**	**0,33**	**0,27**	**0,28**	**0,31**	**0,28**	**0,28**	**0,28**	**0,28**	**0,28**
x2*	**0,11**	**0,11**	**0,11**	**0,12**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**
x3	0,05	0,04	0,04	0,05	0,04	0,04	0,04	0,04	0,04	0,05	0,04	0,04	0,04	0,04	0,04	0,04	0,04	0,04
x4	0,12	0,17	0,17	0,13	0,17	0,17	0,17	0,17	0,17	0,12	0,16	0,17	0,11	0,17	0,17	0,17	0,18	0,17
x5	0,04	0,03	0,03	0,05	0,03	0,03	0,03	0,03	0,03	0,04	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03
x6	0,13	0,12	0,11	0,14	0,11	0,11	0,11	0,11	0,11	0,12	0,12	0,11	0,14	0,11	0,11	0,11	0,11	0,11
x7	0,04	0,03	0,03	0,05	0,03	0,03	0,03	0,03	0,03	0,04	0,03	0,03	0,04	0,03	0,03	0,03	0,03	0,03
x8	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03
x9	0,05	0,04	0,04	0,05	0,04	0,04	0,04	0,04	0,04	0,05	0,04	0,04	0,05	0,04	0,04	0,04	0,04	0,04
x10	0,05	0,04	0,03	0,04	0,04	0,03	0,03	0,03	0,03	0,05	0,04	0,03	0,05	0,04	0,03	0,03	0,03	0,03
x11	0,13	0,10	0,09	0,13	0,09	0,09	0,09	0,09	0,09	0,12	0,10	0,09	0,13	0,09	0,09	0,09	0,09	0,09
x12	0,04	0,03	0,03	0,04	0,03	0,03	0,03	0,03	0,03	0,04	0,03	0,03	0,04	0,03	0,03	0,03	0,03	0,03
x13	0,04	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,04	0,03	0,03	0,03	0,03	0,03
x14	0,10	0,07	0,07	0,10	0,07	0,07	0,07	0,07	0,07	0,10	0,08	0,07	0,11	0,07	0,07	0,07	0,07	0,07
x15	0,04	0,03	0,03	0,05	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,04	0,03	0,03	0,03	0,03	0,03
x16	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03
x17	0,10	0,04	0,04	0,08	0,04	0,04	0,04	0,04	0,04	0,10	0,05	0,04	0,09	0,04	0,04	0,04	0,04	0,04
x18*	**0,12**	**0,11**	**0,11**	**0,12**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**	**0,12**	**0,11**	**0,11**	**0,11**	**0,11**	**0,11**
x19	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03
x20	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,04	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03
x21	0,11	0,07	0,06	0,09	0,07	0,06	0,06	0,06	0,06	0,11	0,07	0,06	0,10	0,07	0,06	0,06	0,06	0,06
x22	0,08	0,05	0,05	0,06	0,05	0,05	0,05	0,05	0,05	0,07	0,05	0,05	0,07	0,05	0,05	0,05	0,05	0,05
x23	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,04	0,03	0,03	0,03	0,03	0,03
x24	0,10	0,05	0,05	0,07	0,05	0,05	0,05	0,05	0,05	0,10	0,05	0,05	0,09	0,05	0,05	0,05	0,05	0,05
x25	0,08	0,04	0,04	0,07	0,04	0,04	0,04	0,04	0,04	0,09	0,04	0,04	0,08	0,04	0,04	0,04	0,04	0,04
x26	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,04	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03
x27	0,26	0,24	0,26	0,28	0,26	0,26	0,26	0,25	0,26	0,32	0,25	0,26	0,29	0,26	0,26	0,26	0,26	0,25
x28	0,06	0,05	0,05	0,07	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,05	0,06	0,05	0,05	0,05	0,05	0,05
x29	0,03	0,03	0,03	0,04	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03	0,03
x30*	**0,13**	**0,17**	**0,18**	**0,13**	**0,18**	**0,18**	**0,18**	**0,18**	**0,18**	**0,13**	**0,16**	**0,18**	**0,12**	**0,18**	**0,18**	**0,18**	**0,18**	**0,18**

Appendix C: Posterior Tables – Adaptive Sampling (100 Samples)

Table C1: Average Marginal inclusion Probabilities (100 samples)

	Initial Probabilities: Uniform								Initial Probabilities: P-value Calibration							
	AIC	BIC	g-prior	ZS-Null	ZS-Full	hyper-g	Laplace	EB-Local	EB-Global	AIC	BIC	g-prior	ZS-Null	ZS-Full	hyper-g	Laplace	EB-Local	EB-Global
x1*	0,97	0,95	0,96	0,94	0,97	0,95	0,93	0,94	0,94	0,97	0,95	0,94	0,95	0,95	0,96	0,96	0,94	0,95
x2*	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x3	0,35	0,10	0,09	0,08	0,25	0,08	0,07	0,08	0,08	0,33	0,11	0,09	0,08	0,29	0,08	0,08	0,08	0,08
x4	0,23	0,12	0,11	0,09	0,22	0,09	0,10	0,09	0,10	0,26	0,11	0,11	0,10	0,22	0,10	0,10	0,10	0,10
x5	0,24	0,09	0,08	0,07	0,19	0,07	0,07	0,07	0,07	0,21	0,09	0,08	0,08	0,19	0,08	0,08	0,08	0,08
x6	0,26	0,10	0,09	0,09	0,25	0,08	0,08	0,08	0,09	0,25	0,10	0,09	0,09	0,21	0,09	0,08	0,09	0,09
x7	0,26	0,12	0,10	0,10	0,21	0,11	0,11	0,11	0,11	0,28	0,13	0,11	0,11	0,23	0,10	0,11	0,11	0,11
x8	0,30	0,12	0,11	0,10	0,24	0,10	0,11	0,10	0,11	0,31	0,13	0,12	0,11	0,24	0,11	0,11	0,11	0,11
x9	0,27	0,09	0,08	0,07	0,22	0,07	0,07	0,07	0,07	0,26	0,10	0,08	0,07	0,23	0,07	0,08	0,07	0,07
x10	0,24	0,08	0,07	0,07	0,22	0,07	0,07	0,07	0,07	0,25	0,09	0,08	0,07	0,19	0,07	0,08	0,07	0,08
x11	0,26	0,10	0,08	0,08	0,22	0,09	0,08	0,09	0,08	0,27	0,11	0,10	0,09	0,24	0,09	0,09	0,09	0,09
x12	0,28	0,11	0,09	0,09	0,23	0,09	0,09	0,09	0,08	0,28	0,11	0,10	0,09	0,23	0,09	0,09	0,09	0,09
x13	0,30	0,09	0,08	0,07	0,25	0,07	0,08	0,07	0,08	0,30	0,10	0,09	0,08	0,25	0,08	0,08	0,08	0,08
x14	0,26	0,12	0,09	0,09	0,21	0,09	0,09	0,09	0,10	0,27	0,12	0,11	0,10	0,23	0,10	0,10	0,10	0,10
x15	0,28	0,12	0,10	0,09	0,27	0,10	0,09	0,10	0,10	0,30	0,13	0,11	0,10	0,27	0,10	0,10	0,10	0,10
x16	0,22	0,10	0,09	0,08	0,21	0,08	0,08	0,08	0,08	0,23	0,09	0,09	0,09	0,18	0,09	0,08	0,09	0,09
x17	0,33	0,12	0,09	0,09	0,30	0,09	0,10	0,09	0,10	0,35	0,13	0,11	0,10	0,29	0,10	0,10	0,10	0,10
x18*	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
x19	0,25	0,08	0,07	0,06	0,21	0,06	0,06	0,07	0,06	0,26	0,09	0,08	0,07	0,22	0,07	0,07	0,07	0,07
x20	0,27	0,09	0,08	0,07	0,20	0,07	0,07	0,07	0,07	0,24	0,09	0,09	0,08	0,21	0,07	0,08	0,07	0,08
x21	0,34	0,09	0,09	0,08	0,27	0,08	0,08	0,07	0,08	0,31	0,10	0,09	0,08	0,25	0,08	0,09	0,08	0,08
x22	0,27	0,10	0,09	0,08	0,20	0,08	0,08	0,08	0,08	0,24	0,10	0,09	0,08	0,20	0,08	0,09	0,08	0,08
x23	0,23	0,07	0,06	0,06	0,17	0,06	0,06	0,06	0,06	0,22	0,08	0,07	0,06	0,19	0,07	0,07	0,07	0,07
x24	0,31	0,11	0,09	0,09	0,28	0,09	0,08	0,08	0,08	0,33	0,11	0,10	0,09	0,29	0,09	0,09	0,09	0,09
x25	0,29	0,10	0,09	0,09	0,24	0,09	0,08	0,09	0,09	0,29	0,11	0,10	0,09	0,24	0,09	0,09	0,09	0,09
x26	0,29	0,08	0,07	0,06	0,21	0,06	0,06	0,06	0,06	0,25	0,09	0,08	0,07	0,22	0,07	0,06	0,07	0,07
x27	0,27	0,14	0,13	0,13	0,26	0,12	0,14	0,13	0,13	0,27	0,13	0,13	0,12	0,25	0,11	0,11	0,13	0,12
x28	0,30	0,10	0,08	0,08	0,20	0,08	0,08	0,08	0,09	0,27	0,10	0,09	0,09	0,23	0,08	0,08	0,08	0,09
x29	0,28	0,09	0,08	0,08	0,24	0,08	0,08	0,08	0,08	0,25	0,09	0,09	0,08	0,22	0,08	0,08	0,09	0,08
x30*	0,98	0,95	0,95	0,96	0,96	0,95	0,95	0,96	0,95	0,98	0,96	0,95	0,95	0,97	0,95	0,95	0,95	0,95

Table C2 : Std Deviation of Average Marginal inclusion Probabilities (100 samples)

	Initial Probabilities: Uniform								Initial Probabilities: P-value Calibration							
	AIC	BIC	g-prior	ZS-Null	ZS-Full	hyper-g	Laplace	EB-L	EB-Gl	AIC	BIC	g-prior	ZS-Null	ZS-Full	hyper-g	Laplace	EB-L	EB-G
x1*	0,14	0,17	0,14	0,20	0,14	0,19	0,22	0,19	0,20	0,13	0,20	0,20	0,18	0,17	0,16	0,17	0,20	0,19
x2*	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
x3	0,38	0,14	0,12	0,11	0,32	0,11	0,09	0,09	0,11	0,36	0,15	0,12	0,10	0,33	0,11	0,11	0,11	0,11
x4	0,29	0,22	0,19	0,15	0,29	0,17	0,19	0,16	0,18	0,32	0,20	0,19	0,19	0,29	0,18	0,18	0,19	0,19
x5	0,31	0,17	0,14	0,13	0,25	0,14	0,13	0,13	0,14	0,29	0,17	0,15	0,14	0,27	0,14	0,14	0,15	0,14
x6	0,35	0,18	0,14	0,14	0,31	0,14	0,13	0,12	0,15	0,33	0,17	0,15	0,15	0,30	0,14	0,14	0,15	0,15
x7	0,35	0,22	0,20	0,20	0,32	0,20	0,21	0,20	0,20	0,35	0,23	0,21	0,20	0,32	0,20	0,21	0,21	0,20
x8	0,35	0,21	0,19	0,17	0,33	0,17	0,18	0,17	0,18	0,36	0,23	0,20	0,19	0,33	0,18	0,19	0,18	0,18
x9	0,33	0,15	0,13	0,11	0,30	0,12	0,11	0,12	0,12	0,31	0,16	0,12	0,12	0,28	0,12	0,12	0,12	0,12
x10	0,31	0,15	0,12	0,13	0,29	0,12	0,13	0,12	0,13	0,32	0,16	0,14	0,13	0,28	0,13	0,13	0,13	0,14
x11	0,35	0,17	0,11	0,12	0,29	0,13	0,11	0,11	0,13	0,34	0,17	0,14	0,13	0,31	0,12	0,13	0,12	0,13
x12	0,34	0,19	0,16	0,15	0,30	0,14	0,15	0,14	0,15	0,34	0,19	0,17	0,15	0,31	0,15	0,15	0,15	0,15
x13	0,33	0,13	0,09	0,09	0,30	0,09	0,10	0,09	0,10	0,35	0,15	0,11	0,10	0,30	0,10	0,11	0,11	0,09
x14	0,35	0,21	0,16	0,16	0,29	0,15	0,15	0,15	0,17	0,34	0,21	0,18	0,17	0,31	0,16	0,16	0,16	0,17
x15	0,34	0,21	0,17	0,16	0,34	0,16	0,16	0,18	0,16	0,34	0,21	0,18	0,17	0,31	0,17	0,17	0,16	0,17
x16	0,31	0,20	0,16	0,17	0,28	0,16	0,15	0,15	0,17	0,29	0,19	0,18	0,18	0,25	0,17	0,16	0,17	0,17
x17	0,37	0,17	0,13	0,14	0,34	0,13	0,14	0,13	0,14	0,37	0,18	0,15	0,14	0,35	0,14	0,14	0,14	0,14
x18*	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
x19	0,32	0,12	0,11	0,09	0,30	0,10	0,10	0,11	0,09	0,31	0,16	0,11	0,09	0,28	0,10	0,09	0,10	0,09
x20	0,32	0,15	0,11	0,10	0,25	0,11	0,11	0,11	0,12	0,30	0,17	0,13	0,12	0,28	0,10	0,12	0,11	0,12
x21	0,36	0,15	0,13	0,12	0,32	0,13	0,12	0,12	0,12	0,33	0,16	0,14	0,13	0,31	0,12	0,13	0,13	0,13
x22	0,33	0,18	0,17	0,15	0,27	0,15	0,15	0,15	0,15	0,31	0,19	0,16	0,16	0,27	0,15	0,16	0,15	0,16
x23	0,30	0,12	0,10	0,11	0,25	0,09	0,11	0,09	0,09	0,30	0,13	0,11	0,09	0,26	0,10	0,10	0,10	0,11
x24	0,36	0,16	0,12	0,11	0,32	0,12	0,09	0,10	0,10	0,34	0,15	0,12	0,12	0,30	0,12	0,12	0,12	0,12
x25	0,33	0,16	0,14	0,13	0,31	0,12	0,13	0,13	0,14	0,34	0,18	0,15	0,14	0,29	0,13	0,14	0,13	0,14
x26	0,35	0,12	0,09	0,07	0,29	0,07	0,08	0,09	0,09	0,33	0,15	0,11	0,09	0,29	0,09	0,09	0,08	0,09
x27	0,33	0,23	0,20	0,22	0,29	0,21	0,24	0,23	0,23	0,33	0,25	0,24	0,21	0,31	0,19	0,20	0,23	0,21
x28	0,34	0,18	0,14	0,14	0,29	0,13	0,14	0,13	0,16	0,35	0,16	0,16	0,15	0,31	0,14	0,14	0,14	0,16
x29	0,33	0,17	0,16	0,16	0,29	0,16	0,15	0,16	0,16	0,31	0,17	0,17	0,16	0,27	0,16	0,16	0,17	0,16
x30*	0,14	0,19	0,19	0,18	0,17	0,19	0,20	0,18	0,19	0,12	0,19	0,19	0,19	0,14	0,19	0,19	0,20	0,19

Table C3: Average Posterior Means (100 samples)

		Initial Probabilities: Uniform									Initial Probabilities: P-value Calibration								
	TRUE	AIC	BIC	g-prior	ZS-Null	ZS-Full	hyper-g	Laplace	EB-L	EB-G	AIC	BIC	g-prior	ZS-Null	ZS-Full	hyper-g	Laplace	EB-L	EB-G
x1*	2,3	2,2	2,2	2,2	2,2	2,2	2,2	2,1	2,2	2,2	2,3	2,2	2,2	2,2	2,2	2,2	2,2	2,2	2,2
x2*	-1,5	-1,5	-1,5	-1,5	-1,5	-1,5	-1,5	-1,5	-1,5	-1,5	-1,5	-1,5	-1,5	-1,5	-1,5	-1,5	-1,5	-1,5	-1,5
x3	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x4	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x5	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x6	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x7	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x8	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x9	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x10	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x11	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x12	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x13	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x14	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x15	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x16	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x17	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x18*	-0,8	-0,8	-0,8	-0,8	-0,8	-0,8	-0,8	-0,8	-0,8	-0,8	-0,8	-0,8	-0,8	-0,8	-0,8	-0,8	-0,8	-0,8	-0,8
x19	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x20	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x21	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x22	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x23	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x24	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x25	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x26	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x27	0	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1
x28	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
x29	0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0

Table C4: Std Deviation of Average Posterior Means (100 samples)

	Initial Probabilities: Uniform									Initial Probabilities: P-value Calibration								
	AIC	BIC	g-prior	ZS-Null	ZS-Full	hyper-g	Laplace	EB-L	EB-G	AIC	BIC	g-prior	ZS-Null	ZS-Full	hyper-g	Laplace	EB-L	EB-G
x1	0,63	0,52	0,44	0,52	0,58	0,49	0,57	0,52	0,52	0,65	0,56	0,54	0,48	0,65	0,45	0,46	0,54	0,5
x2	0,17	0,13	0,12	0,13	0,15	0,12	0,12	0,12	0,13	0,17	0,13	0,13	0,13	0,16	0,12	0,12	0,12	0,13
x3	0,12	0,04	0,03	0,03	0,09	0,03	0,02	0,03	0,03	0,12	0,04	0,03	0,03	0,1	0,03	0,03	0,03	0,03
x4	0,09	0,08	0,07	0,06	0,1	0,07	0,07	0,06	0,07	0,1	0,07	0,07	0,07	0,1	0,07	0,07	0,07	0,07
x5	0,08	0,05	0,04	0,04	0,06	0,04	0,04	0,03	0,04	0,07	0,04	0,04	0,04	0,07	0,04	0,04	0,04	0,04
x6	0,08	0,05	0,04	0,04	0,08	0,04	0,04	0,03	0,04	0,08	0,05	0,04	0,04	0,08	0,04	0,04	0,04	0,04
x7	0,1	0,07	0,06	0,06	0,09	0,07	0,07	0,07	0,07	0,1	0,07	0,07	0,06	0,09	0,06	0,07	0,07	0,07
x8	0,1	0,06	0,05	0,05	0,09	0,05	0,05	0,05	0,05	0,1	0,07	0,06	0,05	0,1	0,05	0,05	0,05	0,05
x9	0,08	0,04	0,04	0,04	0,07	0,04	0,04	0,04	0,04	0,08	0,04	0,04	0,04	0,07	0,04	0,04	0,04	0,04
x10	0,08	0,04	0,04	0,04	0,07	0,04	0,04	0,04	0,04	0,08	0,05	0,04	0,04	0,07	0,04	0,04	0,04	0,04
x11	0,09	0,04	0,03	0,03	0,07	0,03	0,03	0,03	0,03	0,09	0,04	0,04	0,03	0,08	0,03	0,03	0,03	0,03
x12	0,09	0,05	0,04	0,04	0,07	0,04	0,04	0,04	0,04	0,09	0,05	0,04	0,04	0,08	0,04	0,04	0,04	0,04
x13	0,08	0,03	0,02	0,02	0,07	0,02	0,02	0,02	0,03	0,09	0,04	0,03	0,02	0,08	0,03	0,03	0,03	0,02
x14	0,09	0,06	0,05	0,05	0,08	0,05	0,05	0,05	0,05	0,09	0,06	0,06	0,05	0,09	0,05	0,05	0,05	0,05
x15	0,09	0,06	0,04	0,04	0,08	0,04	0,04	0,05	0,04	0,09	0,06	0,05	0,04	0,08	0,05	0,04	0,04	0,05
x16	0,09	0,07	0,06	0,06	0,08	0,05	0,05	0,05	0,06	0,09	0,06	0,06	0,06	0,08	0,06	0,05	0,05	0,06
x17	0,1	0,05	0,04	0,04	0,09	0,04	0,04	0,04	0,04	0,1	0,05	0,04	0,04	0,09	0,04	0,04	0,04	0,04
x18	0,12	0,11	0,11	0,11	0,12	0,11	0,11	0,11	0,11	0,12	0,12	0,11	0,11	0,12	0,11	0,11	0,11	0,11
x19	0,07	0,03	0,03	0,02	0,06	0,02	0,02	0,03	0,02	0,07	0,04	0,03	0,02	0,06	0,03	0,02	0,03	0,02
x20	0,07	0,04	0,03	0,03	0,06	0,03	0,03	0,03	0,03	0,07	0,04	0,03	0,03	0,07	0,03	0,03	0,03	0,03
x21	0,09	0,04	0,04	0,03	0,08	0,04	0,03	0,03	0,03	0,09	0,04	0,04	0,04	0,08	0,04	0,04	0,04	0,04
x22	0,08	0,06	0,05	0,05	0,07	0,05	0,05	0,05	0,05	0,08	0,06	0,05	0,05	0,07	0,05	0,05	0,05	0,05
x23	0,07	0,03	0,02	0,03	0,06	0,02	0,03	0,02	0,02	0,07	0,03	0,03	0,02	0,06	0,02	0,03	0,03	0,03
x24	0,09	0,04	0,03	0,03	0,07	0,03	0,02	0,03	0,02	0,08	0,04	0,03	0,03	0,08	0,03	0,03	0,03	0,03
x25	0,08	0,04	0,04	0,03	0,08	0,03	0,03	0,04	0,04	0,09	0,05	0,04	0,04	0,08	0,04	0,04	0,04	0,04
x26	0,08	0,03	0,02	0,02	0,07	0,02	0,02	0,02	0,02	0,08	0,03	0,02	0,02	0,07	0,02	0,02	0,02	0,02
x27	0,61	0,5	0,42	0,5	0,57	0,48	0,56	0,51	0,51	0,63	0,55	0,52	0,46	0,63	0,43	0,45	0,53	0,48
x28	0,15	0,09	0,07	0,08	0,13	0,07	0,07	0,06	0,08	0,15	0,08	0,08	0,08	0,14	0,07	0,07	0,07	0,08
x29	0,11	0,06	0,06	0,06	0,09	0,05	0,05	0,05	0,05	0,12	0,06	0,06	0,05	0,1	0,06	0,06	0,06	0,05
x30	0,14	0,14	0,14	0,13	0,14	0,14	0,14	0,13	0,14	0,13	0,14	0,14	0,14	0,14	0,14	0,14	0,14	0,14

References

Agresti A. (2002), Categorical Data Analysis, 2nd edition, Wiley.

Aitkin, M. (1991), 'Posterior Bayes factors', *Journal of the royal statistical society. Series B (Methodological)* **53** (1), 111-142.

Andrieu, C. and Thoms, J. (2008), 'A tutorial on adaptive MCMC', *Statistics and computing* **18** (4), 343-373.

Atchade, Y.F. and Rosenthal, J.S. (2003), 'On adaptive Markov Chain Monte Carlo algorithms', *Bernoulli* **5**, 759-948.

Baragatti, M. and Pommeret, D. (2012), 'A study of variable selection using g-prior distribution with ridge parameter', *Computational statistics and data analysis* **56** (6), 1920-1934.

Barbieri, M. and Berger, J.O. (2004), 'Optimal predictive model selection', *The annals of statistics* **32** (3), 870-897.

Bartlett, M.S. (1957), 'A comment on D.V. Lindley's statistical paradox', *Biometrica* **44** 533-534.

Berger, J.O. and Pericchi, L.R. (1996), 'The intrinsic Bayes factor for model selection and prediction', *Journal of the american statistical association* **91** (433), 109-122.

Berger, J.O. and Pericchi, L.R. (1998), 'Accurate and stable model selection: The median intrinsic Bayes factor', *Sankhya: The Indian journal of statistics, Series B* 60 (1) 1-18.

Bernardo, J.M. (1979), 'Reference posterior distributions for Bayesian inference', *Journal of the royal statistical society, Series B (Methodological)* **41**, 113-147.

Bernardo, J.M. and Smith, A.F.M. (1994), 'Bayesian Theory', Chichester: Wiley.

Bottolo, L. and Richardson, S. (2010), 'Evolutionary search for Bayesian model exploration', *Bayesian analysis* **5** (3), 583-618.

Brooks, S.P, Yanan, F. and Rosenthal, J.S. 'Perfect forward simulation via simulated tempering', *Research report*, Cambridge University

Brown, P., Vanucci, M. and Fearn, T. (1998), 'Multivariate Bayesian variable selection and prediction', *Journal of the royal statistical society, Series B (Methodological)* **60**, 627-641.

Carlin, B.P. and Chib, S. (1995), 'Bayesian model choice via Markov Chain Monte Carlo methods', *Journal of the royal statistical society, Series B (Methodological)* **157**, 473-484.

Celeux, G., Anbari, M.E., Marin, J.M. and Robert, C. (2010), 'Regularization in regression: Comparing Bayesian and frequentist methods in a poorly informative situation', *Bayesian*

analysis **7** (2), 477-502.

Chib, S. (1995), 'Marginal likelihood from the Gibbs output', *journal of the american statistical association* **90,** 1313-1321.

Chib, S. and Jeliaskov, I. (2001), 'Marginal likelihood from the Metropolis-Hastings output', *Journal of the american statistical association* **96**, 270-281.

Clyde, M.A. and George, E.I., (2000), 'Flexible empirical Bayes estimation for wavelets', *Journal of the royal statistical society, Series B (Methodological)* **62**, 681-698.

Clyde, M.A., Ghosh, J. and Littman, M.L. (2011), 'Bayesian adaptive sampling for variable selection and model averaging', *Journal of computational and graphical statistics* **20** (1).

Dellaportas, P., Forster, J.J. and Ntzoufras, I. (2002), 'On Bayesian model and variable selection using MCMC', *Statistics and computing* **12** (1), 27-36.

De Santis, F. and Spezzaferi, F. (1997), 'Alternative Bayes factors for model selection', *The Canadian journal of statistics/La revue Canadiene de statistique* **25** (4), 503-515.

Dobson, A. J. (2002), An introduction to generalized linear models second edition, Chapman & Hall/CRC

Efroymson, M. (1960), 'Multiple regression analysis, mathematical methods for digital computers', *Statistics and computing* **1**, 191-203.

Evans, M. and Swartz, T. (1995), 'Methods for approximating integrals in statistics with special emphasis on Bayesian intergration problems' *Statistical science* **10** (3), 254-272.

Fernandez, C., Ley, E. and Steel, M.F. (2001), 'Model uncertainty in cross-country growth regressions', *Journal of applied econometrics* **16** (5), 563-576.

Foster, D. and George, E. (1994), 'The risk inflation criterion for multiple regression', *Annals of statistics*, **22**, 1947-1975.

Fouskakis., D., Ntzoufras, I. and Draper, D. (2009), 'Bayesian variable selection using cost-adjusted BIC with application to cost-effective measurement of quality of health care', *The annals of applied statistics* **3** (2), 663-690.

Fouskakis., D., Ntzoufras, I. and Draper, D. (2009), 'Population-based reversible jump Markov chain Monte Carlo methods for Bayesian variable selection and evaluation under cost limit restrictions', *Journal of the royal statistical society, Series C (Applied statistcs)* **58**, 383-403.

Fouskakis., D., Ntzoufras, I. and Draper, D. (2015), 'Power-expected-posterior priors for variable selection in Gaussian linear models', B*ayesian Analysis*, **10**, 75-107.

Fouskakis., and D., Ntzoufras, I. (2015), 'Power-conditional-expected priors: Using g-priors with random imaginary data dor variable selection' JCGS (accepted)

Gelfand, A.E. and Dey, D.K. (1994), 'Bayesian model choice: Asymptotics and exact calculations', *Journal of the royal statistical society. Series B (Methodological)* **56** (3), 501-514.

Gelman, A. and Meng, X.L. (1998), 'Simulating normalizing constants: From importance sampling to bridge sampling to path sampling', *Statistical Science* **13**, 163-185.

Geman, S. and Geman, D. (1984), 'Stochastic relaxaation, Gibbs distributions and the Bayesian restoration of images', IEEE *transactions on pattern analysis and machine intelligence* **6**, 721-741.

Genz, A. and Kass R.E. (1993), 'Subregion-adaptive integration of functions having a dominant peak', *Journal of computational and graphical statistics* **6** (1) 92-111.

George, E. and Foster, D. (2000), 'Calibration and empirical Bayes variable selection', *Biometrika* **87** (4), 731-747.

George, E.I. and McCulloch, R.E. (1993), 'Variable selection via Gibbs sampling', *Journal of the american statistical association* **88** (423), 881-889.

George, E.I. and McCulloch, R.E. (1997), 'Approaches for Bayesian variable selection', *Statistica sinica* **7**, 339-374

Gilks, W.R., Roberts, G.O. and Sahu, S.K. (1998), 'Adaptive Markov Chain Monte Carlo thruogh regeneration', *Journal of the american statistical association* **93** (443), 1045-1054.

Claeskens G. and Hjort N. L. , 2008, Model Selection and Model Averaging, Cambridge University Press

Godsill, S.J. (2001), 'On the relationship between Markov Chain Monte Carlo methods for model uncertainty', *Journal of computational and graphical statistics* **10** (2), 230-248.

Green, P.J. (1995), 'Reversible Jump Markov Chain Monte Carlo computation and Bayesian model determination', *Biometrika* **82** (4), 711-732.

Guo, R. and Speckman, P.L. (1999), 'Bayes factor consistency in linear models' *in the 2009 international workshop on objective Bayes methodology, Philadelphia, June 5-9, 2009*.

Gupta, M. and Ibrahim, J. G. \An Information Matrix Prior for Bayesian Analysis in Generalized Linear Models with High Dimensional Data." Statistics Sinica, 19:1641{1663 (2009)

Haario, H., Saksman, E. and Tamminen J. (2001), 'An adaptive Metropolis algorithm', *Bernoulli* **7** (2), 223-242.

Han, C. and Carlin, B.P. (2000), 'MCMC methodsfor computing Bayes factors: A comparative review', *Biometrica* **82** (4), 711-732.

Hans, C., Dobra, A. and Wesst, M. (2007), 'Shotgun stochastic search for 'large p' regression', *Journal of the american statistical association* **102** (478), 507-516.

Hansen, M.H. and Yu, B. (2001), 'Model selection and the principle of minimum description length', *Journal of the american statistical association* **96**, 746-774.

Hansen M.H. and Yu B. (2003), 'Minimum Description Length selection criteria for Generalized Linear Models', Lecture Notes-Monograph Series, **40**, 145-163.

Hanson, T. E., Branscum, A. J., and Johnson, W. O. \Informative g-Priors for Logistic Regression." Bayesian Analysis, 9(3):597{612 (2014).

Hartman, B.M, and Hart, J.D. (2009), 'Using reversible jump MCMC to account for model uncertainty', *Actuarial research clearing house.*

Hastings, W. (1970), 'Monte Carlo sampling methods using Markov chains and their applications', *Biometrica* **57** (1), 97-109.

Jasra, A. Stephens, D.A. And Holmes, C.C. (2007), 'On population-based simulation for static inference', *Statistics and computing* **17** (3), 263-279.

Jeffreys, H. (1946), 'An invariant form for the prior probability in estimation problems' *Proceedings of the royal statistical society in London, Series A* **186**, 453-461.

Jeffreys, H. (1961), 'Theory of probability', New York: Oxford University Press.

Ji, C.S. and Schmidler, S.C. (2009), 'Adaptive Markov Chain Monte Carlo for Bayesian variable selection', *Journal of computational and graphical statistics*, **22** (3), 708-728.

Johnson V.E. and Rossell D. (2010), 'On the Use of Non-Local Prior Densities in Bayesian Hypothesis Tests', *Journal of the Royal Statistical Society*, *Series B*, **72**,143–170.

Johnson V.E. and Rossell D (2012). 'Bayesian Model selection in high-dimensional settings', *Journal of the American Statistical Association*, **107**, 649–660.

Kass, R.E. and Raftery, A.E. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90** (430), 773-795.

Kass, R.E. and Vaidyanathan S.K. (1992), 'Approximate Bayes factors and orthogonal parameters, with application to testing equality of two Binomial proportions', *Journal of the royal statistical society. Series B (Methodological)* **54** (1), 129-144.

Kass, R.E. and Wasserman, L. (1995), 'Bayes factors', *Journal of the american statistical association* **52** (2), 93-100.

Kass, R.E. and Wasserman, L. (1996), 'The selection of prior distributions by formal rules', J*ournal of the american statistical association* **65**, 356-369.

Kuo, L. and Mallick, B. (1998), 'Variable selection for regression models', *Sankhya: The Indian journal of statistics, Series B* **60**, 65-81.

Krishna, A., Bondell, H.D. and Ghosh, S.K. (2009), 'Bayesian variable selection using an adaptive powered correlation prior', *Journal of statistical planning and inference*, **139** (8), 2665-2674.

Lamnisos, D., Griffin, J.E. and Steel, M.F.J. (2012), 'Adaptive Monte Carlo for Bayesian variable selection in regression models', *Journal of computational and graphical statistics* **22** (3).

Liang, F., Paulo, R., Molina, G., Clyde, M. and Berger, J. (2008), 'Mixtures of g-priors for Bayesian variable selection', *Journal of the american statistical association* **103**, 410-423.

Lindley, D.V. (1957), 'A statistical paradox', *Biometrica* **44** (½), 187-192.

Liu, J.S. (2001), 'Monte Carlo strategies in scientific computing', *Springer series in statics.*

Madigan, D. and York, J. (1995), 'Bayesian graphical models for discrete data', *International statistical review* 215-232.

Marin, J. and Robert, C. (2007), 'Bayesian core: A practical approach to computational Bayesian statistics' Springer-Verlag, New York.

Maruyama, Y. and George, E.I. (2011), 'gBF: A fully Bayes factor with a generalized g-prior', *The annals of statistics*, **39** (5), 2243-2794.

Mavridis, D. and Ntzoufras I. (2014). Stochastic Search Item Selection for Factor Analytic Models. *British Journal of Mathematical and Statistical Psychology,* **67***,* 284–303.

McCulloch, R. and Nelder, J.A. (1989), 'Generalized linear models (Monographs on statistics and applied probability 37)', Chapman Hall, London.

Meng, X.L. and Wong H.W. (1996), 'Simulating ratios of normalizing constants via a simple identity: A theoretical exploration', *Statistica sinica* **6**, 831-860.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E. (1953), 'Equation of state calculations by fast computing machines', *The journal of chemical physics* **21**, 1087

Miller, A. (2002), 'Subset selection in regression', Boca Raton, Florida, U.S.A.: Chapman & Hall/CRC.

Mitchell, T.J. and Beauchamp, J.J. (1988), 'Bayesian variable selection in linear regression', *Journal of the american statistical association* **83** (404), 1023-1032.

Newton, M. and Raftery, A.E. (1994), Approximate Bayesian inference with the weighted likelihood bootstrap', *Journal of the royal statistical society. Series B (Methodological)* **56** , 3-48.

Nott, D.J and Kohn, R. (2005), 'Adaptive sampling for Bayesian variable selection', *Biometrica* **92** (4), 747-763.

Ntzoufras, I. (1999), 'Aspects of Bayesian model and variable selection using MCMC, Athens: University of economics and business'

Ntzoufras, I. (2009), Bayesian modeling using WinBUGS, Wiley.

Oh, C., Ye, K., He, Q. and Mendell, N. (2003), 'Locating disease genes using Bayesian variable selection with the Haseman-Elston method', *BMC Genetics* **4**, Supl.1-S9, available at http://www.biomedcentral.com/1471-2156/4/s1/S69.

O' Hagan, A. (1995), 'Fractional Bayes factors', *Journal of the royal statistical society. Series B (Methodological)* **57** (1), 99-138.

O' Hara, R.B. And Sinalpaa, M.J. (2009), 'A review of Bayesian variable selection methods: What, how and which', *Bayesian analysis* **4** (1), 85-117.

Pasarica, C. and Gelman, A. (2010), 'Adaptively scaling the Metropolis algorithm using expected squared jumped distance', *Statistica sinica* **20**, 0000-0000.

Petralias, A. and Dellaportas, P. (2012), 'An MCMC model search algorithm for regression problems', *Journal of statistical computation and simulation* (ahead-of-print), 1-19.

Raftery, A.E. (1996), 'Approximate Bayes factors and accounting for model uncertainty in generalised linear models', *Biometrica* **83 (**2), 251-266.

Raftery, A.E., Madigan, D. and Hoeting, J.A. (1997), 'Bayesian model averaging for linear regression models', *Journal of the American Statistical Association* **92** (437), 179-191.

Roberts, G.O. and Rosenthal, J.S. (2009), 'Examples of adaptive MCMC', *Journal of Computational and Graphical Statistics* **18** (2), 349-367.

Shafer, G. (1982), 'Lindley's paradox', *Journal of the American Statistical Association* **77** (378), 325-334.

Smith, M. and Kohn, R. (1996), 'Nonparametric regression using Bayesian variable selection', *Journal of Econometrics* **75** (2), 317-343.

Spiegelhalter, D.J and Smith, A.F. (1982), 'Bayes factors for linear and log-linear models with vague prior information', *Journal of the royal statistical society. Series B (Methodological)* **44** (3), 377-387.

Tierney, L. and Kadane, B. (1986), 'Accurate approximations for posterior moments and marginal densities', *Journal of the american statistical association* **81** (393), 82-86.

Tierney, L. and Mira, A. (1999), 'Some adaptive Monte Carlo methods for Bayesian inference', Statistics in medicine 18, 2507-2515.

Yang A.J. and Song, X.Y. (2010), 'Bayesian variable selection for disease classification using gene expression data', *Bioinformatics* **26** (2), 215-222.

Yi, N., George, V. and Allison, D.B. (2003), 'Stochastic search variable selection for identifying multiple quantitative trait loci', *Genetics* **167**, 967-975.

Yu J.Z. and Tanner M.A. (1999), 'An analytical study of several Markov Chain Monte Carlo estimators of the marginal likelihood', *Journal of computational and graphical statistics* **4**, 839-853.

Zellner, A. (1971), 'An introduction to Bayesian inference in econometrics', New York: John Wiley.

Zellner, A. (1983), 'Applications of Bayesian analysis in econometrics', *The statistician* **32**.

Zellner, A. (1986), 'On assessing prior distributions and Bayesian regression analysis with g-prior distributions', *in Bayesian inference and decision techniques: Essays in Honor of Bruno de Finetti, North-Holland/Elsevier*, 233-243.

Zellner, A. and Siow, A. (1980), 'Posterior odds ratios for selected regression hpotheses', i*n Bayesian statistics: Proceedings of the first international meeting held in Valencia (Spain)*, 585-603.