Journal of Computational Statistics and Modeling, Vol. 1, No. 1, 165-180, Spring & Summer 2019

Bayesian Variable Selection in Regression Models using The Laplace Approximation

Sima Naghizadeh *[1]

[1] Assistant professor, National Organization for Educational Testing, Tehran, Iran.

Recieved: 16/6/2019 Accepted:23/6/2019

Abstract: The Bayesian variable selection analysis is widely used as a new methodology in air quality control trials and generalized linear models. One of the important and, of course, controversial topics in this area is the selection of prior distribution for unknown model parameters. The aim of this study is presenting a substitution for mixture of priors which besides preservation of benefits and computational efficiencies obviate the available paradoxes and contradictions. In this research we pay attention to two points of view; empirical and fully Bayesian. Especially, a mixture of priors and its theoretical characteristics is introduced. Finally, the proposed model is illustrated with a real example.

Keywords: Generalized Linear Models, Bayesian Variable Selection, Mixture of Priors, Bartlett's Paradox, Information Paradox, Empirical Bayesian analysis, Air Pollution

Mathematics Subject Classification (2010): 62G.

*Corresponding author: s-naghizadeh@yahoo.com

1. Introduction

The Bayesian variable selection approach have been developed by a number of authors, for example, Skene and Wakeeld (1990),Carlin (1992) andSmith et al. (1995). Chung and Dunson (2007) maintain that these approaches are simple and efficient methods for computing posteriors in a mixture of prior distributions. Related approaches with the Bayesian variable selection, have been discussed by Muller et al. (2004). Dominici and Parmigiani (2001), and Carota and Parmigiani (2002) have also focused on semi-parametric Bayesian variable selection approach for count data, although in a different settings than considered here. In the context of Bayesian variable selection, one can use covariates at the study level which could explain the differences among studies. Thompson (1994) argued that heterogeneity can be regarded as an asset rather than a problem. In Bayesian variable selection the trial characteristics are considered as covariates in a regression analysis with the estimated treatment effect of the trial as the dependent variable. Ideally, the covariates used in such an analysis should be specified in advance to reduce the risk of post hoc conclusions prompted by inspecting the available data. Otherwise, there is a danger of false positive results. There are two key ideas in this paper. The first idea is the method used to construct a class of mixture of priors having support close to hyper-parameter of ϕ. The second idea is the computational technique used to find the posterior distributions. It is intended to extend the results of Liang et al. (2008) to Bayesian variable selection approach. However, prior distribution selection of model parameters is an issue which causes motivation and main necessity for this research. In fact, using improper priors in model selection topics is not acceptable, because these priors cause unknown Bayesian factors and posterior probabilities and so application of proper priors in this field is suggested. Here, we pay attention to Bayesian model selection in general linear models or in other words, Bayesian variable selection in these models. In Section 2, we present our proposed model based on conditional distributions. In section 3, we compute the posterior distributions. Since the posterior distributions do not have closed forms, they are approximated by simulation techniques. Gibbs sampling and Metropolis-Hastings algorithm are used to obtain the estimates of the parameters in a Bayesian manner. In the end of section 4, the efficiency of our proposed method is examined on a three simulated data set. Finally, in Section 5, an illustrative example is given.

2. The Model

The random component of generalized linear models consists of a response variable $\mathbf{Y}$ with independent observations $\mathbf{Y} = (Y_1, \cdots, Y_n)^{'}$, from a distribution in the

natural exponential family. This family has probability density function or mass function of form

$$f(y_i;\mu_i) = a(\mu_i)b(y_i)exp(y_iQ(\mu_i)), \tag{2.1}$$

Now, suppose for y, $(y = 1, 2, \cdots, p)$, that $\boldsymbol{\beta}_\gamma = (\beta_{\gamma 0}, \cdots, \beta_{\gamma q})$ is the vector of regression parameters corresponding to the of q covariates $\mathbf{x}_i' = (1, x_{i1}, \cdots, x_{iq})$. Using the link function $g(\mu_i)$, the assumed structure for the link function is $g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}_\gamma$, $i = 1, \cdots, n$. Then, for a response variable Y, and a collection of predictive variables, $\mathbf{X} = (X_1, \cdots, X_q)^{'}$, we would like to choose the best relating model among all the proposed generalized linear models. So, mean of response variable is vector $\mu = (\mu_1, \cdots, \mu_n)^{'}$, and variance-covariance matrix is Σ. We assume that errors are independent and have multivariate distribution with mean vector of zero and variance-covariance matrix of Σ. Bayesian approach in model selection and model uncertainty consists of determination of prior distribution of unknown parameters, $\theta_\gamma = (\beta_\gamma, \Sigma) \in \theta_\gamma$ for correction of prior probabilities of models M and then finding posterior probabilities of each candidate models as following;

$$\begin{aligned} P(M_\gamma|y) =& \frac{P(M)P(YM_\gamma)}{\Sigma(P(M_\gamma)P(YM)} \\ &= \frac{P(M_\gamma)\int P(Y|\theta, M_\gamma)P(\theta|M_\gamma)d\theta}{\Sigma\int P(M_\gamma)P(Y|\theta, M_\gamma)P(\theta|M_\gamma)d\theta} \end{aligned} \tag{2.2}$$

Determining prior distribution of unknown parameters, in model selection of general linear models(GLM) is on the basis of using prior distributions like both symmetric and non-symmetric priors. We use both symmetric and non-symmetric priors families as (SN-priors), for selection of prior distribution (β_γ, Σ), in general linear model, $g(y) = X\beta + \delta$, as follows; $E(\delta) = 0$, $Var(\delta) = \Sigma$

$$P(\Sigma) \propto \Sigma^{-1}, \tag{2.3}$$

and

$$\beta|\Sigma \sim N(\eta, \pi^*) \tag{2.4}$$

where $\pi^* = \prod_i^p (X_i^{'} X_i)$. One Bayesian approach for this problem is to use a prior distribution for $\boldsymbol{\beta}_\gamma$. Since each individual covariate cannot be expected to exert much leverage on the response, one may choose $\boldsymbol{\mu}$ in the vicivity of zero and σ a small number, at most to reflect the prior belief on $\boldsymbol{\beta}_\gamma$. Therefore, this prior for the sake of simplicity, being intelligible and computational efficiency has been adaptive in SN-prior, extensively. One of the benefits of using priors is that their users are only obliged to specify parameter π and it is clear that the relevant inferences are also impressed by that selection. Many papers are presented in how

should be behaved with this parameter and until recently using constant values for π has been considered. In recent years, George and Foster (2000) noticed at the policy of empirical Bayes for specifying the parameter π but for the contradictions in using constant values and also, for the unfavorable points of view of many statisticians toward empirical Bayes, fully Bayes is a common substitution. In this approach, a suitable prior for parameter π is used so this way, our inferences will be more robust. In this research by study of fully Bayes approach, we introduce and evaluate a mixture of symmetric and non-symmetric priors. Noting that the computation of marginal likelihoods using a mixture of symmetric and non-symmetric priors (SN-priors) only includes a one-dimensional integral, In addition to guaranteeing robustness in regard to misspecification of parameter π in mixture of symmetric and non-symmetric priors (SN-priors), this point of view keeps some interesting computational benefits of primary SN-priors. Variable selection in General linear models is a multiple hypothesis testing problem which ends up in many non-nested comparisons. So, for the possibilities of comparison among models, using a base model is being noticed. The model posterior probability M_γ in 2.2 using Bayes factors can be rewritten as follows;

$$P(M_\gamma|y) = \frac{P(M_\gamma)BF(M_\gamma : M_b)}{\Sigma_\gamma P(M_\gamma)BF(M_\gamma : M_b)} \tag{2.5}$$

So that the Bayes factor $BF(M_\gamma : M_b)$ is obtained form the proportion of marginal likelihood of model, M_γ, to the base model M_b. In the approach which uses zero model as base model for computation of Bayes factors and posterior probabilities of models, each model, M_γ, is compared to zero model, M_b, through hypotheses $H_0 : \beta_\gamma = 0$ and $H_1 : \beta_\gamma \neq 0$. Besides these assumptions, we assume that columns of design matrix have been centralized and then $|X'X| = 0$. This status is based on discussions related to the transformation stability of location and scale, and orthogonal parameterizations which results in using priors;

$$P(\Sigma_\gamma|M_\gamma) \propto \Sigma_\gamma^{-1}, \quad (\beta_\gamma|\Sigma_\gamma, M_\gamma) \sim N(\mu_\gamma, \Phi_\gamma^*) \tag{2.6}$$

as a SN-prior of parameters, $\theta_\gamma = (\beta_\gamma, \Sigma_\gamma) \in \theta_\gamma$ under model M_γ. Now, Bayesian inference is based on the analysis of the posterior distribution. In general, this posterior will not have a known closed form rather it will have a complicated high dimensional density only known up to the normalizing constant which makes direct inferences almost impossible. Markov Chain Monte Carlo (MCMC) methods are techniques that have been developed to resolve this kind of problem. Thus, we employ the MCMC method to compute the posterior distribution, the posterior expectation of some function of β_γ and the marginal likelihood. Important benefit of SN-priors is their computational efficiency which ends up in a closed form for marginal likelihoods and then provides accurate interpretation for Bayes fac-

tors. We have to compute $m(y|M_\gamma)$ as the marginal likelihood and $\pi(\beta_\gamma, \Sigma_\gamma|y)$ as the joint posteriors distribution of β_γ and Σ_γ. Estimation of marginal likelihood and the posterior distribution is quite simple. First, we estimate the marginal distribution. The marginal distribution can be computed from the n realizations of the Gibbs sequence. For $j = 1, \ldots, K$, if we draw a large number of values $(\beta_{1\gamma}, \cdots, \beta_{n\gamma})$ and $(\Sigma_{1\gamma} \cdots, \Sigma_{n\gamma})$ from the density 2.4, then, from 2.6 we shall have:

$$\begin{aligned} m(y|M_\gamma) &= E_{\pi(\beta,\Sigma)}(L(\beta_{1\gamma}, \cdots, \beta_{n\gamma}, \Sigma_{1\gamma}, \cdots, \Sigma_{n\gamma})|y) \\ &\propto \frac{1}{n}\Sigma_{i=1}^{n} L(\beta_{i.1\gamma}, \cdots, \beta_{i.n\gamma}, \Sigma_{i.1\gamma}, \cdots, \Sigma_{i.n\gamma}|y) \end{aligned} \tag{2.7}$$

The estimator 2.7 is unstable when the priors are diffuse or the likelihood is much more concentrated than the priors. In such cases the simulation will be inefficient since most of the simulated values will have low likelihood values and therefore the estimator will be dominated by few large values. Moreover, the variance of the estimator 2.7 will be large and convergence of the estimator to its true value will be very slow. An alternative way to approximate the marginal distribution is the Laplace approximation. This method has been used by Kadane and Lazar (2004)

$$log(m(y|M_\gamma)) \approx .5d * log(2) + .5log|H^*| + log(L(\beta^*, \Sigma^*|y)\pi(\beta^*, \Sigma^*)) \tag{2.8}$$

where β_γ^* and Σ_γ^* is the vector of posterior mode estimate of β_γ and Σ_γ , and * is the inverse of the Hessian matrix $\partial^2 h(\beta_\gamma, \Sigma_\gamma)/\partial\beta_\gamma\partial\beta_\gamma'$ of

$$h(\beta_\gamma, \Sigma_\gamma) = log(L(\beta_\gamma^*, \Sigma_\gamma^*|y)\pi(\beta_\gamma^*, \Sigma_\gamma^*))$$

evaluated at β^* and Σ^*. Usually, the Bayes factor is used for models comparison. If we apply the above approximation by expanding the numerator and denominator of the Bayes factor, we would get an approximation of the Bayes factor. For the models M_γ and $M_{(\gamma\prime)}$, one has

$$BF_{(\gamma\gamma')}[M_\gamma, M_(\gamma')] \approx \frac{L(\beta_\gamma^*, \Sigma_\gamma^*|y)}{L(\beta_{\gamma'}^*, \Sigma_{\gamma'}^*|y)} \left| \frac{H(\beta_\gamma^*, \Sigma_\gamma^*)}{H(\beta_{\gamma'}^*, \Sigma_{\gamma'}^*)} \right|^{0.5} \frac{1}{2\pi}^{(d_{\gamma'} - d_\gamma)/2} \tag{2.9}$$

or

$$log(BF_{(\gamma\gamma')}[M_\gamma, M_(\gamma')] \approx log(\lambda_n) + S((\beta_\gamma^*, \Sigma_\gamma^*), (\beta_{(\gamma\prime)}^*, \Sigma_{(\gamma\prime)}^*)) \tag{2.10}$$

Where λ_n is the standard likelihood ratio for the comparison of models M_γ and $M_{\gamma'}$ and $S((\beta_\gamma^*, \Sigma_\gamma^*), (\beta_{(\gamma\prime)}^*, \Sigma_{(\gamma\prime)}^*))$, denote the remainder term. When M_γ is a sub-model$M_{\gamma'}$, the remainder term $S((\beta_\gamma^*, \Sigma_\gamma^*), (\beta_{(\gamma\prime)}^*, \Sigma_{(\gamma\prime)}^*))$, is $o(1)$. To compare various models by the Bayes factor, we use the Bayes factor approximation 2.10. Now, 2.10 along with its interpretations provided by Kass and Raftery (1995),

is used to choose the best model for an specific example , i.e. in Tehran's Air Pollution as capital of Iran. Constant Values for parameter π in SN-priors is like a dimensionality penalty parameter. So, its selection is crucial and important. Until recently, it was emphasized to use a constant value for parameter π, and some selection methods were suggested, which we refer to some of them. Berger and Perichi (2001) suggested some selections for priors so that the amount of information about the parameter equals to the average amount of information contained in one observation. In the discussion of Normal-Linear models, unit information prior corresponds to the selection of $\pi = n$. George and Foster (2000) indicated that Bayesian model selection using this selection corresponds to the selection of the BIC criterion. George and Foster (2000) calibrated some priors based on RIC criterion for model selection and suggested $\pi = p^2$ based on a minimax point of view.
Carota and Parmigiani (2002) had an extensive study on different and possible selections of parameter π so that these selections depend on the sample size n and the dimension of dependent model, p. They showed the results of their studies with suggest $\pi = max(n, p^2)$. We will refer their benchmark prior which relates RIC and BIC criterion as BRIC.

2.1 Paradoxes of Constant SN-priors

It can be concluded that, Bayes factors for model selection using constant selections for parameter π may have some unfavorable features. For inference under a given model even when we consider a very large amount for π, again the posterior probability of this model is logical and justifiable but for constant n, when parameter π goes to infinity,$\pi \longrightarrow \infty$, Bayes factor 2.8, for comparison between two models, and will go to zero. So, high dispersal of prior distribution which is done with non-informative choice of π, results in a way that Bayes factor always behaves for confirmation of zero model, regardless of available information in data. This issue is not always true, actually, it is a paradox. This contradiction is Bartlett Paradox.

3. Empirical Bayes SN-priors

The idea behind local empirical Bayes strategy is that under each of these models, we estimate parameter π separately. Using 2.5, an empirical Bayes estimation of this parameter is a maximum likelihood estimation (marginal) which is conditional on not being negative. The marginal distribution can be computed from the n realization of the Gibbs sequence. So, equivalently, we can present empirical Bayes estimator of π by $\pi_\gamma(EB)$. It will be the local empirical Bayes estimation of pa-

rameter Π so that $\pi_{\gamma}^{EB} = Max(\Sigma \int P(M_{\gamma})P(Y|\theta, M_{\gamma})P(\theta|M_{\gamma})d\theta)$ is an ordinary statistic of Distribution function for testing .

3.1 Global Empirical Bayes:

Global empirical Bayes point of view for assigning parameter π sets an estimation of a common value among each models in which this common value will be obtained from maximizing weighted mean of marginal likelihoods and using model priors as a weighted mean. So, global empirical Bayes estimation of parameter π is; $\pi_{\gamma}(GEB) = Mean(Max(\Sigma \int P(M_{\gamma})P(Y|\theta, M_{\gamma})P(\theta|M_{\gamma})d\theta))$. The last expression is not flexible and does not provide a closed form for $\pi_{\gamma}(GEB)$ however numerical methods can be useful. We can expand an EM algorithm for computation of the above expression. $\pi_{\gamma}(GEB)$ is recalled as MLE type II.
It can be shown that for constant n and $p, p < n$, under each of two local and global empirical Bayes estimators, whenever the probability goes to 1, Bayes factor 2.8 for comparison of M and M_b goes to infinity and therefore, the information paradox which results in using constant value for parameter g in SN-priors will be resolved. In addition to using empirical Bayes estimation for obtaining g, a common substitution is a closed-form marginal likelihood under a proper prior on this parameter.

3.2 A Mixture of SN-Prior

Suppose $P(g)$ (It is possible to be dependent on n) represent prior distribution of hyper parameter g in Zellner SN-priors. So, in the approach based on zero model, marginal likelihood of data, , will be proportionate to the following Bayes factor:

$$BF(M_{\gamma} : M_N) = \int_0^{\infty} (1+g)^{((n+p)/2)}(1+g(1-R_{\gamma}^2))^{(-(n-1)/2)}P(g)dg$$

According to relation 2.9, a similar expression for the approach based on complete model is available. Regarding relation 2.7 and under the selection of a model, the posterior mean of is as follows;

$$E(\beta_{\gamma}|Y, M_{\gamma}) = E(E(\beta_{\gamma}|Y, M_{\gamma}, g)) = E(\frac{g}{1+g}|Y, M_{\gamma})\beta_{\gamma}^{OLS}$$

So that β_{γ}^{OLS} is ordinary least squares estimator of β under M_{γ} model. Under constant G-priors, posterior mean of β_{γ} in a selected model, M_{γ}, is a linear shrinkage estimator with constant shrinkage factor of $\frac{g}{1+g}$ whereas a mixture of SN-priors provides the possibility of dependency of shrinkage estimator on data, adaptively. We investigate two cases of SN-priors mixtures here; Zellner and Siow Cauchy multivariate prior and hyper SN-prior.

3.3 Zellner- Siow prior

Zellner and Siow (1980) in a hypothesis testing related to the mean of one-variable Normal Distribution, rejected Normal priors and suggested Cauchy priors. After that Kass and Raftery (1995) paid attention to multivariate Cauchy priors for regression coefficients which in fact, it was a generalization of Jeffryes work to the multivariate normal mean problem. For comparison between two models, if one of them is considered as nested into the other one, then Zellner and Siow point of view for assigning the priors of unknown parameters is a flat prior on common parameters of two models and also a multivariate Cauchy prior on remaining parameters. For example, in the approach based on zero model, Zellner and Siow prior is as follows;

$$P(\alpha, \phi|M_\gamma) \propto \frac{1}{\phi}$$

and

$$P(\beta_\gamma|\phi, M_\gamma) \propto \frac{\Gamma(p_\gamma/2)}{\pi^{p_\gamma/2}} \left|\frac{X_\gamma' X_\gamma}{n/\phi}\right|^{1/2} \left(1 + \beta_\gamma' \frac{X_\gamma' X_\gamma}{n/\phi}\beta_\gamma\right)^{-\frac{p_\gamma}{2}}$$

This does not offer a closed form for marginal likelihoods but Zellner and Siow got some approximation for marginal likelihoods so that whatever the dimension of model (p_γ) increases, the accuracy of these approximations decreases. But one of the significant features of multivariate Cauchy distribution is the possibility of presenting it by a mixture of Normal random variables densities. Then Zellner and Siow priors can be represented via this method and with setting an Inverse Gamma prior distribution, $IG(.5, n/2)$ on parameter g. This means;

$$P(\beta_\gamma|\phi, M_\gamma) \propto \int N(\beta_\gamma|0, \frac{g}{\phi}(X_\gamma' X_\gamma)^{-1})P(g)dg$$

In which;

$$P(g) = \frac{(n/2)^{0.5}}{\Gamma(0.5)} g^{-1.5} exp(-n/(2g))$$

The benefit of this representation is; by using $P(g)$ in relation 2.10 for computation of Bayes factor, we only face a one-dimensional integral on g which is independent of model dimension and then the accuracy of approximations will not differ with increase in the model dimension.

3.4 Hyper SN- priors

As a substitution for Zellner-Siow priors in Bayesian variable selection problems, we introduce another family of priors named Hyper G-prior for parameter g as follows;

$$P(g) = \frac{a-2}{2}(1+g)^{-a/2} \tag{3.11}$$

This prior is a special case of Inverse Beta prior Zellner and Siow (1980) which has the following density function;

$$P(g) = \frac{\Gamma(b+c)}{\Gamma(b)\Gamma(c)} g^{b-1}(1+g)^{-(b+c)}, \quad g > 0$$

With $b = 1$ and $c = \frac{a}{2} - 1$, prior 3.11 is obtained. Hyper G-prior for $a > 2$ is a proper prior but for $a < 2$, it is an improper one. Also, $a = 2$ corresponds to Jeffreys and reference priors. Using the values of $a \leq 2$ for Bayesian factors causes contradictions, so we ignore this case. In this field, each selection in range of $2 < a \leq 4$ is logical. So, with presence of shrinkage factor, $\frac{g}{1+g}$, in marginal posterior distribution of β_γ , mostly the prior corresponding to this factor is paid attention. If we consider the prior in relation 3.11 for g, then this shrinkage factor has Beta distribution with mean, $\frac{2}{a}$. The benefit of Hyper G-prior is; for each model M_γ, the posterior density of parameter g is available as closed form. According to relation 2.5, we have;

$$\begin{aligned} P(g|Y, M_\gamma) &\propto P(Y|g, M_\gamma)P(g) \\ &= \frac{P_\gamma + a - 2}{F((n-1)/2, 1; (p_\gamma + a)/2; R_\gamma^2)} * \frac{(1+g)^{(n-1-p_\gamma)/2}}{(1+g(1-R_\gamma^2))^{(n-1)/2}} \end{aligned}$$

In this relation $F(a, b; c; d)$ is an indicator of Gaussian Hyper Geometric function (See the appendix). Also with using this function many necessary quantities can be computed. As an example, by using Hyper SN-priors and integral form of Gaussian Hyper Geometric function, the Bayes factor based on zero model and according to relation 3.11 is as follows;

$$\begin{aligned} BF[M_\gamma : M_N] &= \frac{a-2}{2} \int_0^\infty (1+g)^{(n-1-p_\gamma-a)/2}(1+g(1-R_\gamma^2))^{-(n-1)/2} dg \\ &= \frac{a-2}{p_\gamma + a - 2} * F((n-1)/2, 1; (p_\gamma + a)/2; R_\gamma^2) \end{aligned}$$

In addition, expected value of shrinkage factor, $\frac{g}{1+g}$, which is used for computation of posterior mean of β_γ , under each of M models is as follows;

$$E(\frac{g}{1+g}|M_\gamma, Y) = \frac{2}{p_\gamma + a} \frac{F((n-1)/2, 2; (p_\gamma + a)/2 + 1; R_\gamma^2)}{F((n-1)/2, 1; (p_\gamma + a)/2; R_\gamma^2)}$$

4. Efficiency Evaluation

In this section for efficiency evaluation of a small sample in the discussed approaches, in the form of a simulation study and investigation of a real sample, we compare Zellner-Siow Cauchy priors and Hyper G-prior with other important points of view. Description of comparing approaches –under comparison- is mentioned in Table (1).

Table 1: Some explanation related to approaches compared in simulation Comparison Via Simulation.

AIC	Akaike Information Criterion
HG-3	Hyper SG -prior with a=3
EBG	Global Empirical Bayes Estimation of Parameter g in G-prior
ZS-F	Base Model-Complete Model, Multivariate Cauchy Prior forβ and Flat Prior ϕ and β_γ
ZS-N	Base Model-Zero Model, Multivariate Cauchy Prior forβ and Flat Prior ϕ and β_γ
BRIC	Constant G-prior of George and Foster (2000) with $g = max(n, p^2)$
EBL	Local Empirical Bayes Estimation of Parameter g in G-prior
BIC	Bayesian Information Criterion

4.1 Comparison Via Simulation

In this simulation we produce response variable data, Y, from a Linear-Normal model,$Y = 1_n\alpha + X\beta + \delta$, so that errors, δ , follow a multivariate normal distribution with mean vector of zero and variance-covariance matrix, $\frac{I_n}{\phi}$. In the process of producing data, we select $\alpha = 2$ and $\phi = 1$, also we suppose sample size, n, equals to 100. Following, Cui and George (2008) and for speeding up in computation, we focus on a special case of orthogonal design matrix ($X = I_n$) and consider the number of predictive variables as $p = 7$. For a model with p_γ variables the regression multiples, β_γ , are produced from a multivariate normal distribution, $N_{p_\gamma}(0, g/\phi I_{p_\gamma})$, and also, we consider that remaining elements, β, equal to zero; $\beta_{p_\gamma+1}, \cdots, \beta_p = 0$. In this simulation as George and Foster (2000), we investigate two values for $g = 5, 25$, and consider squared error loss criterion for comparison of different methods efficiency for each method of m as follows;

$$MSE(m) = (X\beta - X\hat{\beta}^{(m)})'(X\beta - X\hat{\beta}^{(m)})$$

In this quantity $\hat{\beta}^{(m)}$ is an estimator of β under method m. Under each of these methods and for comparison among different models, we considered highest posterior probability index of model. In this situation, β estimator is its posterior mean under the selected model. For BIC the marginal likelihood logarithm of model M_γ is defined as follows;

$$log(P(Y|M_\gamma)) = -\frac{1}{2}(nlog(\hat{\sigma}^2_\gamma) + p_\gamma log(n)). \tag{4.12}$$

In which $\hat{\sigma}^2_\gamma = \frac{RSS_\gamma}{n}$ is maximum likelihood estimator of σ^2 under model M_γ . For AIC this quantity is obtained with replacement of $2p_\gamma$ with $log(n)$ in 4.12.

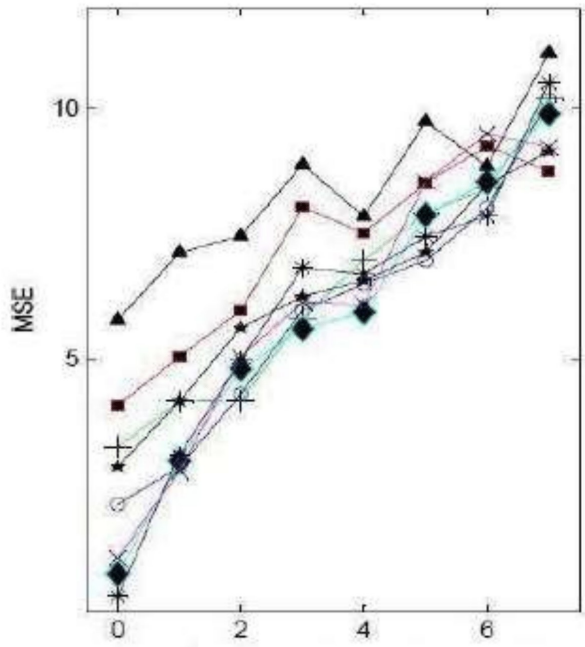

Figure 1: Graphs of mean of MSE for each method simulation using AIC(∗),BIC(*),BIRC(SN-prior)(◁) EB-local(▷),EB-global(.),hyper-π(○), ZS-full(×),and ZS-null,π=5,a=3

These marginal likelihoods are applied for the computations related to posterior probability of models. Under BIC and AIC, β_γ estimator (posterior mean) will be ordinary least squares estimator of ($\hat{\beta}_\gamma^{ols}$). Also in this section, we consider the uniform prior probabilities for models which mean $P(M_\gamma) = 2^{-p}$. For each value of g and $p_\gamma = 0, 1, \cdots, 7$, we produce the response variables, Y, as mentioned, and for getting posterior mean of β_γ under each eight methods, we compute $MSE(m)$. For each three-compound of g, p_γ and method m, we repeat this data production and computation of $MSE(m)$, for 500 times, and then get mean of $MSE(m)$ for each combination and mentioned repetitions. In figure (1), mean of MSE against the number of predictive variables of accepted model has been drawn. For the results of MSE, for local empirical Bayes (LEB), global empirical Bayes (GEB), Zellner-Siow prior based on zero model (ZS-N) and Hyper SN-prior (HG-3) are approximately the same and for models with different sizes, these methods are preferable to the other ones. In this figure, for simplicity only Hyper SN-prior with $a = 3$ is drawn but we cannot judge about other hyper SN-priors based on this case. When a complete model is fit to data, the results show that a hyper-SN prior with $a = 3$ has a better function than other approaches of fully Bayes.

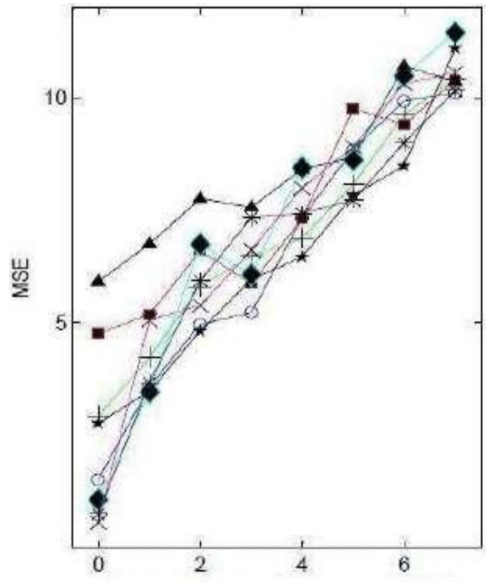

Figure 2: Graphs of mean of MSE for each method simulation using AIC(),BIC(*),BIRC(SN-prior)(◁) EB-local(▷),EB-global(.),hyper-π(○), ZS-full(×),and ZS-null,π=5,a=25

The difference between fully Bayes methods and other ones when a zero model is fit to data is more obvious. When zero model is accepted, the global empirical Bayes (EBG) has the best function and the reason is that in EBG, the g estimator will get its power from all models and in this situation, it tends more towards estimator $\pi = 0$. We observe that a mixture of fully Bayes SN-priors except in zero model acts like EBG. While Cui and George (2008), with a similar simulation, found that EBG has better function than fully Bayes standpoints. In this simulation, we have used a uniform prior on model space but Cui and George used Bernoulli prior probabilities for prior inclusion probability of variables. So, selection of model prior probabilities can affect the results.

5. Surveying a Real Example

Some researchers which work on air quality control testify that Carbon Monoxide (CO) pollutant has the most portion in Tehran's air pollution (Capital of Iran). In recent years by imposing some policies such as using gas-fuel cars, traffic control plans, technical examination of cars, substitution of old cars with new ones and etc., there has been an effort to reduce the density of this pollutant but, this pollutant again has a significant portion in air pollution. In year 2008, with the

Table 2: : Posterior inclusion probabilities for each variable under different priors in Carbon Monoxide data.

Var	BRIC	ZS-N	ZS-F	HG3	EBL	EBG	AIC	BIC
A1	0.57	1.00	1.00	1.00	1.00	1.00	1.00	1.00
A2	0.45	1.00	1.00	1.00	1.00	1.00	1.00	1.00
A3	0.16	0.99	1.00	0.99	0.99	1.00	1.00	1.00
A4	0.06	0.99	0.99	0.99	0.99	0.99	1.00	1.00
A5	0.09	0.97	0.99	0.97	0.98	0.98	0.99	0.99
H1	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
H2	0.66	1.00	1.00	1.00	1.00	1.00	1.00	1.00
H3	0.68	0.99	0.99	0.99	0.99	0.99	0.99	0.99
H4	0.69	0.35	0.52	0.35	0.34	0.34	0.68	0.58
H5	0.66	0.95	0.97	0.95	0.95	0.96	0.99	0.99
H6	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00
H7	0.66	1.00	1.00	1.00	1.00	1.00	1.00	1.00
WS	0.36	0.06	0.61	0.06	0.06	0.06	0.88	0.76
ZI	0.81	0.71	0.58	0.70	0.71	0.70	0.60	0.56

aim of surveying Carbon Monoxide density in Tehran's air and in 6 air pollution evaluation stations(Fatemi, Bazzar, Aqdasyeh, Geophysic, Tehransar and Shahre Rey), the density of this pollutant in different hours of days and nights was measured. Also besides measuring CO density, other variables such as wind speed and vertical profile of temperature were measured too. Here, we pay attention to spring observations and compute the mean of CO density and weather variables in each estimation and in 8 different hours(00:30, 3:30, 6:30, 9:30, 12:30, 15:30, 18:30, 21:30), and use the above data for evaluating the influence of G-priors on posterior probabilities and estimation of multiples. In regard to the number of stations and surveyed hours, five dummy variables for different stations and 7 dummy variables for different hours in the regression model are used. The aim is to survey factors of region, hour, wind speed and vertical profile of temperature index in Carbon Monoxide density. Table 2 shows the influence of a mixture of SN-priors and other approaches on marginal posterior inclusion probabilities of predictive variables of Carbon Monoxide density.

These marginal probabilities for i-th variable are defined as follow;

$$P(\beta_i \neq 0|Y) = \sum_{\gamma_i} P(M_\gamma|Y)$$

Posterior inclusion probabilities are used for median probability model which is mostly similar to the model with the highest posterior probability. Median proba-

bility model includes predictive variables which have posterior inclusion probability greater than 0.5. It means $P(\beta_i \neq 0|Y) > 0.5$. The constant benchmark SN-prior of Carota and Parmigiani (2002) in this example corresponds to $RIC\pi = 256$. A mixture of ZS-N and HG3 SN- priors ends up with marginal posterior inclusion probabilities similar to data-dependent approaches of global and local empirical Bayes and so it causes the same median probability model. Table 1 shows the posterior mean and standard deviation of multiples under a mixture of SN- priors. In this table, it is clear from comparison of 6 stations that the most CO density in spring season is related to Fatemi station, also in different hours, the most CO density belongs to hour 18:30. Although, under a mixture of SN-priors, the relation between CO density and wind speed (WS) is inverse but this factor in spring season does not have a significant effect on CO density. CO density also has an inverse relation with vertical profile of temperature index (ZI) and as this index increases CO density decreases but unlike the wind speed, the influence of this index is significant. In the above table, dummy variables $A1, \cdots, A5$ present the station effects and dummy variables $H1, \cdots, H7$ are related to different hours effects.

5.1 Conclusion and Discussion

In this research; for Bayesian variable selection problem and prior distribution determination of unknown parameters, we review SN-priors. It is specified that using constant values for parameter has some inconsistencies. As a result, we introduce empirical Bayes SN-prior and a mixture of SN-priors as substitutions for constant SN-priors. In the simulated example, it was specified that a mixture of SN- prior behaves better than, or at least, as good as other supposed points of view. The investigation of Tehran's air pollution data also indicates that each three compound of SN-priors has the same results and using each of them does not change the results. Under zero model, although the global empirical Bayes approach has better function than other approaches, it is necessary to mention that whenever the number of models is high, the estimation of parameter g faces many challenges and in this situation a mixture of SN-priors such as Zellner-Siow Cauchy prior and hyper SN-priors are very good substitutions, in adaption and robustness, rather than misspecification of parameter ,and accelerate the computations related to marginal likelihoods. This feature is necessary for investigation of model spaces with high dimensions. The prior distribution on model space is a vital case in discussions related to Bayesian model selection and needs a special attention. In many studies in Bayesian variable selection fields, Bernoulli priors are used for prior inclusion probabilities of variables. Corresponds to uniform prior on model space that we use this prior for the probabilities of model priors. Mutually, we

can hold a hierarchical point of view towards the model space, and with setting a prior distribution, apply the fully or empirical Bayes for specifying the model priors.

Appendix

Gaussian Hyper Geometric Function Suppose for $r > 0$;

$$(\alpha)_r = \frac{\Gamma(\alpha + r)}{\Gamma(\alpha)} = \alpha(\alpha + 1) \cdots (\alpha + r - 1)$$

Then the hyper geometric function, is defined as follows;

$$F(\alpha_1, \cdots, \alpha_m, \beta_1, \cdots, \beta_n; Z) = \sum_{r=0}^{\infty} \frac{(\alpha_1)_r, \cdots, (\alpha_m)_r}{(\beta_1)_r, \cdots, (\beta_n)_r} \frac{z^r}{r!}$$

If we put $m = 2$, $n = 1$ in this definition, then the Gaussian Hyper Geometric function will be obtained. The integral presentation of Gaussian Hyper Geometric function $F_1(a, b; c; z)$ is as follows;

$$F_1(a, b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c - b)} \int_0^1 \frac{t^{b-1}(1 - t)^{c-b-1}}{(1 - tz)^a} dt$$

This relation is obtained easily through definition. The above integral for real value of $|z| < 1$ and the condition $c > b > 0$, is convergence and for values of $z = 1$ or $z = -1$, only under the conditions $c > b + a$ and $b > 0$ is convergence. In different statistical software package such as BAS package and under R software, some special instructions for computation of values of Gaussian Hyper Geometric values are defined.

Note: For accessing to real sample data and the program written under R software for the simulation example, please refer to the authors.

References

Skene, A. M. and Wakeeld, J. C. (1990). Hierarchical Modelsfor Multicenter Binary Response Studies. *Statistics in Medicine*, 9, 919-929.

Carlin, J.B. (1992). Meta-Analysis for 22 Tables:A Bayesian Approach. *Journal of Statistics in Medicine*, 11, 141-158.

Smith, T. C., Spiegelhalter, D. J., and Thomas, A. (1995). Bayesian Approaches to Random eects Meta-Analysis: A Comparative Study. *Journal of Statistics in Medicine*, 14, 2685-2699.

Chung, Y. and Dunson, D.B. (2007). Local Dirichlet Process. *working paper*, Duke University.

Muller, P., Quintana, F. and Rosner. G.L.(2004). A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society B*, 66, 735-749.

Dominici, F. and Parmigiani, G. (2001). Bayesian Semi-parametric Analysis of Developmental Data. *Biometrics*, 57, 150-157.

Carota, C. and Parmigiani, G. (2002). Semi-parametric regression for count data. *Biometrika*, 89, 265-281.

Thompson, S.G. (1994). Why Sources of heterogeneity in meta-analysis shouldbe investigated. *British Medical Journal*, 309, 1351-1355.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008), Mixtures of g-Priors for Bayesian Variable Selection, *Journal of the American Statistical Association*,103,410-422.

George, E. I., and Foster, D. P. (2000), Calibration and Empirical Bayes Variable Selection, *Biometrika*, 87, 731–742.

Kadane, J. and Lazar, N.(2004). Methods and criteria for model selection. *Journal of the American Statistical Association*, 99, 279-290.

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

Berger, J.O., and Perichi, L. (2001), *Objective Bayesian Methods for Model Selection: Introduction and Comparison in Model Selection*, ed. P. Lahiri,Hayward, CA: Institute of Mathematical Statistics, 135–192.

Zellner, A. and Siow, A. (1980), Posterior Odds Ratios for Selected Regression Hypotheses,in Bayesian Statistics: *Proceedings of the First International Meeting Held in Valencia*, Spain: University of Valencia Press, 585–60.

Cui, W. and George, E. I. (2008), Empirical Bayes vs. Fully Bayes Variable Selection, *Journal of the Statistical Planning and Inference*,138:4,888-900.